



# SARS-COV-2: FROM GENETIC VARIABILITY TO VACCINE DESIGN

EDITED BY: Indrajit Saha, Dariusz Plewczynski and Nimisha Ghosh  
PUBLISHED IN: Frontiers in Genetics



# frontiers

## Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-83250-176-4

DOI 10.3389/978-2-83250-176-4

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [frontiersin.org/about/contact](https://frontiersin.org/about/contact)



# SARS-COV-2: FROM GENETIC VARIABILITY TO VACCINE DESIGN

Topic Editors:

**Indrajit Saha**, National Institute of Technical Teachers' Training and Research, Kolkata, India

**Dariusz Plewczynski**, Warsaw University of Technology, Poland

**Nimisha Ghosh**, Siksha O Anusandhan University, India

**Citation:** Saha, I., Plewczynski, D., Ghosh, N., eds. (2022). SARS-CoV-2: From Genetic Variability to Vaccine Design. Lausanne: Frontiers Media SA.  
doi: 10.3389/978-2-83250-176-4

# Table of Contents

## **05 Editorial: SARS-CoV-2: From Genetic Variability to Vaccine Design**

Indrajit Saha, Nimisha Ghosh and Dariusz Plewczynski

## **PREDICTION**

### **08 COVID-DeepPredictor: Recurrent Neural Network to Predict SARS-CoV-2 and Other Pathogenic Viruses**

Indrajit Saha, Nimisha Ghosh, Debasree Maity, Arjit Seal and Dariusz Plewczynski

### **20 Detecting the Multiomics Signatures of Factor-Specific Inflammatory Effects on Airway Smooth Muscles**

Yu-Hang Zhang, Zhandong Li, Tao Zeng, Lei Chen, Hao Li, Tao Huang and Yu-Dong Cai

## **MUTATION**

### **34 Deciphering the Subtype Differentiation History of SARS-CoV-2 Based on a New Breadth-First Searching Optimized Alignment Method Over a Global Data Set of 24,768 Sequences**

Qianyu Lin, Yunchuanxiang Huang, Ziyi Jiang, Feng Wu and Lan Ma

### **44 Hotspot Mutations in SARS-CoV-2**

Indrajit Saha, Nimisha Ghosh, Nikhil Sharma and Suman Nandi

### **60 Structural and Drug Screening Analysis of the Non-structural Proteins of Severe Acute Respiratory Syndrome Coronavirus 2 Virus Extracted From Indian Coronavirus Disease 2019 Patients**

Nupur Biswas, Krishna Kumar, Priyanka Mallick, Subhrangshu Das, Izaz Monir Kamal, Sarpita Bose, Anindita Choudhury and Saikat Chakrabarti

### **79 Structural Insights on the SARS-CoV-2 Variants of Concern Spike Glycoprotein: A Computational Study With Possible Clinical Implications**

Marni E. Cueno and Kenichi Imai

## **VACCINE AND THERAPEUTICS**

### **89 Perspectives About Modulating Host Immune System in Targeting SARS-CoV-2 in India**

Sreyashi Majumdar, Rohit Verma, Avishek Saha, Parthasarathi Bhattacharyya, Pradipta Maji, Milan Surjit, Manikuntala Kundu, Joyoti Basu and Sudipto Saha

### **110 AI Aided Design of Epitope-Based Vaccine for the Induction of Cellular Immune Responses Against SARS-CoV-2**

Giovanni Mazzocco, Iga Niemiec, Alexander Myronov, Piotr Skoczylas, Jan Kaczmarczyk, Anna Sanecka-Duin, Katarzyna Gruba, Paulina Król, Michał Drwał, Marian Szczepanik, Krzysztof Pyrc and Piotr Stępnik

### **128 A Peptide Vaccine Candidate Tailored to Individuals' Genetics Mimics the Multi-Targeted T Cell Immunity of COVID-19 Convalescent Subjects**

Eszter Somogyi, Zsolt Csiszovszki, Levente Molnár, Orsolya Lőrincz, József Tóth, Sofie Pattijn, Jana Schockaert, Aurélie Mazy, István Miklós, Katalin Pántya, Péter Páles and Enikő R. Tőke

**147** *Structural Analysis of SARS-CoV-2 ORF8 Protein: Pathogenic and Therapeutic Implications*

Antonio Valcarcel, Antonio Bensussen, Elena R. Álvarez-Buylla and José Díaz

**155** *Correlation Between SARS-Cov-2 Vaccination, COVID-19 Incidence and Mortality: Tracking the Effect of Vaccination on Population Protection in Real Time*

Kiyoshi F. Fukutani, Mauricio L. Barreto, Bruno B. Andrade and Artur T. L. Queiroz

**160** *ORF8 and Health Complications of COVID-19 in Down Syndrome Patients*

Antonio Bensussen, Antonio Valcarcel, Elena R. Álvarez-Buylla and José Díaz



# Editorial: SARS-CoV-2: From Genetic Variability to Vaccine Design

Indrajit Saha<sup>1\*</sup>, Nimisha Ghosh<sup>2,3</sup> and Dariusz Plewczynski<sup>4,5</sup>

<sup>1</sup>Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, <sup>2</sup>Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India, <sup>3</sup>Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland, <sup>4</sup>Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, <sup>5</sup>Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland

**Keywords:** correlation, COVID-19, mutations, prediction, SARS-CoV-2, vaccine

## Editorial on the Research Topic

## SARS-CoV-2: From Genetic Variability to Vaccine Design

## 1 INTRODUCTION

The whole world has been at a standstill for more than 2 years now due to the pandemic of COVID-19, the disease caused by the SARS-CoV-2 virus. The first case of COVID-19 was detected in Wuhan, China in December 2019, and the rest, as they say, is history. The disease has claimed more than 6 million lives worldwide. SARS-CoV-2 is a positive-stranded RNA virus with a length of about 30 kb encompassing non-structural and structural proteins. Spike glycoprotein, a structural protein present on the virus surface plays an important role in binding with human ACE2 and other receptors. Since its detection in Wuhan, the virus has mutated several times and has given way to variants such as B.1.1.7 (Alpha), B.1.351 (Beta), B.1.525 (Eta), B.1.427/B.1.429 (Epsilon), B.1.526 (Iota), B.1.617.1 (Kappa), B.1.617.2 (Delta), C.37 (Lambda), P.1 (Gamma), P.2 (Zeta), P.3 (Theta), and B.1.1.529 (Omicron).

In the initial days of the pandemic, there was little to no knowledge of this deadly virus. Thus, to understand the virus, whole genome analysis, and viral protein-based comparisons were carried out which concluded that SARS-CoV-2 is mostly related to bat SARS-like coronaviruses. Though there have been viruses like SARS-CoV-1 and MERS-CoV which belong to the same family of Coronaviridae just like SARS-CoV-2, outbreaks were sporadic and they did not cause global pandemics. Moreover, since the virus shared similarities with other viruses, its prediction was yet another challenge that the research community faced. Also, phylogenetic analyses were carried out by different researchers around the world to understand the virus mutations which mostly take place in the Spike glycoprotein. In fact, tools like Nextstrain have been used to visualize the virus evolution as well. These efforts by the researchers helped in a lot of ways to understand the virus's spread and its mutations. However, the studies are mostly focused on the structural proteins, especially Spike glycoprotein of SARS-CoV-2 while research on non-structural proteins is still underway. Such proteins can be investigated further to understand the virus and its mutations better.

The efforts of the researchers have also paved the way for the development of vaccines to fight against this deadly virus. There are several vaccines like Oxford-AstraZeneca, Pfizer-BioNTech, Moderna, Novavax, Covaxin, Sputnik V, and Johnson & Johnson which have been developed to date by scientists around the world. However, the developed vaccines are primarily designed to generate neutralizing antibodies against Spike glycoprotein. Moreover, due to the waning antibody response

## OPEN ACCESS

### Edited and reviewed by:

Richard D. Emes,  
University of Nottingham,  
United Kingdom

### \*Correspondence:

Indrajit Saha  
indrajit@nittrkol.ac.in

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 02 June 2022

**Accepted:** 20 June 2022

**Published:** 26 August 2022

### Citation:

Saha I, Ghosh N and Plewczynski D  
(2022) Editorial: SARS-CoV-2: From  
Genetic Variability to Vaccine Design.  
Front. Genet. 13:960107.  
doi: 10.3389/fgene.2022.960107

and some emerging variants like Omicron being somewhat resistant to the antibody response evoked by these vaccines, the long-term sustainability of these vaccines is a bit questionable. In this regard, T-cell responses against coronaviruses can last for a very long time which has been demonstrated by SARS and MERS viruses as well. All these factors have motivated us to have an issue on “SARS-CoV-2: From Genetic Variability to Vaccine Design” to benefit the scientific community. The articles covered in this issue are discussed in the subsequent section.

## 2 RESEARCH TOPIC ORGANIZATION

This Research Topic is divided into three main sections: two papers discuss the prediction of the SARS-CoV-2 virus, four papers cover the virus mutations, and six papers discuss the various vaccines and therapeutics for COVID-19.

In the first part, we have focussed on the prediction of the virus by using machine learning and deep learning techniques. We believe that this section will appeal to researchers working in the field of artificial intelligence. This section is especially interesting as while one paper has worked to predict SARS-CoV-2 by using genomic information, the other one has used machine learning to reveal pathological factors for diseases associated with airway smooth muscle inflammation on multi-omics levels.

The second part encompasses the mutations in the virus. Understanding the virus mutation is very important as the mutations lead to the various variants of the virus. The works in this part mostly deal with multiple sequence alignment (MSA) to reveal the virus mutations. What makes this section non-trivial is the fact that MSA has been performed with a huge number of SARS-CoV-2 sequences by all the contributions.

The third and final section discusses the various vaccines and therapeutics that can be used to fight against SARS-CoV-2. As discussed earlier, though there are several vaccines already approved by different medical agencies, their sustainability is not known till now. Thus, apart from the vaccine host immune system modulation can also be considered to find alternative solutions. Also, epitope-based vaccines and other therapeutics can be taken into account. Furthermore, vaccination, COVID-19 incidence, and mortality have also been explored in one work in this section. Moreover, Spike glycoprotein and ORF8 protein of SARS-CoV-2 are also analyzed to provide clinical and therapeutic implications.

### 2.1 Prediction

In Saha et al., deep learning based predictor viz. COVID-DeepPredictor has been proposed to predict unknown sequences of SARS-CoV-2 as well as other pathogens like SARS-CoV-1, MERS-CoV, Ebola, Dengue, and Influenza. COVID-DeepPredictor uses Long Short Term Memory as Recurrent Neural Network where  $k$ -mer technique is used to generate Bag-of-Unique-Descriptors. COVID-DeepPredictor achieves 100% prediction accuracy on validation datasets while on test datasets, the accuracy is as high as 99%.

Zhang et al. explore SARS-CoV-2 infection in airway smooth muscles which may play an important role in several other inflammatory diseases as well. They have used machine-learning-based computational approaches to identify specific regulatory factors that contribute to the activation and simulation of airway smooth muscles. This will lead to the identification of potential regulatory mechanisms linking airway smooth muscle tissues and inflammatory factors which will eventually help in identifying specific pathological factors for diseases associated with airway smooth muscle inflammation on multi-omics levels.

### 2.2 Mutation

In Lin et al., multiple sequence alignment using a conserved sequence search algorithm has been optimized to align 24,768 sequences from the GISAID dataset. This will help in conserved sequence searches to segment long sequences as well as make large-scale multisequence alignment possible, thereby facilitating comprehensive gene mutation analysis.

In Saha et al., multiple sequence alignment of 71,038 SARS-CoV-2 genomes from 98 countries have been performed to identify hotspot mutations in SARS-CoV-2. This has led to the identification of 45 unique hotspot mutations. Such mutations include L452R, T478K, E484Q, and N501Y.

In Biswas et al., database DbNSP InC has been reported which provides information on the NSPs of SARS-CoV-2 extracted from patients in India. It provides functional information, mutations observed in samples of Indian patients, primary and secondary structural analyses, strain and mutation analyses as well as mutations observed in the deceased, mild, and asymptomatic patients samples along with the distribution of mutations across different Indian states and phylogenetic analysis.

In Cueno et al., the authors have generated spike models of endemic HCoVs (HCoV 229E, HCoV OC43, HCoV NL63, HCoV HKU1, SARS CoV, MERS CoV), original SARS-CoV-2 and variants of concern (Alpha, Beta, Gamma, and Delta). They propose that structural similarities among the pathogens may help ascertain immune cross-reactivity while differences may result in viral infection.

### 2.3 Vaccine and Therapeutics

In Majumdar et al., the differences in COVID-19 death and infection ratio between the urban and rural population in India have been explored to discuss the role of the immune system, comorbidities, and associated nutritional status that may play role in the death rate of COVID-19 patients in such populations. Furthermore, they have also focussed on strategies for developing masks, vaccines, and other diagnostics to combat COVID-19.

In Mazzocco et al., a novel *in-silico* approach based on artificial intelligence and bioinformatics methods have been put forth to support the design of epitope-based vaccines. Their methods have also been evaluated for predicting the immunogenicity of epitopes. They have also discussed the potential applicability of such epitopes for the development of a vaccine eliciting cellular immunity for COVID-19.



In Somogyi et al., selection of immunoprevalent SARS-CoV-2-derived T cell epitopes using an *in-silico* cohort of HLA-genotyped individuals with different ethnicities has been considered. The results of this work are significant for the development of highly efficient epitope-based vaccines against various pathogens and diseases as well.

In Valcarcel et al., the focus is on analyzing structural similarities of ORF8 protein of SARS-CoV-2 with immunological molecules such as IL-1, thereby contributing to the immunological deregulation observed in COVID-19.

In Fukutani et al., the association between vaccine implementations, the occurrence of new cases, and mortality rate have been tracked. They have used CaVaCo (Cases, Vaccinations, and COVID-19) tool to retrieve the COVID-19 cases as well as the deaths and vaccination data to compare and correlate vaccination coverage of the countries with other parameters.

In Bensussen et al., a minimal mathematical model of the effect of the extra copy of TMPRSS2 on ORF8 production and persistence in the infected cells of a Down syndrome patient having COVID-19 disease has been proposed. Their results support the hypothesis that people with Down syndrome have a high susceptibility to COVID-19 due to the overproduction of TMPRSS2.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## FUNDING

This work has been partially carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship Program awarded to NG. This work has also been supported by a CRG short-term research grant on COVID-19 (CVD/2020/000991) from the Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India. This work is co-funded by Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) program. This work has also been co-supported by Polish National Science Centre (2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757).

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Saha, Ghosh and Plewczynski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# COVID-DeepPredictor: Recurrent Neural Network to Predict SARS-CoV-2 and Other Pathogenic Viruses

Indrajit Saha<sup>1\*†</sup>, Nimisha Ghosh<sup>2†</sup>, Debasree Maity<sup>3</sup>, Arijit Seal<sup>4</sup> and Dariusz Plewczynski<sup>5,6</sup>

<sup>1</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, <sup>2</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to Be University), Bhubaneswar, India, <sup>3</sup> Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, India, <sup>4</sup> Cognizant Technology Solutions Pvt. Ltd., Kolkata, India, <sup>5</sup> Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, <sup>6</sup> Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland

## OPEN ACCESS

### Edited by:

Xian-Tao Zeng,  
Wuhan University, China

### Reviewed by:

Sarath Chandra Janga,  
Indiana University, Purdue University  
Indianapolis, United States  
Xue-Qun Ren,  
Henan University, China

### \*Correspondence:

Indrajit Saha  
indrajit@nittrkol.ac.in

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

Received: 03 June 2020

Accepted: 13 January 2021

Published: 11 February 2021

### Citation:

Saha I, Ghosh N, Maity D, Seal A and  
Plewczynski D (2021)  
COVID-DeepPredictor: Recurrent  
Neural Network to Predict  
SARS-CoV-2 and Other Pathogenic  
Viruses. *Front. Genet.* 12:569120.  
doi: 10.3389/fgene.2021.569120

The COVID-19 disease for Novel coronavirus (SARS-CoV-2) has turned out to be a global pandemic. The high transmission rate of this pathogenic virus demands an early prediction and proper identification for the subsequent treatment. However, polymorphic nature of this virus allows it to adapt and sustain in different kinds of environment which makes it difficult to predict. On the other hand, there are other pathogens like SARS-CoV-1, MERS-CoV, Ebola, Dengue, and Influenza as well, so that a predictor is highly required to distinguish them with the use of their genomic information. To mitigate this problem, in this work COVID-DeepPredictor is proposed on the framework of deep learning to identify an unknown sequence of these pathogens. COVID-DeepPredictor uses Long Short Term Memory as Recurrent Neural Network for the underlying prediction with an alignment-free technique. In this regard,  $k$ -mer technique is applied to create Bag-of-Descriptors (BoDs) in order to generate Bag-of-Unique-Descriptors (BoUDs) as vocabulary and subsequently embedded representation is prepared for the given virus sequences. This predictor is not only validated for the dataset using  $K$ -fold cross-validation but also for unseen test datasets of SARS-CoV-2 sequences and sequences from other viruses as well. To verify the efficacy of COVID-DeepPredictor, it has been compared with other state-of-the-art prediction techniques based on Linear Discriminant Analysis, Random Forests, and Gradient Boosting Method. COVID-DeepPredictor achieves 100% prediction accuracy on validation dataset while on test datasets, the accuracy ranges from 99.51 to 99.94%. It shows superior results over other prediction techniques as well. In addition to this, accuracy and runtime of COVID-DeepPredictor are considered simultaneously to determine the value of  $k$  in  $k$ -mer, a comparative study among  $k$  values in  $k$ -mer, Bag-of-Descriptors (BoDs), and Bag-of-Unique-Descriptors (BoUDs) and a comparison between COVID-DeepPredictor and Nucleotide BLAST have also been performed. The code, training, and test datasets used for COVID-DeepPredictor are available at <http://www.nittrkol.ac.in/indrajit/projects/COVID-DeepPredictor/>.

**Keywords:** long-short term memory, SARS-CoV-2, sequence analysis, virus prediction, genomic information

# 1. INTRODUCTION

The first case of COVID-19 surfaced in Wuhan, China in December 2019 (Huang et al., 2020; Meng et al., 2020; Yan L. et al., 2020). In no time it spread to 212 countries and territories (Worldometer, 2021) worldwide creating a pandemic in its wake. SARS-CoV-2 falls in the same family as SARS-CoV and MERS-CoV (all belong to the family of coronavirus) and mainly targets the respiratory system (Zhou et al., 2020). As of 8th January 2021, over 885 million cases of COVID-19 have been reported worldwide, with more than 1,906 thousand cases of death and 63.6 million cases of recovery (Worldometer, 2021).

SARS-CoV-2 is defined as an enveloped, positive-sense, single-stranded RNA virus with a genome of around 30 kilobases in length (Weiss and Navas-Martin, 2005; Su et al., 2016; Cui et al., 2019). RNA viruses generally have very high mutation rates (Jenkins et al., 2002; Woo et al., 2009). Genetic mutation can occur infrequently between viruses of the same species but of divergent lineages. The resulting mutated viruses may sometimes cause an outbreak of infection in humans e.g., the case of SARS-CoV-2. Coronavirus results from zoonotic transmission to human and shows symptoms of pneumonia, fever, and breathing difficulties (Guan et al., 2003; Alagaili et al., 2014). Human to human transmission has also been confirmed for SARS-CoV-2 (Chan et al., 2020; Huang et al., 2020). Next-generation sequencing using metagenomic analysis has recently been used to identify the genetic features of SARS-CoV-2 (Zhou et al., 2020).

There have been several analysis regarding SARS-CoV-2. This include whole genome analysis of a virus and viral protein-based comparisons which have resulted in the conclusion that SARS-CoV-2 is mostly related to two bat SARS-like coronaviruses (Chan et al., 2020; Lu et al., 2020). Phylogenetic analysis of full genome alignment and similarity plot show that SARS-CoV-2 has high similarity with bat coronavirus *RaTG13* (Paraskevis et al., 2020). Furthermore, another study (Wan et al., 2020) has shown that spike protein receptor-binding domain (RBD) of SARS-CoV-2 binds with host receptor angiotensin-converting enzyme 2 (ACE2), just like other *Sarbecovirus* strains, thus making the claim that SARS-CoV-2 originated from bat very likely (Letko et al., 2020; Liu and Wang, 2020).

As the genomic structure of SARS-CoV-2 is similar to the other viruses of the same family, and it shows similar symptoms like them, the early prediction of SARS-CoV-2 is a very challenging task. Ozturk et al. (2020) have used deep neural networks with X-ray images for automated detection of SARS-CoV-2 cases. The results show that the method has a prediction accuracy of 98.08% for binary classes (COVID vs. No-Findings) and 87.02% for multiple classes (COVID vs. No-Findings vs. Pneumonia). Another work (Yan Q. et al., 2020) where deep learning has been used to predict age-related macular degeneration (AMD) which is a leading cause of blindness among the elderly population. The results show an average area under the curve (AUC) value of 0.85. On the other hand, the authors in Koohi-Moghadam et al. (2019) have used deep learning approach to predict disease-associated mutation of metal-binding sites in proteins. The prediction

results depict AUC as 0.90 and an accuracy of 0.82. These encouraging results show that deep learning has the potential for highly accurate prediction. This led us to devise a predictor based on deep learning which uses genomic sequences of pathogenic viruses. In this work, a deep learning technique, viz. COVID-DeepPredictor based on Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997; Tang et al., 2019) is developed. Though, LSTM has been profusely used in many works for text classification (Jin et al., 2019; Liu et al., 2019; Zhang et al., 2019), to the best of the authors' knowledge, this is the first attempt to use LSTM for the prediction of SARS-CoV-2 using genomic sequences of virus considering alignment-free approach. For this purpose, *k*-mer technique is used to generate Bag-of-Descriptors (BoDs) and consequently Bag-of-Unique-Descriptors (BoUDs) as vocabulary. Subsequently embedded representation is prepared for the given virus sequences using BoDs and BoUDs. It is worth mentioning that, though SARS-CoV-2 is a single-stranded RNA virus, the genomic information of a virus is captured in the form of DNA sequence. These DNA sequences are used in this work to predict SARS-CoV-2 and other pathogenic viruses viz. SARS-CoV-1, MERS-CoV, Ebola, Dengue, and Influenza. COVID-DeepPredictor achieves 100% prediction accuracy on validation dataset while on test datasets, the accuracy ranges from 99.51 to 99.94%. COVID-DeepPredictor also shows superior results over the existing prediction techniques based on Linear Discriminant Analysis, Random Forests, and Gradient Boosting Method. Moreover, apart from prediction accuracy, critical analysis like the choice of *k* in *k*-mer by considering the accuracy and runtime of COVID-DeepPredictor simultaneously, a comparative study of Bag-of-Descriptors (BoDs) and Bag-of-Unique-Descriptors (BoUDs) for different values of *k* and a comparison between an alignment-based technique viz. Nucleotide Basic Local Alignment Search Tool (BLASTN) and COVID-DeepPredictor as alignment-free technique.

# 2. MATERIALS AND METHODS

In this section, description of dataset preparation that has been used in this work are elucidated, a brief description of Long-Short Term Memory (LSTM) and the detailed discussion of proposed COVID-DeepPredictor are put forth.

## 2.1. Data Preparation

The datasets of SARS-CoV-1, MERS-CoV, Ebola, Dengue, and Influenza have been downloaded from NCBI (National Center for Biotechnology Information)<sup>1</sup>. Dataset for SARS-CoV-2 has been downloaded from NCBI and GISAID (Global Initiative on Sharing All Influenza Data)<sup>2</sup>. The total number of complete or near-complete genomic sequences of all the pathogenic viruses amounted to 4,643, named as Initial dataset. Additionally, the recent complete or near-complete SARS-CoV-2 sequences of 3,030 during January 2020 to August 2020 are taken from NCBI whereas 2,410 (from February 2020 to July 2020) and 4,000 (from June 2020 to December 2020) sequences are considered

<sup>1</sup><https://www.ncbi.nlm.nih.gov/genome/viruses>

<sup>2</sup><https://gisaid.org/CoV2020>

from GISAID. For our training purpose, 1,500 samples from 4,643 sequences are taken randomly for training. To ensure that representatives from all the six pathogenic viruses are available and to avoid imbalance class problem, 250 samples from each pathogenic viruses are taken in the training dataset. In order to perform testing, five different test datasets are created and named as Testdata-1, Testdata-2, Testdata-3, Testdata-4, and Testdata-5. It is important to mention that Testdata-1 consists of the remaining 3,143 sequences out of 4,643 sequences, while Testdata-2 contains 200 sequences each for MERS-CoV, SARS-CoV-2, Ebola, Dengue, and Influenza and 90 sequences of SARS-CoV-1 from different sources. Moreover, Testdata-3, Testdata-4, and Testdata-5 comprise of recent SARS-CoV-2 sequences from NCBI and GISAID respectively along with other pathogenic viruses. The statistics of Initial dataset as well as training and testing datasets are given in **Table 1**. It is worth mentioning that in this work more than 10,000 SARS-CoV-2 genomic sequences have been used from January 2020 to December 2020 considering different sources in order to develop COVID-DeepPredictor.

All the experiments are performed with the training and testing datasets as mentioned in **Table 1**. For the visualization of all the virus sequences (SARS-CoV-1, MERS-CoV, SARS-CoV-2, Ebola, Dengue, and Influenza), t-distributed Stochastic Neighbor Embedding (tSNE) (Hinton and Roweis, 2003) is applied on 4,643 sequences after generating the count vector (Khattak et al., 2019) using *k*-mer technique (Manekar and Sathe, 2018; Solis-Reyes et al., 2018). In this regard, the number of clusters known apriori is six and such embedded representation of virus sequences is shown in **Figure 1A** along with the distribution of initial SARS-CoV-2 sequences in 56 countries in **Figure 1B**. It is to be noted that COVID-DeepPredictor is developed in MATLAB R2020a.

## 2.2. Long-Short Term Memory

Long-Short Term Memory (LSTM) is a type of recurrent neural network (sub-branch of deep learning) which is capable of learning order dependence in sequence prediction problems. The main components of an LSTM network are sequence input layer and an LSTM layer. A sequence input layer provides text as an input into the LSTM network. An LSTM layer learns long-term association between steps of sequence data. Elaborately speaking, an LSTM network acquires a context vector from previous time step and an input vector from the given data. This is used to calculate the next context and gate vectors to control memory cell state vector (Kim et al., 2018). With an input data at time  $t$  and a context vector  $h$ , a raw cell vector and input vectors for each gate are created by one hidden layer. At the input gate, the cell vector is then multiplied by the input vector. The cell input is added to given previous cell vector weighted by the forget vector. Then the resultant vector is controlled by the output vector. The update of the cell is controlled by the control gate. LSTM is mainly trained using Back-propagation Through Time and mitigates the vanishing gradient problem that is quite rampant in RNN. In LSTM, the memory cells and the gates can store time and thus can eliminate old observations overcoming vanishing gradient problem.

To sum up, LSTM consists of four gates, input gate ( $i_t$ ), forget gate ( $f_t$ ), control gate ( $C_t$ ), and output gate ( $o_t$ ). Considering a sentence  $S = x_1, x_2, \dots, x_K$ , where  $K$  is the length of a sentence, the equations for LSTM can be depicted as:

$$i_t = \text{sigm}(W_i \times [h_{t-1}, x_t] + b_i) \quad (1)$$

$$f_t = \text{sigm}(W_f \times [h_{t-1}, x_t] + b_f) \quad (2)$$

$$\begin{aligned} \tilde{C}_t &= \text{tanh}(W_c \times [h_{t-1}, x_t] + b_c) \\ C_t &= f_t \times C_{t-1} + i_t \times \tilde{C}_t \end{aligned} \quad (3)$$

$$\begin{aligned} o_t &= \text{sigm}(W_o \times [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \text{tanh}(C_t) \end{aligned} \quad (4)$$

Here,  $W$  are weight matrices,  $h_{t-1}$  is the hidden layer which is used updated by the output layer and is also responsible for updating the output and *tanh* and *sigm*, respectively represent the tanh-activation and sigmoid-activation functions.

## 2.3. COVID-DeepPredictor

The main objective of COVID-DeepPredictor is to correctly predict the virus classes based on the given genomic sequences of the different pathogenic viruses using an alignment-free technique. To achieve this, the entire genomic sequence is initially divided into descriptors of sequences called as Bag-of-Descriptors (BoDs) using the popular *k*-mer technique. Here, descriptors are patterns of the genomic sequences of length  $k$ . Thereafter, Bag-of-Unique-Descriptors (BoUDs) as vocabulary are created using such BoDs. With the use of BoDs and BoUDs, an embedded representation is created of size  $N \times M$  where  $N$  is the number of genomic sequences and  $M$  is the indices of the descriptors in vocabulary. This embedded representation is then used to train COVID-DeepPredictor. Since we have divided the genomic sequences into descriptors and represented in the form of tokens, they behave like texts, thus boiling down to a text classification problem. The pipeline of the proposed COVID-DeepPredictor is shown in **Figure 2**.

## 3. RESULTS

To validate COVID-DeepPredictor, experiments are conducted on genomic sequences of different pathogenic viruses. In this regard, MATLAB R2020a is used on an Intel Core i5-8250U CPU @ 1.80 GHz machine with 8 GB RAM and Windows 10 operating system. The parameters of the underlying predictor, LSTM of COVID-DeepPredictor have been set experimentally. In this regard, the number of hidden units for LSTM layer is set as 80. Next, to use the LSTM layer for a sequence-to-label prediction problem, the output mode is set to "last." Finally, a fully connected layer with the same size as the number of classes, a softmax layer and a prediction layer are added as well. Mini-batch gradient descent is used to train LSTM. The mini-batch size is specified as 16 and the gradient threshold is set to 2. The COVID-DeepPredictor is compared

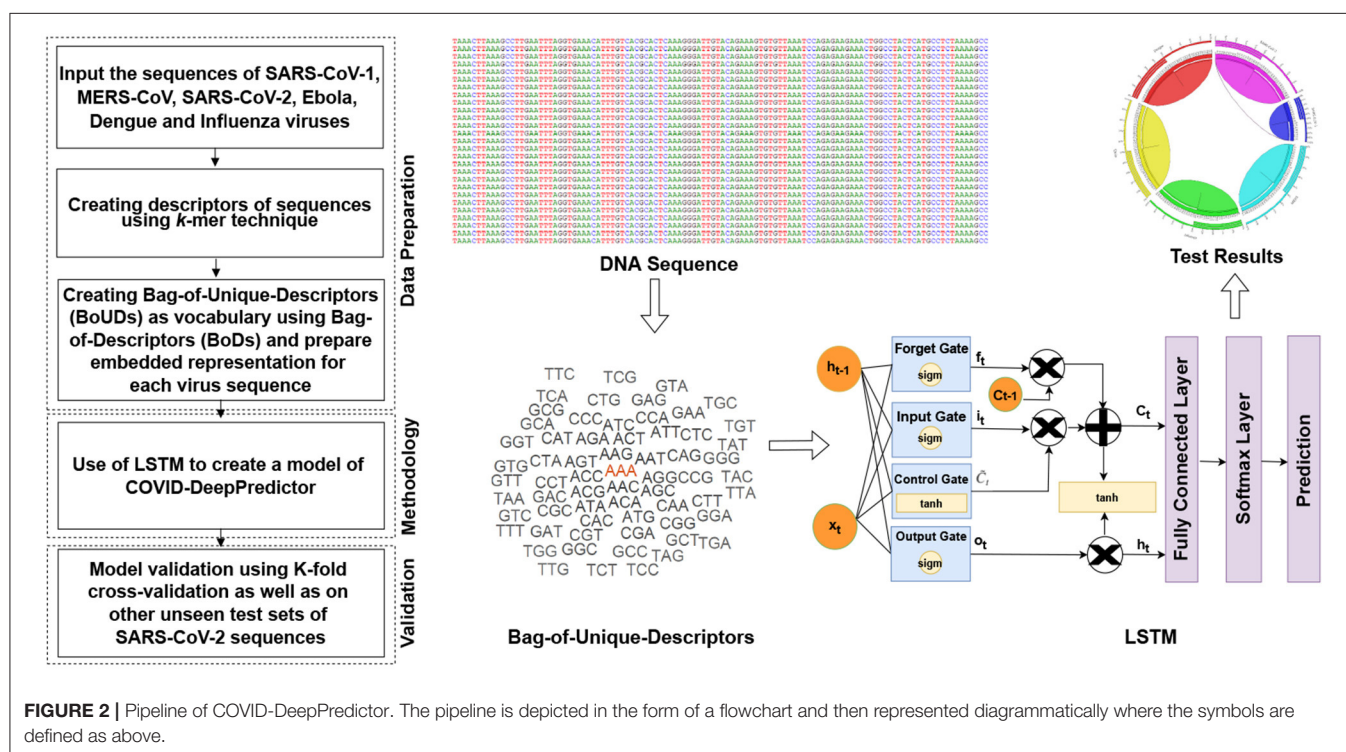
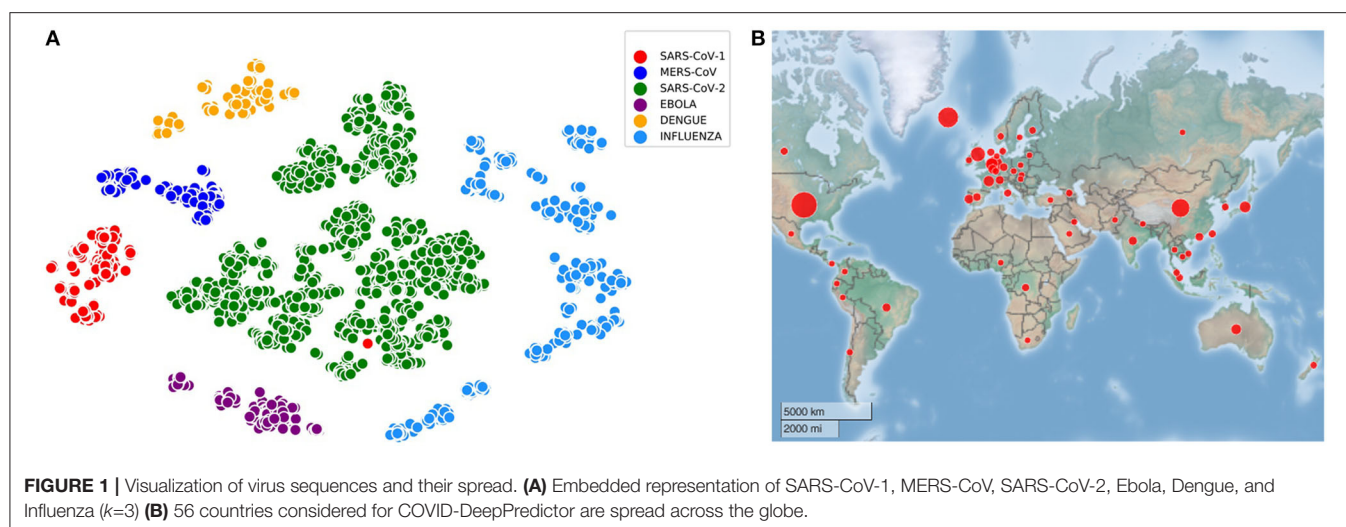
**TABLE 1** | Description of initial, training, and test datasets.

Dataset	Virus name	Number of sequences	Max. length of sequence	Avg. length of sequence	Source of sequence
Initial dataset	SARS-CoV-1	340	30,311	29,515	NCBI-SARS-CoV-1
	MERS-CoV	291	30,150	29,983	NCBI-MERS-CoV
	SARS-CoV-2	2,402	29,986	29,507	GISAID-SARS-CoV-2
	Ebola	300	19,897	18,976	NCBI-Ebola
	Dengue	300	11,195	10,746	NCBI-Dengue
	Influenza	1,010	2,347	2,322	NCBI-Influenza
Training dataset	SARS-CoV-1	250	29,765	29,520	NCBI-SARS-CoV-1
	MERS-CoV	250	30,123	29,999	NCBI-MERS-CoV
	SARS-CoV-2	250	29,927	29,334	GISAID-SARS-CoV-2
	Ebola	250	19,897	18,979	NCBI-Ebola
	Dengue	250	11,195	10,748	NCBI-Dengue
	Influenza	250	2,347	2,333	NCBI-Influenza
Testdata-1	SARS-CoV-1	90	30,311	29,494	NCBI-SARS-CoV-1
	MERS-CoV	41	30,150	29,887	NCBI-MERS-CoV
	SARS-CoV-2	2,152	29,986	29,527	GISAID-SARS-CoV-2
	Ebola	50	19,034	18,964	NCBI-Ebola
	Dengue	50	10,764	10,737	NCBI-Dengue
	Influenza	760	2,341	2,318	NCBI-Influenza
Testdata-2	SARS-CoV-1	90	30,311	29,494	NCBI-SARS-CoV-1
	MERS-CoV	200	30,423	29,066	NCBI-MERS-CoV
	SARS-CoV-2	200	29,855	29,850	GISAID-SARS-CoV-2
	Ebola	200	18,798	18,762	NCBI-Ebola
	Dengue	200	10,731	10,692	NCBI-Dengue
	Influenza	200	2,341	2,323	NCBI-Influenza
Testdata-3	SARS-CoV-1	90	30,311	29,494	NCBI-SARS-CoV-1
	MERS-CoV	220	30,423	29,162	NCBI-MERS-CoV
	SARS-CoV-2	3,030	29,903	29,780	NCBI-SARS-CoV-2
	Ebola	220	18,871	18,850	NCBI-Ebola
	Dengue	220	10,690	10,677	NCBI-Dengue
	Influenza	220	2,341	2,323	NCBI-Influenza
Testdata-4	SARS-CoV-1	90	30,311	29,494	NCBI-SARS-CoV-1
	MERS-CoV	250	30,423	29,277	NCBI-MERS-CoV
	SARS-CoV-2	2,410	30,423	29,726	GISAID-SARS-CoV-2
	Ebola	250	18,871	18,852	NCBI-Ebola
	Dengue	250	10,757	10,538	NCBI-Dengue
	Influenza	250	2,316	2,316	NCBI-Influenza
Testdata-5	SARS-CoV-1	90	30,311	29,494	NCBI-SARS-CoV-1
	MERS-CoV	250	30,423	29,277	NCBI-MERS-CoV
	SARS-CoV-2	4,000	29,903	29,798	GISAID-SARS-CoV-2
	Ebola	200	18,798	18,762	NCBI-Ebola
	Dengue	220	10,690	10,677	NCBI-Dengue
	Influenza	250	2,316	2,316	NCBI-Influenza

with other predictors based on Linear Discriminant Analysis (LDA), Random Forests (RFs), and Gradient Boosting Method (GBM). For LDA, the discriminant type is considered to be pseudo-linear, for Random Forests, the number of trees taken are 50 and for GBM the maximum depth of the tree is 10 and maximum iterations are taken as 100. All these parameters are set experimentally.

Each predictor has been evaluated using  $\mathcal{K}$ -fold cross-validation ( $\mathcal{K} = 10$ ) technique followed by further validation on unseen test datasets. The cross-validation partition uses random non-stratified sampling method which is applied to prepare the training and validation datasets resulting in a total of 1,500 samples. The training and validation datasets consist of all the pathogenic virus classes; SARS-CoV-1, MERS-CoV,





SARS-CoV-2, Ebola, Dengue, and Influenza. For each predictor, the descriptors of the sequences of the viruses are created using  $k$ -mer method. Thereafter to train the COVID-DeepPredictor and the other compared predictors, an embedded matrix of size  $N \times M$  is created with the use of BoDs and BoUDs.

To determine the performance of COVID-DeepPredictor and the other predictors, *Confusion Matrix* (Luque et al., 2019) is considered. In confusion matrix, **True Positives** (TP) refer to a data being correctly identified and they are represented

by the diagonal elements. The remaining predictions lead to an error  $\epsilon$ . Moreover, **False Positives** (FP) for a particular class refer to the sum of the values in the corresponding column, excluding the TP and **False Negatives** (FN) for a class is the sum of the values in the corresponding row, excluding the TP. Lastly, **True Negatives** (TN) for a class is the sum of all columns and row, barring the one for itself. To evaluate the results of COVID-DeepPredictor, metrics like *Accuracy*, *Precision*, *Recall*, and *G-Mean* have been considered

which can be deduced from a confusion matrix. They can be calculated as:

**Accuracy:**

$$\frac{TP + TN}{TN + FP + FN + TP} \quad (5)$$

**Precision:**

$$\frac{TP}{FP + TP} \quad (6)$$

**Recall:**

$$\frac{TP}{TP + FN} \quad (7)$$

**G-mean:**

$$\frac{TP}{\sqrt{(TP + FP)(TP + FN)}} \quad (8)$$

Different existing state-of-the-art predictors based on Linear Discriminant Analysis (LDA), Random Forests (RFs), and Gradient Boosting Method (GBM) are used in this work for comparison purposes. LDA is a very popular machine learning tool for prediction. In LDA, each dependent variable is expressed as a linear combination of other features. RFs are ensemble learning methods which build numerous decision trees during training and as an output produces the class that is the mode of the classes. GBM is also an ensemble learning model which produces a prediction model in the form of an ensemble weak prediction models, usually decision trees.

For conducting the experiments, first and foremost, we need to determine the value of  $k$  in  $k$ -mer. In order to do this, the experiments have been conducted on five test datasets as mentioned in section 2. The results are shown in **Figures 3A–E**, where  $k$  is varied from 3 to 15 with accuracy and running time of COVID-DeepPredictor. It can be seen from figures that the accuracy is higher at  $k = 3$  for all the five test datasets. Although, the same accuracy can be found for other  $k$  values as well, e.g., in **Figure 3A**  $k = 9, 11$ , and  $13$  show the same accuracy, as we increase the  $k$ -mer value, the run time increases. This trend of increasing time with the increasing value of  $k$ -mer has also been shown in Solis-Reyes et al. (2018). Keeping this in context, we have taken the value of  $k$  in  $k$ -mer to be 3 as with this value, the run time is least. For the compared predictors based on LDA, RF, and GBM, the  $k$  values are similarly determined as 13, 4, and 4, respectively. In this work,  $\mathcal{K}$ -fold cross-validation with  $\mathcal{K} = 10$  is used. The average results in terms of accuracy for the test datasets are shown in **Figure 4A**. Moreover, apart from accuracy, different metrics such as precision, recall and g-mean have also been computed for the test datasets and reported in **Table 2**. As can be seen from the results of **Figure 4A**, for COVID-DeepPredictor the accuracy ranges from 99.51 to 99.94%. Thus, the experiments establish the fact that COVID-DeepPredictor can detect SARS-CoV-2 with a very high accuracy. The confusion matrices as circo plots for Testdata-1 and Testdata-2 ( $k = 3$ ) are shown in **Figures 4B,C**. It can be seen from **Figures 4B,C** that there is only one misprediction, where SARS-CoV-1 has been wrongly predicted as SARS-CoV-2. The confusion matrices

for Testdata-3, Testdata-4, and Testdata-5 ( $k = 3$ ) are shown in **Supplementary Figure 2**.

COVID-DeepPredictor is performed on a validation dataset as well. Accuracy, precision, recall, and G-mean values of the prediction for the validation dataset are 100, 100, 100, and 1%, respectively ( $k=3$ ). As we have used  $\mathcal{K}$ -fold cross-validation with  $\mathcal{K} = 10$ , ten convergence plots of COVID-DeepPredictor are generated. One of the corresponding convergence plots for COVID-DeepPredictor is given in **Figure 4D**. The blue line indicates the training accuracy and the black line is the validation accuracy. All the convergence plots are shown in **Supplementary Figure 1**. The Bag-of-Unique-Descriptors of the six virus classes, SARS-CoV-1, MERS-CoV, SARS-CoV-2, Ebola, Dengue, and Influenza are shown in **Figures 4E–J** for  $k=3$ .

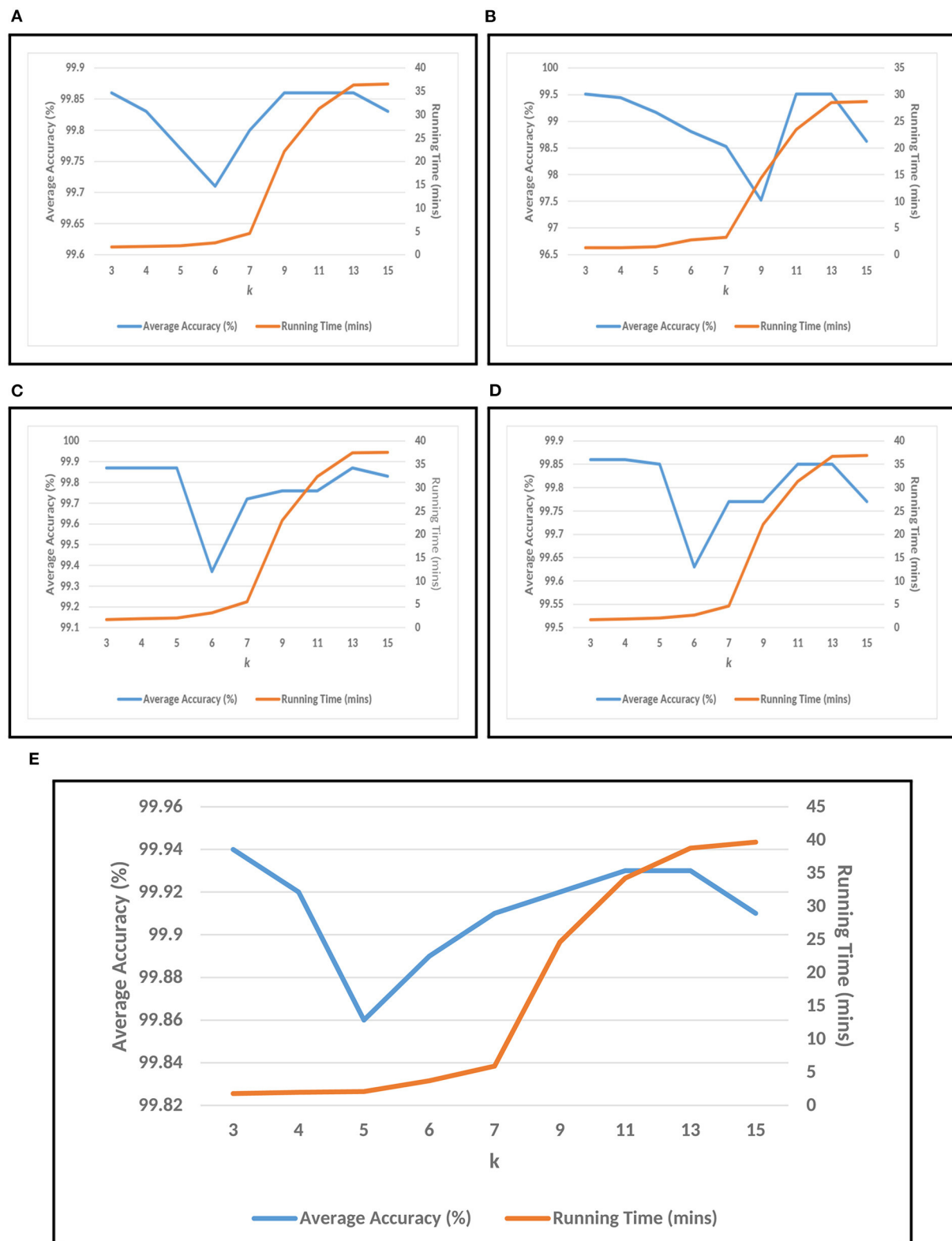
## 4. DISCUSSION

SARS-CoV-2 is a global pandemic and since human to human transmission (Chan et al., 2020; Huang et al., 2020) is confirmed for SARS-CoV-2, the need for its early prediction has become imperative. Viral outbreaks of this kind call for timely and prompt analysis of the genomic sequences to help the prediction of the virus in its early stages. COVID-DeepPredictor can be used by pathogen laboratories for the prediction of SARS-CoV-2 very quickly and as concluded from the results, most accurately. It is worth mentioning over here that for COVID-DeepPredictor to be effective, there must be at least two virus classes present in the training input sequences.

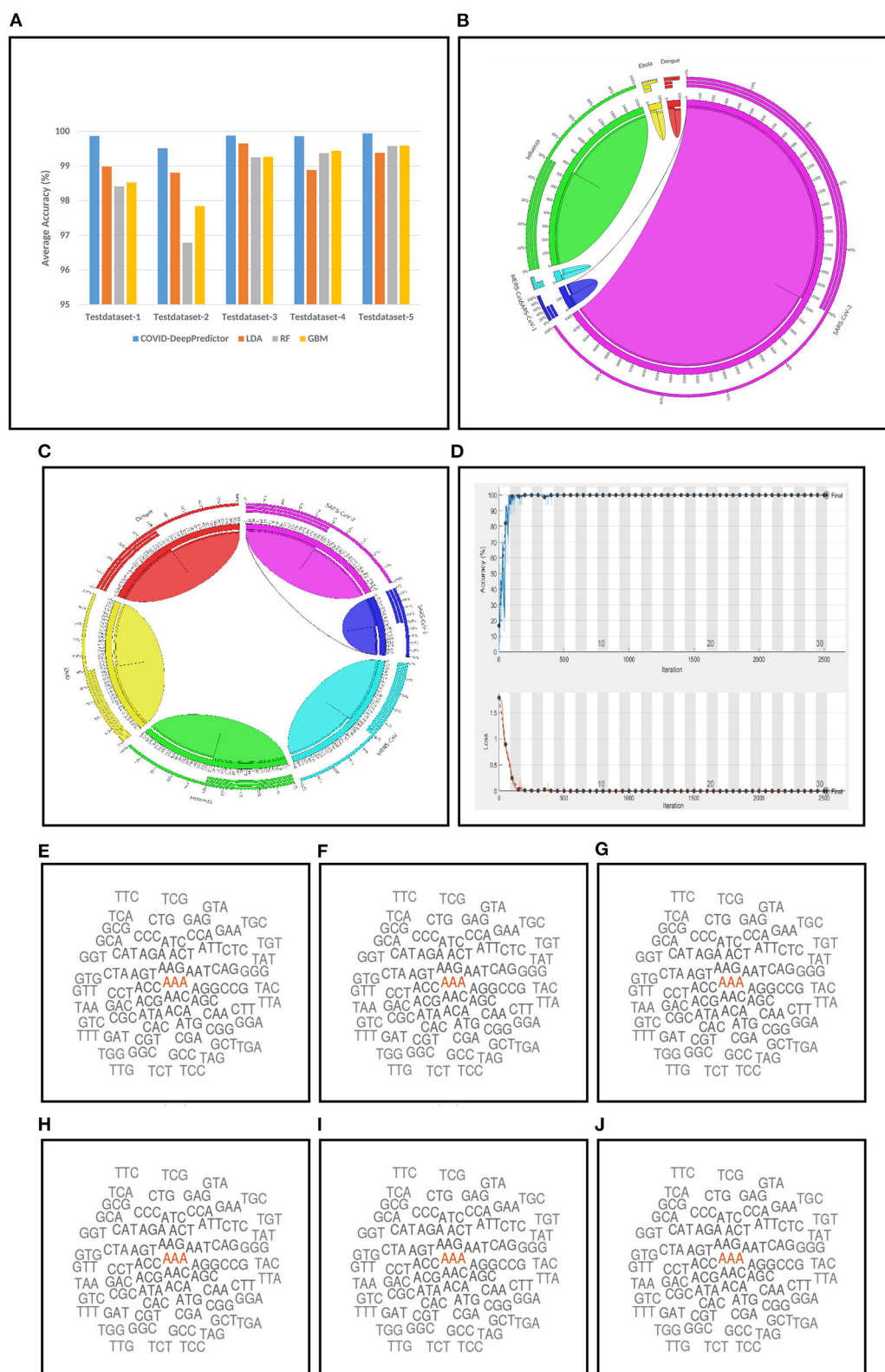
COVID-DeepPredictor has two functions for: (a) training, testing, and accordingly saving an LSTM model [COVIDdeepPredictor()] and (b) loading a pre-trained LSTM model for testing on unseen test dataset [COVIDdeepPredictorLoad()]. There is a main code COVIDmain.m which loads both COVIDdeepPredictor() and COVIDdeepPredictorLoad(). If users want to have their own training model and also get the results for a test dataset, they need to use only COVIDdeepPredictor() and disable COVIDdeepPredictorLoad(). On the other hand, if they want to use a pre-trained model, they can disable COVIDdeepPredictor() and run only COVIDdeepPredictorLoad() to get the results for test datasets.

For ease of users, training and testing files are provided to make them acquainted with the functionalities of COVIDdeepPredictor(). Trainingdata.csv is the input file for training and any one of the test files among Testdata-1.csv, Testdata-2.csv, Testdata-3.csv, Testdata-4.csv, and Testdata-5.csv can be used for testing. The results of the prediction will have the sequence ID, predicted virus name, along with its sequence which will be stored in Results.csv.

On the other hand, in case of COVIDdeepPredictorLoad(), only any one of the test files needs to be provided to get the results in Results.csv. Similarly, new training and test datasets can be prepared by the users after following the same structures of the training and testing files as provided. This is important so that new training models of COVID-DeepPredictor can be prepared for different set of viruses or similar kind of



**FIGURE 3 |** Choosing  $k$  value of  $k$ -mer for COVID-DeepPredictor based on accuracy and running time. **(A)** Testdata-1, **(B)** Testdata-2, **(C)** Testdata-3, **(D)** Testdata-4, **(E)** Testdata-5.



**FIGURE 4 |** Results related to COVID-DeepPredictor. **(A)** Prediction performance of COVID-DeepPredictor and other compared methods in terms of average accuracy for the five test datasets. Circos plots of confusion matrix for COVID-DeepPredictor ( $k=3$ ) for **(B)** Testdata-1 **(C)** Testdata-2. **(D)** Convergence plot of COVID-DeepPredictor. Word cloud of  $k$ -mer descriptors ( $k=3$ ) of genome sequences for **(E)** SARS-CoV-1 **(F)** MERS-CoV **(G)** SARS-CoV-2 **(H)** Ebola **(I)** Dengue **(J)** Influenza.



**TABLE 2 |** Prediction performance of COVID-DeepPredictor and other compared methods on test datasets.

Method	DataSet	k-mer	Average accuracy	Average precision	Average Recall	Average G-Mean
COVID-DeepPredictor	Testdata-1	3	<b>99.867</b>	<b>99.914</b>	<b>99.336</b>	<b>0.996</b>
LDA		13	98.981	91.845	98.015	0.948
RF		4	98.409	97.577	90.024	0.937
GBM		4	98.524	97.611	90.121	0.937
COVID-DeepPredictor	Testdata-2	3	<b>99.513</b>	<b>99.527</b>	<b>99.423</b>	<b>0.994</b>
LDA		13	98.807	98.814	98.925	0.988
RF		4	96.788	96.981	97.264	0.971
GBM		4	97.844	97.542	97.991	0.977
COVID-DeepPredictor	Testdata-3	3	<b>99.877</b>	<b>99.595</b>	<b>99.686</b>	<b>0.996</b>
LDA		13	99.650	98.981	99.162	0.989
RF		4	99.250	97.727	98.440	0.981
GBM		4	99.265	97.728	98.891	0.983
COVID-DeepPredictor	Testdata-4	3	<b>99.860</b>	<b>99.637</b>	<b>99.682</b>	<b>0.996</b>
LDA		13	98.885	97.281	97.648	0.974
RF		4	99.371	98.414	99.325	0.988
GBM		4	99.441	98.922	99.444	0.991
COVID-DeepPredictor	Testdata-5	3	<b>99.940</b>	<b>99.766</b>	<b>99.808</b>	<b>0.997</b>
LDA		13	99.380	97.467	97.927	0.976
RF		4	99.580	98.519	99.371	0.989
GBM		4	99.590	98.956	99.763	0.993

The results highlighted in bold show that COVID-DeepPredictor has superior performance as compared to the other predictors.

tasks. It is to be noted that the pre-trained model is provided in **Supplementary Material**, where the value of  $k$  for  $k$ -mer is 3. The choice of  $k$  has been done experimentally as it takes computationally less amount of time and provides higher accuracy. Sample files for training, testing, pre-trained models for COVID-DeepPredictor and the code of the software are available in **Supplementary Material** for re-usability<sup>3</sup>.

Setting the appropriate value of  $k$  in  $k$ -mer is very important to achieve the desired results in a text classification problem. As this work is based on the underlying concept of text classification,  $k$ -mer has a very important role to play. Thus, to determine the value of  $k$  in  $k$ -mer, extensive experiments have been performed. It can be observed from **Figures 3A–E** that with the increasing value of  $k$ , the run time of COVID-DeepPredictor is also on the rise. Therefore, to choose the appropriate value of  $k$ , apart from the accuracy, the run time of COVID-DeepPredictor also needs to be taken into account. For Testdata-1, at  $k = 9, 11$ , and  $13$ , the accuracy is same as at  $k = 3$ . Similarly, for Testdata-2, Testdata-3, Testdata-4, and Testdata-5, similar accuracies can be observed at  $k = 3, 11, 13, k = 3, 4, 5, 13, k = 3, 4$ , and  $k = 3, 13$ , respectively. Although, the accuracies are same at these  $k$ -mer values, run time is increasing as can be seen from **Figures 3A–E**. Thus, the smallest  $k$ -mer value has been chosen without compromising on the accuracy. From **Table 2** and **Figure 4A**, it is quite evident that with  $k = 3$ , COVID-DeepPredictor shows the best results among all the compared predictors.

To understand the relation among  $k$ -mer, size of BoDs and BoUDs, **Table 3** is reported. From this table, we can see that the sizes of both BoDs and BoUDs increase with the increase in  $k$ -mer for each virus class. In the table, “All” represents all the six virus classes taken together. For example, at  $k = 15$  for training dataset of all virus classes, the sizes of BoDs and BoUDs are 30193594 and 518372, respectively for 1,500 sequences while for the same  $k$ , for Testdata-1, the sizes of BoDs and BoUDs are 70595908 and 581774 respectively for 3,143 sequences. On the other hand, for  $k = 3$ , less number of BoDs and BoUDs are generated. Here, as expected, the BoD values for “All” are the summation of the BoDs of the individual virus classes. On the contrary, BoUD is less than the summation of the BoUDs of the six virus classes. This can be attributed to the relatedness between different virus classes. For example, SARS-CoV-1, MERS-CoV, and SARS-CoV-2 are more related and thus they may share unique descriptors (BoUDs) resulting in the intersection of the BoUDs when all the virus classes are considered together. Apart from this, BoDs and BoUDs for the varying  $k$  have also an impact on the accuracy and run time of COVID-DeepPredictor as well which can be observed by combining **Figure 3** and **Table 3**.

The main advantage of COVID-DeepPredictor is that it uses  $k$ -mer technique which is an alignment-free technique. Most analysis based works attempted so far have used alignment based techniques. Although, they are highly successful in detecting similarities in sequences of viruses, they take a lot of computational time. Also, alignment based techniques have the underlying constraint of homologous sequences which may not be the case every time. To mitigate these problems of alignment

<sup>3</sup><http://www.nitttrkol.ac.in/indrajit/projects/COVID-DeepPredictor/>



**TABLE 3 |** Bag-of-Descriptors and Bag-of-Unique-Descriptors for each virus class.

<i>k</i> -mer	Virus Name	Training dataset		Testdata-1		Testdata-2		Testdata-3		Testdata-4		Testdata-5	
		BoD	BoUD	BoD	BoUD	BoD	BoUD	BoD	BoUD	BoD	BoUD	BoD	BoUD
3	SARS-CoV-1	16000	64	5760	64	5760	64	5760	64	5760	64	5760	64
	MERS	16000	64	2642	81	12831	90	14083	67	16003	67	16003	67
	SARS-CoV-2	16000	64	138336	181	12800	64	193920	64	154240	64	256000	64
	Ebola	16000	64	3200	64	14248	125	14741	125	16661	125	14248	125
	Dengue	16000	64	3212	75	12827	82	14080	64	16496	138	14080	64
	Influenza	16000	64	48688	90	12803	67	14080	64	16000	64	16000	64
	All	96000	64	201838	181	71269	125	256664	125	225160	141	322091	125
5	SARS-CoV-1	255723	1024	92053	1024	92053	1024	92053	1024	92053	1024	92053	1024
	MERS	256000	1024	42012	1054	204674	1081	225101	1029	255821	1029	255821	1029
	SARS-CoV-2	255578	1024	2202055	1446	204592	1023	3099528	1024	2465318	1024	4091752	1024
	Ebola	255966	1024	51195	1024	208766	2461	227294	2104	257985	2104	208766	2461
	Dengue	253210	1024	50659	1044	202616	1054	222741	1024	253923	1493	222741	1024
	Influenza	200176	1022	608293	1093	159407	1020	175272	1015	201513	1007	201513	1007
	All	1476653	1024	3046267	1548	1072108	2555	4041989	2106	3526613	2293	5072646	2452
7	SARS-CoV-1	2804578	15151	1008955	15813	1008955	15813	1008955	15813	1008955	15813	1008955	15813
	MERS	2928952	12897	479752	12526	2293586	15184	2528113	15111	2879852	15113	2879852	15113
	SARS-CoV-2	2649879	12330	22899216	15728	2137492	11100	32349863	14073	25724590	12971	42685988	14211
	Ebola	2443931	13407	490077	14109	1947490	18116	2143135	17557	2435668	17562	1947490	18116
	Dengue	1681474	15764	337983	13206	1332951	15733	1454478	14773	1650576	16470	1454478	14773
	Influenza	513627	10642	1545260	9627	407434	8175	447771	8253	510118	6824	510118	6824
	All	13022441	16365	26761243	17235	9127908	20509	39932315	18815	34209759	19521	50486881	20334
9	SARS-CoV-1	6628103	74045	2384098	99891	2384098	99891	2384098	99891	2384098	99891	2384098	99891
	MERS	6789715	36574	1109206	32462	5266196	68377	5811335	68421	6628997	68503	6628997	68503
	SARS-CoV-2	6477353	39782	56109728	87633	5264550	29600	79603531	62655	63327698	47111	105111057	65922
	Ebola	4441121	38632	888076	42449	3510149	52268	3873871	69072	4403894	69127	3510149	52268
	Dengue	2552607	85437	510925	39245	2032038	84400	2230265	59231	2503617	83849	2230265	59231
	Influenza	576353	25781	1736059	20908	458593	15572	504138	15921	571045	11662	571045	11662
	All	27465252	170456	62738092	176102	18915624	190230	94407238	191127	79819349	194988	120435611	188263
11	SARS-CoV-1	7307627	107764	2628669	164654	2628669	164654	2628669	164654	2628669	164654	2628669	164654
	MERS	7433338	43970	1214507	37565	5761632	93236	6358646	93410	7254330	93587	7254330	93587
	SARS-CoV-2	7255552	50534	62870692	146218	5905735	34664	89280255	94001	71036334	64429	117924347	100857
	Ebola	4708196	47084	940996	50512	3714237	64927	4101614	91849	4663098	91945	3714237	64927
	Dengue	2670007	136386	534074	51172	2126694	135576	508411	19304	2619237	132259	508411	19304
	Influenza	580256	33741	1752556	26635	462340	18759	2334852	85407	576053	13648	576053	13648
	All	29954976	385098	69941518	425910	20599307	465475	105212447	491662	88777721	504060	132606047	448483
13	SARS-CoV-1	7368667	122008	2650637	191450	2650637	191450	2650614	191438	2650614	191438	2650614	191438
	MERS	7491153	47114	1223927	39330	5806329	101342	6408044	101572	7310682	101788	7310682	101788
	SARS-CoV-2	7320339	54634	63432818	171269	5959215	36016	90082938	109721	71677120	72455	118991161	117831
	Ebola	4733413	51117	946000	53465	3732476	70736	4122922	100238	4687630	100355	3732476	70736
	Dengue	2679746	163142	535989	57918	2134755	162571	2344694	99695	2629120	157280	2344694	99695
	Influenza	580108	39489	1752528	30935	462251	21020	508334	21713	15101	575971	15101	575971
	All	30173426	466701	70541899	523579	20745663	569846	106117546	607703	88970267	622408	135044728	578236
15	SARS-CoV-1	7374678	133005	2652762	211394	2652762	211394	2652739	211383	2652739	211383	2652739	211383
	MERS	7495710	49814	1224682	40764	5809898	106916	6412005	107189	7315184	107441	7315184	107441
	SARS-CoV-2	7326267	57890	63484669	189982	5964143	37021	90156005	123153	71735232	79346	119088222	132184
	Ebola	4737182	54589	946752	55755	3735616	75678	4126489	106635	4691704	106770	3735616	75678
	Dengue	2680123	185342	536022	63800	2135271	184450	2345444	111850	2629444	177592	2345444	111850
	Influenza	579634	44695	1751021	34958	461851	23061	507894	23904	575480	16471	575480	16471
	All	30193594	518372	70595908	581774	20759541	630327	106200576	673741	89599783	689040	135712685	642812

**TABLE 4 |** Runtime comparison of COVID-DeepPredictor and BLASTN.

Number of sequences of SARS-CoV-2	Alignment-free technique	Alignment-based technique
	COVID-DeepPredictor ( $k=3$ ) [Training and Testing (in min)]	BLASTN
50	1.26	1 h 15 min
100	1.27	2 h 40 min
200	1.28	4 h 30 min
300	1.29	6 h 35 min
400	1.31	9 h 10 min

based techniques, alignment-free techniques (Kari et al., 2015) can be used. Alignment-free techniques are meant to be fast and can work with a large number of sequences. To prove the advantage of COVID-DeepPredictor over BLASTN<sup>4</sup>, which is an alignment-based technique, **Table 4** is reported where different input sequences of size 50, 100, 200, 300, and 400 of SARS-CoV-2 are taken. For 50 sequences, BLASTN takes 1 h 15 min to align the sequences and to produce the subsequent results. Thereafter, such results are further required to be analyzed by machine intelligence technique to predict the virus class which takes some additional time as well. On the contrary, COVID-DeepPredictor successfully completes the job of training and testing, which involves prediction, in just 1.26 min. Similar results are also seen for the other varying sequences as well. Thus, we can conclude that an alignment-free technique is significantly faster than an alignment based technique.

## 5. CONCLUSION

In the current scenario of global pandemic, it has become very important to predict SARS-CoV-2 as early as possible as both the affected and the number of death cases are increasing exponentially everyday. However, polymorphic nature of SARS-CoV-2 allows it to adapt and sustain in different kinds of environment which makes SARS-CoV-2 very hard to predict. In such scenarios, the proposed COVID-DeepPredictor can be very useful for predicting SARS-CoV-2 and other kinds of pathogenic viruses based on their genomic information very quickly as it uses an alignment-free technique. The results for COVID-DeepPredictor are highly encouraging as it shows prediction accuracy in the range of 99.51 to 99.94% for test datasets. Human health being the main concern of this work, the code for COVID-DeepPredictor along with the pre-trained model are also provided so that the scientific community can reap as much benefit as possible from it. Apart from SARS-CoV-2, COVID-DeepPredictor can also be used by pathogen laboratories to recognize the other five pathogenic viruses (SARS-CoV-1, MERS-CoV, Ebola, Dengue, and Influenza) very easily and accurately

from a given genomic sequence. To achieve good performance, data preprocessing and the experiments are carried out on real-life datasets. Moreover, comparisons with popular existing prediction methods based on Linear Discriminant Analysis, Random Forests, and Gradient Boosting Method are also performed to show the superiority of COVID-DeepPredictor. Additionally, accuracy and runtime of COVID-DeepPredictor are taken together to determine the value of  $k$  in  $k$ -mer, comparison among  $k$  values in  $k$ -mer, Bag-of-Descriptors (BoDs) and Bag-of-Unique-Descriptors (BoUDs) is considered along with a comparative study between COVID-DeepPredictor and Nucleotide BLAST.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

IS designed the research. IS, NG, DM, AS, and DP analyzed data and wrote the manuscript. NG performed the experiments and collected results. All authors reviewed and approved the final version of the manuscript.

## 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. In addition to this, this work has been supported by Polish National Science Centre (2019/35/O/ST6/02484), Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (TEAM to DP). This research was co-funded by IDUB against COVID-19 project granted by Warsaw University of Technology under the program Excellence Initiative: Research University (IDUB).

## ACKNOWLEDGMENTS

We thank all those who have contributed sequences to GISAID and NCBI databases. We are also thankful to the reviewers for providing valuable comments to improve the paper.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.569120/full#supplementary-material>

<sup>4</sup>[https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\\_TYPE=BlastSearch&BLAST\\_SPEC=&LINK\\_LOC=blasttab&LAST\\_PAGE=blastn](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&BLAST_SPEC=&LINK_LOC=blasttab&LAST_PAGE=blastn)

## REFERENCES

- Alagaili, A. N., Briese, T., Mishra, N., Kapoor, V., Sameroff, S. C., de Wit, E., et al. (2014). Middle east respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. *MBio* 5:e00884-14. doi: 10.1128/mBio.01002-14
- Chan, J. F.-W., Yuan, S., Kok, K.-H., Kai-Wang, K., Chu, H., Yang, J., et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 395, 514–523. doi: 10.1016/S0140-6736(20)30154-9
- Cui, J., Li, F., and Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi: 10.1038/s41579-018-0118-9
- Guan, Y., Zheng, B., He, Y., Liu, X., Zhuang, Z., Cheung, C., et al. (2003). Isolation and characterization of viruses related to the sars coronavirus from animals in southern China. *Science* 302, 276–278. doi: 10.1126/science.1087139
- Hinton, G. E., and Roweis, S. T. (2003). “Stochastic neighbor embedding,” in *Advances in Neural Information Processing Systems* (Vancouver, BC), 857–864.
- Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735
- Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506. doi: 10.1016/S0140-6736(20)30183-5
- Jenkins, G. M., Rambaut, A., Pybus, O. G., and Holmes, E. C. (2002). Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.* 54, 156–165. doi: 10.1007/s00239-001-0064-3
- Jin, Y., Luo, C., Guo, W., Xie, J., Wu, D., and Wang, R. (2019). Text classification based on conditional reflection. *IEEE Access* 7, 76712–76719. doi: 10.1109/ACCESS.2019.2921976
- Kari, L., Hill, K. A., Sayem, A. S., Karamichalis, R., Bryans, N., Davis, K., et al. (2015). Mapping the space of genomic signatures. *PLoS ONE* 10:e119815. doi: 10.1371/journal.pone.0119815
- Khattak, F. K., Jebblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., and Rudzicz, F. (2019). A survey of word embeddings for clinical text. *J. Biomed. Informatics* 4:100057. doi: 10.1016/j.yjbinx.2019.100057
- Kim, K., Kim, D., Noh, J., and Kim, M. (2018). Stable forecasting of environmental time series via long short term memory recurrent neural network. *IEEE Access* 6, 75216–75228. doi: 10.1109/ACCESS.2018.2884827
- Koohi-Moghadam, M., Wang, H., Wang, Y., Yang, X., Li, H., Wang, J., et al. (2019). Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. *Nat. Mach. Intell.* 1, 561–567. doi: 10.1038/s42256-019-0119-z
- Letko, M., Marzi, A., and Munster, V. (2020). Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage b betacoronaviruses. *Nat. Microbiol.* 5, 562–569. doi: 10.1038/s41564-020-0688-y
- Liu, J., Xia, C., Yan, H., Xie, Z., and Sun, J. (2019). Hierarchical comprehensive context modeling for Chinese text classification. *IEEE Access* 7, 154546–154559. doi: 10.1109/ACCESS.2019.2949175
- Liu, X., and Wang, X.-J. (2020). Potential inhibitors against 2019-nCoV coronavirus m protease from clinically approved medicines. *J. Genet. Genomics* 47, 119–121. doi: 10.1016/j.jgg.2020.02.001
- Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395, 565–574. doi: 10.1016/S0140-6736(20)30251-8
- Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recogn.* 91, 216–231. doi: 10.1016/j.patcog.2019.02.023
- Manekar, S., and Sathe, S. (2018). A benchmark study of k-mer counting methods for high-throughput sequencing. *Gigascience* 7:giy125. doi: 10.1093/gigascience/giy125
- Meng, Y., Wu, P., Lu, W., Liu, K., Ma, K., Huang, L., et al. (2020). Sex-specific clinical characteristics and prognosis of coronavirus disease-19 infection in Wuhan, China: a retrospective study of 168 severe patients. *PLoS Pathol.* 16:e1008520. doi: 10.1371/journal.ppat.1008520
- Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., and Rajendra Acharya, U. (2020). Automated detection of covid-19 cases using deep neural networks with x-ray images. *Comput. Biol. Med.* 121:103792. doi: 10.1016/j.compbiomed.2020.103792
- Paraskevis, D., Kostaki, E., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G., and Tsiodras, S. (2020). Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infect. Genet. Evol.* 79:104212. doi: 10.1016/j.meegid.2020.104212
- Solis-Reyes, S., Avino, M., Poon, A., and Kari, L. (2018). An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLoS ONE* 13:e206409. doi: 10.1371/journal.pone.0206409
- Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C., Zhou, J., et al. (2016). Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol.* 24, 490–502. doi: 10.1016/j.tim.2016.03.003
- Tang, B., Pan, Z., Yin, K., and Khateeb, A. (2019). Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* 10:214. doi: 10.3389/fgene.2019.00214
- Wan, Y., Shang, J., Graham, R., Baric, R. S., and Li, F. (2020). Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J. Virol.* 94:e00127-20. doi: 10.1128/JVI.00127-20
- Weiss, S., and Navas-Martin, S. (2005). Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiol. Mol. Biol. Rev.* 4, 635–664. doi: 10.1128/MMBR.69.4.635-664.2005
- Woo, P. C., Lau, S. K., Huang, Y., and Yuen, K.-Y. (2009). Coronavirus diversity, phylogeny and interspecies jumping. *Exp. Biol. Med.* 234, 1117–1127. doi: 10.3181/0903-MR-94
- Worldometer (2021). *Coronavirus Disease 2019 (COVID-19) Cases*. Available online at: <https://www.worldometers.info/coronavirus> (accessed January 8, 2021).
- Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020). An interpretable mortality prediction model for covid-19 patients. *Nat. Mach. Intell.* 2, 283–288. doi: 10.1038/s42256-020-0180-7
- Yan, Q., Weeks, D. E., Xin, H., Swaroop, A., Y. E. Chew, E., Huang, H., et al. (2020). Deep-learning-based prediction of late age-related macular degeneration progression. *Nat. Mach. Intell.* 2, 141–150. doi: 10.1038/s42256-020-0154-9
- Zhang, Y., Zheng, J., Jiang, Y., Huang, G., and Chen, R. (2019). A text sentiment classification modeling method based on coordinated CNN-LSTM-attention model. *Chinese J. Electron.* 28, 120–126. doi: 10.1049/cje.2018.11.004
- Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

**Conflict of Interest:** AS was employed by company Cognizant Technology Solutions.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Saha, Ghosh, Maity, Seal and Plewczynski. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Detecting the Multiomics Signatures of Factor-Specific Inflammatory Effects on Airway Smooth Muscles

Yu-Hang Zhang<sup>1,2†</sup>, Zhandong Li<sup>3†</sup>, Tao Zeng<sup>4</sup>, Lei Chen<sup>5</sup>, Hao Li<sup>3</sup>, Tao Huang<sup>6\*</sup> and Yu-Dong Cai<sup>1\*</sup>

<sup>1</sup> School of Life Sciences, Shanghai University, Shanghai, China, <sup>2</sup> Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, <sup>3</sup> College of Food Engineering, Jilin Engineering Normal University, Changchun, China, <sup>4</sup> Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China, <sup>5</sup> College of Information Engineering, Shanghai Maritime University, Shanghai, China, <sup>6</sup> Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China

## OPEN ACCESS

### Edited by:

Richard D. Emes,  
University of Nottingham,  
United Kingdom

### Reviewed by:

Bernard Fongang,  
The University of Texas Health  
Science Center at San Antonio,  
United States  
Roney Santos Coimbra,  
Oswaldo Cruz Foundation (Fiocruz),  
Brazil

### \*Correspondence:

Tao Huang  
tohuangtao@126.com  
Yu-Dong Cai  
cai\_yud@126.com

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 28 August 2020

**Accepted:** 14 December 2020

**Published:** 13 January 2021

### Citation:

Zhang Y-H, Li Z, Zeng T, Chen L,  
Li H, Huang T and Cai Y-D (2021)  
Detecting the Multiomics Signatures  
of Factor-Specific Inflammatory  
Effects on Airway Smooth Muscles.  
Front. Genet. 11:599970.  
doi: 10.3389/fgene.2020.599970

Smooth muscles are a specific muscle subtype that is widely identified in the tissues of internal passageways. This muscle subtype has the capacity for controlled or regulated contraction and relaxation. Airway smooth muscles are a unique type of smooth muscles that constitute the effective, adjustable, and reactive wall that covers most areas of the entire airway from the trachea to lung tissues. Infection with SARS-CoV-2, which caused the world-wide COVID-19 pandemic, involves airway smooth muscles and their surrounding inflammatory environment. Therefore, airway smooth muscles and related inflammatory factors may play an irreplaceable role in the initiation and progression of several severe diseases. Many previous studies have attempted to reveal the potential relationships between interleukins and airway smooth muscle cells only on the omics level, and the continued existence of numerous false-positive optimal genes/transcripts cannot reflect the actual effective biological mechanisms underlying interleukin-based activation effects on airway smooth muscles. Here, on the basis of newly presented machine learning-based computational approaches, we identified specific regulatory factors and a series of rules that contribute to the activation and stimulation of airway smooth muscles by IL-13, IL-17, or the combination of both interleukins on the epigenetic and/or transcriptional levels. The detected discriminative factors (genes) and rules can contribute to the identification of potential regulatory mechanisms linking airway smooth muscle tissues and inflammatory factors and help reveal specific pathological factors for diseases associated with airway smooth muscle inflammation on multiomics levels.

**Keywords:** smooth muscles, multiomics signatures, Monte Carlo feature selection, machine learning, rule learning

## INTRODUCTION

Smooth muscles are a specific muscle subtype that is widely identified in the tissues of internal passageways, such as vessels, and internal organs, including the lungs and intestines. This type of muscle has the capacity for controlled or regulated contraction and relaxation. Various types of smooth muscles are distributed all over the human body. Airway smooth



muscle is a unique smooth muscle type that constitutes the effective, adjustable, and reactive wall covering most of the entire airway from the trachea to lung tissues (Chung, 2000; Lam et al., 2019). Similar to that of other smooth muscles, the coupling of excitation and contraction is the basic approach of airway smooth muscles to realize their unique basic biological function: maintaining the normal and effective ventilation of the lungs (Cieri, 2019).

Airway smooth muscle is regulated by various internal and external factors to maintain the balance required for pulmonary oxygen exchange (Dahl et al., 2018; Reyes-García et al., 2018). Cytokines, such as IL-13 and IL-17, have been confirmed to participate in the regulation of airway smooth muscles (Pascoe et al., 2017; Ba et al., 2018; Zhang et al., 2019; Koziol-White et al., 2020). A systematic analysis of human airway smooth muscle cells (ASMCs) has confirmed that interleukins, including IL-13 and IL-4, participate in the regulation of the hypo-responsiveness of smooth muscle subtypes (Koziol-White et al., 2020). IL-17 has been confirmed to participate in the typical inflammatory reactions of ASMCs (Bexiga et al., 2018; Thompson et al., 2018). The identification of IL-17 together with multiple interleukins as candidate regulators validates the specific contributions of interleukins to the actions of ASMCs.

As discussed above, interleukins, such as IL-13 and IL-17, are functionally correlated with the biological processes of ASMCs, and interactions between interleukins and ASMCs may also be correlated with various diseases. Asthma is a typical respiratory inflammatory disease that has been widely reported to be functionally correlated with airway smooth muscles in an inflammatory environment (Bousquet et al., 2000; Salter et al., 2017; Ramakrishnan et al., 2019; Tliba and Panettieri, 2019). For example, the migration of human airway smooth muscles has been confirmed to be regulated by cytokines, including IL-13 and IL-17, and further contribute to the pathogenesis of asthma (Salter et al., 2017). Moreover, infection with SARS-CoV-2, which caused the worldwide COVID-19 pandemic, involves airway smooth muscles and their surrounding inflammatory environment (Frohman et al., 2020; Sungnak et al., 2020). Therefore, airway smooth muscles and related inflammatory factors (like interleukins) may play an irreplaceable role in the initiation and progression of several severe diseases. Studies on the interactions between airway smooth muscles and related interleukins and the detailed contributions of interleukins to the biological or pathological activation of ASMCs may contribute to the explanation of the detailed pathogenesis of inflammatory pulmonary diseases and help the identification of potential effective biomarkers for drug discovery and treatment improvement.

Many previous studies have attempted to reveal the potential relationships between interleukins and ASMCs at different omics levels. Recently, a specific study on the relationships between asthma-promoting cytokines (IL-13 and IL-17) and ASMCs tried to identify key regulatory factors on the transcriptomics and epigenetics levels. Researchers identified 225 genes around differentially methylated regions by using independent IL-13 and IL-17 and combined interleukins and 2014 differentially expressed transcripts by comparing different cytokine-stimulated

groups (Thompson et al., 2019). However, the continued existence of numerous false-positive optimal genes/transcripts cannot reflect the actual effective biological mechanisms underlying interleukin-based activation effects on airway smooth muscles. In this study, on the basis of newly presented computational approaches based on machine learning, we first identified specific regulatory factors (genes) that contribute to the activation and stimulation of airway smooth muscles by IL-13, IL-17, or the combination of both interleukins on the epigenetic and/or transcriptional levels. Next, we also established a series of rules based on essential genes that contribute to distinguishing quiescent and interleukin (either independent or combined)-activated ASMCs in a quantitative manner. Our results, including detected discriminative genes and quantitative rules, corresponding to different patterns, can contribute to the identification of potential regulatory mechanisms underlying interactions between airway smooth muscle tissues and inflammatory factors (IL-13 and IL-17) and help reveal specific pathological factors for diseases associated with airway smooth muscle inflammation on multiomics levels.

## MATERIALS AND METHODS

### Data

In March 2020, researchers from the University of Chicago released the gene methylation and expression data of ASMCs under the stimulation of multiple inflammatory factors to the Gene Expression Omnibus database (GSE146377) with more than 500 samples (either transcriptomics or methylation data). All the transcriptomics and gene methylation data were generated from the primary cultured ASMCs. In this study, we aimed at interpreting the biological significance of lung smooth muscle and related inflammatory factors during the initiation and progression of multiple diseases like COVID-19 which has ravaged all over the world recently. Following the goal, we downloaded the methylation and gene expression profiles of primary cultured ASMCs exposed to IL-13, IL-17, IL-13 + IL-17, and vehicle from the Gene Expression Omnibus database under the accession number of GSE146377. Only samples with methylation and gene expression data were analyzed. Each of the IL-13, IL-17, IL-13 + IL-17, and vehicle groups had 64 samples. Methylation data were generated with Infinium MethylationEPIC and included 786,326 probes. The expression levels of 18,279 genes were profiled with Illumina HumanHT-12 V4.0 expression beadchip. We aimed to investigate the responsive genes of ASMCs to IL-13, IL-17, and IL-13 + IL-17.

### Monte Carlo Feature Selection

The methylation and gene expression profiles of ASMCs have much more features than samples. The Monte Carlo feature selection (MCFS) (Dramiński et al., 2007) was deemed to be excellent in tackling such type of dataset. It is a powerful and widely used feature selection technology.

To evaluate the importance of features, MCFS generally includes the following steps: (i) the selection of random feature subsets with  $m$  features from the original whole  $M$  features



( $m \ll M$ ); (ii) the learning of a classification model on the bootstrap dataset for each feature subset, which generates  $p$  decision trees from classification model; (iii) the production of  $p \times t$  decision trees by repeating the above steps  $t$  times; and (iv) the calculation of the relative importance score (RI) for each feature. Among the constructed  $p \times t$  decision trees, a given feature may occur in some of them. The split on a node using such feature in each of these decision trees can reflect its importance, which can be measured by the information gain achieved by such split. Furthermore, the classification ability of the decision tree should also be included. Thus, the contribution of a feature in a decision tree can be determined by the information gain achieved by the split, the number of samples in the split node and the classification ability of the tree. The RI value of a feature  $f$  can be the sum of contributions on all constructed decision trees, which is defined as

$$RI_f = \sum_{\tau=1}^{pt} (wAcc)^u \sum_{n_f(\tau)} IG(n_f(\tau)) \left( \frac{\text{no. in } n_f(\tau)}{\text{no. in } \tau} \right)^v \quad (1)$$

where  $wAcc$  is the weighted accuracy, and  $n_f(\tau)$  is a node of feature  $f$  in the decision tree  $\tau$ . The information gain of  $n_f(\tau)$  is expressed as  $IG(n_f(\tau))$ , and  $(\text{no. in } n_f(\tau))$  is the number of training samples in  $n_f(\tau)$ .  $u$  and  $v$  are two weighting factors, which is suggested to one.

After all investigated features are assigned the RI values, a feature list is produced by the decreasing order of RI values of features. In this study, we adopted the MCFS program downloaded from <http://www.ipipan.eu/staff/m.draminski/mcfs.html>. For convenience, default parameters were used.

## Incremental Feature Selection

Incremental feature selection (IFS) (Liu and Setiono, 1998) is an iterative feature selection approach, which can find the best number of features for a given classification algorithm. For a feature list (e.g., a list produced by the MCFS method), IFS always generates lots of feature subsets, each of which contains some top features in the list. For example, the first feature subset contains the top one feature in the list, the second feature subset consists of the top two features, and so forth. Then, for each feature subset, a classifier can be built based on a given classification algorithm and samples represented by features in the subset. Finally, all constructed classifiers are evaluated by a cross-validation method (e.g., 10-fold cross-validation) (Kohavi, 1995). The classifier with the best performance is extracted, which were called the optimum classifier in the study. Furthermore, the corresponding feature subset was termed as the optimum feature subset.

## Classification Algorithm

As mentioned in section “Incremental Feature Selection,” a powerful classification algorithm is necessary for the IFS method. This study tried four classification algorithms: random forest (RF) (Breiman, 2001), support vector machine (SVM) (Cortes and Vapnik, 1995), k-nearest neighbor (kNN) (Cover and Hart, 1967), and repeated incremental pruning to produce error reduction (RIPPER) (Cohen, 1995). Their brief descriptions are as follows.

## Random Forest

Random forest (Breiman, 2001) is an assemble classification model that is based on multiple decision tree classifiers. Each decision tree is constructed using randomly selected samples and features. Although decision tree is a relative weak classification algorithm, RF is much power and always an important choice for building different classification models (Tang et al., 2018; Baranwal et al., 2019; Zhao et al., 2019; Jia et al., 2020; Liang et al., 2020). The predicted sample label of RF is obtained on the basis of the aggregated votes of decision tree classifiers. The subtle difference among decision trees in RF causes the potential overfitting of learned models. Thus, RF usually adopts the final consensus results in accordance with the average of all decision trees' predictions. This study adopted the tool “RandomForest” in Weka (Frank et al., 2004; Witten and Frank, 2005), which implements the RF algorithm. The major parameter, number of decision trees, was set to 10.

## Support Vector Machine

Support vector machine (Cortes and Vapnik, 1995) is a statistical learning-based classification algorithm. Similar to RF, SVM is another essential candidate for constructing classification models (Sang et al., 2020; Zhou et al., 2020a,b). It first transforms original data from a low-dimensional space to a high-dimensional space by using a kernel function and then divides the data samples of each label in accordance with the principle of data interval maximization in high-dimensional space. It further predicts the new samples' label in accordance with the interval to which this new sample belongs to. In this work, the tool “SMO” in Weka software (Frank et al., 2004; Witten and Frank, 2005) was employed to construct the SVM classifier. The training procedures are optimized by the sequential minimal optimization algorithm (Platt, 1998). The kernel was a polynomial function and the parameter  $C$  was set to 1.0.

## k-Nearest Neighbor Classification

k-nearest neighbor is another classification model with a voting scheme (Theilhaber et al., 2002; Zhang and Srihari, 2004; Yu et al., 2016; Chen et al., 2017a). Given a query sample and one training dataset, kNN includes several computation steps to determine its class: (1) the calculation of the sample distance between the query sample and training samples; (2) the ranking of training samples on the basis of their distances to the query sample; (3) the selection of  $k$  training samples with the least distance to the query sample (i.e., kNNs, and  $k$  usually ranges from 1 to 10); (4) the estimation of the label distribution of such  $k$  nearest training samples; and (5) the prediction of labels for the query sample by using the class label with the highest distribution frequency. In this work, the tool “IBk” in Weka (Frank et al., 2004; Witten and Frank, 2005) was used to build the kNN classifier. The distance between samples was defined as the Euclidean distance.

## Rule Learning

In addition to the above black-box classification algorithms, we also applied a rule learning algorithm, RIPPER (Cohen, 1995), to generate classification rules for enhancing model interpretation. This algorithm starts to generate rules for the class containing

least samples. When a rule is produced, covered samples are removed. Other rules are yielded on the rest samples. Each rule generated by RIPPER is represented by an IF-ELSE statement. For instance, If (GPR44  $\geq$  7.200) and (ZC3H12A  $\leq$  8.211), THEN class = IL-13. Rules in such form can provide human-readable predictions for new samples. In this study, tool “JRip” in Weka (Frank et al., 2004; Witten and Frank, 2005) was utilized to learn RIPPER rules.

## Performance Evaluation

The Matthew correlation coefficient (MCC) (Matthews, 1975; Chen et al., 2017a,b; Zhao et al., 2018), a widely used evaluation measurement, was applied to evaluate the performance of the classification model through 10-fold cross-validation (Kohavi, 1995). MCC ranges from  $-1$  to  $+1$ . The classification model with an MCC of  $+1$  has the best performance. Our analyzed data were organized into four categories. Thus, the multiclass version of the MCC (Gorodkin, 2004) was calculated as follows:

$$\text{MCC} = \frac{\text{cov}(X, Y)}{\sqrt{\text{cov}(X, X)\text{cov}(Y, Y)}} \quad (2)$$

where  $X$  is a 0–1 matrix indicating the predicted class of each sample,  $Y$  is a 0–1 matrix representing the actual classes of all samples, and  $\text{cov}(\cdot, \cdot)$  represents the covariance of two matrixes.

In addition, the accuracy on each category and overall accuracy (ACC) were also calculated to fully indicate the performance of each model.

## RESULTS

In this study, we employed several computational methods to investigate the methylation and gene expression profiles of ASMCs. Samples were divided into four groups: the control group, IL-13 stimulation group, IL-17 stimulation group, and combined (IL-13 and IL-17) stimulation group. We organized the data into three types: the methylation data of the four groups, the expression data of the four groups, and the combined data of the four groups. For each type of data, we utilized a similar analytical pipeline. The entire procedures are illustrated in **Figure 1**.

### Results for Methylation Data

For methylation data, we first used MCFS to evaluate each feature, obtaining a feature list, which is available in **Supplementary Table 1**. Due to the huge number of methylation features, IFS only constructed the top 5000 feature subsets. A RF, SVM, or kNN classifier was built on each feature subset, which was further evaluated by 10-fold cross-validation. The performance of each classifier, including accuracies on four categories, ACC and MCC, is provided in **Supplementary Table 2**. For an easy observation, a curve with MCC as  $Y$ -axis and number of used features as  $X$ -axis was plotted for each classification algorithm, as shown in **Figure 2**. The SVM exhibited the best performance and had the MCC of 0.831 when top 4940 features were used. For RF and kNN, the best MCC was 0.710 and 0.182, respectively, which was based on top 629 and 4 methylation features. Accordingly, the optimum SVM, RF, and

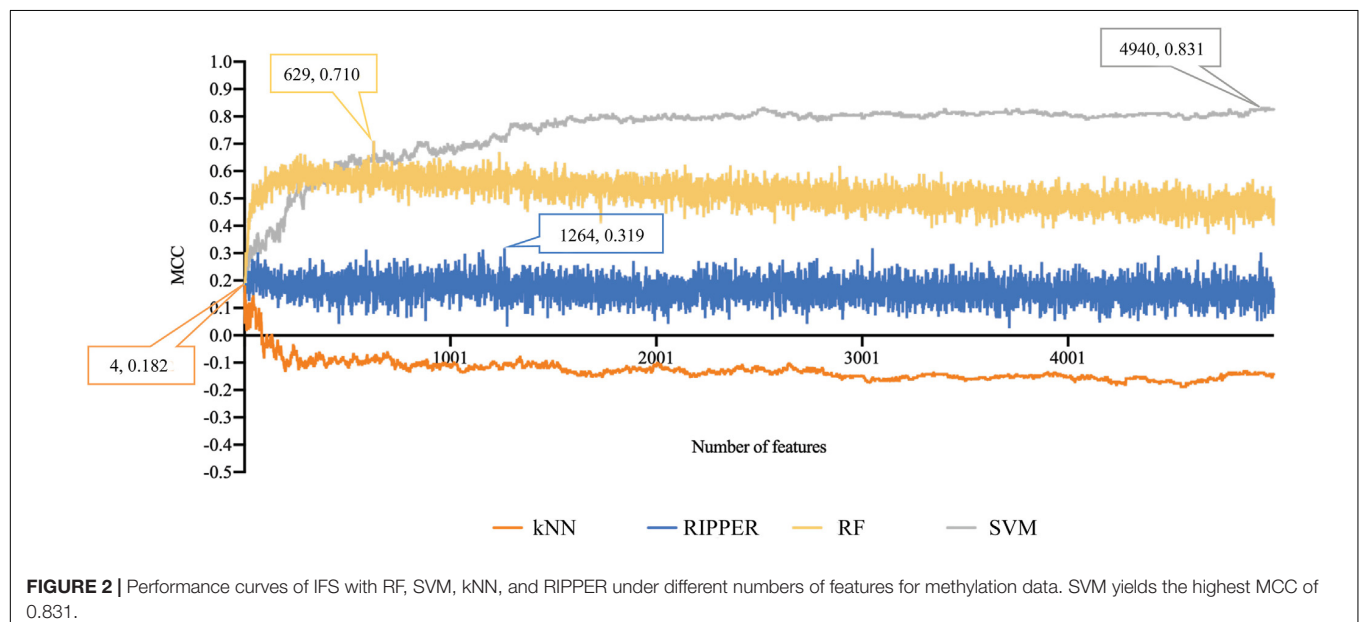
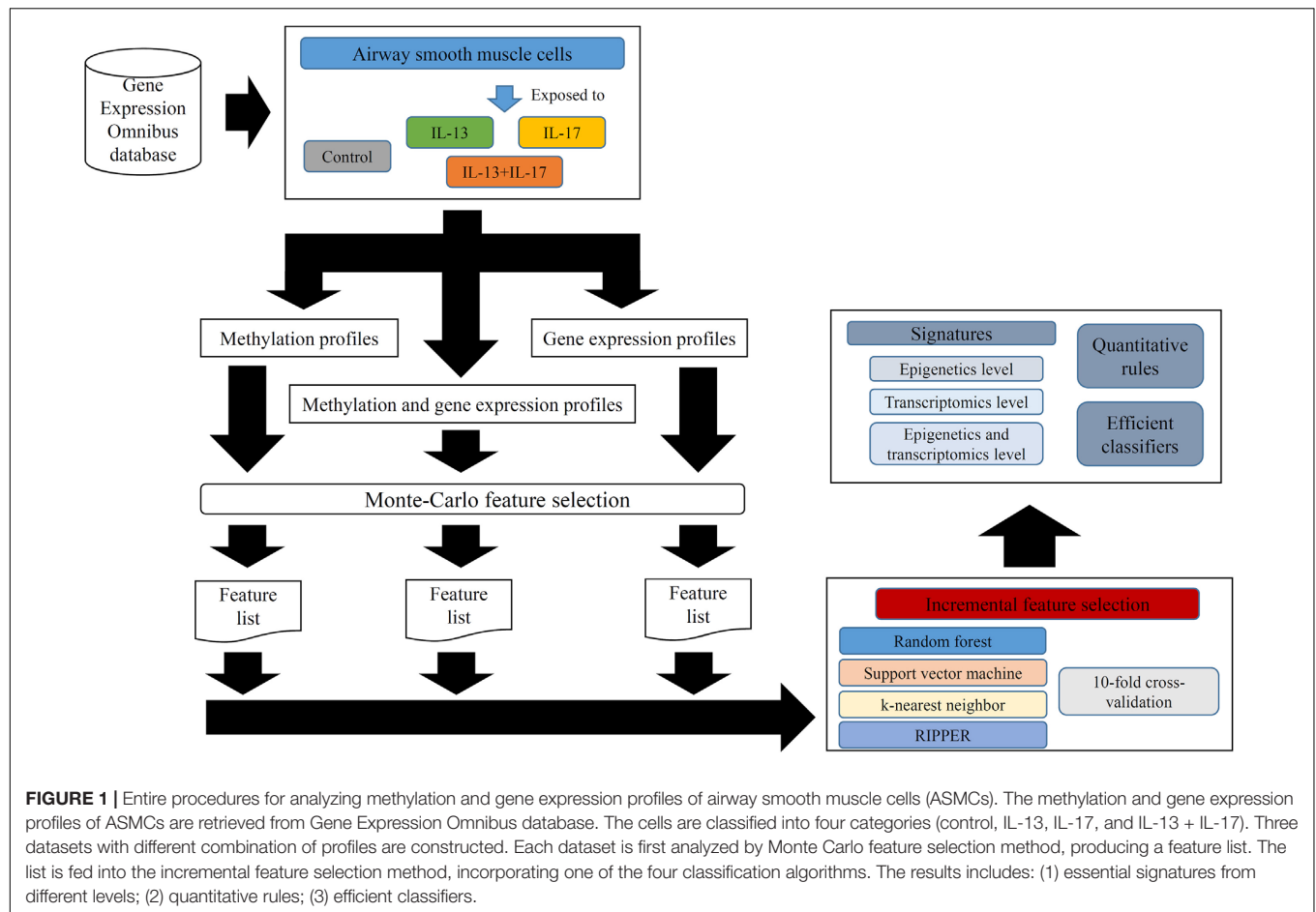
kNN classifiers were built using corresponding optimum feature subsets. The ACCs of these classifiers are listed in **Table 1** and the accuracies on four categories are illustrated in **Figure 3A**. Besides the black-box classifiers, we also tried the rule learning algorithm, RIPPER, in IFS method. Similarly, we still considered the top 5000 feature subsets. The performance of RIPPER classifiers is provided in **Supplementary Table 2** and the corresponding curve is shown in **Figure 2**. The optimum RIPPER classifier yielded the MCC of 0.319 when top 1264 features were used, the corresponding ACC was 0.488 (**Table 1**). **Figure 3A** shows the four accuracies on four categories yielded by such classifier. This performance was insufficiently satisfactory for such a rule-based approach.

### Results for Gene Expression Data

The similar analytical pipeline was applied on the gene expression data. A feature list was first obtained according to the results of MCFS, which are provided in **Supplementary Table 3**. Then, we applied IFS with 1 as an interval to build classifiers with one of the four classification algorithms. To save time, we still considered top 5000 features. Each classifier was evaluated by 10-fold cross-validation. Obtained measurements are listed in **Supplementary Table 4**. The corresponding curves were plotted in **Figure 4**, from which we can see that the four optimum classifiers with different classification algorithms yielded the MCC of 0.870, 0.928, 0.990, and 0.897, respectively, and adopted the top 24, 40, 3440, and 794 features, respectively. The corresponding ACCs are listed in **Table 1** and accuracies on four categories are shown in **Figure 3B**. Similar to the results on the methylation data, the optimum SVM classifier was still best (MCC = 0.990). As for the optimum RIPPER classifier, its performance was much better than that for the methylation data. It produced the MCC of 0.897 and ACC of 0.922 (**Table 1**). This performance was sufficiently satisfactory. Accordingly, we used top 794 features, which was adopted to build such classifier, to construct rules with RIPPER, obtaining seven rules, where three rules were for IL-13, two rules for control, one rule for both of other two categories. These rules are listed in **Table 2**. A further analysis would be given in section “Optimal Rules for Distinguishing the Different Statuses of ASMCs.”

### Results for Combined Data

Finally, for combined data, we did the same test. The feature list yielded by the MCFS method is provided in **Supplementary Table 5**. The IFS method was applied on such list using one of the four classification algorithms. Also, only top 5000 features were considered. The accuracies on four categories, ACCs and MCCs for each classification algorithm are listed in **Supplementary Table 6** and a curve for each algorithm was plotted in **Figure 5** to show the trends of the performance. It can be observed that SVM consistently achieved the best performance among all algorithms. Its MCC was 0.969 when 3103 top features were used. The ACC was 0.977 (**Table 1**) and accuracies on all categories are shown in **Figure 3C**. The performance of other optimum classifiers are listed in **Table 1** and **Figure 3C**. The optimum RIPPER classifier also provided good performance of MCC = 0.891, which used top 42 features. In view of this,



we obtained seven rules, listed in **Table 3**, based on these 42 features and RIPPER. Among these seven rules, two rules were for IL-17, three rules were for IL-13, and one rule was

for both of other two categories. We would analyze them in section “Optimal Rules for Distinguishing the Different Statuses of ASMCs.”

**TABLE 1** | Performance of the best classification model on three datasets with different classification algorithms.

Dataset	Classification algorithm	Number of features	ACC	MCC
Methylation	kNN	4	0.387	0.182
	RF	629	0.781	0.710
	SVM	4940	0.871	0.831
	RIPPER	1264	0.488	0.319
Gene expression	kNN	24	0.902	0.870
	RF	40	0.945	0.928
	SVM	3440	0.992	0.990
	RIPPER	794	0.922	0.897
Methylation+ gene expression	kNN	4	0.883	0.844
	RF	96	0.938	0.917
	SVM	3103	0.977	0.969
	RIPPER	42	0.918	0.891

The results for RIPPER indicated that datasets containing only epigenetic data with RIPPER and the MCC of 0.319 might be unacceptable for further analyses and that the use of methylation data might be ineffective for constructing reliable quantitative rule-based models for distinguishing the different statuses of ASMCs. Expression data and combined data could provide an optimal RIPPER MCC of approximately 0.900, validating the reliability and efficacy of the features and rules learned from the two datasets.

## Enrichment Results

For all three datasets, the best optimum classifiers all used SVM as the classification algorithm. In detail, for methylation data, the optimum SVM classifier adopted top 4940 features, while the optimum SVM classifiers used top 3440 and 3103 features, respectively, for other two datasets. Their corresponding genes were called optimum signatures (genes) for the corresponding dataset. To reveal the potential biological functions that optimum genes are correlated with, we performed GO enrichment analyses using R package (*topGO* v2.38.1) on them. The results are provided in **Supplementary Table 7**. Of the optimum genes on epigenetic and transcriptomics levels, they enriched five and 39 GO terms, respectively, while 68 GO terms were enriched by the optimum genes on both epigenetic and transcriptomics levels. An analysis would be performed in section “Go Enrichment Analyses for Optimal Signatures for Distinguishing the Different Statuses of ASMCs.”

## DISCUSSION

We applied multiple machine learning models to identify potential multi-omics signatures on the epigenetic and transcriptomic levels. By using our newly presented computational methods, we not only identified a group of effective signatures (genes) that were remarkably correlated with the interactions between interleukins (IL-13, IL-17, or their combination) and ASMCs, but also established specific rules to distinguish four ASMC statuses: quiescent, IL-13 activated, IL-17 activated, and IL-13–IL-17 combined activated.

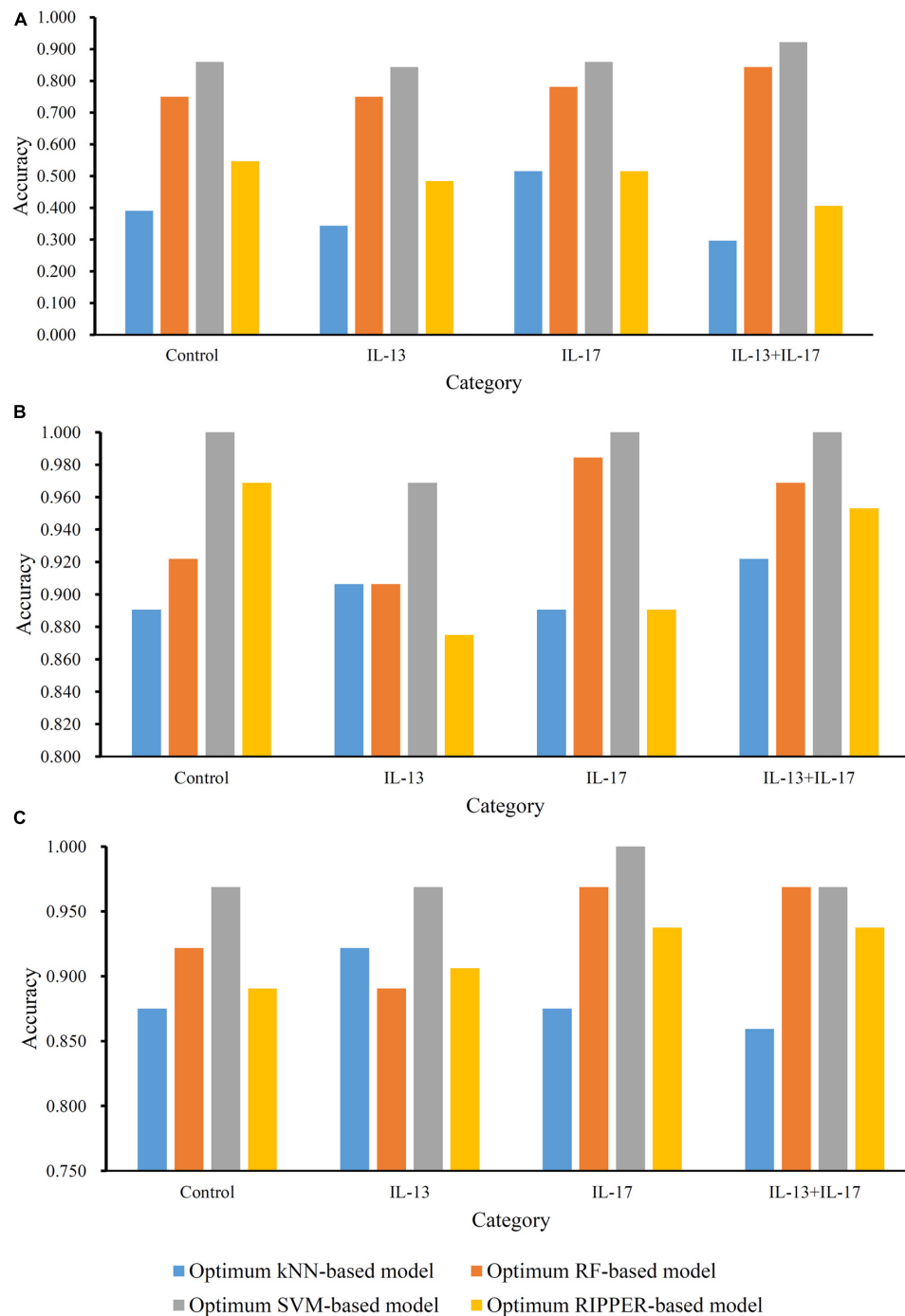
Similar signature analyses have been validated under three conditions, i.e., single transcriptomics level, single epigenetic level, and combined transcriptomics and epigenetic levels. All the identified signatures and rules were validated on the basis of recent publications, indicating the efficacy and accuracy of our prediction. Given the limitation of this manuscript's length, we only chose several typical genes for introduction. The detailed discussion on the signatures and rules is given below.

## Optimal Signatures for Distinguishing the Different Statuses of ASMCs

### Signatures on the Epigenetics Level

The top-ranked gene in our prediction list obtained from the epigenetic dataset is *BEND6* with specific methylation alterations on the first exon (**cg08811259**). *BEND6* has been widely reported to be functionally correlated with the Notch signaling pathway (Dai et al., 2013). Early in 2008, the Notch signaling pathway was confirmed to regulate the hyper-responsiveness and inflammation of ASMCs (Okamoto et al., 2008) via multiple interleukins, including IL-13 (Lee et al., 2001) and IL-17 (Plé et al., 2015). Therefore, given that the methylation alteration of *BEND6* has been validated to affect the Notch signaling pathway, this methylation probe together with its target gene *BEND6* are potential biomarkers for distinguishing ASMCs with or without interleukin stimulation.

The next probe (**cg26074603**) targets the 5' UTR of *KCNC2*. This gene is a core regulator of the voltage-gated potassium channel and has been confirmed to participate in the pathogenesis of multiple diseases, including extratemporal epilepsy (Vetri et al., 2020) and spinocerebellar ataxia (Rajakulendran et al., 2013). Moreover, *KCNC2* has been reported to participate in pulmonary neutrophilic inflammation in the lungs and airway; this condition can involve local smooth muscles (Nadadur et al., 2005). Although direct evidence confirming that interleukins may affect the contribution of *KCNC2* to the inflammation of airway smooth muscles does not exist, previous studies have confirmed that *KCNC2* indeed interacts with multiple interleukins, including, IL-13 and IL-1 (Haas et al., 1993), partially validating our prediction.

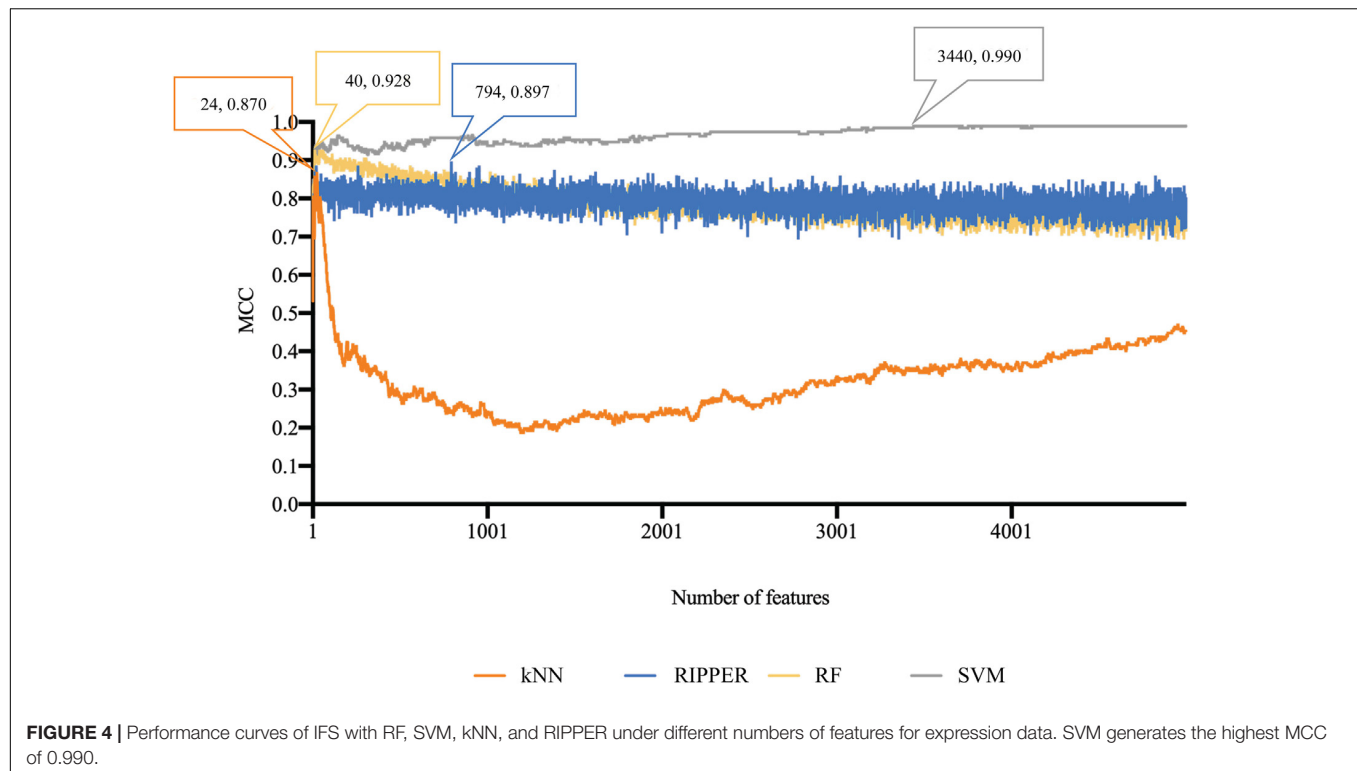


**FIGURE 3 |** Accuracies on all categories yielded by the optimum classifiers on three datasets. **(A)** Methylation data; **(B)** gene expression data; **(C)** combined data.

The next optimal gene on the methylation level is *MAST4*, which is targeted by the probe **cg06040990**. This gene is a microtubule-associated protein kinase (Sun et al., 2006) that has been widely reported to participate in multiple inflammatory-associated biological processes (Gongol et al., 2017; Cortes et al., 2020). *MAST4* is a part of the PTEN signaling pathway (Valiente et al., 2005; Sotelo et al., 2012), which

has been confirmed to mediate the IL-13-induced stimulation, hyper-responsiveness, and inflammation, of airway smooth muscles, thus validating this predicted target gene (Jiang et al., 2012). Similar conclusions have also been further validated in later studies (Hu et al., 2014; Khalifeh-Soltani et al., 2018). Therefore, *MAST4* is definitely correlated with the interleukin-mediated stimulation of airway smooth muscles.





**TABLE 2 |** Rules by RIPPER on expression data.

Index	Condition	Result
1	(GPR44 $\geq$ 7.200) and (ZC3H12A $\leq$ 8.211)	IL-13
2	(SEMA3A $\leq$ 9.623) and (NFKBIZ $\leq$ 10.612)	IL-13
3	(MYOM1 $\geq$ 7.527) and (MAP3K8 $\leq$ 9.269)	IL-13
4	(NFKBIZ $\leq$ 10.483) and (MAP3K8 $\leq$ 8.234)	Control
5	LSS $\leq$ 10.932	Control
6	CCL26 $\leq$ 9.291	IL-17
7	Others	IL-13 and IL-17

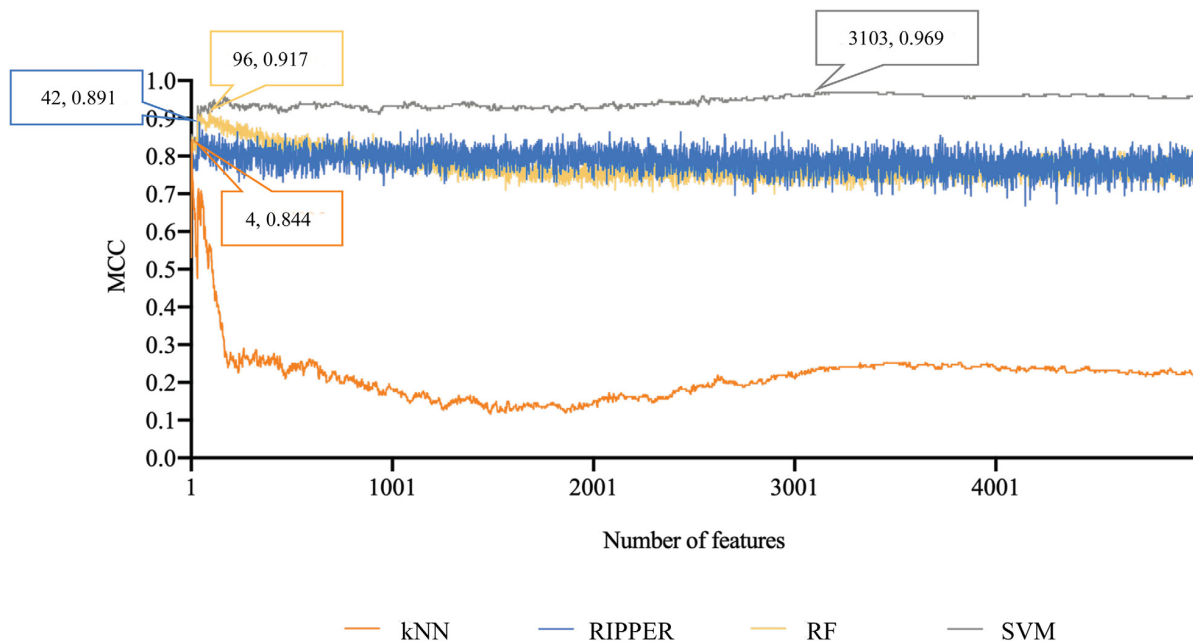
### Signatures on the Transcriptomics Level

Similar to the analyses based on the methylation-level dataset, our other analyses also identified a group of genes (transcripts) that contributes to distinguishing the different statuses of ASMCs. All such genes/transcripts have also been further validated to be effective in accordance with recent publications.

The first gene in our prediction list is *MAP3K8*, a member of the serine/threonine protein kinase family. *MAP3K8* has been confirmed to be associated with typical differential expression levels in systematic neutrophilic inflammation involving airway tissues (Fu et al., 2013). Although no direct reports have confirmed the regulatory roles of IL-13 and IL-17 in *MAP3K8*-mediated airway inflammation responses, *MAP3K8* has been widely reported to perform an interleukin-dependent inflammatory regulatory role during multiple biological or pathological processes (Glossop and Cartmell, 2009; Kim et al., 2014; Sánchez et al., 2017), implying the specific role of such a gene in the different statuses of ASMCs.

The next gene that contributes to cell classification on the transcriptomics level is *CCL26*, a functional secretory factor that contributes to immune regulatory and inflammatory processes in human bodies (Sangaphunchai et al., 2020). This gene has also been reported to be differentially expressed in airway tissues and participates in the inflammatory response in lung and airway tissues during the pathogenesis of asthma (Sangaphunchai et al., 2020). It has been directly reported to be functionally correlated with IL-13 (Higham et al., 2020; Min et al., 2020) and IL-17 (Kamijo et al., 2020; Mamber et al., 2020) in focal regions surrounding airway smooth muscles at the transcriptomics level and is further pathologically correlated with several chronic lung diseases, including chronic obstructive pulmonary diseases (Min et al., 2020). Therefore, given its functional correlation with the potential regulatory effects of IL-13 and IL-17 on airway smooth muscle inflammation, the predicted gene *CCL26* is definitely an effective signature for cell classification on the transcriptomics level.

*CISH*, also known as *SOCS*, is predicted to be important for the classification of ASMCs with different interleukin stimulation statuses. *CISH* is present at specific expression levels in Treg cells in allergic-associated airway inflammation (Zheng et al., 2020), implying the specific regulatory role of *CISH* in airway regional inflammation on the transcriptional level. *SOCS* can also participate in the regulation of human monocyte inflammatory responses involving IL-13 and IL-4 (Wolde et al., 2020), confirming its potential classification capacity at the gene expression level. Summarizing the specific biological regulatory role of *CISH* in airway tissues reveals that *CISH* is a potential regulatory factor of interactions



**FIGURE 5 |** Performance curves of IFS with RF, SVM, kNN, and RIPPER under different numbers of features for combined data. SVM produces the highest MCC of 0.969.

**TABLE 3 |** Rules by RIPPER on combined data.

Index	Condition	Result
1	(COL17A1 $\leq 7.226$ ) and (NFKBIZ $\geq 10.687$ )	IL-17
2	(CCDC86 $\leq 9.636$ ) and (NFKBIZ $\geq 10.454$ )	IL-17
3	(DTNA $\leq 7.176$ ) and (NFKBIZ $\leq 10.496$ )	IL-13
4	(CCL11 $\geq 12.814$ ) and (MAP3K8 $\leq 9.294$ ) and (SLIT2 $\leq 11.423$ )	IL-13
5	PPFIBP2 $\geq 9.582$	IL-13
6	MAP3K8 $\geq 8.957$	IL-13 and IL-17
7	Others	Control

between interleukins and airway smooth muscles on the transcriptomics level.

### Combinatory Signatures on the Epigenetic and Transcriptomics Level

Epigenetic- and transcriptomics-level data may be applicable for distinguishing different ASMC statuses on the basis of combinatory signatures. Here, we integrated epigenetic- and transcriptomics-level data to identify specific signatures at the dual-omics levels by using our presented computational method. In accordance with the prediction list, most of the top-ranked features are the same as the features identified through the above transcriptomics-only analyses. Therefore, we further discussed the epigenetic contribution of the top three genes that have already been discussed on the transcriptomics level to provide wide and solid literature support.

As discussed above, *MAP3K8* has been validated to be a transcriptomic regulator that can be used to distinguish

different stimulation statuses. The abnormal methylation status of this gene is correlated with multiple chronic pathological conditions, such as lung adenocarcinoma (Tsay et al., 2015) and autoimmune lung injuries (Diaconu et al., 2010; Xie et al., 2018). Although no direct evidence has shown that the methylation alteration of *MAP3K8* is functionally correlated with interleukins, such as IL-17, in the inflammation of airway smooth muscles, a recent publication on colorectal cancer has indicated that the methylation of *MAP3K8* controls focal inflammatory responses via the regulation of related interleukins (Hartley, 2020). Therefore, in addition to its unique contribution on the transcriptomics level, *MAP3K8* is an effective epigenetic regulator of interleukin-mediated airway smooth muscle activation.

*CCL26*, the next predicted gene, is ranked second on the transcriptomic level but fourth on the epigenetic level. It is also associated with specific methylation status in lung- and respiratory-related tissues under various pathological conditions, including lung adenocarcinoma (Dong et al., 2020) and asthma (Kim et al., 2020). *CCL26* has been validated to be regulated by specific interleukins, such as IL-13 (Lyles and Rothenberg, 2019), and further studies have validated that the methylation status of *CCL26* is greatly altered during the inflammatory responses of ASMCs under either pathological or physical conditions (Grozdanovic et al., 2019). Therefore, *CCL26* can also be regarded as a methylation signature of interleukin-mediated inflammation involving ASMCs in addition to its role as an effective transcriptomics signature.

Recently, in correspondence with our prediction, a review of the inflammation profiling of asthma involving airway smooth muscles identified *CISH* as a potential methylation biomarker for airway regional inflammation. Furthermore, *CISH* has been

reported to exhibit different methylation patterns in different asthma statuses with different interleukin profiles (Vermeulen et al., 2020), validating the specific role of *CISH* in inflammatory lung diseases on the methylation level.

Collectively, all the optimal signatures have been validated even at the dual-omics level by recent publications. Summarizing the classification model of datasets on different levels revealed that the optimal features on transcriptomics level are similar to those based on combinations but different from those on the methylation level, indicating that transcriptomics-level datasets may perform better than other datasets in indicating the different activation statuses of airway smooth muscles under interleukin stimulation.

## Optimal Rules for Distinguishing the Different Statuses of ASMCs

In addition to the above specific signatures for distinguishing the different statuses of ASMCs, we established a group of effective quantitative rules for cell classification by using the RIPPER computational method. In accordance with the above discussion, we focused on the quantitative rules obtained by using transcriptomics-level data and the dataset combining transcriptomics- and epigenetic-level data.

### Rules on the Transcriptomics Level

We identified seven rules to distinguish the four groups of cells on the transcriptomics level. The first three rules are defined to identify groups under only *IL-13* stimulation and involved genes *GPR44*, *ZC3H12A*, *SEMA3A*, *NFKBIZ*, *MYOM1*, and *MAP3K8*. We have already analyzed the specific role of *MAP3K8* in *IL-13*- or *IL-17*-stimulated inflammation involving ASMCs (Hartley, 2020). For other quantitative parameters, we took *GPR44* and *MYOM1* as two typical examples. *GPR44* encodes a receptor for prostaglandin D2. *IL13* participates in the activation of Th2 cells, on which our target gene *GPR44* is expressed. Therefore, *GPR44* can be reasonably predicted to have a greater expression level in the group under *IL-13* stimulation (Huang et al., 2016) or at least greater expression than that in the controls and *IL-17* stimulation. Another parameter of *MYOM1* is increased expression level in the *IL-13* stimulated group, and we found some evidence confirming that *MYOM1* is up-regulated during *IL-13*-mediated interleukin stimulation under inflammatory conditions, (Campbell and Hardman, 2020) although few publications have shown potential correlations between *MYOM1*- and *IL-13*-mediated stimulation.

Similar to the rules identified for *IL-13* stimulation group, the specific gene *MAP3K8* remains important for low expression levels in control group. A unique parameter, *LSS*, is down-regulated in controls but up-regulated in all activated ASMCs. *LSS* has been widely associated with nonspecific inflammation in human beings (Vykhovanets et al., 2006; Qin et al., 2013; Li et al., 2016). Therefore, interleukin-mediated airway smooth muscle activation can definitely trigger the up-regulation of *LSS*, indicating that the down-regulated expression of *LSS* may be an effective signature for controls without inflammatory reactions on any levels. *CCL26* in the unique rule identifying *IL-17* stimulation group is the only quantitative parameter for

identifying the *IL-17* stimulated group. As analyzed above, on the transcriptomics level, *CCL26* has been already confirmed to be up-regulated under stimulation by *IL-13* (Wolde et al., 2020). Therefore, the low expression level of *CCL26* may be used to further distinguish samples under only *IL-17* stimulation from samples under combined stimulation. Finally, the remaining samples are reasonably stimulated by *IL-13* and *IL-17*, thus validating the efficacy and accuracy of our quantitative predictive rules.

### Rules on Epigenetic and Transcriptomics Levels

By combining epigenetic and transcriptomics data, we also obtained a group of combined signatures with specific quantitative thresholds that reflect expression or methylation tendency. In accordance with the combined rules and in correspondence with our above discussion on the comparison of the contributions of methylation and transcription features to cell classification, all the optimal parameters are simply transcriptomics features. The detailed discussion is provided below.

The first two rules identify *IL-17* stimulation group. Both rules include the up-regulation of *NFKBIZ*, a specific regulator of interleukin-mediated immune responses (Garg et al., 2015). Previous studies have already connected the up-regulation of *NFKBIZ* with the stimulation of *IL-17* (Göransson et al., 2009; Chapman et al., 2010). This connection corresponds with our prediction. Another effective parameter, *CCDC86*, is positively correlated with *IFNG* and *IL-13*. Therefore, the low expression level of *CCDC86* may indicate that a group may not be stimulated by *IL-13* and further confirms that a group is stimulated by only *IL-17* but not the combination of interleukins. *NFKBIZ* remains one of the most significant parameters for *IL-17* stimulation group. The up-regulation of *NFKBIZ* indicates the stimulation of *IL-17*. Therefore, the down-regulation of *NFKBIZ* may distinguish this group from the combined stimulation and *IL-17* stimulation groups. In addition, the high expression of *PPFIBP2* is correlated with *IL-13*-associated inflammatory immune responses, and samples not fitting all the above rules can definitely be classified as control group.

## GO Enrichment Analyses for Optimal Signatures for Distinguishing the Different Statuses of ASMCs

As several GO terms were extracted for different datasets, we selected some of them for analysis.

### GO Enrichment Analyses for Signatures on the Epigenetics Level

As shown in **Supplementary Table 7**, we only identified five enriched GO terms of different clusters. We chose **GO:0046872 (metal ion binding)** for detailed discussion. For metal ion binding, correlated with ciliary base, calcium ion binding has been shown to regulate the ASMCs related inflammatory reactions via regulating the function of ciliary base (Aisenberg et al., 2016), validating accuracy of the optimum signatures on the epigenetics level.

## GO Enrichment Analyses for Signatures on the Transcriptomics Level

For specific GO enrichment generated from signatures on transcriptomics level, only 39 enriched GO terms were identified. The detailed results can be seen in **Supplementary Table 7**. Here, we chose two terms as for detailed discussion: (1) **GO:0005925 (focal adhesion)** and (2) **GO:0001666 (response to hypoxia)**. Early in 2014, a systematic network analyses on the transcriptomics profiling of airway smooth muscle tissue confirmed that focal adhesion associated pathways play irreplaceable role for physical or pathological inflammatory effects like asthma related inflammation (Yick et al., 2014). As for another enriched GO term named as hypoxia, similar with focal adhesion, hypoxia has also been shown to be correlated with the inflammatory activation of ASMCs. Based on related transcriptomics studies (Ricciardi et al., 2008; Yang et al., 2014), hypoxia has been confirmed to be directly correlated with dendritic cell mediated inflammatory responses. Therefore, it is also reasonable for us to enrich our optimum genes at transcriptomics level in such GO term.

## GO Enrichment Analyses for Signatures on the Epigenetic and Transcriptomics Levels

As shown in **Supplementary Table 7**, we identified 68 enriched GO terms of different clusters. We chose **GO:0051301 (cell division)** and **GO:0017147 (Wnt-protein binding)** for detailed analyses. For multi-omics data, the GO term seems to be more general. Cell division has been widely shown to be correlated with inflammatory responses in the airway related tissues (McWilliam et al., 1996; Lambrecht et al., 2000; Grausenburger et al., 2010). Therefore, it is reasonable for potential biomarkers distinguishing different ASMCs inflammatory status to enrich in such GO term. As for Wnt-protein binding, WNT and beta-catenin signaling pathway, which involves multiple WNT proteins, has been widely reported to be correlated with the inflammatory responses of ASMC (DiRenzo et al., 2016; Kumawat et al., 2016; Koopmans, 2017).

All in all, as we have discussed above, for the first time, we recognized the functional enrichment pattern of multi-omics biomarkers. Biologically, we identified multi-omics level regulation associated biological entity (functions, processes, or cellular components), laying a foundation for fully demonstration on the inflammatory factor mediated regulations on ASMCs. Methodologically, we confirmed that the application of multi-omics biomarkers for GO enrichment analyses may improve the efficacy and accuracy for disease associated function exploration, providing an alternative approach for pathological studies.

## CONCLUSION

Via multiple machine learning models, we identified a group of signatures for the different statuses of ASMCs on the transcriptomics, epigenetic, or dual-omics level and established several quantitative rules on the multiomics level for the classification of cells with different biological/pathological

statuses. All the qualitative signatures and quantitative rules have been validated by recent publications, confirming the efficacy and accuracy of our analyses. By summarizing the results, we conclude that the use of transcriptomics data may be more appropriate than that of epigenetic data to classify ASMCs under different activation conditions. Moreover, we conclude that the combined use of transcriptomics and epigenetic data is highly effective and accurate for cell classification.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE146377>.

## AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. Y-HZ, ZL, and TZ performed the experiments. Y-HZ, LC, and HL analyzed the results. Y-HZ, ZL, TZ, and HL wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

## FUNDING

This research was funded by the National Key R&D Program of China (2017YFC1201200), Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Key R&D Program of China (2018YFC0910403), National Natural Science Foundation of China (31701151), Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.599970/full#supplementary-material>

**Supplementary Table 1** | MCFS-ranked features for methylation data.

**Supplementary Table 2** | Performance of IFS with RF, SVM, kNN, and RIPPER under different numbers of features for methylation data.

**Supplementary Table 3** | MCFS-ranked features for expression data.

**Supplementary Table 4** | Performance of IFS with RF, SVM, kNN, and RIPPER under different numbers of features for expression data.

**Supplementary Table 5** | MCFS ranked features for combined data.

**Supplementary Table 6** | Performance of IFS with RF, SVM, kNN, and RIPPER under different numbers of features for combined data.

**Supplementary Table 7** | GO enrichment analyses results for signatures on the epigenetic level, transcriptomics level, epigenetic and transcriptomics level.



## REFERENCES

- Aisenberg, W. H., Huang, J., Zhu, W., Rajkumar, P., Cruz, R., Santhanam, L., et al. (2016). Defining an olfactory receptor function in airway smooth muscle cells. *Sci. Rep.* 6:38231.
- Ba, M., Rawat, S., Lao, R., Grous, M., Salmon, M., Halayko, A. J., et al. (2018). Differential regulation of cytokine and chemokine expression by MK2 and MK3 in airway smooth muscle cells. *Pulm. Pharmacol. Ther.* 53, 12–19. doi: 10.1016/j.pupt.2018.09.004
- Baranwal, M., Magner, A., Elvati, P., Saldinger, J., Violi, A., and Hero, A. O. (2019). A deep learning architecture for metabolic pathway prediction. *Bioinformatics* 36, 2547–2553. doi: 10.1093/bioinformatics/btz954
- Bexiga, N., Lam, H., Alencar, A., Stephano, M., and An, S. (2018). “Direct effects of interleukins on airway smooth muscle cell functions,” in *Proceedings of the A29. Novel Mechanisms for Airway Smooth Muscle Contraction and Relaxation: Potential Targets for Modulation*, (New York, NY: American Thoracic Society), A1212.
- Bousquet, J., Jeffery, P. K., Busse, W. W., Johnson, M., and Vignola, A. M. (2000). Asthma: from bronchoconstriction to airways inflammation and remodeling. *Am. J. Respir. Crit. Care Med.* 161, 1720–1745. doi: 10.1164/ajrccm.161.5.9903102
- Breiman, L. (2001). Random forests. *Machine Learn.* 45, 5–32.
- Campbell, L., and Hardman, M. J. (2020). “Differential estrogen receptor-mediated gene profiles define cutaneous homeostasis and injury,” in *The Role of Estrogen and Inflammation in Cutaneous Wound Healing*, ed. B. Andrew (Ann Arbor, MI: ProQuest), 55.
- Chapman, S. J., Khor, C. C., Vannberg, F. O., Rautanen, A., Segal, S., Moore, C. E., et al. (2010). NFKBIZ polymorphisms and susceptibility to pneumococcal disease in European and African populations. *Genes Immun.* 11, 319–325. doi: 10.1038/gene.2009.76
- Chen, L., Chu, C., Zhang, Y.-H., Zheng, M.-Y., Zhu, L., Kong, X., et al. (2017a). Identification of drug-drug interactions using chemical interactions. *Curr. Bioinform.* 12, 526–534.
- Chen, L., Wang, S., Zhang, Y. H., Li, J., Xing, Z. H., Yang, J., et al. (2017b). Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* 5, 26582–26590. doi: 10.1109/access.2017.2775703
- Chung, K. F. (2000). Airway smooth muscle cells: contributing to and regulating airway mucosal inflammation? *Eur. Respir. J.* 15, 961–968. doi: 10.1034/j.1399-3003.2000.15e26.x
- Cieri, R. L. (2019). Pulmonary smooth muscle in vertebrates: a comparative review of structure and function. *Integr. Comp. Biol.* 59, 10–28. doi: 10.1093/icb/icz002
- Cohen, W. W. (1995). “Fast Effective Rule induction,” in *Proceedings of the Twelfth International Conference on Machine Learning*, (San Francisco, CA).
- Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learn.* 20, 273–297.
- Cortes, L. M. P., Ginger, R. S., Gunn, D. A., Nijsten, T. E. C., Sanders, M. G. H., and Smith, A. M. (2020). Prevention and/or treatment of inflammatory skin disease. *Google Patents*
- Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27.
- Dahl, M. J., Veneroni, C., Lavizzari, A., Pillow, J., Yoder, B., and Albertine, K. (2018). Lung mechanics, airway reactivity, and muscularization are altered in former mechanically ventilated preterm lambs. *Eur. Respir. J.* 52: OA303.
- Dai, Q., Andreu-Agullo, C., Insolera, R., Wong, L. C., Shi, S.-H., and Lai, E. C. (2013). BEND6 is a nuclear antagonist of Notch signaling during self-renewal of neural stem cells. *Development* 140, 1892–1902. doi: 10.1242/dev.087502
- Diaconu, C. C., Neagu, A. I., Lungu, R., Tardei, G., Alexiu, I., Bleotu, C., et al. (2010). Plasticity of regulatory T cells under cytokine pressure. *Roum. Arch. Microbiol. Immunol.* 69, 190–196.
- DiRenzo, D. M., Chaudhary, M. A., Shi, X., Franco, S. R., Zent, J., Wang, K., et al. (2016). A crosstalk between TGF- $\beta$ /Smad3 and Wnt/ $\beta$ -catenin pathways promotes vascular smooth muscle cell proliferation. *Cell. Signal.* 28, 498–505. doi: 10.1016/j.cellsig.2016.02.011
- Dong, Y.-M., Li, M., He, Q.-E., Tong, Y.-F., Gao, H.-Z., Zhang, Y.-Z., et al. (2020). Epigenome-Wide tobacco-related methylation signature identification and their multilevel regulatory network inference for lung adenocarcinoma. *BioMed. Res. Int.* 2020:2471915.
- Dramiński, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2007). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486
- Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481. doi: 10.1093/bioinformatics/bth261
- Frohman, E. M., Cruz, R. A., Longmuir, R., Steinman, L., Zamvil, S. S., Villemarette-Pittman, N. R., et al. (2020). Part II. High-dose methotrexate with leucovorin rescue for severe COVID-19: an immune stabilization strategy for SARS-CoV-2 induced ‘PANIC’ attack. *J. Neurol. Sci.* 415:116935. doi: 10.1016/j.jns.2020.116935
- Fu, J.-J., Baines, K. J., Wood, L. G., and Gibson, P. G. (2013). Systemic inflammation is associated with differential gene expression and airway neutrophilia in asthma. *Omic J. Integr. Biol.* 17, 187–199. doi: 10.1089/omi.2012.0104
- Garg, A. V., Amaty, N., Chen, K., Cruz, J. A., Grover, P., Whibley, N., et al. (2015). MCP1P1 endoribonuclease activity negatively regulates interleukin-17-mediated signaling and inflammation. *Immunity* 43, 475–487. doi: 10.1016/j.immuni.2015.07.021
- Glossop, J. R., and Cartmell, S. H. (2009). Effect of fluid flow-induced shear stress on human mesenchymal stem cells: differential gene expression of IL1B and MAP3K8 in MAPK signaling. *Gene Expr. Patterns* 9, 381–388. doi: 10.1016/j.gexp.2009.01.001
- Gongol, B., Marin, T. L., Jeppson, J. D., Mayagoitia, K., Shin, S., Sanchez, N., et al. (2017). Cellular hormetic response to 27-hydroxycholesterol promotes neuroprotection through AICD induction of MAST4 abundance and kinase activity. *Sci. Rep.* 7:13898.
- Göransson, M., Andersson, M. K., Forni, C., Ståhlberg, A., Andersson, C., Olofsson, A., et al. (2009). The myxoid liposarcoma FUS-DDIT3 fusion oncoprotein deregulates NF- $\kappa$ B target genes by interaction with NFKBIZ. *Oncogene* 28, 270–278. doi: 10.1038/nc.2008.378
- Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comp. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006
- Grausenburger, R., Bilic, I., Boucheron, N., Zupkovitz, G., El-Housseiny, L., Tschisnarov, R., et al. (2010). Conditional deletion of histone deacetylase 1 in T cells leads to enhanced airway inflammation and increased Th2 cytokine production. *J. Immunol.* 185, 3489–3497. doi: 10.4049/jimmunol.0903610
- Grozdanovic, M., Laffey, K. G., Abdelkarim, H., Hitchinson, B., Harijith, A., Moon, H.-G., et al. (2019). Novel peptide nanoparticle-biased antagonist of CCR3 blocks eosinophil recruitment and airway hyperresponsiveness. *J. Allergy Clin. Immunol.* 143, 669–680.
- Haas, M., Ward, D., Lee, J., Roses, A., Clarke, V., D’eustachio, P., et al. (1993). Localization of Shaw-related K<sup>+</sup> channel genes on mouse and human chromosomes. *Mamm. Genome* 4, 711–715. doi: 10.1007/bf00357794
- Hartley, A.-V. A. (2020). *Regulation of Protein Arginine Methyl Transferase 5 by Novel Serine 15 Phosphorylation in Colorectal Cancer*. Bloomington, IN: Indiana University.
- Higham, A., Wolosianka, S., Beech, A., Jackson, N., Long, G., Kolsum, U., et al. (2020). “Type 2 inflammation in eosinophilic chronic obstructive pulmonary disease,” in *Proceedings of the C31. Copd Basic Mechanisms*, (New York, NY: American Thoracic Society), A4731.
- Hu, R., Pan, W., Fedulov, A. V., Jester, W., Jones, M. R., Weiss, S. T., et al. (2014). MicroRNA-10a controls airway smooth muscle cell proliferation via direct targeting of the PI3 kinase pathway. *FASEB J.* 28, 2347–2357. doi: 10.1096/fj.13-247247
- Huang, T., Hazen, M., Shang, Y., Zhou, M., Wu, X., Yan, D., et al. (2016). Depletion of major pathogenic cells in asthma by targeting CRTh2. *JCI Insight* 1:e86689.
- Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-based machine learning model for predicting the metabolic pathways of compounds. *IEEE Access* 8, 130687–130696. doi: 10.1109/access.2020.3009439
- Jiang, H., Xie, Y., Abel, P. W., Toews, M. L., Townley, R. G., Casale, T. B., et al. (2012). Targeting phosphoinositide 3-kinase  $\gamma$  in airway smooth muscle cells to suppress interleukin-13-induced mouse airway hyperresponsiveness. *J. Pharmacol. Exp. Ther.* 342, 305–311. doi: 10.1124/jpet.111.189704
- Kamijo, H., Miyagaki, T., Hayashi, Y., Akatsuka, T., Watanabe-Otobe, S., Oka, T., et al. (2020). Increased IL-26 expression promotes T helper type 17 and T helper type 2-associated cytokine production by keratinocytes in atopic dermatitis. *J. Invest. Dermatol.* 140, 636–644. doi: 10.1016/j.jid.2019.07.713



- Khalifeh-Soltani, A., Gupta, D., Ha, A., Podolsky, M. J., Datta, R., and Atabai, K. (2018). The Mfge8- $\alpha$ 8 $\beta$ 1-PTEN pathway regulates airway smooth muscle contraction in allergic inflammation. *FASEB J.* 32, 5927–5936. doi: 10.1096/fj.201800109r
- Kim, K., Kim, G., Kim, J.-Y., Yun, H. J., Lim, S.-C., and Choi, H. S. (2014). Interleukin-22 promotes epithelial cell transformation and breast tumorigenesis via MAP3K8 activation. *Carcinogenesis* 35, 1352–1361. doi: 10.1093/carcin/bgu044
- Kim, S., Forno, E., Zhang, R., Yan, Q., Boutaoui, N., Acosta-Perez, E., et al. (2020). Expression quantitative trait methylation analysis reveals methylomic associations with gene expression in childhood asthma. *Chest* 158, 1841–1856. doi: 10.1016/j.chest.2020.05.601
- Kohavi, R. (1995). “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the International joint Conference on artificial intelligence*, (Mahwah, NJ: Lawrence Erlbaum Associates Ltd), 1137–1145.
- Koopmans, T. (2017). *WNT and  $\beta$ -Catenin Signalling in Airway Smooth Muscle: Emerging Concepts for Asthma*. Groningen: University of Groningen.
- Koziol-White, C., Parikh, V., Chupp, G., and Panettieri, R. (2020). “Blocking YKL-40 (Chitinase-Like Protein Chitinase 3) Reverses IL-13/IL-4-induced hyporesponsiveness to bronchodilators in human small airways and in human airway smooth muscle cells,” in *Proceedings of the A30. Contract and Relax: What's New in Airway Smooth Muscle Mechanisms*, (New York, NY: American Thoracic Society), A1252.
- Kumawat, K., Koopmans, T., Menzen, M. H., Prins, A., Smit, M., Halayko, A. J., et al. (2016). Cooperative signaling by TGF- $\beta$ 1 and WNT-11 drives sm- $\alpha$ -actin expression in smooth muscle via Rho kinase-actin-MRTF-A signaling. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 311, L529–L537.
- Lam, M., Lamanna, E., and Bourke, J. E. (2019). Regulation of airway smooth muscle contraction in health and disease. *Adv. Exp. Med. Biol.* 1124, 381–422.
- Lambrecht, B. N., De Veerman, M., Coyle, A. J., Gutierrez-Ramos, J.-C., Thielemans, K., and Pauwels, R. A. (2000). Myeloid dendritic cells induce Th2 responses to inhaled antigen, leading to eosinophilic airway inflammation. *J. Clin. Invest.* 106, 551–559. doi: 10.1172/jci8107
- Lee, J. H., Kaminski, N., Dolganov, G., Grunig, G., Koth, L., Solomon, C., et al. (2001). Interleukin-13 induces dramatically different transcriptional programs in three human airway cell types. *Am. J. Respir. Cell Mol. Biol.* 25, 474–485. doi: 10.1165/ajrcmb.25.4.4522
- Li, B., Zhang, J., Wang, Z., and Chen, S. (2016). Ivabradine prevents low shear stress induced endothelial inflammation and oxidative stress via mTOR/eNOS pathway. *PLoS One* 11:e0149694. doi: 10.1371/journal.pone.0149694 doi: 10.1371/journal.pone.0149694
- Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of drug side effects with a refined negative sample selection strategy. *Comp. Math. Methods Med.* 2020:1573543.
- Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230.
- Lyles, J., and Rothenberg, M. (2019). Role of genetics, environment, and their interactions in the pathogenesis of eosinophilic esophagitis. *Curr. Opin. Immunol.* 60, 46–53. doi: 10.1016/j.coi.2019.04.004
- Mamber, S. W., Gurel, V., Lins, J., Ferri, F., Beseme, S., and McMichael, J. (2020). Effects of cannabis oil extract on immune response gene expression in human small airway epithelial cells (HSAEPc): implications for chronic obstructive pulmonary disease (COPD). *J. Cannabis Res.* 2:5.
- Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA) Protein Struct.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9
- McWilliam, A. S., Napoli, S., Marsh, A. M., Pemper, F. L., Nelson, D. J., Pimm, C. L., et al. (1996). Dendritic cells are recruited into the airway epithelium during the inflammatory response to a broad spectrum of stimuli. *J. Exp. Med.* 184, 2429–2432. doi: 10.1084/jem.184.6.2429
- Min, J. Y., Schleimer, R., and Tan, B. (2020). Inhibition of the non-gastric H<sup>+</sup>/K<sup>+</sup>-ATPase (ATP12A) by ilaprazole and vonoprazan decreased IL-13-stimulated eotaxin-3 expression in airway epithelial cells. *J. Allergy Clin. Immunol.* 145:AB184.
- Nadadur, S. S., Costa, D. L., Slade, R., Silbjörns, R., and Hatch, G. E. (2005). Acute ozone-induced differential gene expression profiles in rat lung. *Environ. Health Perspect.* 113, 1717–1722. doi: 10.1289/ehp.7413
- Okamoto, M., Takeda, K., Joetham, A., Ohnishi, H., Matsuda, H., Swasey, C. H., et al. (2008). Essential role of Notch signaling in effector memory CD8<sup>+</sup> T cell-mediated airway hyperresponsiveness and inflammation. *J. Exp. Med.* 205, 1087–1097. doi: 10.1084/jem.20072200
- Pascoe, C. D., Ragheb, M., Stelmack, G., Jha, A., and Halayko, A. J. (2017). “Oxidized phosphatidylcholine induces COX2 gene expression and cytokine secretion by human airway smooth muscle cells,” in *Proceedings of the C108. Getting Inflamed: Markers Of Lung Injury And Remodelling*, (New York, NY: American Thoracic Society), A6944.
- Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14. Microsoft Research, Redmond, United States.
- Plé, C., Fan, Y., Yahia, S. A., Vornig, H., Everaere, L., Chenivesse, C., et al. (2015). Polycyclic aromatic hydrocarbons reciprocally regulate IL-22 and IL-17 cytokines in peripheral blood mononuclear cells from both healthy and asthmatic subjects. *PLoS One* 10:e0122372. doi: 10.1371/journal.pone.0122372
- Qin, W.-D., Wei, S.-J., Wang, X.-P., Wang, J., Wang, W.-K., Liu, F., et al. (2013). Poly (ADP-ribose) polymerase 1 inhibition protects against low shear stress induced inflammation. *Biochim. Biophys. Acta (BBA) Mol. Cell Res.* 1833, 59–68.
- Rajakulendran, S., Roberts, J., Koltzenburg, M., Hanna, M. G., and Stewart, H. (2013). Deletion of chromosome 12q21 affecting KCNC2 and ATXN7L3B in a family with neurodevelopmental delay and ataxia. *J. Neurol. Neurosurg. Psychiatry* 84, 1255–1257.
- Ramakrishnan, R. K., Al Heialy, S., and Hamid, Q. (2019). Role of IL-17 in asthma pathogenesis and its implications for the clinic. *Expert Rev. Respir. Med.* 13, 1057–1068.
- Reyes-García, J., Flores-Soto, E., Carbajal-García, A., Sommer, B., and Montaña, L. M. (2018). Maintenance of intracellular Ca<sup>2+</sup> basal concentration in airway smooth muscle. *Int. J. Mol. Med.* 42, 2998–3008.
- Ricciardi, A., Elia, A. R., Cappello, P., Puppo, M., Vanni, C., Fardin, P., et al. (2008). Transcriptome of hypoxic immature dendritic cells: modulation of chemokine/receptor expression. *Mol. Cancer Res.* 6, 175–185.
- Salter, B., Pray, C., Radford, K., Martin, J. G., and Nair, P. (2017). Regulation of human airway smooth muscle cell migration and relevance to asthma. *Respir. Res.* 18:156.
- Sánchez, Á., Relación, C., Carrasco, A., Contreras-Jurado, C., Martín-Duce, A., Aranda, A., et al. (2017). Map3k8 controls granulocyte colony-stimulating factor production and neutrophil precursor proliferation in lipopolysaccharide-induced emergency granulopoiesis. *Sci. Rep.* 7:5010.
- Sang, X., Xiao, W., Zheng, H., Yang, Y., and Liu, T. (2020). HMMPred: accurate prediction of DNA-binding proteins based on HMM Profiles and XGBoost feature selection. *Comp. Math. Methods Med.* 2020:1384749.
- Sangaphunchai, P., Todd, I., and Fairclough, L. C. (2020). Extracellular Vesicles and Asthma: a review of the literature. *Clin. Exp. Allergy* 50, 291–307.
- Sotelo, N. S., Valiente, M., Gil, A., and Pulido, R. (2012). A functional network of the tumor suppressors APC, hDlg, and PTEN, that relies on recognition of specific PDZ—domains. *J. Cell. Biochem.* 113, 2661–2670.
- Sun, L., Gu, S., Li, X., Sun, Y., Zheng, D., Yu, K., et al. (2006). Identification of a novel human MAST4 gene, a new member of the microtubule associated serine-threonine kinase family. *Mol. Biol.* 40, 808–815.
- Sungnak, W., Huang, N., Bécavin, C., Berg, M., Queen, R., Litvinukova, M., et al. (2020). SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* 26, 681–687.
- Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation Markers. *Bioinformatics* 34, 398–406.
- Theilhaber, J., Connolly, T., Roman-Roman, S., Bushnell, S., Jackson, A., Call, K., et al. (2002). Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome Res.* 12, 165–176.
- Thompson, E., Dang, Q., Mitchell-Handley, B., Rajendran, K., Ram-Mohan, S., Solway, J., et al. (2018). “Mapping human airway smooth muscle cell transcriptional and epigenetic responses to asthma-promoting cytokines reveals enrichments for asthma-associated SNPs,” in *Proceedings of the D54. Interplay of Diverse Cellular and Molecular Pathways in Asthma and Airway Disease*, (New York, NY: American Thoracic Society), A7181.
- Thompson, E., Dang, Q., Mitchell-Handley, B., Rajendran, K., Ram-Mohan, S., Solway, J., et al. (2019). “Primary airway smooth muscle cells from subjects

- with and without asthma reveal distinct differences in contractile, epigenetic, and transcriptional responses to the asthma-promoting cytokines IL-13+ IL-17,” in *Proceedings of the B62. Asthma Mechanisms*, (New York, NY: American Thoracic Society), A3822.
- Tliba, O., and Panettieri, R. A. Jr. (2019). Paucigranulocytic asthma: uncoupling of airway obstruction from inflammation. *J. Allergy Clin. Immunol.* 143, 1287–1294.
- Tsay, J.-C. J., Li, Z., Yie, T.-A., Wu, F., Segal, L., Greenberg, A. K., et al. (2015). Molecular characterization of the peripheral airway field of cancerization in lung adenocarcinoma. *PLoS One* 10:e0118132. doi: 10.1371/journal.pone.0118132
- Valiente, M., Andrés-Pons, A., Gomar, B., Torres, J., Gil, A., Tapparel, C., et al. (2005). Binding of PTEN to specific PDZ domains contributes to PTEN protein stability and phosphorylation by microtubule-associated serine/threonine kinases. *J. Biol. Chem.* 280, 28936–28943.
- Vermeulen, C. J., Xu, C.-J., Vonk, J. M., Ten Hacken, N. H., Timens, W., Heijink, I. H., et al. (2020). Differential DNA methylation in bronchial biopsies between persistent asthma and asthma in remission. *Eur. Respir. J.* 55:1901280.
- Vetri, L., Cali, F., Vinci, M., Amato, C., Roccella, M., Granata, T., et al. (2020). A de novo heterozygous mutation in KCNC2 gene implicated in severe developmental and epileptic encephalopathy. *Eur. J. Med. Genet.* 63:103848.
- Vykhovanets, E. V., Resnick, M. I., and Marengo, S. R. (2006). Intraprostatic lymphocyte profiles in aged wistar rats with estradiol induced prostate inflammation. *J. Urol.* 175, 1534–1540.
- Witten, I. H., and Frank, E. (eds) (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann.
- Wolde, M., Laan, L. C., Medhin, G., Gadissa, E., Berhe, N., and Tsegaye, A. (2020). Human monocytes/macrophage inflammatory cytokine changes following in vivo and in vitro *Schistosoma mansoni* infection. *J. Inflamm. Res.* 13:35.
- Xie, B., Laxman, B., Hashemifar, S., Stern, R., Gilliam, T. C., Maltsev, N., et al. (2018). Chemokine expression in the early response to injury in human airway epithelial cells. *PLoS One* 13:e0193334. doi: 10.1371/journal.pone.0193334
- Yang, W., Ramachandran, A., You, S., Jeong, H., Morley, S., Mulone, M. D., et al. (2014). Integration of proteomic and transcriptomic profiles identifies a novel PDGF-MYC network in human smooth muscle cells. *Cell Commun. Signal.* 12:44.
- Yick, C., Zwinderman, A., Kunst, P., Grünberg, K., Mauad, T., Chowdhury, S., et al. (2014). Gene expression profiling of laser microdissected airway smooth muscle tissue in asthma and atopy. *Allergy* 69, 1233–1240.
- Yu, Z., Chen, H., Liuxs, J., You, J., Leung, H., and Han, G. (2016). Hybrid k -nearest neighbor classifier. *IEEE Trans. Cybern.* 46, 1263–1275.
- Zhang, B., and Srihari, S. N. (2004). Fast k-nearest neighbor classification using cluster-based trees. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 525–528.
- Zhang, W., Wu, Y., Huang, Y., and Gunst, S. (2019). “The proprotein convertase furin suppresses the inflammatory responses of airway smooth muscle (ASM) Tissues to IL-13 by modulating integrin-mediated signaling pathways,” in *Proceedings of the B29. Mechanisms for Airway Hyperresponsiveness: From Cell to Organism*, (New York, NY: American Thoracic Society), A2847.
- Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinform.* 14, 709–720.
- Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144.
- Zheng, H., Wu, X., Wu, D., Jiang, R.-L., Castillo, E. F., Chock, C. J., et al. (2020). Treg expression of CIS suppresses allergic airway inflammation through antagonizing an autonomous TH2 program. *Mucosal Immunol.* 13, 293–302.
- Zhou, J.-P., Chen, L., and Guo, Z.-H. (2020a). iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. *Bioinformatics* 36, 1391–1396.
- Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020b). iATC-FRAKEL: a simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only. *Bioinformatics* 36, 3568–3569.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Zhang, Li, Zeng, Chen, Li, Huang and Cai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Deciphering the Subtype Differentiation History of SARS-CoV-2 Based on a New Breadth-First Searching Optimized Alignment Method Over a Global Data Set of 24,768 Sequences

## OPEN ACCESS

### Edited by:

Indrajit Saha,  
National Institute of Technical  
Teachers' Training and Research,  
India

### Reviewed by:

Rachel Graham,  
University of North Carolina at Chapel  
Hill, United States  
Manoj Kumar,  
Institute of Microbial Technology  
(CSIR), India  
Marina Muñoz,  
Universidad del Rosario, Colombia

### \*Correspondence:

Lan Ma  
malan@sz.tsinghua.edu.cn

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 13 August 2020

**Accepted:** 04 December 2020

**Published:** 11 January 2021

### Citation:

Lin QY, Huang Y, Jiang Z, Wu F  
and Ma L (2021) Deciphering  
the Subtype Differentiation History  
of SARS-CoV-2 Based on a New  
Breadth-First Searching Optimized  
Alignment Method Over a Global Data  
Set of 24,768 Sequences.  
Front. Genet. 11:591833.  
doi: 10.3389/fgene.2020.591833

Qianyu Lin<sup>1</sup>, Yunchuanxiang Huang<sup>1</sup>, Ziyi Jiang<sup>2</sup>, Feng Wu<sup>2</sup> and Lan Ma<sup>1,2,3\*</sup>

<sup>1</sup> Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, China, <sup>2</sup> Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, <sup>3</sup> Shenzhen Bay Laboratory, Shenzhen, China

SARS-CoV-2 has caused a worldwide pandemic. Existing research on coronavirus mutations is based on small data sets, and multiple sequence alignment using a global-scale data set has yet to be conducted. Statistical analysis of integral mutations and global spread are necessary and could help improve primer design for nucleic acid diagnosis and vaccine development. Here, we optimized multiple sequence alignment using a conserved sequence search algorithm to align 24,768 sequences from the GISAID data set. A phylogenetic tree was constructed using the maximum likelihood (ML) method. Coronavirus subtypes were analyzed via t-SNE clustering. We performed haplotype network analysis and t-SNE clustering to analyze the coronavirus origin and spread. Overall, we identified 33 sense, 17 nonsense, 79 amino acid loss, and 4 amino acid insertion mutations in full-length open reading frames. Phylogenetic trees were successfully constructed and samples clustered into subtypes. The COVID-19 pandemic differed among countries and continents. Samples from the United States and western Europe were more diverse, and those from China and Asia mainly contained specific subtypes. Clades G/GH/GR are more likely to be the origin clades of SARS-CoV-2 compared with clades S/L/V. Conserved sequence searches can be used to segment long sequences, making large-scale multisequence alignment possible, facilitating more comprehensive gene mutation analysis. Mutation analysis of the SARS-CoV-2 can inform primer design for nucleic acid diagnosis to improve virus detection efficiency. In addition, research into the characteristics of viral spread and relationships among geographic regions can help formulate health policies and reduce the increase of imported cases.

**Keywords:** SARS-CoV-2, multiple sequence alignment, phylogenetic tree, t-SNE, haplotype network analysis

## INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus, which is the etiologic agent of the disease, coronavirus disease 2019 (COVID-19) (Li et al., 2020). SARS-CoV-2 emerged in late 2019 in Hubei Province, China (Chen et al., 2020; Zhou et al., 2020), and spread worldwide with incredible rapidity, resulting in a global pandemic. As of October 13, 2020, more than 38 million people have been infected worldwide with approximately 1,090,000 deaths. The number of newly diagnosed cases has increased dramatically with tens of thousands confirmed daily. In Europe, the case fatality rate exceeded 7%, and those in France and Belgium have reached unprecedented levels at 24.5 and 33.4%, respectively (World Health Organization, 2020; Worldometer, 2020).

Analysis of virus mutation sites is necessary for applications, including vaccine development (Ma et al., 2020) and primer design for virus nucleic acid detecting. It is reported that conserved sequence-based mRNA vaccines (Frey et al., 2020) and peptide vaccines (Herrera-Rodriguez et al., 2018) have successfully made the vaccinated generate immunity to multistrains of the same virus. The conserved sequences have great potential in long-acting vaccine design. Multiple sequence alignment (MSA) methods are invariably used for automated identification of mutation sites and widely used in SARS-CoV-2 sequence analysis (Lai et al., 2020; Wu et al., 2020) in the early stage of the pandemic. With the fast increase of SARS-CoV-2 sequencing data, it is significant to improve the efficiency of the current MSA algorithms to fit the large-scale data set. We developed a new method for conserved sequence searching. Large data sets containing long sequences, such as the SARS-CoV-2 data set, can be optimized by pruning conserved sequences to fit current MSA algorithms. MSA methods invariably detect conserved sequences, and, using our approach, conserved sequence identification is independent of MSA.

Using data from the GISAID database (Elbe and Buckland-Merrett, 2017), we analyzed COVID-19 strains from around the world on an unprecedented scale. All the mutations in SARS-CoV-2, including 33 sense, 17 nonsense, and amino acid loss/insertion mutations, were identified using MSA. Further, based on the results of MSA, we constructed phylogenetic trees and used the t-SNE method to cluster SARS-CoV-2 subtypes. Our findings demonstrate the characteristics of viral spread and uncover relationships among countries and continents.

## MATERIALS AND METHODS

### Data Source and Data Selection

The SARS-CoV-2 sequences used in this study were all collected from the GISAID (Elbe and Buckland-Merrett, 2017) database and were download on May 14, 2020.

To identify mutations in full-length sequences and determine global spread relationships, the download parameters were set as, “complete(>29,000 bp)” and a total of 24,768 sequences were retrieved. According to codon table and DNA translation rules, sequences were compared with annotations of NC\_045512.2

from NCBI (Benson et al., 2018) and high-quality open reading frame (ORF) regions with no degenerate bases (including N) translated into amino acid sequences for each record. The number of sequences for each ORF are shown in **Supplementary Table 1**. Further, 9,308 sequences with 12 full-length, high-quality ORF regions and a clear collection date were available for use in building phylogenetic trees and t-distributed stochastic neighbor embedding (t-SNE). The coding language used was Matlab (R2020a for windows).

### Conserved Sequence Searching

A new strategy to evaluate conserved sequences was developed based on the breadth-first search algorithm. The search queue “q” was initiated using 20 one-length protein segments (single amino acid). Considering  $q_i$  as the  $i$ th string in the queue and  $d_j$  as the  $j$ th sample sequence of the data set,  $p_i$ , Eq. 1 was used to evaluate the probability of conservation:

$$p_i = \frac{\sum_{j=1}^n f(q_i, d_j)}{n}, f(q_i, d_j) = \begin{cases} 1 & \text{if } q_i \text{ and } d_j \text{ match} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The conserved sequence search algorithm (pseudo code in **Supplementary Data 1**) is shown in **Figure 1**.

The match function, “strcmp,” was applied in Matlab. In another coding language, the KMP (Knuth et al., 1977) algorithm can improve the string matching speed. In this approach, the conserved sequences in queue  $q$  with length >5 can avoid ectopic repeats.

### Multiple Sequence Alignment (MSA) and Mutant Site Analysis

Multiple sequence alignment was conducted for each ORF data set. Identification of conserved sequences can efficiently separate long sequences into short segments, which can greatly reduce the alignment time cost so that it reaches a tolerable level. The MSA function used was “multialign” in Matlab. Statistical analysis of mutant sites was conducted directly using aligned sequences.

### Phylogenetic Tree Construction and Haplotype Network Analysis

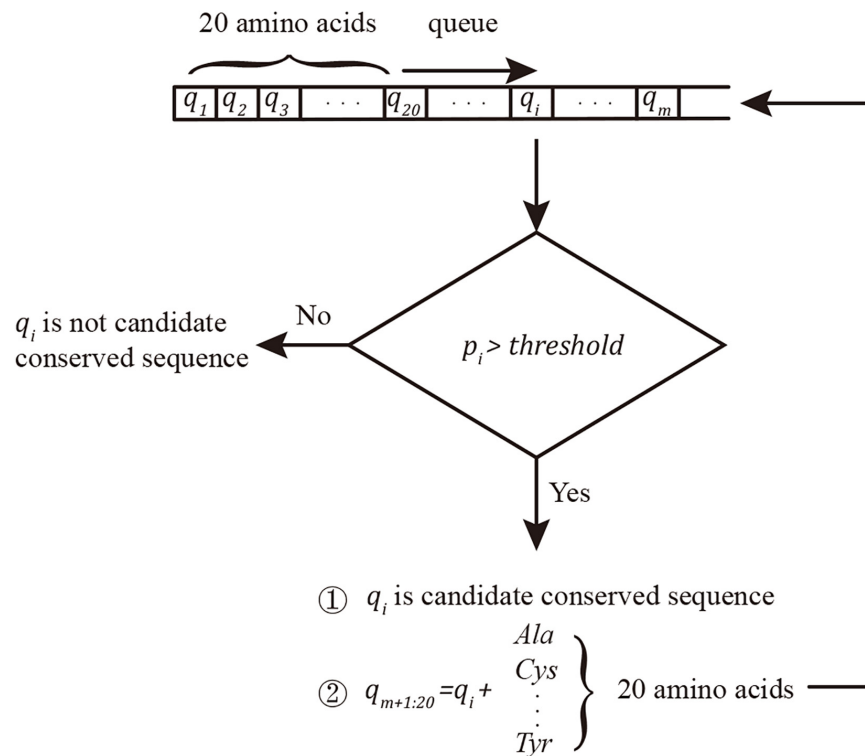
We constructed the maximum likelihood (ML) tree using Raxml-ng (Kozlov et al., 2019) (v.0.8.0 BETA) software.

Sample sequences ( $n = 9,308$ ) with full-length translated ORF and clear collection date were selected; however, these were still too long [length > 7,000 nucleotides (nt)] for use in MSA; in general, samples with long sequences and large data sets lead to excessively high time costs for phylogenetic tree construction. Therefore, sequence pruning was necessary. According to Shannon’s information theory (Shannon, 1948), the information entropy of any segment in sequence can be calculated by Eq. 2:

$$H_i = p_i \times \log(p_i) \quad (2)$$

Clearly, if conserved sequence  $q_i$  with  $p_i \rightarrow 100\%$  has information entropy  $H_i \rightarrow 0$ , it can be easily proven that





**FIGURE 1 |** Conserved sequence searching algorithm. The queue was initiated by 20 amino acids.  $q_i$  was considered a conserved sequence when  $p_i \geq \text{threshold}$ ; in this project, the threshold was set as 100%. Twenty new conserved sequence candidates were built by adding new amino acids to the end of the confirmed candidate conserved sequences and added to the end of the searching queue  $q$ .

deleting  $q_i$  from all sequences will not substantially influence the results of phylogenetic analysis.

Considering the limitations of phylogeny performance time cost and visualization, we pruned sequences by deleting highly conserved bases with  $p_i > 99.9\%$ . Pruned sequences with exactly the same sequence were reduced to 1 as a representative, resulting in a final selected 1,291 samples.

We performed 50 tree searches using 25 random and 25 parsimony-based starting trees on each DNA data substitution matrix in Raxml-ng, and we got a ML and lowest AIC/AICs/BIC score with GTR + GA model. One thousand bootstrap replicates with seed 2020 were conducted and the transfer bootstrap expectation (tbe) metric was calculated to map onto the best-scoring ML tree to generate proportional support values.

Phylogenetic trees were visualized using iTOL (Letunic and Bork, 2019). Larger clade naming rules refer from GISAID (S: C8782T, T28144C; L: C241, C3037, A23403, C8782, G11083, G25563, G26144, T28144, G28882; V: G11083T, G26144T; G: C241T, C3037T, A23403G; GH: C241T, C3037T, A23403G, G25563T; GR: C241T, C3037T, A23403G, and G28882A). It is worth noting that the marker variant C241T for clade identification is not included in ORF region. We count the C241T base frequency in each haplotype and give the C241T base information lost samples an inferred subtype if 100% frequency base exist; otherwise, the samples will be labeled as “Other.”

The haplotype map with median-joining network (Bandelt et al., 1999) was created by PopART (version 1.7) (Leigh and Bryant, 2015), and 9,308 full ORF region sequenced samples are identified into 300 haplotypes by 77 variant sites ( $p_i < 99.5\%$ ). For better visualization and clearer topology of the haplotype network, we deleted the haplotypes with a single case, and in total, 153 haplotypes are used in haplotype network construction.

## t-Distributed Stochastic Neighbor Embedding (t-SNE)

Samples with 12 high-quality, full-length ORF regions, and a clear collection date ( $n = 9,308$ ) were subjected to t-SNE unsupervised clustering. The t-SNE function used was “tsne” in Matlab. Results were visualized using the “gscatter” function in Matlab. The distance function used was the PAM250 matrix (for amino acid sequence) and BLOSUM45 matrix (for nucleotide sequence). For each aligned amino acid sequence, each amino acid was considered as a dimension of the sample. The distance between sample  $S_i$  and  $S_j$  was calculated using Eqs 3 and 4:

$$\text{Distance}_{i,j} = \frac{\sum_{k=1}^{\text{length}(S)} \text{PAM250}(S_i(k), S_j(k)) \wedge 2}{\text{length}(S)} \quad (3)$$

$$\text{Distance}_{i,j} = \frac{\sum_{k=1}^{\text{length}(S)} \text{BLOSUM45}(S_i(k), S_j(k)) \wedge 2}{\text{length}(S)} \quad (4)$$



In Eq. 3,  $S_i(k)$  is the  $k$ th amino acid of the amino acid sequence  $S_i$ . In Eq. 4,  $S_i(k)$  is the  $k$ th base of the nucleotide sequence  $S_i$ . To reduce overlap, the final coordinate of each sample was adjusted to a short radius from the origin position, which did not influence cluster information.

## RESULTS

### Mutations of SARS-CoV-2 ORF Regions

SARS-CoV-2 full-length nucleotide sequences ( $n = 24,768$ ) were collected from GISAID up to May 14, 2020, and translated into amino acid sequences. Due to the presence of degenerate bases, the size of the available high-quality amino acid sequence data set for analysis of mutant sites was  $<24,768$ . The final sequence set size was as follows: ORF1a ( $n = 16,863$ ), ORF1b ( $n = 14,252$ ), S ( $n = 16,851$ ), ORF3a ( $n = 23,390$ ), E ( $n = 24,344$ ), M ( $n = 23,513$ ), ORF6 ( $n = 24,199$ ), ORF7a ( $n = 21,690$ ), ORF7b ( $n = 21,953$ ), ORF8 ( $n = 24,288$ ), N ( $n = 23,176$ ), and ORF10 ( $n = 24,043$ ).

According to the MSA of each ORF region (available in **Supplementary Data 2**), 50 mutation sites with frequencies  $>1\%$  were detected, including 33 sense and 17 nonsense mutation sites (**Supplementary Tables 2, 3** and **Figure 2**).

Mutations were present in the ORF1ab, S, ORF3a, M, ORF8, and N regions with more than half of mutations in the ORF1ab region. Given the differences in length of ORFs, ORFs with a higher proportion of mutations (number of mutation sites/ORF length) were ORF8 (2.48%), ORF3a (1.09%), and N (1.67%). ORF6, ORF7a, and ORF10 were completely conserved across the entire length of the ORF region. Notably, the S region, which is the SARS-CoV-2 antigen recognition protein, contained only one sense mutation site D614G encoded by A23403G; hence, current data indicate that the S region is highly conserved. Although the D614G spike protein variant has proved it is more infectious than D614 strains (Korber et al., 2020; Yurkovetskiy et al., 2020), it is equally sensitive to neutralization by monoclonal antibodies targeting the receptor-binding domain (Yurkovetskiy et al., 2020).

Amino acid loss and insertion mutations are listed in **Supplementary Tables 4, 5**, respectively. The location referred to in these tables is based on the NC\_045512.2 nucleotide sequence as a prototype. As shown, the majority of amino acid loss and insertion mutations only occurred in a single sample although loss mutation No. 7 and insertion mutation No. 1 had higher frequencies than other mutations of this type.

### Phylogenetic Trees and Haplotype Analysis

The ML phylogenetic tree is shown in **Figure 3**. The complete phylogenetic tree in normal format with bootstrap support value and leaf labels is shown in **Supplementary Figure 1**. The large clade branches have high tbe-supported values ( $>0.75$ ). Deeper branches' tbe-supported values are sometimes lower. We have similar main group results as the neighbor-joining (NJ) tree in GISAID; what is different from GISAID's NJ tree is that our results show clades G/GH/GR are closer to the root than clades S/L/V.

**Figure 4** depicts the median-joining network haplotype result. Haplotypes L1, S6, V3, G1, GR1, and GH1 are the biggest haplotypes of their belonging clades. Clade S and clade G play important roles in coronavirus strain differentiation. Compared with big haplotypes in clades L/S/V, big haplotypes in clades G/GH/GR have more connections to other haplotypes, consistent with the fact that wider spreading will inevitably provide more mutant opportunities and, thus, lead to more sub-haplotypes and haplotype connections.

### t-SNE Unsupervised Clustering Reveals International Spread Relationships

The t-SNE method is widely used in single-cell RNA sequencing investigations to cluster different cell types. In this study, t-SNE was used to cluster coronavirus gene subtypes according to amino acid sequence.

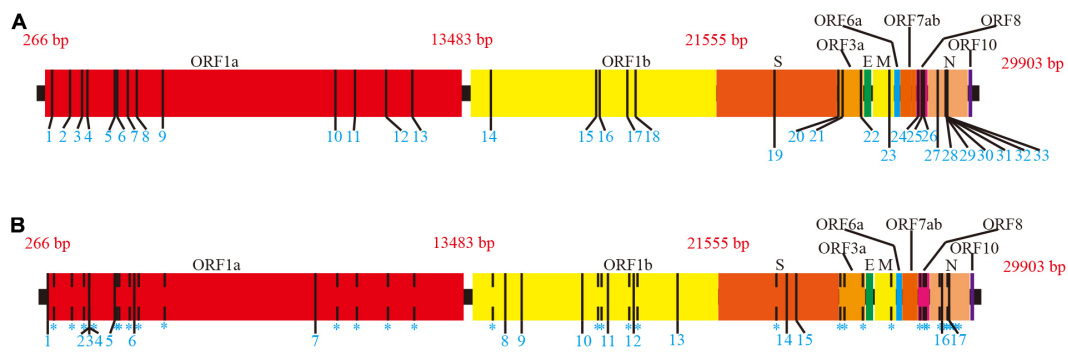
Each cluster can be considered as a subtype, and labeling the samples according to current clade naming rules, t-SNE clustering does have good performance in sequence subtype identification, at both the nucleotide (**Figure 5B**) and amino-acid levels (**Figure 5D**). Samples in the same subtype are more closely related in terms of spread characteristics. Labeling the samples by geographical information (**Figures 5A,C**), as the figures show, cases from China were mainly concentrated in cluster a, and cases from the United States were present in all main clusters. Most of the smaller clusters as well as most cases in cluster b were from the United States. Clustering of western European cases coincided with those from the United States, indicating that their spread relationships were closer than those of others. Compared with other countries/continents, cases in the United States and western Europe appeared to include more clusters, indicating more sources of spread or a longer history of mutation accumulation.

Cluster development and the process of COVID-19 spread in recent months are shown in **Figure 6**. Cluster b contained only cases from the United States in the early stage of the pandemic, and it also contained other North American and Oceania cases in subsequent months. Cases in cluster a showed a limited increase, and those in clusters b, c, d, and e have grown rapidly. Compared with the early stages of the pandemic, the number of clusters has not increased substantially with the main mutations in SARS-CoV-2 occurring before March 18, 2020.

## DISCUSSION

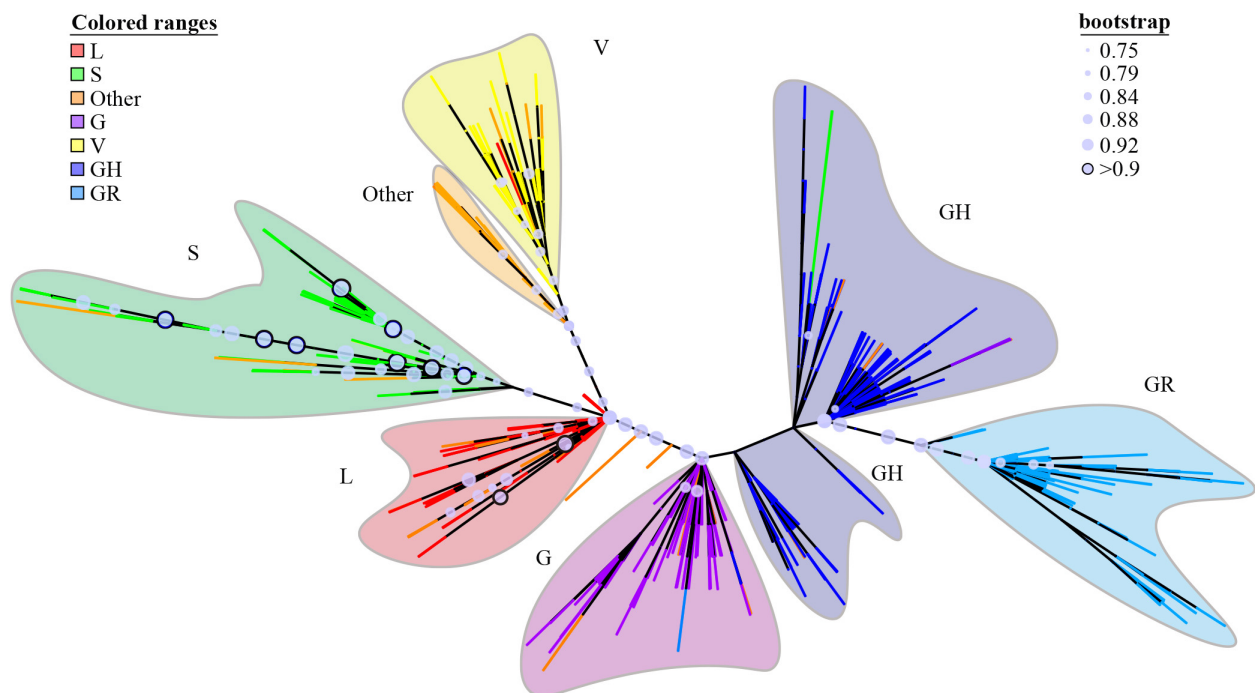
### Conserved Sequence Searching and MSA Optimization

Traditional conserved sequence analyses rely on MSA tools, such as ClustalW (Larkin et al., 2007), MUSCLE (Edgar, 2004), and T-coffee (Di Tommaso et al., 2011). ClustalW calculates a distance matrix by pairwise alignment, builds a guide tree, and makes progressive alignment based on the guide tree. It is the most widely used tool for MSA, but it is also the slowest. T-coffee generates more accurate results than other methods; however, it is more applicable for small data sets



**FIGURE 2 |** Mutations of full length on SARS-CoV-2 ORF region. The NC\_045512.2 SARS-CoV-2 sequence from NCBI was used as the reference. Adjacent ORF areas are distinguished by different colors. **(A)** Sense mutation sites. Mutation sites are marked with a thin black vertical line. Details of sense mutant sites are presented in **Supplementary Table 2**. **(B)** Nonsense mutation sites. Sense mutation sites are indicated by dashed lines for comparison. Nonsense mutation sites are marked as thin, black vertical lines. Details of nonsense mutant sites are presented in **Supplementary Table 3**.

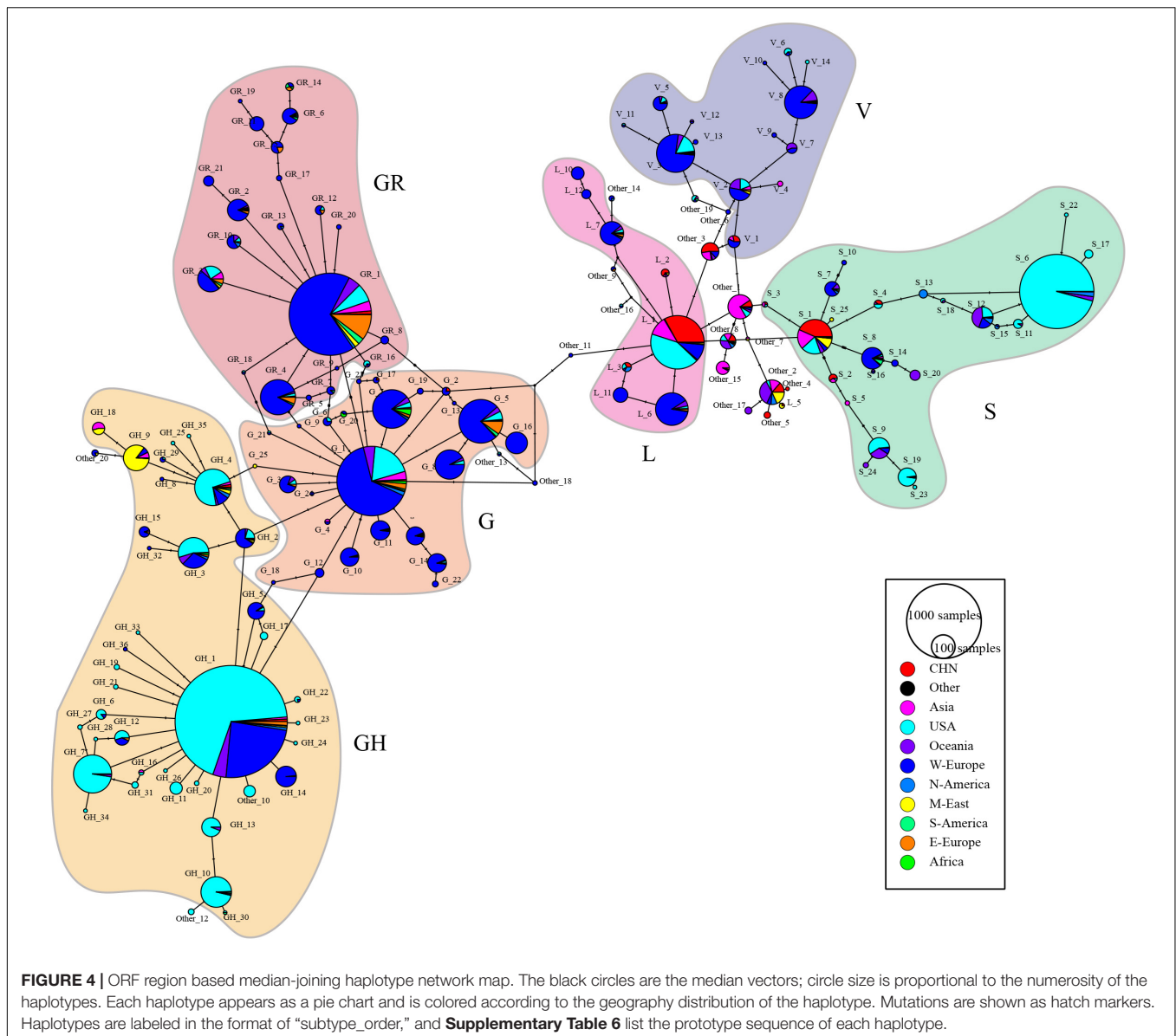
Tree scale: 0.01



**FIGURE 3 |** ORF region based maximum likelihood (ML) phylogenetic tree. The ML tree is displayed in unrooted mode; the deepest branches are colored to represent different subtypes. Major lineages are colored and named. Bootstrap support values are indicated by circles on nodes for support of 0.75 and above. The circles with bootstrap support values over 0.9 are highlighted by a black border. Label information is present in **Supplementary Figure 1**, and in Newick format in **Supplementary Data 3**. Configuration files for iTOL visualization are also in **Supplementary Data 3**.

( $n < 100$  advised) and short-length sequence data, and the alignment speed is inadequate. MUSCLE is faster than the first two methods; however, it has a high memory requirement and cannot match long sequences. Compared with these current methods, our method using conserved sequence searching runs more rapidly in large data sets under time complexity described by  $O(NL_{\text{sample}}L_{\text{longest conserved sequence}})$ . Taking conserved segments as anchors to separate a long sequence into several short

sequences can effectively improve the efficiency of traditional MSA methods, which makes them feasible for long sequences and large-scale data sets. Our method can only be applied for conserved sequence searching; however, its implementation can assist in application of other MSA methods for phylogenetic tree construction although it cannot perform this function directly. Further, the optimization approach is best applied to intraspecific data sets as there may be insufficient conserved sequences for



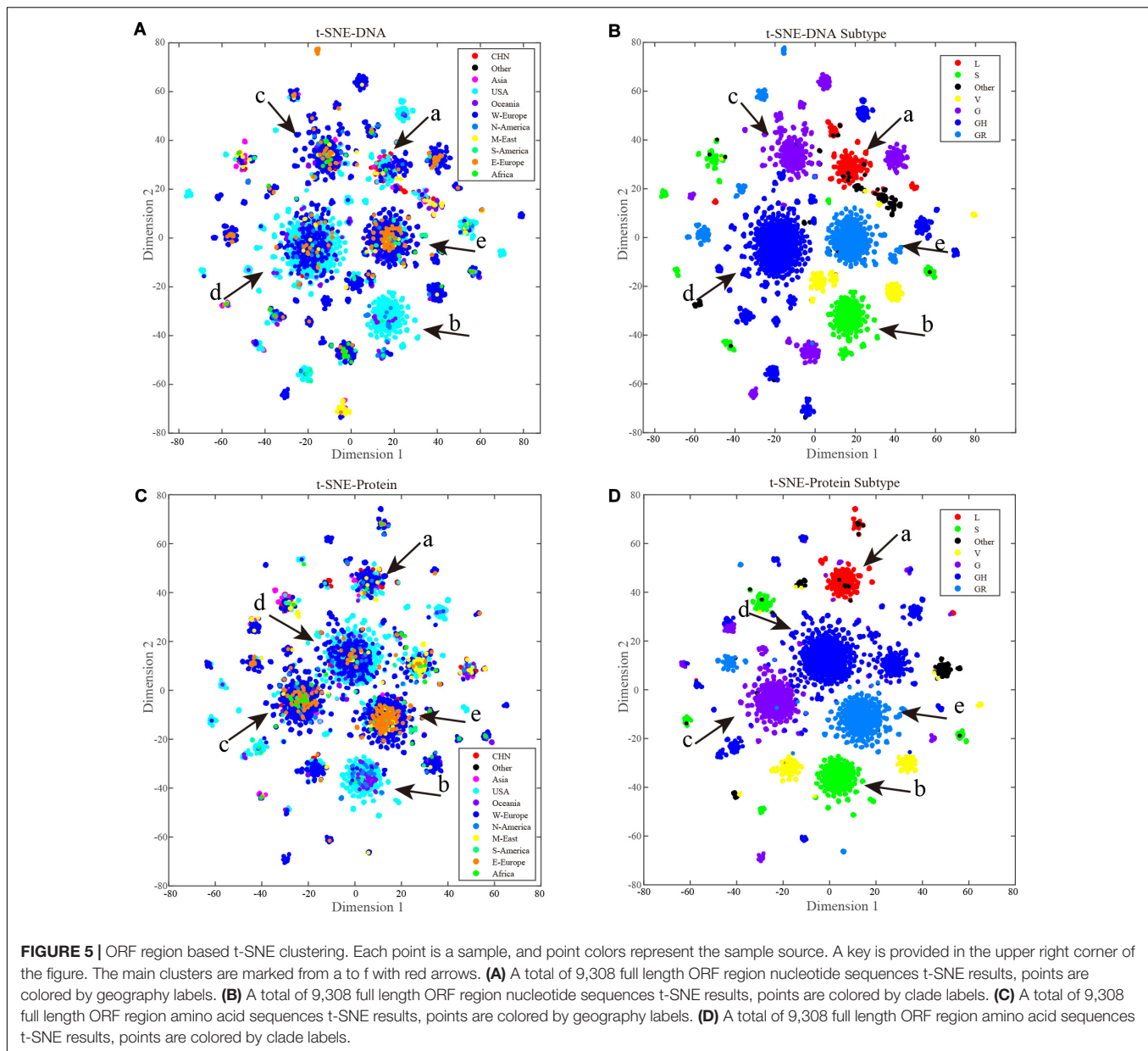
pruning in other analyses; however, for applications such as primer design, our method of conserved sequence searching has unparalleled advantages compared with other MSA approaches.

## Necessity of Long-Term Conserved Regions Analysis

In our results statistics, the sense and nonsense mutations in SARS-CoV-2 have occurred up to April 5, 2020. Some research (Zhou et al., 2020; Yu et al., 2020) discusses the relationship between SARS-CoV-2 and bat SARSr-CoVs although we only focus on intraspecific subtype differentiation and mutations of SARS-CoV-2. Compared with other research on SARS-CoV-2 mutations (Kim et al., 2020; Ugurel et al., 2020), we had the same results in the main mutations, we list more rare mutation sites in order to have a more comprehensive presentation, but

we did not statistical analyze the mutations on a non-coding region, such as C241T.

The COVID-19 pandemic is lasting; however, its duration is relatively short compared with other viral epidemics, and this epidemic may become a long-term public health event (Kissler et al., 2020). New mutations occurring in currently conserved sequences, even conserved ORF regions, remain possible and will bring new challenges for nucleic acid-based diagnosis and vaccine development. Nucleotide mutations in the coronavirus may result in failures of detection. Therefore, it is necessary to avoid frequently mutated areas when designing primers for nucleic acid diagnosis, and primers should be updated in real time, according to mutations in the viral nucleic acid. Therefore, MSA is important for updating primers used for nucleic acid-based diagnosis and improving detection rates. Hence, continuous MSA analyses of new sequencing data are



necessary. The influence of rare mutations prone to phylogeny and haplotype analysis.

To improve the efficiency of phylogenetic tree construction, bases conserved in more than 99.9% of samples were pruned, and although rare mutations may possibly be technical artifacts rather than biological mutations (De Maio et al., 2020), the resolution of the tree is still influenced. In haplotype analysis, we pruned more rare mutations ( $p_i < 0.5\%$ ) for better visualization, which may lose some subtype connections linked by these rare haplotypes. In addition, ignoring the non-coding region is also another kind of sequence over pruning. Because we did not use the non-coding region in phylogenetic tree construction, we would lose information from some important variants, such as C241T on 5'-UTR even though we used many more haplotypes in the phylogenetic tree and haplotype network

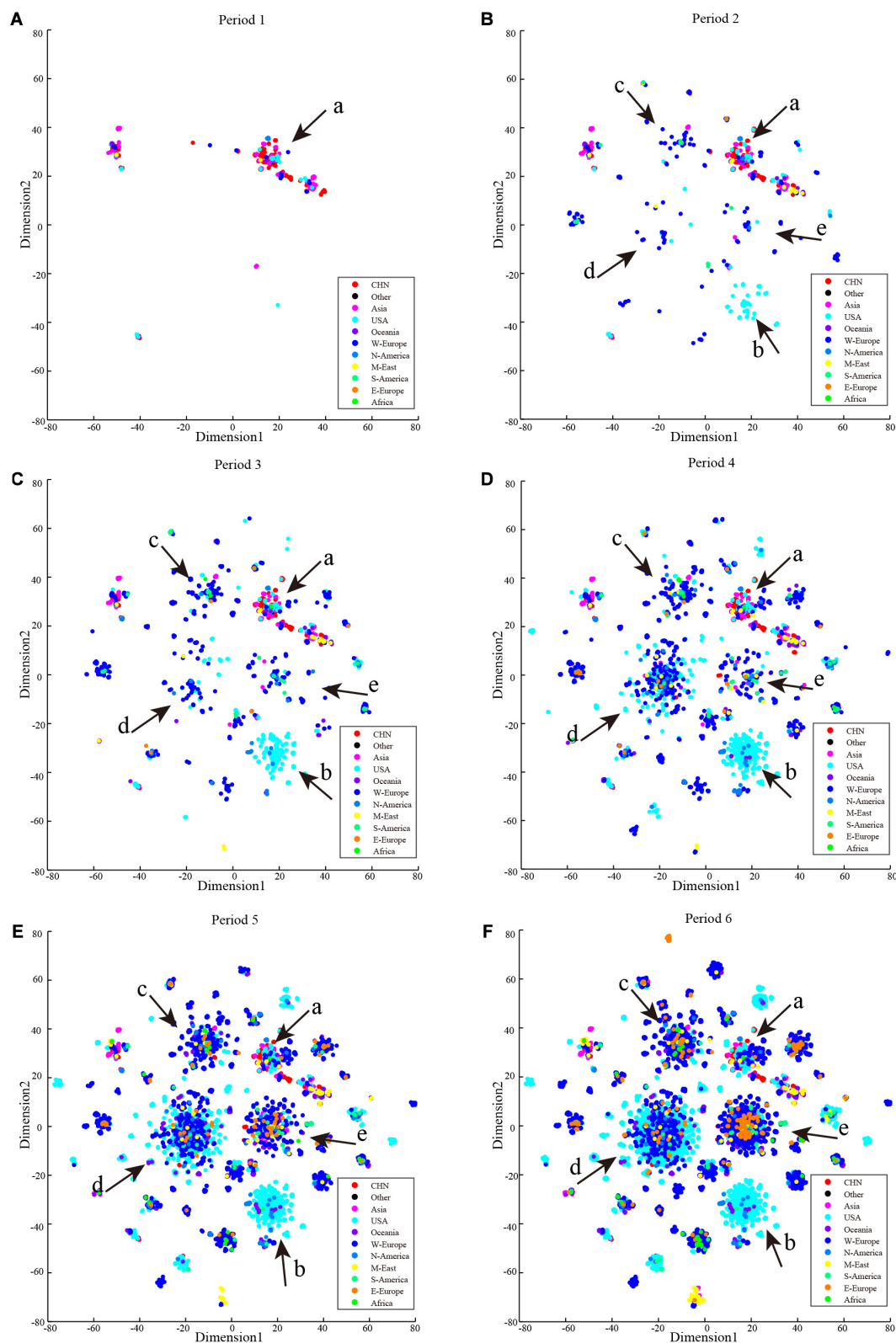
construction than other research and provided more details in SARS-CoV-2 subtype differentiation in the early stage of the pandemic.

### t-SNE Clustering in Sequence Analysis

The t-SNE method provides a new perspective for sequence data analysis. The comparison (Figures 5B,D) between t-SNE clustering results and current clade identification results prove the good performance of t-SNE in sequence-based subtype identification.

Our t-SNE results clearly demonstrate the relationships among countries/continents in the pandemic (Figure 6); however, the cases that occurred in the early period of the pandemic do not tell the origins of their belonging subtypes. One subtype strain may have already spread





**FIGURE 6 |** Period t-SNE result. t-SNE results according to collection date order. (A) Collection date from December 24, 2019 to February 2, 2020. (B) Collection date from December 24, 2019 to March 4, 2020. (C) Collection date from December 24, 2019 to October 3, 2020. (D) Collection date from December 24, 2019 to March 16, 2020. (E) Collection date from December 24, 2019 to March 25, 2020. (F) Collection date from December 24, 2019 to April 5, 2020.



widely in another region but not be detected due to limited testing ability. From this perspective, providing universal viral nucleic acid detection capability remains highly desirable for analysis of SARS-CoV-2 and requires international cooperation and information sharing.

## CONCLUSION

In this research, we developed a breadth-first search-based conserved sequence searching method for MSA optimizing and applied it on GISAID's SARS-CoV-2 data set for sequence analyzing. Our phylogenetic tree and haplotype network results show that clade S and clade G play important roles in SARS-CoV-2 subtype differentiation history. In addition, we show the feasibility of t-SNE clustering in sequence data-based subtype classification. Overall, our research provides new ideas for sequence analysis, which can provide benefits for SARS-CoV-2 sequence-based researches.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author/s.

## REFERENCES

- Bandelt, H. J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48. doi: 10.1093/oxfordjournals.molbev.a026036
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., et al. (2018). GenBank. *Nucleic Acids Res.* 46, D41–D47.
- Chen, J., Qi, T., Liu, L., Ling, Y., Qian, Z., Li, T., et al. (2020). Clinical progression of patients with COVID-19 in Shanghai. *China. J. Infect.* 80, e1–e6.
- Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., et al. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* 39, W13–W17.
- Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. doi: 10.1186/1471-2105-5-113
- Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* 1, 33–46. doi: 10.1002/gch2.1018
- Frey, A. W., Ramos da Silva, J., Rosado, V. C., Bliss, C. M., Pine, M., Mui, B. L., et al. (2020). A multi-targeting, nucleoside-modified mRNA influenza virus vaccine provides broad protection in mice. *Mol. Ther.* 28, 1569–1584. doi: 10.1016/j.ymthe.2020.04.018
- Herrera-Rodriguez, J., Meijerhof, T., Niesters, H. G., Stjernholm, G., Hovden, A.-O., Sørensen, B., et al. (2018). A novel peptide-based vaccine candidate with protective efficacy against influenza A in a mouse model. *Virology* 515, 21–28. doi: 10.1016/j.virol.2017.11.018
- Kim, J.-S., Jang, J.-H., Kim, J.-M., Chung, Y.-S., Yoo, C.-K., and Han, M.-G. (2020). Genome-wide identification and characterization of point mutations in the SARS-CoV-2 genome. *Osong. Public Health Res. Perspect.* 11, 101–111. doi: 10.24171/j.phrp.2020.11.3.05
- Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H., and Lipsitch, M. (2020). Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* 368, 860–868. doi: 10.1126/science.abb5793

## AUTHOR CONTRIBUTIONS

LM conceived, designed, and supervised this study. QL designed the study, coded all the programs, collected, and analyzed the data. QL, ZJ, and FW drafted and checked the manuscript. YH visualized the data and format all pictures and tables. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

## FUNDING

This work was supported by National Key R&D Plan in China (2016YFD0501100), Shenzhen Strategic Emerging Industry Development special funds (JCYJ20170816143646446), and Shenzhen Bay Laboratory, Shenzhen, China.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.591833/full#supplementary-material>

- Knuth, D. E., Morris, J. M. Jr., and Pratt, V. R. (1977). Fast pattern matching in strings. *SIAM J. Comput.* 6, 323–350.
- Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182, 812.e19–827.e19.
- Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi: 10.1093/bioinformatics/btz305
- Lai, A., Bergna, A., Acciarri, C., Galli, M., and Zehender, G. (2020). Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *J. Med. Virol.* 92, 675–679. doi: 10.1002/jmv.25723
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404
- Leigh, J. W., and Bryant, D. (2015). popart: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210x.12410
- Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259.
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., et al. (2020). Early transmission dynamics in Wuhan, China, of novel Coronavirus-infected pneumonia. *N. Engl. J. Med.* 382, 1199–1207.
- Ma, C., Su, S., Wang, J., Wei, L., Du, L., and Jiang, S. (2020). From SARS-CoV to SARS-CoV-2: safety and broad-spectrum are important for coronavirus vaccine development. *Microbes Infect.* 22, 245–253. doi: 10.1016/j.micinf.2020.05.004
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst Tech J.* 27, 379–423.
- Ugurel, O. M., Ata, O., and Turgut-Balik, D. (2020). An updated analysis of variations in SARS-CoV-2 genome. *Turk. J. Biol.* 44, 157–167. doi: 10.3906/biy-2005-111

- World Health Organization (2020). *Estimating Mortality From COVID-19*. Available online at: <https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19> (accessed October 13, 2020).
- Worldometer (2020). *COVID-19 Coronavirus Pandemic*. Available online at: <https://www.worldometers.info/coronavirus/#countries> (accessed October 13, 2020).
- Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.
- Yu, W.-B., Tang, G.-D., Zhang, L., and Corlett, R. T. (2020). Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zool. Res.* 41, 247–257.
- Yurkovetskiy, L., Wang, X., Pascal, K. E., Tomkins-Tinch, C., Nyalile, T. P., Wang, Y., et al. (2020). Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* 183, 739.e8–751.e8.
- Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Copyright © 2021 Lin, Huang, Jiang, Wu and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Hotspot Mutations in SARS-CoV-2

Indrajit Saha<sup>1\*</sup>, Nimisha Ghosh<sup>2†</sup>, Nikhil Sharma<sup>3</sup> and Suman Nandi<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, <sup>2</sup>Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India, <sup>3</sup>Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, India

Since its emergence in Wuhan, China, severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has spread very rapidly around the world, resulting in a global pandemic. Though the vaccination process has started, the number of COVID-affected patients is still quite large. Hence, an analysis of hotspot mutations of the different evolving virus strains needs to be carried out. In this regard, multiple sequence alignment of 71,038 SARS-CoV-2 genomes of 98 countries over the period from January 2020 to June 2021 is performed using MAFFT followed by phylogenetic analysis in order to visualize the virus evolution. These steps resulted in the identification of hotspot mutations as deletions and substitutions in the coding regions based on entropy greater than or equal to 0.3, leading to a total of 45 unique hotspot mutations. Moreover, 10,286 Indian sequences are considered from 71,038 global SARS-CoV-2 sequences as a demonstrative example that gives 52 unique hotspot mutations. Furthermore, the evolution of the hotspot mutations along with the mutations in variants of concern is visualized, and their characteristics are discussed as well. Also, for all the non-synonymous substitutions (missense mutations), the functional consequences of amino acid changes in the respective protein structures are calculated using PolyPhen-2 and I-Mutant 2.0. In addition to this, SSIPE is used to report the binding affinity between the receptor-binding domain of Spike protein and human ACE2 protein by considering L452R, T478K, E484Q, and N501Y hotspot mutations in that region.

**Keywords:** COVID-19, deletions, entropy, hotspot mutations, SARS-CoV-2 genomes, substitution

## OPEN ACCESS

### Edited by:

Yang Zhang,  
University of Michigan, United States

### Reviewed by:

Xiaoqiang Huang,  
University of Michigan, United States  
Yavuz Oktay,  
Dokuz Eylul University, Turkey

### \*Correspondence:

Indrajit Saha  
indrajit@nittrkol.ac.in

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 04 August 2021

**Accepted:** 07 October 2021

**Published:** 29 November 2021

### Citation:

Saha I, Ghosh N, Sharma N and  
Nandi S (2021) Hotspot Mutations  
in SARS-CoV-2.  
Front. Genet. 12:753440.  
doi: 10.3389/fgene.2021.753440

## 1 INTRODUCTION

COVID-19 caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) was first identified in late December 2019 and has a high transmission rate (Zhu et al., 2020). The WHO declared this outbreak as a pandemic on March 11, 2020 (Cucinotta and Vanelli, 2020). Like other coronaviruses, SARS-CoV-2 is also an enveloped single-stranded RNA virus containing nearly 30 K nucleotide sequences (Alexandersen et al., 2020). SARS-CoV-2 encompasses 11 coding regions, which include ORF1ab, Spike (S), ORF3a, Envelope (E), Membrane (M), ORF6, ORF7a, ORF7b, ORF8, Nucleocapsid (N), and ORF10.

Though the vaccination process has started, the virus is evolving and spreading all across the world, causing fresh waves every few months. Since the virus is mutating frequently, it creates new variant of the original virus. Among several variants, B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), and B.1.617.2 (Delta) are declared as variants of concern (Singh et al., 2021). In this regard, the variant B.1.1.7 was first identified in the United Kingdom, which contains E484K, N501Y, D614G, and P681H mutations in Spike glycoprotein (Tang et al., 2020). In December 2020, the variant

B.1.351 was first detected in South Africa, with mutations such as K417N, E484K, N501Y, D614G, and A701V (Tang et al., 2021). The Brazilian variant P.1 also has almost the same mutations as the B.1.351 variant, but instead of A701V, the P.1 variant has H555Y mutation (Faria et al., 2021). On the other hand, the variant B.1.617.2 was first identified in India with L452R, T478K, D614G, and P681R mutations in Spike glycoprotein (Bernal et al., 2021).

To understand the new variants of SARS-CoV-2, Tiwari and Mishra (2021) have performed phylogenetic analysis of 591 SARS-CoV-2 genomes where they have found 43 synonymous and 57 non-synonymous mutations in 12 protein regions. They found the most prevalent mutations in the Spike protein, followed by NSP2, NSP3, and ORF9. They have also highlighted several distinct SARS-CoV-2 features as compared with other human-infecting viruses. Yuan et al. (2020) have analyzed 11,183 global sequences where they have identified 119 single-nucleotide polymorphisms (SNPs) with 74 non-synonymous and 43 synonymous mutations. The mutational profiling shows that the highest mutation has occurred in Nucleocapsid, followed by NSP2, NSP3, and Spike. From China, India, the United States, and Europe, 570 SARS-CoV-2 genomes are analyzed by Weber et al. (2020), where they have identified 10 individual mutations where most of the mutations altered the amino acids in the replication-relevant proteins. Sarkar et al. (2021) have performed a genome-wide analysis of 837 Indian SARS-CoV-2 genomes, where 33 unique mutations were observed, among which 18 mutations were identified in India in five protein regions (six in Spike, five in NSP3, four in RdRp, two in NSP2, and one in Nucleocapsid). The isolated Indian sequences were classified into 22 groups based on their coexisting mutations. This study highlights several mutations identified in various protein regions, which also help to identify the evolution of virus genome across various geographic locations of India. Saha et al. (2020) have performed phylogenetic analysis of 566 Indian SARS-CoV-2 genomes to identify several mutations. As a result, 933 substitutions, 2,449 deletions, and two insertions have been identified from the aligned sequences. In another study, Saha et al. (2021) have performed genomic analysis of 10,664 SARS-CoV-2 genomes, resulting in 7,209 substitutions, 11,700 deletions, 119 insertions, and 53 SNPs.

Motivated by the aforementioned analysis, in this work, we have performed multiple sequence alignment (MSA) of 71,038 SARS-CoV-2 genomes using MAFFT (Katoh et al., 2002) followed by their phylogenetic analysis using Nextstrain (Hadfield et al., 2018) to visualize the virus evolution. This led to the identification of hotspot mutations as deletions and substitutions in the coding regions based on entropy greater than or equal to 0.3. Furthermore, as a demonstrative example, 10,286 Indian sequences are considered from 71,038 global SARS-CoV-2 sequences. For all the non-synonymous substitutions (missense mutations), the functional consequences of amino acid changes in the respective protein structures are calculated using PolyPhen-2 and I-Mutant 2.0. Finally, SSIPE is used to report the binding affinity between the receptor-binding domain (RBD) of Spike protein and human

ACE2 protein by considering the hotspot mutations in that region.

## 2 METHODS

In this section, the dataset collection for the SARS-CoV-2 genomes is discussed along with the proposed pipeline.

### 2.1 Data Preparation

For MSA and phylogenetic analysis, 71,038 global SARS-CoV-2 genomes are collected from Global Initiative on Sharing All Influenza Data (GISAID)<sup>1</sup>, and the Reference Genome (NC 045512.2)<sup>2</sup> is collected from the National Center for Biotechnology Information (NCBI). The SARS-CoV-2 sequences are mostly distributed from January 2020 to June 2021 globally. Moreover, to map the protein sequences and changes in the amino acid, Protein Data Bank (PDB) is collected from Zhang Lab<sup>3</sup> (Zhang et al., 2020; Wu et al., 2021), and it is then used to show the structural changes. All these analyses are performed on the High Performance Computing facility of NITTTR, Kolkata; and for checking the amino acid changes, MATLAB R2019b is used.

### 2.2 Pipeline of the Work

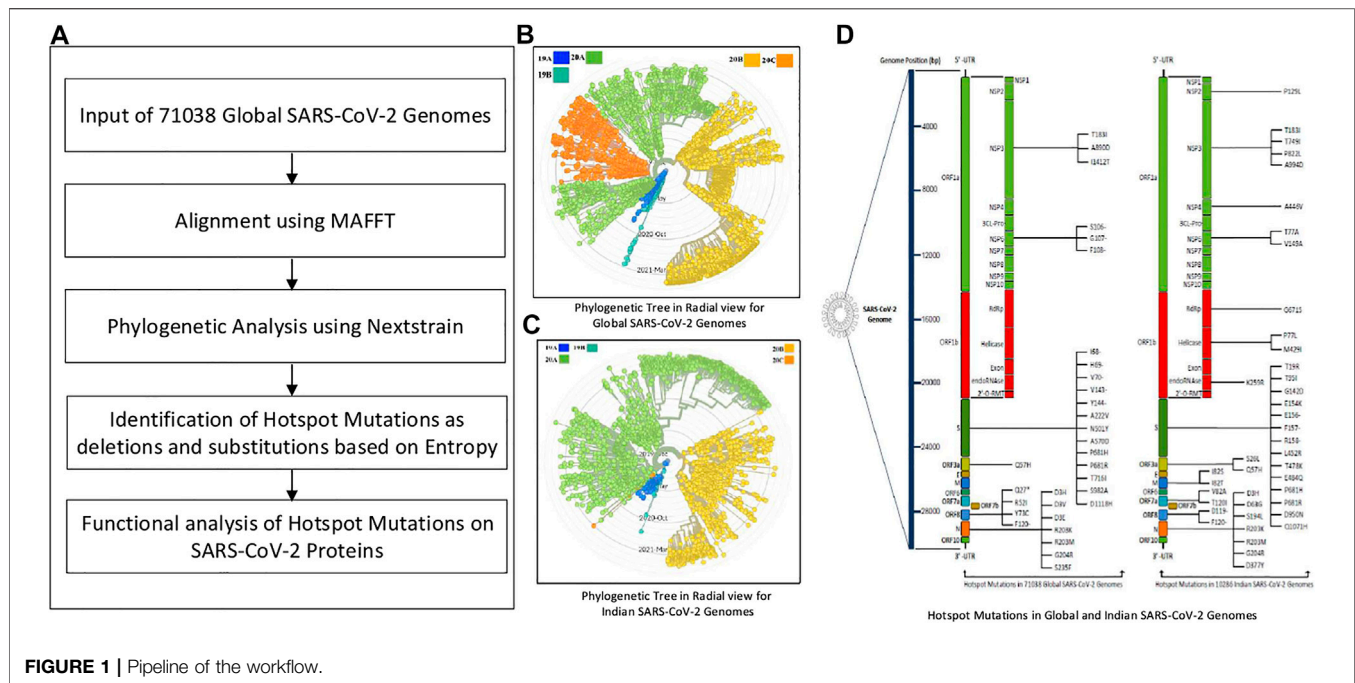
The pipeline of this work is provided in **Figure 1A**. Initially, MSA of 71,038 global SARS-CoV-2 genomes is performed using MAFFT, which is followed by their phylogenetic analysis using Nextstrain. The corresponding phylogenetic tree is shown in **Figure 1B**. MAFFT merges local and global algorithms for MSA, and it uses two different heuristic methods such as progressive (FFT-NS-2) and iterative refinement (FFT-NS-i). To create a provisional MSA, FFT-NS-2 calculates all-pairwise distances from which refined distances are calculated. Thereafter, FFT-NS-i is performed to get the final MSA. As MAFFT uses fast Fourier transform, it scores over other alignment techniques. On the other hand, Nextstrain is a collection of open-source tools, which is useful for understanding the evolution and spread of pathogen, particularly during an outbreak. By taking advantage of this tool, in this work, the evolution and geographic distribution of SARS-CoV-2 genomes are visualized by creating the metadata in our High Performance Computing environment.

Once the alignment and the phylogenetic analysis are completed, hotspot mutations as deletions and substitutions are identified in the coding regions based on entropy greater than or equal to 0.3. Furthermore, 10,286 Indian sequences are considered as an example to identify such mutations as well. The corresponding phylogenetic tree for Indian sequences is shown in **Figure 1C**. Moreover, using the codon table, amino acid changes in the SARS-CoV-2 proteins for the corresponding mutations are highlighted as well. The hotspot mutations are identified considering their entropy values, which are calculated as:

<sup>1</sup><https://www.gisaid.org/>

<sup>2</sup><https://www.ncbi.nlm.nih.gov/nuccore/1798174254>

<sup>3</sup><https://zhanglab.ccmb.med.umich.edu/COVID-19/>



$$\mathcal{E} = \ln 5 + \sum \lambda_{\gamma}^{\delta} [ \ln (\lambda_{\gamma}^{\delta}) ] \quad (1)$$

where  $\lambda_{\gamma}^{\delta}$  represents the frequency of each residue  $\gamma$  occurring at position  $\delta$  and 5 represents the four possible residues as nucleotides plus gap. Thereafter, the amino acid changes in the SARS-CoV-2 proteins for the non-synonymous deletions and substitutions for both global and Indian sequences are graphically visualized as shown in **Figure 1D**. Finally, these changes are also used for the evaluation of their functional characteristics and are visualized in the respective protein structure as well.

### 3 RESULTS

The experiments in this work are carried out according to the pipeline as given in **Figure 1A**. Initially, MSA of 71,038 global SARS-CoV-2 genomes across 98 countries is carried out using MAFFT followed by their phylogenetic analysis using Nextstrain, which revealed five clades: 19A, 19B, 20A, 20B, and 20C. The number of sequences for each country is reported in **Supplementary Table S1**. This resulted in the identification of hotspot mutation points as deletions and substitutions in the coding regions based on entropy. In this regard, only those hotspot mutations are considered whose entropy values are greater than or equal to 0.3. The entropy values for each of the genomic coordinates for both global and Indian sequences are provided in **Supplementary Table S2**. The mutation statistics by considering different threshold values of entropy for each category are reported in **Table 1**. Based on the results in this table, the entropy value of 0.3 is considered as the threshold for choosing the hotspot mutations. It is to be noted that choosing a threshold value as either 0.2 or 0.1 will lead to a huge amount of hotspot mutations, which is not desired. As a consequence

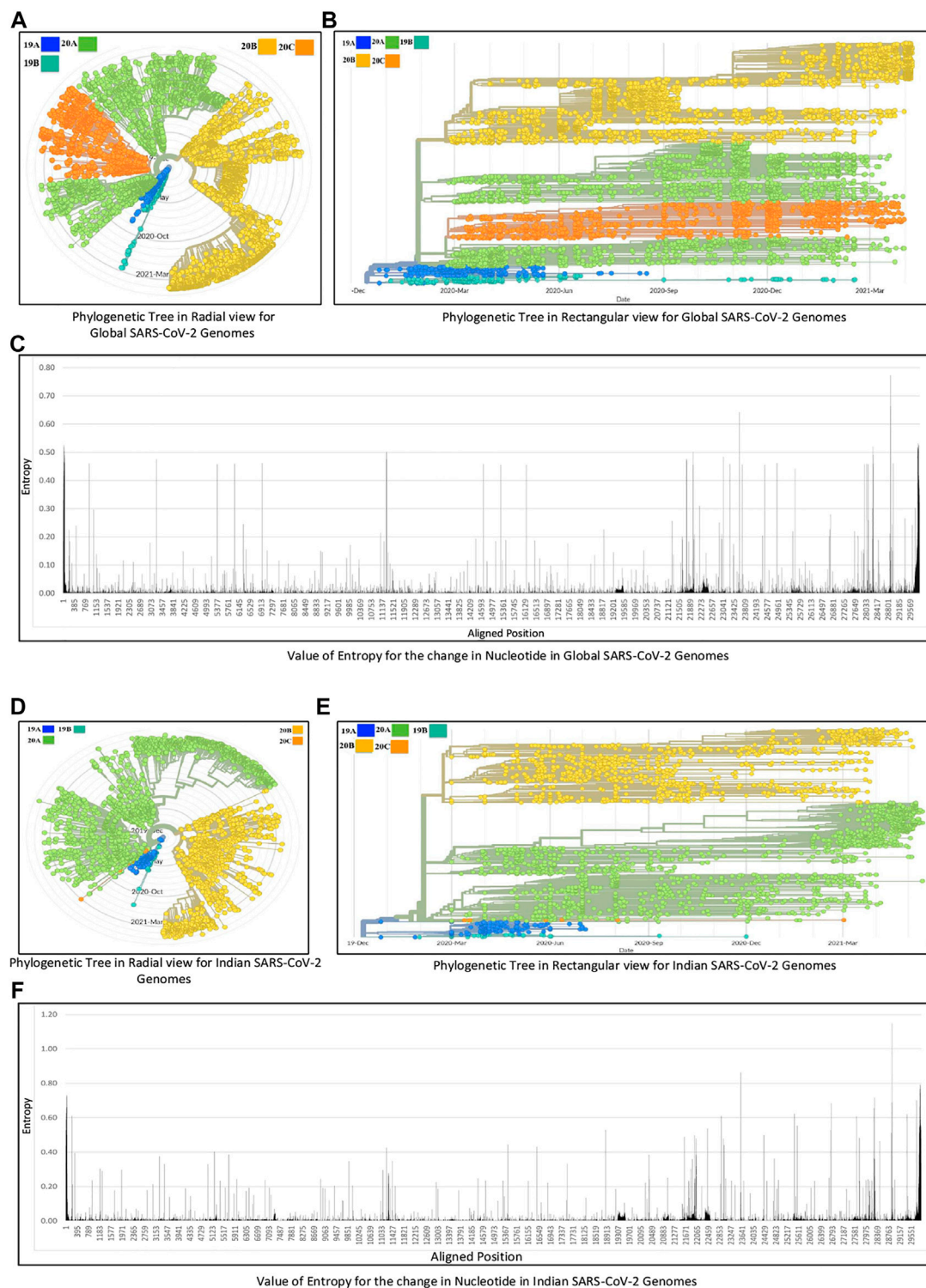
of choosing entropy threshold of 0.3, 45 unique hotspot mutations are identified, which resulted in 39 non-synonymous deletions and substitutions with nine unique deletions and 22 unique amino acid changes. Also, out of the 98 countries that are considered for global analysis, India with 10,286 sequences is taken as an example to demonstrate the mutations for a particular country as well. In this regard, 52 unique hotspot mutations provide 45 non-synonymous deletions and substitutions with five unique amino acid changes for deletions and 36 unique amino acid changes for substitutions. The analysis on other countries with the most number of sequences is provided in the Supplementary Material. The phylogenetic trees in radial and rectangular views considering global analysis are shown in **Figures 2A,B**, respectively, while for Indian sequences, such views are provided in **Figures 2D,E**, respectively. These phylogenetic trees respectively show the evolution of the global and Indian SARS-CoV-2 genomes over the months. For the benefit of the readers, it is important to mention that the number of sequences does not have any direct relationship with the number of hotspot mutations. The number of hotspots is based on the entropy value, which in turn depends on the frequency of mutations at a given genomic coordinate. So even with smaller number of sequences, if the frequency of mutations is higher than that with larger number of sequences, it will produce more hotspot mutations. Thus, with 71,038 global sequences, 45 unique hotspot mutations are identified, while for 10,286 Indian sequences, 52 such mutations are identified.

The list of hotspot mutations for the global and Indian SARS-CoV-2 genomes along with their associated details is respectively provided in **Tables 2** and **3**. For example, in **Table 2**, genomic coordinate 28,881 in Nucleocapsid with nucleotide changes G > A and G > T has the highest entropy value of 0.773655. India also shows the same mutation but with an entropy value of 1.14807 as shown in **Table 3**. Please note that mutations like G28881A and G28883C may have an impact on antigenicity of Nucleocapsid



**TABLE 1** | Mutation statistics of 71,038 global and 10,286 Indian SARS-CoV-2 genomes by considering different threshold values.

Threshold value	Coding regions of global SARS-CoV-2 genomes																									
	NSP1	NSP2	NSP3	NSP4	3CL- Pro	NSP6	NSP7	NSP8	NSP9	NSP10	NSP11	RdRp	Helicase	Exon	endoRNAse	NSP16	Spike	ORF3a	Envelope	Membrane	ORF6	ORF7a	ORF7b	ORF8	Nucleocapsid	ORF10
>=0.60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	0
>=0.50 to < 0.60	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
>=0.40 to < 0.50	0	1	4	0	0	8	0	0	0	0	0	3	0	0	0	0	13	1	0	0	0	0	0	3	4	0
>=0.30 to < 0.40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
>=0.20 to < 0.30	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	6	1	0	3	0	2	0	1	3	1
>=0.10 to < 0.20	1	3	7	4	1	3	0	0	0	0	0	6	3	3	1	1	14	1	0	0	0	0	1	7	8	0
>=0.05 to < 0.10	1	3	18	3	2	2	0	0	3	0	0	5	6	2	4	1	25	7	0	2	1	2	0	2	5	0
Threshold value	Coding regions of Indian SARS-CoV-2 genomes																									
	NSP1	NSP2	NSP3	NSP4	3CL- Pro	NSP6	NSP7	NSP8	NSP9	NSP10	NSP11	RdRp	Helicase	Exon	endoRNAse	NSP16	Spike	ORF3a	Envelope	Membrane	ORF6	ORF7a	ORF7b	ORF8	Nucleocapsid	ORF10
>=0.60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	1	0	1	0	1	4	0
>=0.50 to < 0.60	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0	1	0
>=0.40 to < 0.50	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	9	0	0	0	0	1	0	1	1	0
>=0.30 to < 0.40	1	1	4	1	0	0	0	0	0	0	0	0	1	0	1	0	6	0	0	0	0	0	0	4	1	0
>=0.20 to < 0.30	0	3	4	3	0	5	0	0	1	0	0	4	1	0	2	1	16	1	0	1	0	1	0	4	4	0
>=0.10 to < 0.20	1	12	5	4	0	10	0	0	1	0	0	3	2	3	1	1	8	2	0	2	0	1	2	0	4	0
>=0.05 to < 0.10	0	7	11	4	1	5	0	1	1	0	0	4	1	8	1	3	53	3	0	11	0	0	0	5	7	1



**FIGURE 2 |** Phylogenetic analysis of (A, B, C) global and (D, E, F) Indian SARS-CoV-2 genomes.

protein (Yuan et al., 2020). The entropy values for the corresponding nucleotide changes for global analysis are shown in **Figure 2C**, while for India, the same is shown in **Figure 2F**. It is to be noted that the total number of unique

amino acid changes for deletions and substitutions is less than the number of non-synonymous deletions and substitutions. One of the reasons for this can be that if there are deletions at consecutive genomic coordinates, the corresponding amino acid changes are

**TABLE 2 |** List of hotspot mutations for 71,038 global SARS-CoV-2 genomes along with the protein change.

Genomic coordinate	Overall entropy	Nucleotide change	Amino acid change	Protein coordinate	Gene
28,881	0.773655	G > A, G > T	R > K, R > M	203	Nucleocapsid
28,883	0.663399	G > C	G > R	204	Nucleocapsid
28,882	0.663308	G > A	R > R	203	Nucleocapsid
23,604	0.642160	C > A, C > G	P > H, P > R	681	Spike
11,296	0.502171	T > -	F > -	108	NSP6
21,993	0.500865	A > -	Y > -	144	Spike
11,291	0.499603	G > -	G > -	107	NSP6
28,280	0.491543	G > C	D > H	3	Nucleocapsid
23,063	0.484066	A > T	N > Y	501	Spike
21,770	0.476393	G > -	V > -	70	Spike
3,267	0.475810	C > T	T > I	183	NSP3
11,288	0.474924	T > -	S > -	106	NSP6
11,289	0.472836	C > -	S > -	106	NSP6
21,765	0.471435	T > -	I > -	68	Spike
21,767	0.469881	C > -	H > -	69	Spike
11,290	0.467890	T > -	S > -	106	NSP6
21,766	0.467479	A > -	I > -	68	Spike
21,768	0.467116	A > -	H > -	69	Spike
21,769	0.466151	T > -	H > -	69	Spike
11,293	0.465319	T > -	G > -	107	NSP6
11,292	0.464056	G > -	G > -	107	NSP6
11,294	0.463926	T > -	F > -	108	NSP6
24,914	0.461770	G > C	D > H	1118	Spike
6,954	0.461746	T > C	I > T	1412	NSP3
28,977	0.460661	C > T	S > F	235	Nucleocapsid
21,992	0.460243	T > -	Y > -	144	Spike
913	0.460233	C > T	S > S	36	NSP2
11,295	0.459624	T > -	F > -	108	NSP6
5,986	0.459543	C > T	F > F	1089	NSP3
28,282	0.459253	T > A	D > E	3	Nucleocapsid
28,048	0.458864	G > T	R > I	52	ORF8
14,676	0.458373	C > T	P > P	412	RdRp
23,271	0.458086	C > A	A > D	570	Spike
28,281	0.458038	A > T	D > V	3	Nucleocapsid
27,972	0.457841	C > T	Q > *	27	ORF8
5,388	0.457761	C > A	A > D	890	NSP3
28,111	0.457624	A > G	Y > C	73	ORF8
23,709	0.456643	C > T	T > I	716	Spike
24,506	0.455921	T > G	S > A	982	Spike
15,279	0.455884	C > T	H > H	613	RdRp
16,176	0.455573	T > C	T > T	912	RdRp
21,991	0.455314	T > -	V > -	143	Spike
25,563	0.442049	G > T	Q > H	57	ORF3a
22,227	0.310063	C > T	A > V	222	Spike
28,253	0.300528	C > T, C > -	F > F, F > -	120	ORF8

the same. For example, as can be seen from **Table 2**, at the three consecutive genomic coordinates 11,288, 11,289, and 11,290, deletion has occurred with the amino acid change as S106-. Thus, though the number of non-synonymous deletions is 3, the number of unique amino acid change is 1. This is true for other such changes as well.

The amino acid changes in protein for the non-synonymous deletions and substitutions as reported in **Tables 2** and **3** are visualized in **Figure 1D**; **Supplementary Figure S1**. All the amino acid changes in the protein for the non-synonymous substitutions or missense mutations for the global sequences are shown in **Figure 3**, while the same for the Indian sequences are depicted in **Figure 4**. The month-wise virus evolution in terms of entropy for

both global and Indian genomic sequences is visualized respectively in **Figures 5** and **6**, while the corresponding entropy values are reported in **Supplementary Tables S3** and **S4**. For example, it can be seen from both the figures that both P681H and P681R, which are part of the variant of concerns Alpha or B.1.1.7 and Delta or B.1.617.2, have evolved over time globally and for India as well. It is to be noted that due to the lack of appropriate number of sequences, the data of January and February 2020 have been merged for the global analysis, while for India, such merging is for the months January to March 2020. Also, please note that since the calculation of entropy is performed on aligned sequences, only coding regions are considered for the identification of hotspot mutations, as the non-coding regions

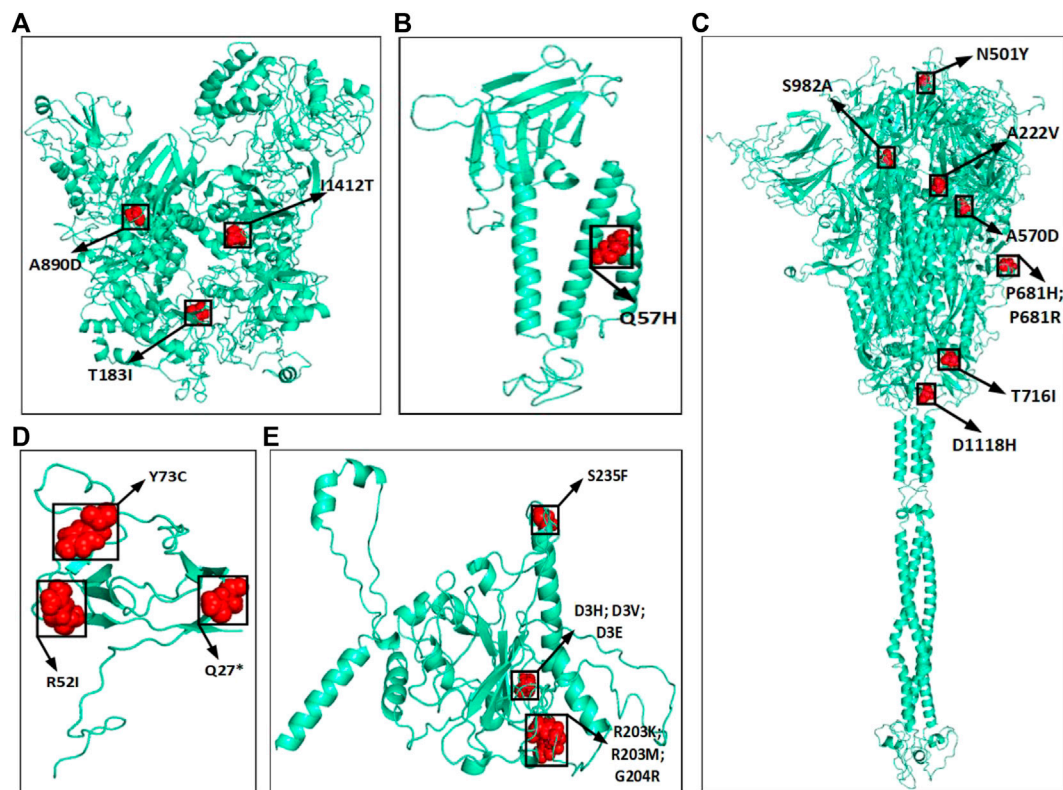
**TABLE 3 |** List of hotspot mutations for 10,286 Indian SARS-CoV-2 genomes along with the protein change.

Genomic coordinate	Overall entropy	Nucleotide change	Amino acid change	Protein coordinate	Gene
28,881	1.14807	G > A, G > T	R > K, R > M	203	Nucleocapsid
23,604	0.8631	C > A, C > G	P > H, P > R	681	Spike
28,882	0.69019	G > A	R > R	203	Nucleocapsid
28,883	0.68846	G > C	G > R	204	Nucleocapsid
26,767	0.68419	T > C, T > G	I > T, I > S	82	Membrane
28,253	0.65534	C > T, C > -	F > F, F > -	120	ORF8
25,469	0.6227	C > T	S > L	26	ORF3a
29,402	0.61955	G > T	D > Y	377	Nucleocapsid
22,917	0.61006	T > G	L > R	452	Spike
27,638	0.60866	T > C	V > A	82	ORF7a
25,563	0.55354	G > T	Q > H	57	ORF3a
22,444	0.53665	C > T	D > D	249	Spike
18,877	0.52834	C > T	L > L	280	Exon
26,735	0.52715	C > T	Y > Y	71	Membrane
28,854	0.51198	C > T	S > L	194	Nucleocapsid
24,410	0.49845	G > A	D > N	950	Spike
21,987	0.49717	G > A	G > D	142	Spike
21,618	0.48836	C > G	T > R	19	Spike
27,752	0.48264	C > T	T > I	120	ORF7a
22,034	0.47915	A > -	R > -	158	Spike
22,995	0.47879	C > A	T > K	478	Spike
28,461	0.46436	A > G	D > G	63	Nucleocapsid
15,451	0.44421	G > A	G > S	671	RdRp
23,012	0.44086	G > C	E > Q	484	Spike
22,033	0.4385	C > -	F > -	157	Spike
16,466	0.43082	C > T	P > L	77	Helicase
22,032	0.42673	T > -	F > -	157	Spike
11,201	0.42554	A > G	T > A	77	NSP6
28,249	0.41704	A > -	D > -	119	ORF8
5,184	0.40139	C > T	P > L	822	NSP3
22,031	0.40074	T > -	F > -	157	Spike
313	0.39475	C > T	L > L	16	NSP1
22,029	0.38676	A > -	E > -	156	Spike
5,700	0.38604	C > A	A > D	994	NSP3
20,396	0.38407	A > G	K > R	259	endoRNAse
3,267	0.37579	C > T	T > I	183	NSP3
22,030	0.3738	G > -	E > -	156	Spike
28,251	0.36694	T > -	F > -	120	ORF8
28,248	0.36497	G > -	D > -	119	ORF8
24,775	0.36197	A > T	Q > H	1071	Spike
21,895	0.35931	T > C	D > D	111	Spike
28,280	0.35905	G > C	D > H	3	Nucleocapsid
28,250	0.35546	T > -	D > -	119	ORF8
28,252	0.351	T > -	F > -	120	ORF8
11,418	0.34861	T > C	V > A	149	NSP6
9,891	0.34766	C > T	A > V	446	NSP4
17,523	0.33196	G > T	M > I	429	Helicase
3,457	0.3314	C > T	Y > Y	246	NSP3
4,965	0.32981	C > T	T > I	749	NSP3
22,022	0.31618	G > A	E > K	154	Spike
1191	0.30404	C > T	P > L	129	NSP2
21,846	0.30253	C > T	T > I	95	Spike

exhibit high entropy values and can be misleading while selecting such mutation points as hotspot mutations. Furthermore, the evolution of the mutation points for global SARS-CoV-2 genomes pertaining to the different variants of concern like Alpha, Beta, Gamma, and Delta as declared by the WHO is also reported respectively in **Figures 7A,B,C,D**. It can be observed from the figures that the popular mutation D614G, which is common in all the variants though predominant in the earlier months of the

pandemic, has waned over time. Also, the mutation T478K, which is unique to the Delta variant, is known to facilitate antibody escape (Planas et al., 2021). Some important hotspot mutations like H69-, V70-, Y144-, A222V, N501Y, A570D, P681H, and P681R identified in this study are associated with the different SARS-CoV-2 variants of concern like Alpha, Beta, Gamma, and Delta.

The unique and common hotspot mutations between global and Indian sequences are represented in the form of Venn diagram



**FIGURE 3** | Highlighted amino acid changes in the protein structures for the non-synonymous substitutions or missense hotspot mutations for global SARS-CoV-2 genomes in (A) NSP3, (B) ORF3a, (C) Spike, (D) ORF8, and (E) Nucleocapsid.

in **Figures 8A,B**, which shows the unique and common non-synonymous hotspot mutations, while the unique and common amino acid changes are shown in **Figure 8C**. As shown in **Figure 8A**, there are 37 and 44 unique mutations in global and Indian sequences, while eight are common in both. For non-synonymous hotspot deletions and substitutions, there are 32 and 38 unique mutations in each category, while the common number of such mutations is seven as reported in **Figure 8B**. For amino acid changes, as shown in **Figure 8C**, these statistics are 22, 32, and nine. The Venn diagram showing the common and unique hotspot mutations for global and Indian sequences with Alpha, Beta, Gamma, and Delta variants of SARS-CoV-2 is reported in **Supplementary Figure S2**. For example, in **Supplementary Figure S2A**, there are four unique mutations in both global sequences and Alpha variant, while there are nine mutations that are common to both.

## 4 DISCUSSION

There are spurts of new waves in almost every country around the globe. India has already gone through the massively catastrophic second wave, and according to the experts, a third wave is imminent. This can be attributed to the fact that the virus is evolving and new strains are getting identified, thereby making the study of this ever-evolving virus all the more important. The functional characteristics

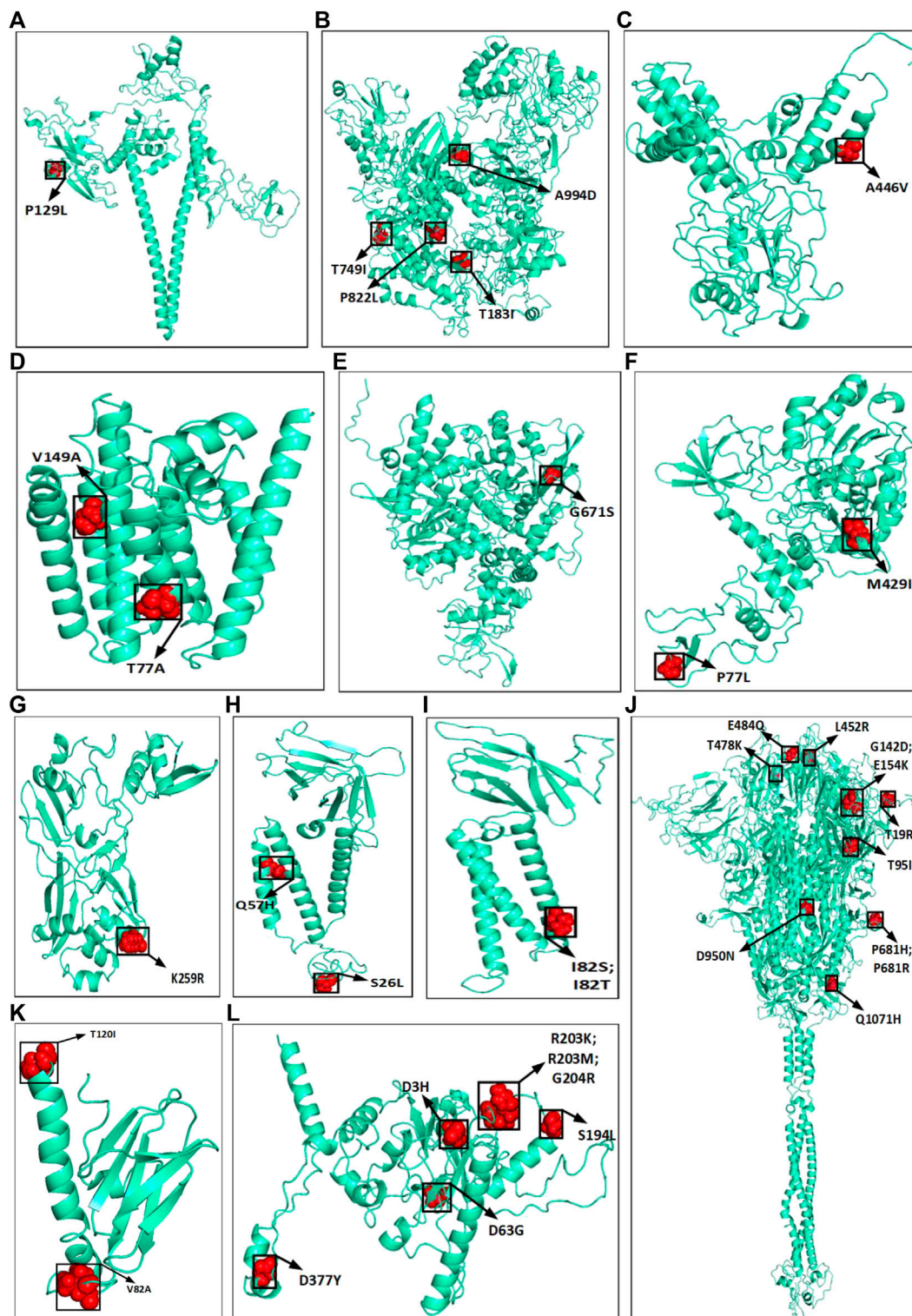
of some important mutations in the global and Indian SARS-CoV-2 genomic sequences are reported in **Table 4**.

Structural changes in amino acid residues may sometimes lead to functional instability in proteins due to change in protein translations. To judge their characteristics, these changes are demonstrated through sequence and structural homology-based prediction for the hotspot deletions and missense mutations for global and Indian sequences in **Table 5**. The tools used for these predictions are PolyPhen-2 (Polymorphism Phenotyping) (Adzhubei et al., 2010) and I-Mutant 2.0 (Capriotti et al., 2005). PolyPhen-2<sup>4</sup> works with sequence, structural, and phylogenetic information of missense mutations, while I-Mutant 2.0<sup>5</sup> uses support vector machine (SVM) for the automatic prediction of protein stability changes upon missense mutations. PolyPhen-2 is used to find the damaging hotspot mutations, and I-Mutant 2.0 determines protein stability. To determine if a mutation is damaging using PolyPhen-2, its score is considered, which lies between 0 and 1. If the score is close to 1, then a mutation is considered to be damaging. It can be concluded from **Table 5** that out of the 22 unique amino acid changes for substitutions in global sequences, 14 are damaging, while for Indian sequences, 24 are damaging out of 36 changes. It is important to note that in case of protein, damaging mostly defines instability. Generally, this is used for

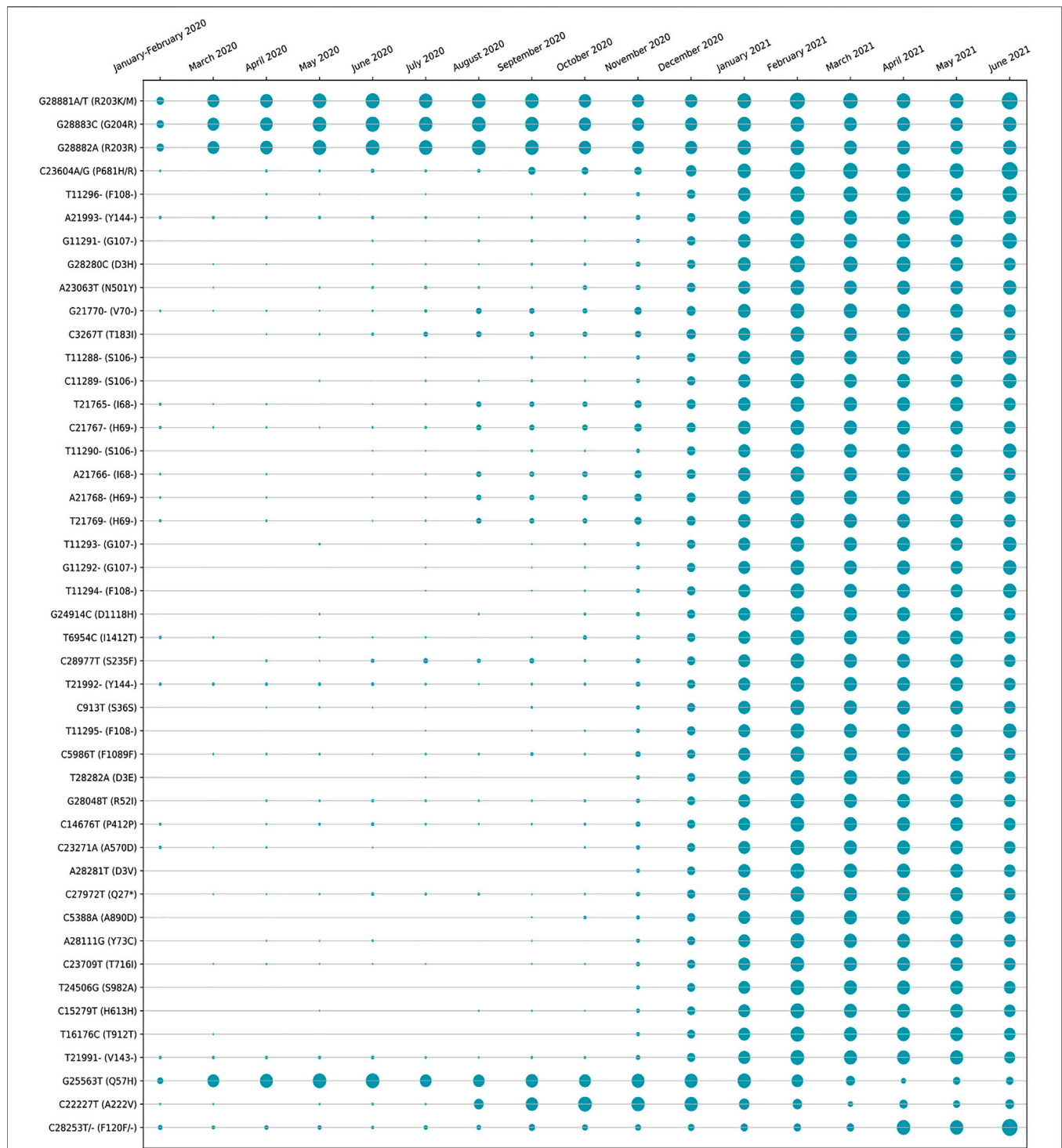
<sup>4</sup><http://genetics.bwh.harvard.edu/pph2/>

<sup>5</sup><https://folding.biofold.org/i-mutant/i-mutant2.0.html>





**FIGURE 4 |** Highlighted amino acid changes in the protein structures for the non-synonymous substitutions or missense hotspot mutations for Indian SARS-CoV-2 genomes in (A) NSP2, (B) NSP3, (C) NSP4, (D) NSP6, (E) RdRp, (F) helicase, (G) endoRNase, (H) ORF3a, (I) Membrane, (J) Spike, (K) ORF7a, and (L) Nucleocapsid.

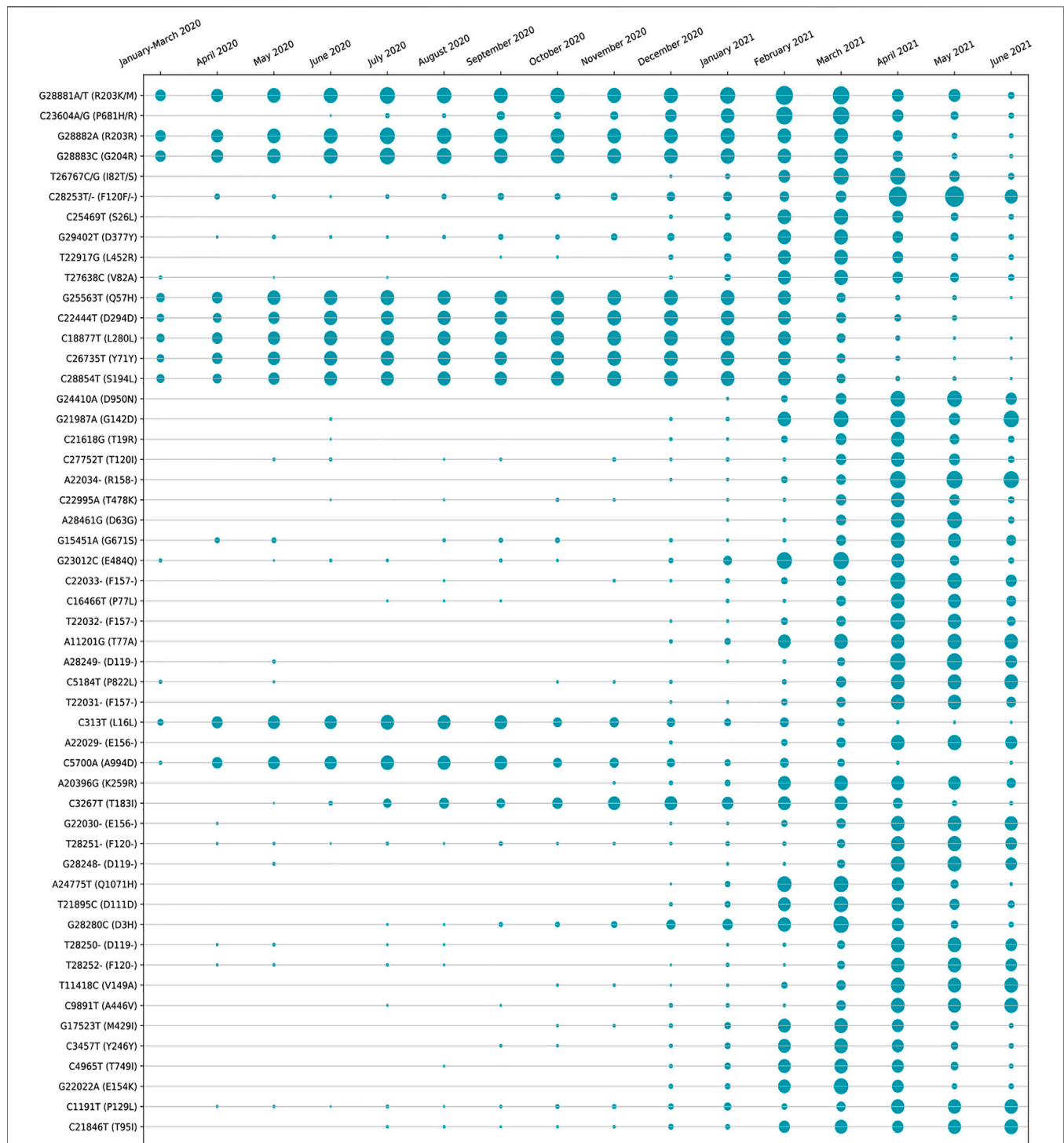


**FIGURE 5 |** Month-wise evolution of global SARS-CoV-2 genomes based on entropy.

human proteins. As a consequence, if the human protein is damaging in nature because of mutations, then the human protein-protein interactions may occur with high or low binding affinity. Now in case of virus, similar consequences may happen, which means that if the virus protein is damaged because of mutations, it may interact with

human proteins with similar binding affinity. As a result, the virus may acquire characteristics like transmissibility and escaping antibodies (Alenquer et al., 2021; Harvey et al., 2021).

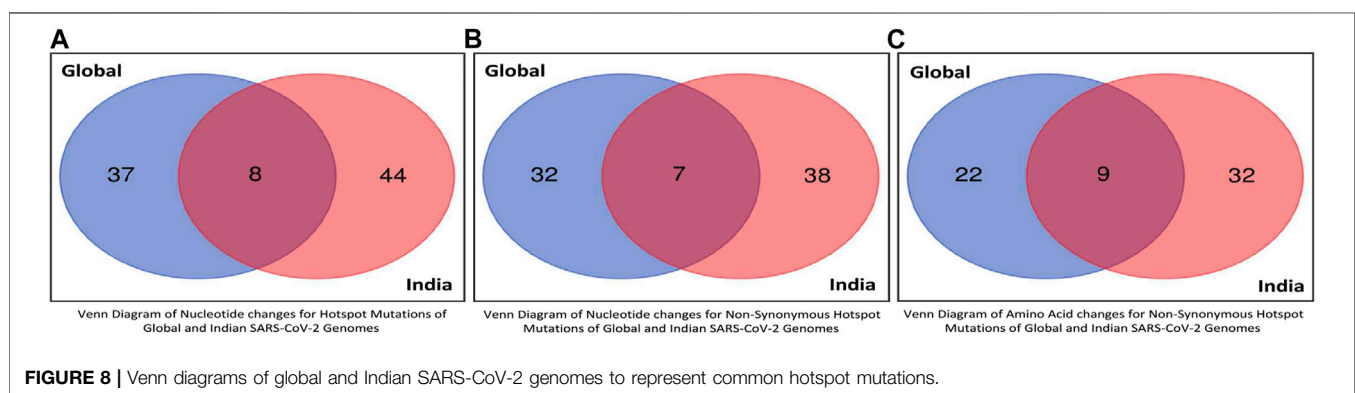
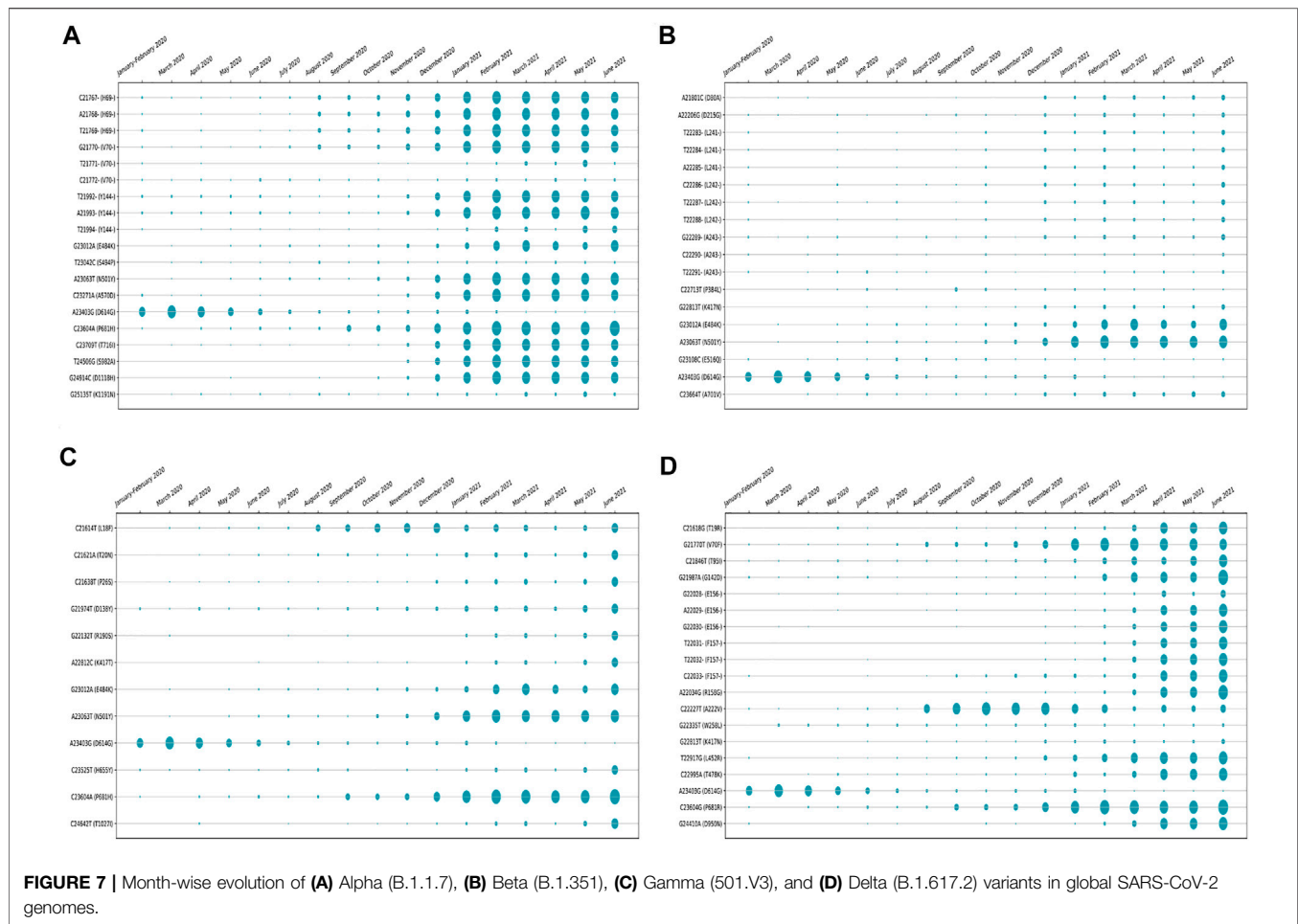
Another important parameter to judge the functional and structural activities of a protein is protein stability, which dictates



**FIGURE 6 |** Month-wise evolution of Indian SARS-CoV-2 genomes based on entropy.

the conformational structure of a protein. Any change in protein stability may cause misfolding, degradation, or aberrant conglomeration of proteins. I-Mutant 2.0 uses free energy change values (DDG (kcal/mol)) to predict the changes in the protein stability wherein a negative value of DDG indicates that the protein has a

decreasing stability, while a positive value indicates an increase in stability. For example, the very low DDG value of G25563T shows that there is a decreased protein stability, thereby resulting in a reduction of virus virulence (Cheng et al., 2021). The results from I-mutant 2.0 show that out of the 14 and 24 unique damaging changes for global and



Indian sequences, 10 and 18 changes respectively decrease the stability of the protein structures. **Figure 9** shows the binding affinity between the RBD of Spike protein and human ACE2 protein performed using SSIPe<sup>6</sup> (Huang et al., 2019) for the four mutations of SARS-CoV-2, viz., L452R, T478K, E484Q, and N501Y, taking place in such domain. The

region marked in red shows the exact positions (471–492) where the binding takes place. To report the binding affinity using SSIPe, initially the RBD region of Spike protein (Woo et al., 2020) is docked with human ACE2 protein<sup>7</sup> using PatchDock<sup>8</sup>. The best docked structure is

<sup>6</sup><https://zhanggroup.org/SSIPe/>

<sup>7</sup><https://www.rcsb.org/structure/1R42>

<sup>8</sup><http://bioinfo3d.cs.tau.ac.il/PatchDock/patchdock.html>



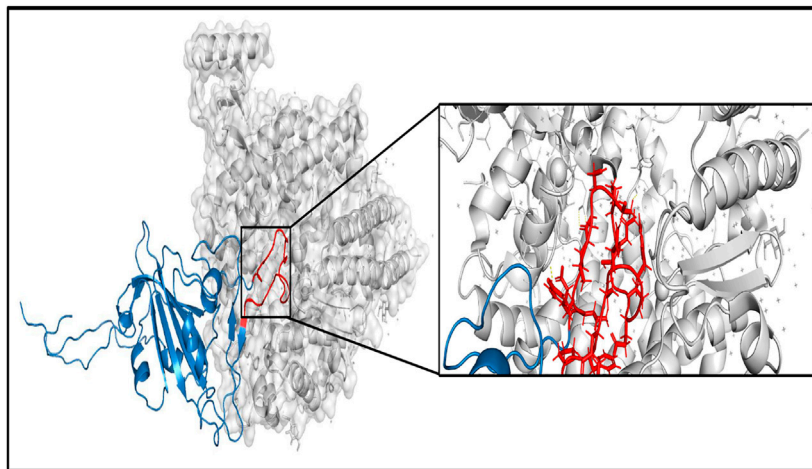
**TABLE 4 |** Functional characteristics of some important mutations.

Mutations	Functional characteristics
H69-	Leads to conformational changes in Spike protein (Meng et al., 2021; McCarthy et al., 2021)
V70-	Leads to conformational changes in Spike protein (Meng et al., 2021; McCarthy et al., 2021)
Y144-	Reduces affinity of antibody binding (McCarthy et al., 2021)
L452R	Increases the binding ability of the ACE2 receptor and can also reduce the attaching capability of vaccine-simulated antibodies with Spike protein (Garcia-Beltran et al., 2021)
T478K	Facilitates antibody escape (Planas et al., 2021)
E484Q	Associated with reduced sera neutralization (Greaney et al., 2021)
N501Y	Highest binding affinity with human receptor cell hACE2 and resistant to neutralization (Luan et al., 2021)
P681H	Near furin cleavage site, may affect transmissibility of the virus (Boehm et al., 2021)
P681R	Near furin cleavage site, may affect transmissibility of the virus (Boehm et al., 2021)

**TABLE 5 |** Sequence and structural homology-based prediction of non-synonymous substitution as hotspot mutations along with their protein structural stability for 71,038 global SARS-CoV-2 genomes.

Change in	Change in	Mapped with	PolyPhen-2		I-mutant 2.0	
Nucleotide	Amino acid	Coding regions	Prediction	Score	Stability	DDG (kcal/mol)
G28881A	R203 K	Nucleocapsid	Probably damaging	0.969	Decrease	-2.26
G28881T	R203M	Nucleocapsid	Probably damaging	0.998	Decrease	-1.52
G28883C	G204R	Nucleocapsid	Probably damaging	1	No change	0
C23604A	P681H	Spike	Not generated	Not generated	Decrease	-0.92
C23604G	P681R	Spike	Not generated	Not generated	Decrease	-0.79
G28280C	D3H	Nucleocapsid	Probably damaging	1	Increase	0.34
A23063T	N501Y	Spike	Benign	0.145	Decrease	-0.34
C3267T	T183I	NSP3	Not generated	Not generated	Decrease	-0.1
G24914C	D1118H	Spike	Probably damaging	0.998	Decrease	-0.1
T6954C	I1412T	NSP3	Benign	0.026	Decrease	-2.78
C28977T	S235F	Nucleocapsid	Probably damaging	0.998	Increase	2.43
T28282A	D3E	Nucleocapsid	Probably damaging	0.997	Decrease	-0.02
G28048T	R52I	ORF8	Probably damaging	1	Decrease	-0.09
C23271A	A570D	Spike	Benign	0.031	Decrease	-1.32
A28281T	D3V	Nucleocapsid	Probably damaging	1	Decrease	-0.22
C5388A	A890D	NSP3	Probably damaging	1	Decrease	-1.09
A28111G	Y73C	ORF8	Probably damaging	0.994	Increase	1.04
C23709T	T716I	Spike	Possibly damaging	0.696	Decrease	-0.95
T24506G	S982A	Spike	Probably damaging	0.996	Decrease	-1.36
C22227T	A222V	Spike	Benign	0.001	Increase	0.48
T26767G	I82S	Membrane	Possibly damaging	0.951	Decrease	-2
C25469T	S26L	ORF3a	Benign	0.017	Increase	0.92
G29402T	D377Y	Nucleocapsid	Probably damaging	1	Increase	0.51
T22917G	L452R	Spike	Benign	0.04	Decrease	-1.4
T27638C	V82A	ORF7a	Possibly damaging	0.732	Decrease	-2.18
G25563T	Q57H	ORF3a	Probably damaging	0.983	Decrease	-1.12
C28854T	S194L	Nucleocapsid	Probably damaging	0.994	Increase	0.45
G24410A	D950N	Spike	Possibly damaging	0.731	Increase	0.15
G21987A	G142D	Spike	Benign	0.051	Decrease	-1.17
C21618G	T19R	Spike	Benign	0.004	Decrease	-0.12
C27752T	T120I	ORF7a	Possibly damaging	0.915	Decrease	-0.26
C22995A	T478K	Spike	Benign	0	Decrease	-0.09
A28461G	D63G	Nucleocapsid	Benign	0	Decrease	-0.57
G15451A	G671S	RdRp	Probably damaging	1	Decrease	-0.29
G23012C	E484Q	Spike	Possibly damaging	0.786	Decrease	-0.48
C16466T	P77L	Helicase	Probably damaging	1	Decrease	-1.03
A11201G	T77A	NSP6	Possibly damaging	0.577	Decrease	-0.7
C5184T	P822L	NSP3	Benign	0.007	Decrease	-0.54
C5700A	A994D	NSP3	Probably damaging	0.972	Decrease	-0.78
A20396G	K259R	endoRNase	Benign	0	Decrease	-0.49
A24775T	Q1071H	Spike	Possibly damaging	0.998	Decrease	-1.19
T11418C	V149A	NSP6	Possibly damaging	0.865	Decrease	-3.43
C9891T	A446V	NSP4	Probably damaging	0.999	Increase	0.64
G17523T	M429I	Helicase	Possibly damaging	0.649	Decrease	-1.26
C4965T	T749I	NSP3	Probably damaging	0.996	Decrease	-0.92
G22022A	E154 K	Spike	Not generated	Not Generated	Decrease	-1.4
C1191T	P129L	NSP2	Possibly damaging	0.888	Decrease	-0.53
C21846T	T95I	Spike	Probably damaging	0.999	Decrease	-1.8





**FIGURE 9 |** Binding between RBD region of Spike protein (specifically 471–492 the region marked in red) and human ACE2 protein. RBD, receptor-binding domain.

**TABLE 6 |** Binding affinity of the mutations in RBD region of Spike protein and human ACE2 protein.

Genomic coordinate	Nucleotide change	Amino acid change	Protein coordinate	DDG (kcal/mol)	SSIPscore	EvoEFscore
22,917	T > G	L > R	452	1.083	2.083	−1.91
22,995	C > A	T > K	478	1.248	1.779	−0.77
23,012	G > C	E > Q	484	−0.769	1.098	−5.22
23,063	A > T	N > Y	501	0.236	0	0.09

Note. RBD, receptor-binding domain.

then provided as an input to SSIPe. **Table 6** further reports the binding affinity values for the four mutations. A strongly favorable mutation is usually defined as the one that has DDG value  $\leq -1.5$  kcal/mol, while a strongly unfavorable mutation is the one that has DDG value  $\geq 1.5$  kcal/mol. The DDG value of  $-0.769$  kcal/mol for E484Q indicates that this is a favorable mutation, while DDG values of 1.083, 1.248, and 0.236 kcal/mol for L452R, T478K, and N501Y indicate that these mutations are somewhat unfavorable. These results corroborate our earlier explanation that because of mutation, virus–human protein–protein interactions may occur with high or low binding affinity.

**Supplementary Figure S3** shows the percentage of nucleotide change and frequency of nucleotide change for hotspot mutations for global and Indian sequences. For example, in **Supplementary Figure S3A**, the occurrence of nucleotide change G > A in 71,038 global sequences is almost 45%, while the number of times it occurs in 45 hotspot mutations is two, as is also evident from **Table 2**. It can also be seen from **Supplementary Figures S3B, S3D** that 10 and 16 out of 39 and 45 non-synonymous mutations are from C to T, thereby representing abundant transition. This transition increases the frequency of codons for hydrophobic amino acids and provides evidence of potential antiviral editing mechanisms driven by host (Yuan et al., 2020). Also, more C-to-T transition means less CpG abundance, indicating rapid adaptation of virus in host. This CpG deficiency, which leads to evasion of host antiviral defense mechanisms, is exhibited the most in SARS-CoV-2 virus (Xia, 2020).

## 5 CONCLUSION

With the imminent third wave, it is very crucial to understand the evolution of SARS-CoV-2. In this regard, MSA of 71,038 SARS-CoV-2 genomes of 98 countries over the period from January 2020 to June 2021 is performed using MAFFT followed by phylogenetic analysis to visualize the evolution of SARS-CoV-2. This resulted in the identification of hotspot mutations as deletions and substitutions in the coding regions based on entropy, which should be greater than or equal to 0.3. Consequently, a total of 45 unique hotspot mutations out of which 39 non-synonymous deletions and substitutions are identified with nine unique amino acid changes for deletions and 22 unique amino acid changes for substitutions. Moreover, 10,286 Indian sequences are considered from 71,038 global SARS-CoV-2 sequences as a demonstrative example, which gives 52 unique hotspot mutations, resulting in 45 non-synonymous deletions and substitutions with five unique amino acid changes for deletions and 36 unique amino acid changes for substitutions. Some important mutations in such sequences pertaining to the Delta variant of SARS-CoV-2 are T19R, G142D, E156-, F157-, L452R, T478K, and P681R. Furthermore, the evolution of the hotspot mutations along with the mutations in variants of concern is visualized, and their characteristics are also discussed. Moreover, for all the missense mutations, the functional consequences of amino acid changes in the respective protein structures are calculated using PolyPhen-2

and I-Mutant 2.0. Finally, SSIPe is used to report the binding affinity between the RBD of Spike protein and human ACE2 protein by considering L452R, T478K, E484Q, and N501Y hotspot mutations in that region.

## DATA AVAILABILITY STATEMENT

The aligned 71038 Global SARS-CoV-2 genomes with the reference sequence and the final results of this work are available at <http://www.nitttrkol.ac.in/indrajit/projects/COVID-Hotspot-Mutation-Global-71K/>. Further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

IS and NG designed the research. IS, NG, NS, and SN analyzed the data and wrote the article. All the authors reviewed and approved the final version of the article.

## REFERENCES

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* 7, 248–249. doi:10.1038/nmeth0410-248
- Alenquer, M., Ferreira, F., Lousa, D., Valério, M., Medina-Lopes, M., Bergman, M.-L., et al. (2021). Signatures in SARS-CoV-2 Spike Protein Conferring Escape to Neutralizing Antibodies. *Plos Pathog.* 17, e1009772. doi:10.1371/journal.ppat.1009772
- Alexandersen, S., Chamings, A., and Bhatta, T. R. (2020). SARS-CoV-2 Genomic and Subgenomic RNAs in Diagnostic Samples Are Not an Indicator of Active Replication. *Nat. Commun.* 11, 1–13. doi:10.1038/s41467-020-19883-7
- Bernal, J. L., Andrews, N., Gower, C., Gallagher, E., Simmons, R., Thelwall, S., et al. (2021). Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *N. Engl. J. Med.* 385 (7), 585–594. doi:10.1056/NEJMoa2108891
- Boehm, E., Kronig, I., Neher, R. A., Eckerle, I., Vetter, P., and Kaiser, L. (2021). Novel SARS-CoV-2 Variants: the Pandemics within the Pandemic. *Clin. Microbiol. Infect.* 27, 1109–1117. doi:10.1016/j.cmi.2021.05.022
- Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure. *Nucleic Acids Res.* 33, W306–W310. doi:10.1093/nar/gki375
- Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., and Zhang, X. (2021). Functional Alterations Caused by Mutations Reflect Evolutionary Trends of SARS-CoV-2. *Brief. Bioinform.* 22, 1442–1450. doi:10.1093/bib/bbab042
- Cucinotta, D., and Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Biomed.* 91, 157–160. doi:10.23750/abm.v91i1.9397
- Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. D. S., Mishra, S., et al. (2021). Genomics and Epidemiology of the P.1 SARS-CoV-2 Lineage in Manaus, Brazil. *Science* 372, 815–821. doi:10.1126/science.abh2644
- García-Beltrán, W. F., Lam, E. C., St. Denis, K., Nitido, A. D., García, Z. H., Hauser, B. M., et al. (2021). Multiple SARS-CoV-2 Variants Escape Neutralization by Vaccine-Induced Humoral Immunity. *Cell* 184, 2372–2383.e9. doi:10.1016/j.cell.2021.03.013
- Greaney, A. J., Starr, T. N., Gilchuk, P., Zost, S. J., Binshtein, E., Loes, A. N., et al. (2021). Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host & Microbe* 29, 44–57.e9. doi:10.1016/j.chom.2020.11.007

## FUNDING

This work has been partially supported by CRG short-term research grant on COVID-19 (CVD/2020/000,991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India. However, it does not provide any publication fees.

## ACKNOWLEDGMENTS

We thank all those who have contributed sequences to GISAID and NCBI databases.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.753440/full#supplementary-material>

- Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics* 34, 4121–4123. doi:10.1093/bioinformatics/bty407
- Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., et al. (2021). SARS-CoV-2 Variants, Spike Mutations and Immune Escape. *Nat. Rev. Microbiol.* 19, 409–424. doi:10.1038/s41579-021-00573-0
- Huang, X., Zheng, W., Pearce, R., and Zhang, Y. (2019). SSIPe: Accurately Estimating Protein-Protein Binding Affinity Change upon Mutations Using Evolutionary Profiles in Combination with an Optimized Physical Energy Function. *Bioinformatics* 36, 2429–2437. doi:10.1093/bioinformatics/btz926
- Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gkf436
- Luan, B., Wang, H., and Huynh, T. (2021). Enhanced Binding of the N501Y-mutated SARS-CoV-2 Spike Protein to the Human ACE2 Receptor: Insights from Molecular Dynamics Simulations. *FEBS Lett.* 595, 1454–1461. doi:10.1002/1873-3468.14076
- McCarthy, K. R., Rennick, L. J., Nambulli, S., Robinson-McCarthy, L. R., Bain, W. G., Haidar, G., et al. (2021). Recurrent Deletions in the SARS-CoV-2 Spike Glycoprotein Drive Antibody Escape. *Science* 371, 1139–1142. doi:10.1126/science.abf6950
- Meng, B., Kemp, S. A., Papa, G., Datir, R., Ferreira, I. A. T. M., Marelli, S., et al. (2021). Recurrent Emergence of SARS-CoV-2 Spike Deletion H69/V70 and Its Role in the Alpha Variant B.1.1.7. *Cell Rep.* 35 (13), 109292. doi:10.1016/j.celrep.2021.109292
- Planas, D., Veyer, D., Baidaliuk, A., Staropoli, I., Guivel-Benhassine, F., Rajah, M. M., et al. (2021). Reduced Sensitivity of SARS-CoV-2 Variant Delta to Antibody Neutralization. *Nature* 596, 276–280. doi:10.1038/s41586-021-03777-9
- Saha, I., Ghosh, N., Maity, D., Sharma, N., Sarkar, J. P., and Mitra, K. (2020). Genome-wide Analysis of Indian SARS-CoV-2 Genomes for the Identification of Genetic Mutation and SNP. *Infect. Genet. Evol.* 85, 104457. doi:10.1016/j.meegid.2020.104457
- Saha, I., Ghosh, N., Pradhan, A., Sharma, N., Maity, D., and Mitra, K. (2021). Whole Genome Analysis of More Than 10 000 SARS-CoV-2 Virus Unveils Global Genetic Diversity and Target Region of NSP6. *Brief. Bioinform.* 22, 1106–1121. doi:10.1093/bib/bbab025
- Sarkar, R., Mitra, S., Chandra, P., Saha, P., Banerjee, A., Dutta, S., et al. (2021). Comprehensive Analysis of Genomic Diversity of SARS-CoV-2 in Different Geographic Regions of India: an Endeavour to Classify Indian SARS-CoV-2 Strains on the Basis of Co-existing Mutations. *Arch. Virol.* 166, 801–812. doi:10.1007/s00705-020-04911-0

- Singh, J., Rahman, S. A., Ehtesham, N. Z., Hira, S., and Hasnain, S. E. (2021). SARS-CoV-2 Variants of Concern Are Emerging in India. *Nat. Med.* 27, 1131. doi:10.1038/s41591-021-01397-4
- Tang, J., Toovey, O., Harvey, K., and Huic, D. (2021). Introduction of the South African SARS-CoV-2 Variant 501Y.V2 into the UK. *J. Infect.* 82, e8. doi:10.1016/j.jinf.2021.01.007
- Tang, J. W., Tambyah, P. A., and Hui, D. S. (2021). Emergence of a New SARS-CoV-2 Variant in the UK. *J. Infect.* 82, e27–e28. doi:10.1016/j.jinf.2020.12.024
- Tiwari, M., and Mishra, D. (2021). Investigating the Genomic Landscape of Novel Coronavirus (2019-nCoV) to Identify Non-synonymous Mutations for Use in Diagnosis and Drug Design. *J. Clin. Virol.* 128, 104441. doi:10.1016/j.jcv.2020.104441
- Weber, S., Ramirez, C., and Doerfler, W. (2020). Signal Hotspot Mutations in SARS-CoV-2 Genomes Evolve as the Virus Spreads and Actively Replicates in Different Parts of the World. *Virus. Res.* 289, 198170. doi:10.1016/j.virusres.2020.198170
- Woo, H., Park, S.-J., Choi, Y. K., Park, T., Tanveer, M., Cao, Y., et al. (2020). Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *J. Phys. Chem. B* 124, 7128–7137. doi:10.1021/acs.jpcb.0c04553
- Wu, S., Tian, C., Liu, P., Guo, D., Zheng, W., Huang, X., et al. (2021). Effects of SARS-CoV-2 Mutations on Protein Structures and Intraviral Protein-Protein Interactions. *J. Med. Virol.* 93, 2132–2140. doi:10.1002/jmv.26597
- Xia, X. (2020). Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. *Mol. Biol. Evol.* 37, 2699–2705. doi:10.1093/molbev/msaa094
- Yuan, F., Wang, L., Fang, Y., and Wang, L. (2020). Global SNP Analysis of 11,183 SARS-CoV-2 Strains Reveals High Genetic Diversity. *Transbound. Emerg. Dis.* doi:10.1111/tbed.13931
- Zhang, C., Zheng, W., Huang, X., Bell, E. W., Zhou, X., and Zhang, Y. (2020). Protein Structure and Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as its Intermediate Host and the Unique Similarity between its Spike Protein Insertions and HIV-1. *J. Proteome Res.* 19, 1351–1360. doi:10.1021/acs.jproteome.0c00129
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi:10.1056/NEJMoa2001017

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Saha, Ghosh, Sharma and Nandi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Structural and Drug Screening Analysis of the Non-structural Proteins of Severe Acute Respiratory Syndrome Coronavirus 2 Virus Extracted From Indian Coronavirus Disease 2019 Patients

## OPEN ACCESS

### Edited by:

Indrajit Saha,  
National Institute of Technical  
Teachers' Training and Research,  
India

### Reviewed by:

Wei-Hua Chen,  
Huazhong University of Science  
and Technology, China  
Md. Zubair Malik,  
Jawaharlal Nehru University, India

### \*Correspondence:

Nupur Biswas  
nupur@csir-iicb.res.in  
Saikat Chakrabarti  
saikat@iicb.res.in

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 06 November 2020

**Accepted:** 26 January 2021

**Published:** 09 March 2021

### Citation:

Biswas N, Kumar K, Mallick P,  
Das S, Kamal IM, Bose S,  
Choudhury A and Chakrabarti S  
(2021) Structural and Drug Screening  
Analysis of the Non-structural  
Proteins of Severe Acute Respiratory  
Syndrome Coronavirus 2 Virus  
Extracted From Indian Coronavirus  
Disease 2019 Patients.  
Front. Genet. 12:626642.  
doi: 10.3389/fgene.2021.626642

**Nupur Biswas\*, Krishna Kumar, Priyanka Mallick, Subhrangshu Das, Izaz Monir Kamal, Sarpita Bose, Anindita Choudhury and Saikat Chakrabarti\***

Structural Biology and Bioinformatics Division, Council for Scientific and Industrial Research (CSIR)-Indian Institute of Chemical Biology (IICB), Kolkata, India

The novel coronavirus 2 (nCoV2) outbreaks took place in December 2019 in Wuhan City, Hubei Province, China. It continued to spread worldwide in an unprecedented manner, bringing the whole world to a lockdown and causing severe loss of life and economic stability. The coronavirus disease 2019 (COVID-19) pandemic has also affected India, infecting more than 10 million till 31st December 2020 and resulting in more than a hundred thousand deaths. In the absence of an effective vaccine, it is imperative to understand the phenotypic outcome of the genetic variants and subsequently the mode of action of its proteins with respect to human proteins and other bio-molecules. Availability of a large number of genomic and mutational data extracted from the nCoV2 virus infecting Indian patients in a public repository provided an opportunity to understand and analyze the specific variations of the virus in India and their impact in broader perspectives. Non-structural proteins (NSPs) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) virus play a major role in its survival as well as virulence power. Here, we provide a detailed overview of the SARS-CoV2 NSPs including primary and secondary structural information, mutational frequency of the Indian and Wuhan variants, phylogenetic profiles, three-dimensional (3D) structural perspectives using homology modeling and molecular dynamics analyses for wild-type and selected variants, host-interactome analysis and viral-host protein complexes, and *in silico* drug screening with known antivirals and other drugs against the SARS-CoV2 NSPs isolated from the variants found within Indian patients across various regions of the country. All this information is categorized in the form of a database named, Database of NSPs of India specific Novel Coronavirus (DbNSP InC), which is freely available at <http://www.hpppi.iicb.res.in/covid19/index.php>.

**Keywords:** SARS-CoV2, COVID-19, non-structural proteins, database, mutation

## INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) is responsible for the global pandemic of coronavirus disease 2019 (COVID-19) (Gorbalenya et al., 2020). The SARS-CoV2 is an enveloped non-segmented positive sense single-stranded RNA virus. It belongs to the Nidovirales order and Coronaviridae family (Fehr and Perlman, 2015). Its genomic length is ~29,900 base pairs, making it one of the largest known RNA virus genomes (Fehr and Perlman, 2015; NC\_045512, 2020). The genomic structure contains a 5' cap structure and 3' poly(A) tail with 11 open reading frames (ORFs). One major characteristic feature of SARS-CoV2 genome is that almost two-thirds of the genome (~20 kb) corresponds to the replicase gene (ORF1ab), which expresses a polyprotein. The remaining part of the genome ~10 kb encodes other structural and accessory proteins. The replicase gene is followed by the ORF2 spike glycoprotein (S), ORF3a, ORF4 envelope (E) gene, ORF5 membrane (M) gene, ORF6, ORF7a, ORF7b, ORF8, ORF9 nucleocapsid phosphoprotein (N), and ORF10 (Wu et al., 2020; Yoshimoto, 2020). Among these, spike, envelope, membrane, and nucleocapsid proteins are the structural proteins, while the rest are accessory proteins. The ORF1ab polyprotein is composed of 16 non-structural proteins (NSPs).

The NSPs of any virus are encoded by the virus genome but are not included in the virus particle. For coronaviruses, NSPs play important roles in RNA synthesis and processing, helping in its survival as well as virulence power (Snijder et al., 2016). For SARS-CoV2, the first NSP (NSP1), also known as the leader protein, binds with 40S ribosomal subunit and plays an inhibitory role in mRNA translation (Narayanan et al., 2020; Thoms et al., 2020). The second NSP, NSP2, binds with host proteins and disrupts host cell environment (Angeletti et al., 2020; Yoshimoto, 2020). The third NSP (NSP3), the longest protein of SARS-CoV2, has 1,945 amino acids and is a papain-like protease. NSP3 plays multiple roles in host cells, including regulation of IRF3 and NF-kappaB signaling (Frieman et al., 2009). NSP3, NSP4, and NSP6 together play a role in host membrane rearrangements necessary for viral replication (Angelini et al., 2013). NSP5 is a 3C-like protease and cleaves at 11 distinct sites of the polyprotein to yield other NSPs (Muramatsu et al., 2016; Yoshimoto, 2020). NSP6 is known to locate at endoplasmic reticulum and generates autophagosomes (Forni et al., 2017; Benvenuto et al., 2020). The NSP7–NSP8 cofactors and NSP12 catalytic subunits create the core polymerase complex (Peng et al., 2020; Wang et al., 2020). Apart from creating complex with NSP7, NSP8 creates complex with accessory protein ORF6 also (Kumar et al., 2007). Both NSP9 and NSP10 are small non-enzymatic proteins and assist in the function of NSP12 (Zhang et al., 2020). NSP10 also interacts with NSP14 and NSP16. The NSP16–NSP10 complex provides protection to the virus from the host's innate immune system (Lin et al., 2020; Viswanathan et al., 2020). NSP11 consists of only 13 amino acids, of which the first nine are identical to the first nine amino acids of NSP12 (Yoshimoto, 2020). NSP12 is the RNA-directed RNA polymerase (RdRp) and is responsible for the replication and transcription of the RNA genome. Several probable drugs, including remdesivir, are

targeted to NSP12 (Shannon et al., 2020). NSP13 is the helicase protein, and its binding with NSP12 enhances helicase activity (Yoshimoto, 2020). NSP13, NSP14, and NSP15 can suppress interferon production and host signaling (Yuen et al., 2020). NSP14 is the guanine-N7 methyltransferase and plays a vital role in the RNA replication process (Romano et al., 2020). NSP15 is the endoribonuclease and is also a probable target of various drugs. NSP16 is the 2'-O-methyltransferase. Both NSP14 and NSP16 play vital roles in creating RNA cap in the viral genome (Krafcikova et al., 2020). Due to their pivotal roles in the replication as well as in the life cycle of SARS-CoV2, it is important to study the frequency, nature, and probable outcomes of the mutations that are being observed at the NSP regions of the virus.

The COVID-19 pandemic has spread in India, the second most populated country in the world. The total number of infected persons is 10,266,674 on 31 December 2020, which resulted in 148,738 deaths (Ministry of Health and Family Welfare Government of India, 2020) along with enormous socioeconomic disturbance (Gopalan and Misra, 2020), and the situation remains alarming to date. In this context, we have focused on the sequences of NSPs of SARS-CoV2 extracted from Indian patients and created a database, Database of NSPs of India specific Novel Coronavirus (DbNSP InC). In this manuscript, we are reporting our database, DbNSP InC, which provides exhaustive information on the NSPs of SARS-CoV2 observed in Indian patients. It provides the functional information; mutations observed in Indian patients samples; comparison of mutations with the Wuhan samples; primary and secondary structural analyses; strain and mutation analyses; and mutations observed in the deceased, mild, and asymptomatic patients samples along with the distribution of mutations across different Indian states and phylogenetic analysis. DbNSP InC is enriched with three-dimensional (3D)/tertiary structures of wild-type (WT) and mutated NSPs. The information on host protein interaction is also provided as interactive interactome networks of NSPs with host proteins and structure of host protein complexes. Molecular dynamics (MD) analysis was also performed in order to investigate the stability of the proposed complexes. *In silico* drug screening with known antiviral and other drugs was performed against the SARS-CoV2 NSPs isolated from the variants found within Indian patients across various regions of the country. The database is freely available at <http://www.hpppi.iicb.res.in/covid19/index.php>.

## MATERIALS AND METHODS

### Sequence and Mutation Data Collection

The protein sequences of SARS-CoV2 virus were collected from the EpiCoV database of GISAID (2020). The database was searched up to 8 October 2020 using keywords “hCoV-19”, “India”, and “human”. It provided 2,338 complete and high-coverage nucleotide sequences. Sequences with genomes > 29,000 bp were considered complete. Sequences with <1% Ns (undefined bases) were considered as high-coverage sequences. Corresponding protein sequences for different NSPs



were extracted. Database specific renaming (code) was done for each sequence based on the Indian state from where it was collected. Additional metadata for the sequences, which include location of sample collection, patient status, and other relevant information, were also collected.

Along with the sequences from India, human coronavirus 2019 (hCoV-19) sequences for samples collected from Wuhan, China, from where the pandemic initiated were also extracted from the GISAID database. Search with keywords “hCoV-19”, “China/Wuhan/”, and “human” yielded 255 sequences, which were used in our analysis. Sequences from different continents (North America, South America, Europe, Africa, Asia, and Oceania) were also collected in a similar fashion from the GISAID database, for comparing frequencies of the most frequent mutations of Indian samples in the global context. National Center for Biotechnology Information (NCBI) reference sequence NC\_405512.2 (NC\_045512, 2020) was considered as a reference sequence for calling the mutations. These sequences (NC\_405512.2) were collected from the human sample in Wuhan, China, in December 2019.

## Alignments, Phylogeny, and Mutation Frequency Calculation

Redundancy filter criteria via CD-HIT server (Fu et al., 2012) were applied to extract unique representative NSP sequences and to exclude redundant sequences, for each NSP of protein family. The number of CD-HIT runs was kept one, with sequence identity cutoff 1.0 (100% identity). It provided clusters of sequences that are less than 100% identical. The cluster representative sequences along with the NCBI reference sequence were aligned using the MUSCLE protein sequence alignment tool (Madeira et al., 2019). MUSCLE also constructed a phylogenetic tree for the cluster representative sequences. The tree files in the *newick* format were further used to construct an interactive phylogenetic tree using javascripts file phylotree.js (Shank et al., 2018). In-house python (version 3.4) codes were used for extracting mutations from alignment data files and calculating mutation frequencies.

## Metadata Analysis

Using the metadata of disease severity status of patients, we analyzed the association of different mutations with disease severity status. Fisher's exact test was performed using the following contingency table (Hoffman, 2019) for deceased samples,

	Mutated	Not mutated	Total
Not deceased	a	b	a + b
Deceased	c	d	c + d
Total	a + c	b + d	a + b + c + d = N

where  $N$  is the total number of sequences. Similar tables were used for mild and asymptomatic samples. The probability of obtaining a given set of result,  $p$ -value, is provided by a

hypergeometric distribution,

$$p = \frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{N}{a+b}} \quad (1)$$

where  $\binom{i}{j}$  denotes binomial coefficient of any given variable  $i$  and  $j$ .

## Strain Specific Mutational Count and Substitution Score Calculation

Distributions of mutation frequencies for Indian sequences were estimated according to their prevalence in various Indian states as the origin of the infected patients. The substitution scores for each cluster representative sequence were calculated using the point accepted mutation (PAM) matrix 250 (Dayhoff, 1969). The substitution scores are displayed as “Strain and mutation analyses” column in the DbNSP InC database. The cells are colored according to the substitution score of the observed mutations. Blank cell means no mutation was observed. All interactive plots were constructed using Google Chart API.

## Primary Structure Analysis and Secondary Structure Prediction

Primary structure analysis was done using the ProtParam tool of ExPASy server (Artimo et al., 2012) where information regarding amino acid sequence, molecular weight, isoelectric point (pI), amino acid composition, number of negatively and positively charged residues, instability index, aliphatic index, and average of hydropathicity of each reference NSP sequence are provided. Additionally, an option is implemented within the module where same information for NSP variants extracted from Indian patients can be retrieved via live search.

Similarly, secondary structure analysis was done using the PSIPRED program (Buchan and Jones, 2019) where the likelihood of each residue forming a helix, strand, or coil is provided along with a confidence score. For each protein, brief functional information, collected from the UniProt (The UniProt Consortium, 2019) was also provided.

## Structure Prediction of Wild-Type and Mutant Non-structural Proteins

SARS-CoV2 WT proteins for which the 3D structures are available were extracted from the Protein Data Bank (PDB) (Burley et al., 2019). 3D structures of WT NSPs for which structures are not available were modeled via homology modeling approach using the MODELER program (Webb and Sali, 2016). WT NSPs models were also collected from the Zhang lab COVID-19 resource (Zhang Lab, 2020) for comparison purposes.

Similarly, 3D models of the mutant (India specific) NSPs were generated using the MODELER. One hundred ensemble model structures were generated for each WT and mutant protein, and the best possible model was selected based on the MODELER DOPE score. All the 3D models were evaluated

using various structure validation tools such as PROCHECK (Laskowski et al., 1993), ERRAT (Colovos and Yeates, 1993), Verify3D (Eisenberg et al., 1997), QMEAN (Benkert et al., 2011), and ProSA (Wiederstein and Sippl, 2007). Images of the protein structures were created by the CHIMERA software (Pettersen et al., 2004).

## Host Protein Interactome Network Analysis

The SARS-CoV2 NSP and human protein–protein interactome (PPI) network (PPIN) was constructed using the interaction data made available by Gordon et al. (2020a,b) and Biogrid (Stark et al., 2006). We have considered only experimentally validated interactions. A total of 802 human interactor proteins were extracted for 15 SARS-CoV2 NSPs. Further, first layer interactors of the human proteins were collected from the STRING (Szklarczyk et al., 2019) database (version 11).

With the use each of this network, a network analysis approach was implemented to identify five types of topologically important nodes (TINs), namely, hubs, central nodes (CNs), bottlenecks (BNs) (Yu et al., 2007), global network perturbing proteins (GNPPs), and local network perturbing proteins (LNPPs) (Bhattacharyya and Chakrabarti, 2015). Network and node indices like degree, betweenness, closeness, and clustering coefficients were calculated from the extracted viral–human PPIN for identifying the TINs. TINs were calculated using previously reported methods and protocols (Bhattacharyya and Chakrabarti, 2015).

A network representation of important nodes of these NSPs and human proteins network is displayed in an interactive 3D network viewer at the DbNSP InC database. Additional functional details about the important network proteins are made available via GeneCards (Stelzer et al., 2016) link embedded within the interaction viewer window. The network is constructed using javascript-based open source technologies (three.js and 3d-force-graph.js).

## Generation of Viral–Host Protein–Protein Interaction Complex

Three-dimensional structures (models) of the selected complexes of SARS-CoV2 NSPs and human proteins (with known 3D structures) were predicted by a widely used protein docking program, PatchDock (Schneidman-Duhovny et al., 2005). PatchDock allows geometric shape complementarity matching with the help of geometric hashing and pose-clustering techniques. The top 100 solutions from PatchDock-based docking score were clustered according to the root mean square deviation (RMSD) in CHIMERA software (Pettersen et al., 2004) to determine the largest docked clusters. The top scoring solution from the largest cluster was selected as representative pose with the assumption that clusters having a higher number of similar frames are more likely to possess the best possible interaction pose.

One hundred and thirteen complex structures were generated using seven known NSP structures and 41 predicted (5 WT and 36 mutant) NSP proteins with 28 human proteins of known

structures. The human proteins were chosen based on the availability of high-quality crystal structures.

PISA software (Krissinel and Henrick, 2007) was used to calculate the structural and chemical properties of the macromolecular interfaces such as interface area, free energy of dissociation, presence of hydrogen bond and salt bridges. The strength of the binding at the interface was estimated via free energy of formation ( $\Delta G_{int}$ ) and solvation energy (SE) gain ( $\Delta G_{solv}$ ). Various types of molecular interactions, such as hydrogen bond and salt bridges, formed by the two interacting chains at the interface were also calculated and provided within the respective window of the complexes at the DbNSP InC database.

Calculation of fraction of conserved native contacts (FNATs) with respect to a reference complex/interface is a standard complex evaluation criterion. FNAT is the number of native (correct) residue–residue contacts in the docked (predicted) complex divided by the number of contacts in the original (known). According to Critical Assessment of PRedicted Interactions (CAPRI) (Lensink et al., 2020) criteria, predicted complexes with  $10\% \leq \text{FNAT} < 30\%$  are regarded as acceptable predictions,  $30\% \leq \text{FNAT} < 50\%$  as medium-quality predictions, and  $\text{FNAT} \geq 50\%$  as high-quality predictions. In this case, we have evaluated the alteration of the interface formed by the mutant NSPs with respect to the WT protein complex via calculation of FNAT. FNAT values of both the chains forming the complex are provided in the DbNSP InC database.

## Molecular Dynamics Analysis

The 3D structures of WT and mutant NSPs as well as complexes of NSPs (WT and mutant) and human proteins were subjected to MD simulation to study the impact of mutation on the structural dynamics by using the Desmond (Bowers et al., 2006) MD simulation package. Further, MD simulations of the NSPs complexed (docked) with antiviral drugs were also performed using the GROMACSv4.5.3 simulation package (Abraham et al., 2015) to understand the structural and energetic stabilities of the proposed protein–drug complexes.

In Desmond (Krissinel and Henrick, 2007) MD simulations, OPLS\_2005 force field parameters (Kaminski et al., 2001) were used to generate the coordinates and topology of the molecules. The system was solvated with TIP3P (Mark and Nilsson, 2001) water, and counter ions were added to neutralize the overall charge of the system. Orthorhombic periodic boundary conditions were defined to specify the shape and size of the simulation box buffered at 10-Å distances from the molecules. A hybrid method combining the steepest decent and the limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm (Saputro and Widyaningsih, 2017) was used to minimize the energy of the system. Further, the system was equilibrated in NVT followed by NPT conditions using default protocol of Desmond. Finally, the production run was performed at 300K temperature and 1 atm pressure with a time step of 2 fs for 200 ns. The temperature and pressure of NPT ensemble were regulated by using Nosé–Hoover chain thermostat (Evans and Holian, 1985) and Martyna–Tobias–Klein barostat (Martyna et al., 1994), respectively. Reversible reference system propagator

algorithms (RESPA) (Tuckerman et al., 1992) was used for integrating the equations of motion. Trajectories were recorded at every 4.8 ps and analyzed by Desmond “simulation analysis tool.” Energy profile during simulation was analyzed by “simulation quality analysis tool” of Desmond package. RMSD and root mean square fluctuations (RMSFs) of the protein residues were analyzed using the “simulation event analysis” module.

Each antiviral drug complexed with SARS-CoV2 NSPs obtained from docking analyses was subjected to MD simulation using the GROMACSv4.5.3 simulation package (Abraham et al., 2015). Coordinates and topology files of receptor molecule were generated with *Amberff99sb* force field (Case et al., 2005). The topology and coordinate files of ligands were generated using ACPYPE (AnteChamber PYthon Parser interface) (Sousa Da Silva and Vranken, 2012). A cubic simulation box was defined and filled with TIP3P water (Mark and Nilsson, 2001) molecules. Two-stage minimization of the system was performed using the steepest-descent (Nocedal and Wright, 2006) and conjugate-gradient (Straeter, 1971) minimization algorithms. The system was equilibrated under NVT (constant number of particles, volume, and temperature) and NPT (constant number of particles, pressure, and temperature) conditions for 500 ps at a temperature of 300K and 1 atm pressure. After equilibration step, final production run was performed under NPT condition for 10 ns at 300K temperature and 1 atm pressure. Trajectories were saved at the interval of 0.02 ps, and a total of 500,000 snapshots were recorded. A total of 100 snapshots, recorded at the interval of 100 ps, were used to calculate the binding free energy using *g\_mmpbsa* tool (Kumari et al., 2014).

## High-Throughput Virtual Screening of Antivirals and Known Drugs Against the Novel Coronavirus 2 Non-structural Proteins

A high-throughput virtual screening (HTVS) technique was employed to identify the efficient binders of NSP structures that may serve as potential inhibitors for various NSPs. In this work, two different small molecule datasets were utilized to identify the potential binders. For the screening of first dataset, all known antiviral drugs (111 compounds) were collected from DrugBank (2020) database, were docked onto the NSP structures (NSP5, NSP12, NSP13, NSP14, NSP15, and NSP16), and were ranked by using all the fitness scores (GoldScore, ChemPLP, Chemscore, and ASP) of GOLD docking software (Jones et al., 1997). The GOLD software optimizes the fitness score of many possible docking solutions using a genetic algorithm. The following parameters were used in the docking cycles: population size (100), selection pressure (1.10), number of operations (100,000), number of islands (5), niche size (2), crossover weight (95), mutation weight (95), and migration weight (10). The docking scores were normalized to 0 to 1 scale by using the following formula:

$$Score_{Normalized} = \frac{(S - S_{min})}{(S_{max} - S_{min})} \quad (2)$$

where  $S$  is raw docking score of a particular molecule, and  $S_{max}$  and  $S_{min}$  are the maximum and minimum docking scores in the top quartile solutions, respectively.

For the screening of second dataset, all the small molecule known drugs and/or drug-like substances available in the DrugBank (2020) database (8,736 compounds) were extracted, and the same strategy used for the screening of antiviral drugs (described above) was followed to identify the potential inhibitors for NSP structures.

Antivirals and known drug molecules commonly appearing (at least in three scoring schemes) among the top 25% solutions of each fitness score were considered as probable inhibitors of the target SARS-CoV2 NSPs. The probable inhibitors were identified and ranked based on the average normalized score. All the probable inhibitors identified from the antiviral drug dataset were subjected to MD simulation followed by binding free energy calculation to check the stability of the protein–ligand complex.

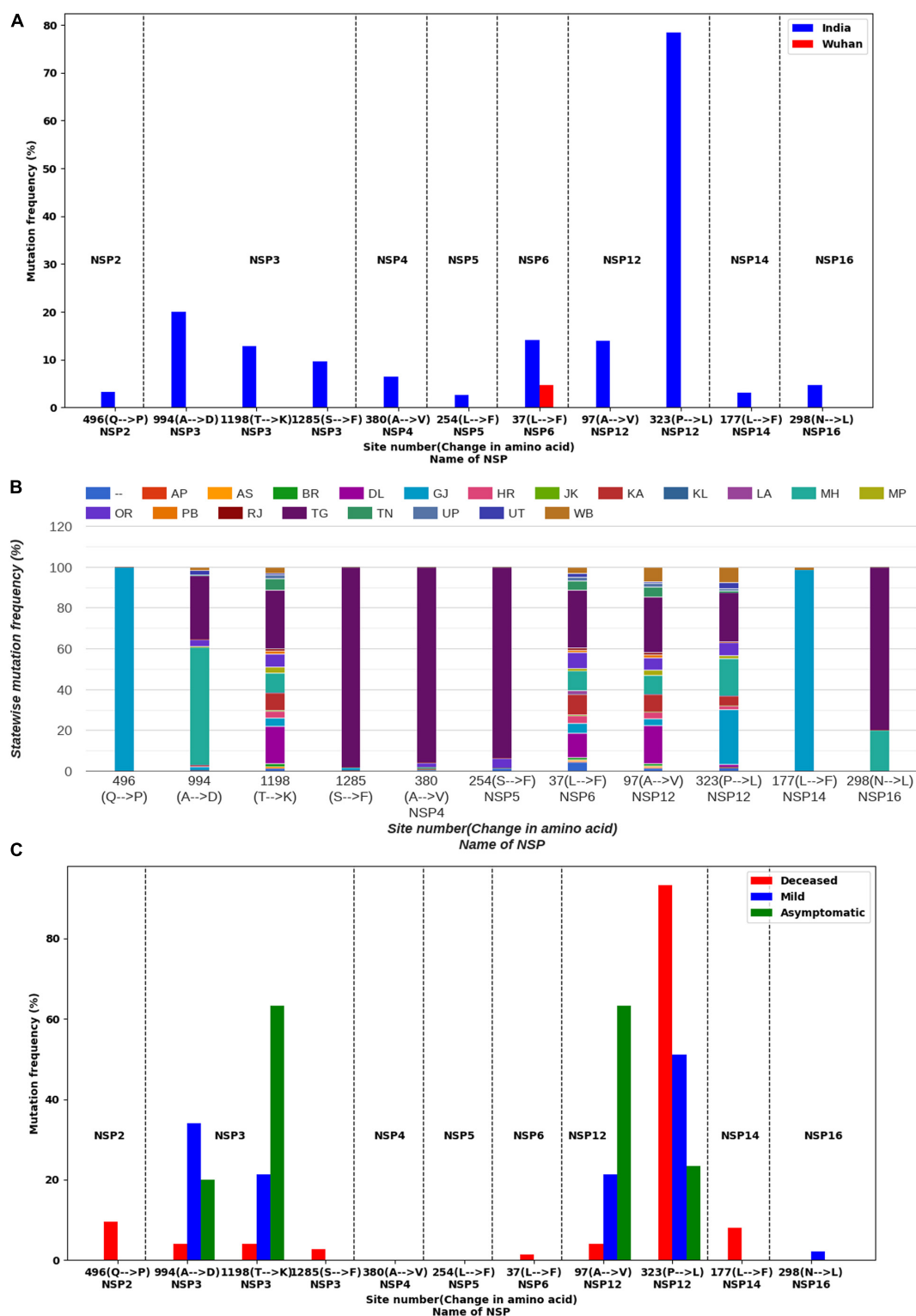
## RESULTS

### Mutational Frequency Analysis of the Indian and Wuhan Novel Coronavirus 2 Variants

Mutations were identified within the sequences of NSPs collected from India and Wuhan, China. The mutation frequencies were calculated, and their distribution plots for each NSP are displayed in the database DbNSP InC under the column “Mutation frequency.” Higher ( $\geq 2.5\%$  of the total 2,338 samples) frequencies of mutations in NSPs from the Indian samples were observed especially for NSP2, NSP3, NSP4, NSP5, NSP6, NSP12, NSP14, and NSP16. On the other hand, NSP1, NSP7, NSP8, NSP9, NSP10, NSP13, and NSP15 show lower mutation frequencies ( $< 2.5\%$ ) for the Indian samples. **Figure 1A** lists the mutations for different NSPs within the Indian population where the mutation frequency is more than 2.5%.

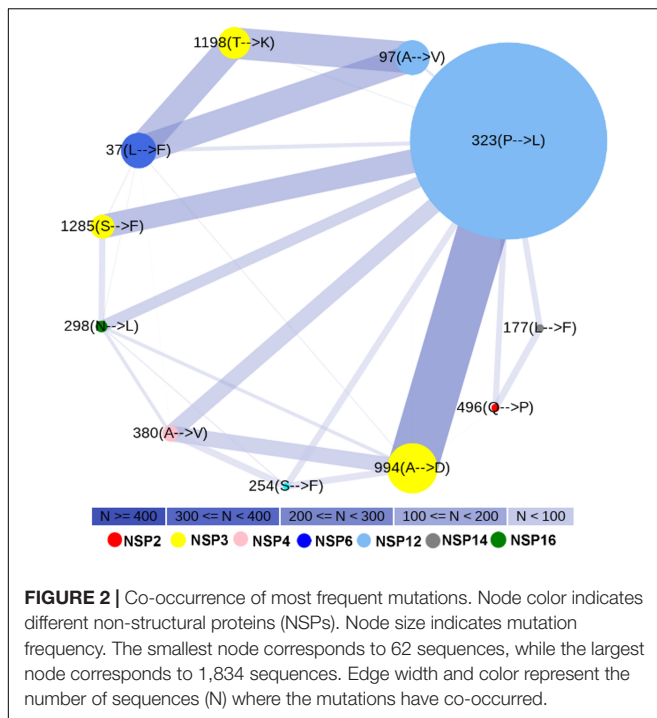
We observed in NSP12 that the RdRp has the most observed mutations at site 323, having a mutation frequency of 78.44% and that the mutation is from amino acid proline (P) to leucine (L). NSP12 sequences possess another mutation at site 97(A→V) having a frequency of 13.9%. NSP3 is the longest NSP and has a maximum number of mutations. The highest mutation frequency (20.02%) observed for NSP3 is at 994(A→D). NSP3 has two more frequently mutated sites, 1198(T→K) having a mutation frequency of 12.75% and 1285(S→F) 9.58% frequency. NSP2 has a mutation at site 496(Q→P) of 3.21% frequency. NSP4 has a mutation at site 380(A→V) with a frequency of 6.42%, while NSP5 has a mutation at site 254(S→F) with a frequency of 2.65%. Similarly, NSP6, NSP14, and NSP16 have mutations at the sites 37(L→F), 177(L→F), and 298(N→L) with mutation frequencies of 14.16, 3.12, and 4.66%, respectively.

We compared the mutations observed in Indian sequences with the mutations observed in Wuhan sequences and found significant differences in these two types of samples (**Figure 1A**). For NSP1, mutation frequencies are low for both the Indian and Wuhan samples. However, for NSP2, site 198(V→I) has been



**FIGURE 1 |** Mutation analysis of non-structural proteins (NSPs) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) sequences having frequencies  $\geq 2.5\%$  in India. **(A)** Comparison of mutation frequencies for different NSPs from samples collected from India (blue) and Wuhan, China (red). Dashed lines are drawn to separate NSPs. **(B)** Distribution of mutations in different Indian states. **(C)** Occurrence of mutations in different types of patients.





mutated in 2.75% of the Wuhan samples and 1.37% of the Indian samples (**Supplementary Figure S1**). No mutation was observed at site 496 of NSP2 for the Wuhan samples, indicating that a mutation at 496(Q→P) is specific to the Indian samples. For NSP3, the Wuhan samples showed a mutation at 1937(T→I), which was not observed in the Indian samples (**Supplementary Figure S1**). However, the Indian samples have shown three highly mutated sites (994, 1198, and 1285) as shown in **Figure 1A**. On the other hand, 230(E→G) site of NSP4 has 3.53% mutation frequency for the Wuhan samples and no mutation for the Indian samples (**Supplementary Figure S1**). Similarly, site 120(G→C) of NSP5 has 3.14% mutations for the Wuhan samples but no mutations for the Indian samples. For NSP6, a mutation at 37(L→F) was observed for both the Indian and Wuhan samples, having frequencies of 14.16 and 4.71%, respectively. NSP7, NSP8, NSP9, and NSP10 appear to have very low mutating sites for both the Indian and Wuhan samples. For the Wuhan samples, NSP12 mutated only at site 415(F→S) with a frequency of 6.67% (**Supplementary Figure S1**). NSP13, NSP14, NSP15, and NSP16 showed a mutation frequency <2.5% for the Wuhan samples.

## State-Wise and Strain-Wise Mutational Analyses of the Indian Variants

We analyzed the presence of mutations across samples collected from different Indian states. The information of the state was not available for some samples, which are marked as “–” in the DbNSP InC database. Other state names are mentioned in an abbreviated form. The abbreviation information is provided at the “Info” page of the database.

We observed marked differences in the mutation frequency across the Indian states, indicating regional accumulation

of certain mutation types. **Figure 1B** shows the state-wise appearances of different mutations. **Figure 2** shows the co-occurrence of mutations across different samples. For example, two major mutating sites, 994(A→D) and 1198(T→K), for NSP3 never co-appeared in the same sample. We also noticed that 57.69% of mutations at 994(A→D) was observed in Maharashtra (MH) state (**Figure 1B**). For mutation 1198(T→K), 28.52% mutations appeared at samples from the state of Telangana (TG) and 18.46% from Delhi (DL). Similar accumulation of certain mutation types was noticed in NSP12 also. The most frequent variant within Indian patients [NSP12: 323(P→L)] has 26.72% representation from the state of Gujarat (GJ), followed by TG (24.21%) and MH (18.21%) (**Figure 1B**). However, for site 97, only 3.38% mutations were observed at samples from GJ and 9.23% for MH. TG has the highest contribution (27.08%) for a mutation at site 97. It indicates that sequences having a mutation at 323 have a tendency of not to be mutated at site 97. However, West Bengal (WB) shares 7.38 and 7.58% of mutations at sites 97 and 323, respectively, indicating a possible co-occurrence of these two mutations. The strain-wise analysis also revealed similar features of the mutual exclusiveness of mutations at sites 97 and 323 for sequences from GJ and TG. We observed 22 sequences have a mutation at both sites 97 and 323. Out of these 22 sequences, 15 are from WB indicated the existence of a variant of NSP12 where both 97 and 323 sites are mutated.

**Figure 2** shows the existence of a broad edge between 994(A→D) of NSP3 and 323(P→L) of NSP12, which is due to their co-occurrence in 19.76% samples. We observed that mutations 1198(T→K) of NSP3 and 97(A→V) of NSP12 occurred simultaneously at 12.49% of samples. Two other broad edges are connected with 37(L→F) of NSP6. These are due to a co-occurrence of 37(L→F) of NSP6 with 1198(T→K) of NSP3 in 10.91% samples and a co-occurrence of 37(L→F) of NSP6 with 97(A→V) of NSP12 in 10.95% samples.

From the PAM 250 matrix (Dayhoff, 1969), we observed that the substitution scores for T→K, A→V, and A→D are 0, indicating that the mutations are tolerable whereas the substitution score of -3 at 323(P→L) mutation (**Supplementary Figure S2**) indicates probable deleterious impact. We observed mutations 323(P→L) of NSP12 and 1285(S→F) of NSP3, both having substitution scores of -3, which co-occurred at 4.94% samples (**Supplementary Figure S2** and **Figure 2**). Mutations 323(P→L) of NSP12 and 298(N→L) of NSP16, both having substitution scores of -3, co-occurred at 4.53% samples. Mutations 496(Q→P) of NSP2 and 380(A→V) of NSP4 have substitution scores of 0. On the other hand, L→F mutation observed at site 37 of NSP6 and at 177 of NSP14 has a substitution score of +2 (**Supplementary Figure S2**).

## Patient Status and Disease Severity-Wise Mutational Analysis

We further analyzed the metadata available with the sequencing data in order to associate the observed mutations with the clinical status/manifestation of the patients. We found, out of 2,338 sequences, that the patient status of 74 sequences was marked as deceased. Forty-seven sequences had patient status “mild,” and 30



were marked as “asymptomatic.” We analyzed the mutations in these samples, and comparative plots of occurrence of mutations for these three types of samples are provided in the DbNSP InC database as “*Mutation in different types of patients*” for different NSPs and are partially reconstructed in **Figure 1C**. We observed that NSP2 mutation 496(Q→P) was present in 9.46% of deceased samples. For NSP3, both mutations 994(A→D) and 1198(T→K) are mostly associated with mild and asymptomatic samples, respectively. Mutation 37(L→F) of NSP6 has a similar trend; 31.91% mild samples and 63.33% of asymptomatic samples showed 37(L→F) mutation, whereas only 1.35% deceased samples had a mutation at 37(L→F). On the contrary, a mutation at 323(P→L) of NSP12 was present in 93.24% of the deceased samples; 51.06% of mild samples and 23.33% of asymptomatic samples have 323(P→L) mutations. Another major mutation of NSP12, 97(A→V) is mostly associated with mild (21.28%) and asymptomatic (63.33%) samples. For NSP14, mutations at 177(L→F) are associated only with deceased (8.11%) samples. These were not observed in the asymptomatic and mild type of samples. We did not find patient status data for NSP4 and NSP5 mutations. Since the number of samples having patient status is quite small, to explore the statistical significance of our observations, we performed Fisher’s exact test. The mutations having  $p\text{-value} \leq 0.05$  in Fisher’s exact test are listed in **Supplementary Table S1** along with their significance level.

## Structural Analysis of the Wild-Type and Mutant Non-structural Proteins

Three-dimensional model structures of 5 WT NSPs and 36 mutant NSPs extracted from Indian patients were

generated, and their structural validations were done using various structure validation tools (**Table 1**). 3D structures were modeled via homology modeling approach using the MODELER program (Webb and Sali, 2016). WT NSP models collected from the Zhang lab COVID-19 resource are also displayed for comparison purposes (Zhang Lab, 2020). 3D coordinates of these models are made available via the DbNSP InC database, and the corresponding links are provided under the “3D/Tertiary structure analysis” analysis column. **Figure 3** shows the structures of the most frequently mutated NSP proteins along with their WT structures.

## Viral–Host Protein–Protein Interaction Network Analysis

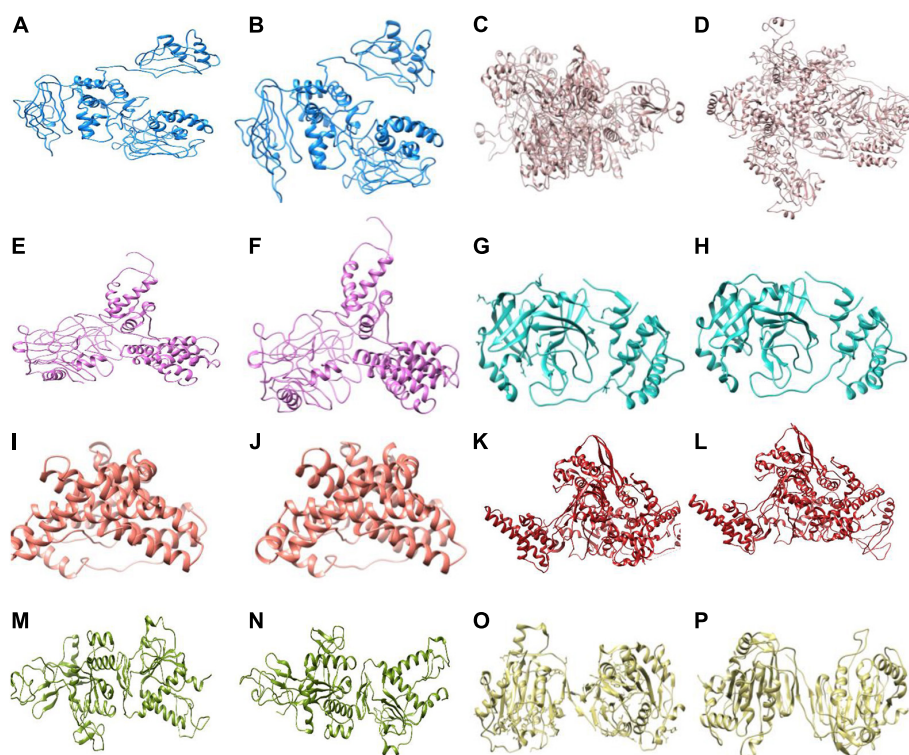
We found a total of 802 human interactor proteins for 15 NSPs. The viral–host PPIN was constructed for each NSP to identify TINs/proteins, namely, hubs, CNs (Bhattacharyya and Chakrabarti, 2015), BNs (Yu et al., 2007), GNPPs, and LNPPs (Bhattacharyya and Chakrabarti, 2015). Further, important interacting proteins (IIPs) were identified using overlap among any two TINs as described in our earlier report (Bhattacharyya and Chakrabarti, 2015). **Table 2** shows the number of IIPs extracted from the SARS-CoV2 and human PPIN. These IIPs may play crucial roles in mediating viral–human interactions. The network representation of these important proteins is displayed in an interactive 3D network viewer at the DbNSP InC database for each NSP. **Figure 4** shows the network for NSPs where different TINs are marked in different colors.

**TABLE 1** | Information of homology models/crystal structures of frequently mutated (mutation frequency  $\geq 2.5\%$  in India) NSPs and corresponding wild type NSPs.

NSP	Mutation	Sequence length	Template/crystal structure pdb id	Verify3D (%) (Eisenberg et al., 1997)	ERRAT (%) (Colovos and Yeates, 1993)	QMEAN (Benkert et al., 2011)
NSP2	Wild type <sup>a</sup>	1–638	NA	76.96	42.05	–13.1
	496(Q→P)	1–638	Wild type	79.47	34.92	–12.34
NSP3	Wild type <sup>a</sup>	1–1945	NA	NA	50.50	–9.46
	994(A→D)	1–1945	Wild type	NA	42.87	–8.71
	1198(T→K)	1–1945	Wild type	NA	43.44	–9.03
NSP4	Wild type <sup>a</sup>	1–500	NA	72.60	49.06	–10.88
	380(A→V)	1–500	Wild type	80.20	42.87	–10.58
NSP5	Wild type <sup>b</sup>	1–306	6w63	93.11	97.24	0.30
	254(S→F)	1–306	6w63	91.83	93.96	–0.74
NSP6	Wild type	1–290	<i>ab initio</i>	83.10	96.44	–2.1
	37(L→F)	1–290	Wild type	87.93	90.78	–2.45
NSP12	Wild type <sup>b</sup>	1–932	6yyt	87.34	96.70	–1.53
	97(A→V)	1–932	6yyt	85.87	73.6	–2.19
	323(P→L)	1–932	6yyt	88.1	75.95	–1.98
NSP14	Wild type <sup>b</sup>	1–527	5c8t	85.39	62.28	–2.6
	177(L→F)	1–527	Wild type	88.05	55.23	–2.91
NSP16	Wild type <sup>b</sup>	1–298	6w75	76.06	90.95	–0.79
	298(N→L)	1–298	Wild type	96.31	84.43	–1.65

<sup>a</sup>Model structure is adopted from Zhang Lab (2020).

<sup>b</sup>Crystal structure.



**FIGURE 3 |** 3D structures (shown in cartoon representation) of the most frequently mutated non-structural proteins (NSPs) along with their wild-type (WT) structure. **(A)** NSP2(WT), **(B)** NSP2[496(Q→P)], **(C)** NSP3(WT), **(D)** NSP3[994(A→V)], **(E)** NSP4(WT), **(F)** NSP4[380(A→V)], **(G)** NSP5(WT), **(H)** NSP5[254(S→F)], **(I)** NSP6(WT), **(J)** NSP6[37(L→F)], **(K)** NSP12(WT), **(L)** NSP12[323(P→L)], **(M)** NSP14(WT), **(N)** NSP14[177(L→F)], **(O)** NSP16(WT), and **(P)** NSP16[298(N→L)].

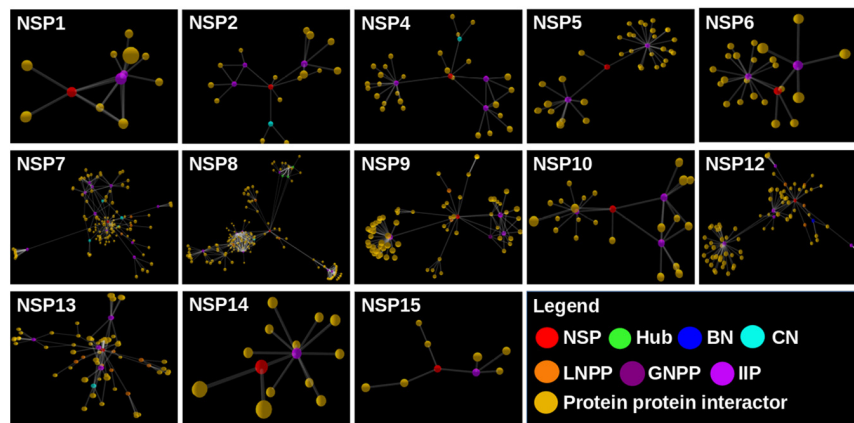
**TABLE 2 |** Number of important interacting proteins (IIPs) for each NSP-human protein interaction network.

NSP	Number of interactors	Number of IIPs
NSP1	12	3
NSP2	20	4
NSP4	33	4
NSP5	37	3
NSP6	25	3
NSP7	133	11
NSP8	232	8
NSP9	76	4
NSP10	26	4
NSP12	102	6
NSP13	83	5
NSP14	14	2
NSP15	9	2

## Generation of 3D Structures of Viral Non-structural Proteins and Human Interacting Proteins

Extensive protein-protein docking approach implemented via PatchDock program was employed to generate 113 complex structures using 7 known NSP structures and 41 predicted (5

WT and 36 mutant) NSP proteins with 28 human proteins with known structures (**Supplementary Table S2**). Further, structural and chemical properties of the predicted interfaces such as interface area, free energy of dissociation, presence of hydrogen bond and salt bridges, free energy of formation ( $\Delta G_{int}$ ), and SE gain ( $\Delta G_{solv}$ ) were calculated to characterize the interfaces (**Supplementary Table S2**). Finally, using FNAT-based criteria, we have evaluated the alteration of the interface formed by the mutant NSPs with respect to the WT protein complex. **Supplementary Table S2** and **Figure 5** show the interfaces that may have altered significantly in complexes formed by the mutant proteins. Almost 45% of the complexes formed by the mutant NSPs show a significant alteration (FNAT  $\leq 50\%$  for both viral and human proteins forming the probable interaction interface) of the binding interface with respect to that formed by their WT counterparts (**Figure 5A**). Thirty-four percent of the complexes formed by the mutant NSPs show a significant alteration of the interface (FNAT  $\leq 50\%$ ) in either viral or human protein partners. However, the complexes formed by the WT and mutant NSPs are found to be energetically stable as shown by relatively low deviation of overall energy of the complexes before and after 100 ns of MD simulations (**Figure 5B**). **Figure 5C** shows one of the examples of a significant alteration of the binding interfaces in NSP12 and human interactor protein, peptidyl-prolyl isomerase like-3 (PPIL3), perhaps due to the mutation at position 323(P→L) of NSP12.



**FIGURE 4 |** Network view of the interactome of non-structural proteins (NSPs) with their human interactor proteins and their first layer of interactors. Different topologically important nodes (TINs) are marked in different colors. Red, NSPs; yellow, protein–protein interactors; green, hubs; blue, bottlenecks; cyan, central proteins; orange, local network perturbing protein (LNPP); purple, global network perturbing proteins (GNPPs); magenta, important interacting proteins (IIPs).

## In silico Drug Screening With Known Antiviral and Other Drugs Against the Novel Coronavirus 2 Non-structural Proteins

A total of 111 antiviral compounds and 8,736 known drugs and/or drug-like substances available in the DrugBank (2020) were screened against the NSP WT structures using the GOLD docking software (Jones et al., 1997) where all the fitness scores (GoldScore, ChemPLP, Chemscore, and ASP) were implemented. Compounds commonly appearing (at least in three scoring schemes) among the top 25% solutions of each fitness score were considered as probable inhibitors and were further ranked based on the average value of normalized fitness scores. **Figure 6** shows the top five antiviral and known drugs that are likely to act as inhibitors for the SARS-CoV2 NSPs. Several antivirals such as indinavir, nelfinavir, inarigivir soproxil, and doravirine were found to be targeting multiple NSPs. Similarly, known drugs like montelukast and GSK-1004723 seem to bind three or more NSPs as probable targets. Interestingly, the types of antiviral drugs and their relative ranks based on the normalized docking score changed significantly with respect to the WT when the screening was performed against the most frequent mutants of the targeted NSPs {NSP5[254(S→F)], NSP12[323(P→L)], NSP13[253(Y→H)], NSP14[177(L→F)], NSP15[109(K→N)], and NSP16[298(N→L)]} (**Figure 7**). These findings indicate that drug sensitivity can get altered due to the mutations in the NSPs.

MD simulations implemented by GROMACS were also undertaken to evaluate the structural and energetic stabilities of the drug–NSP complexes retrieved from the molecular docking-based screening procedure. Drug–NSP complexes with progressive stabilized binding free energy profiles suggest better stability. **Figure 7** shows higher a fraction of the WT complexes that remain stable ( $\pm 20\%$  deviation) or getting more stable ( $> 20\%$  deviation) in terms of binding free energy throughout the duration of the simulation. For most of the NSPs, the highest peaks observed either for no deviation or at positive

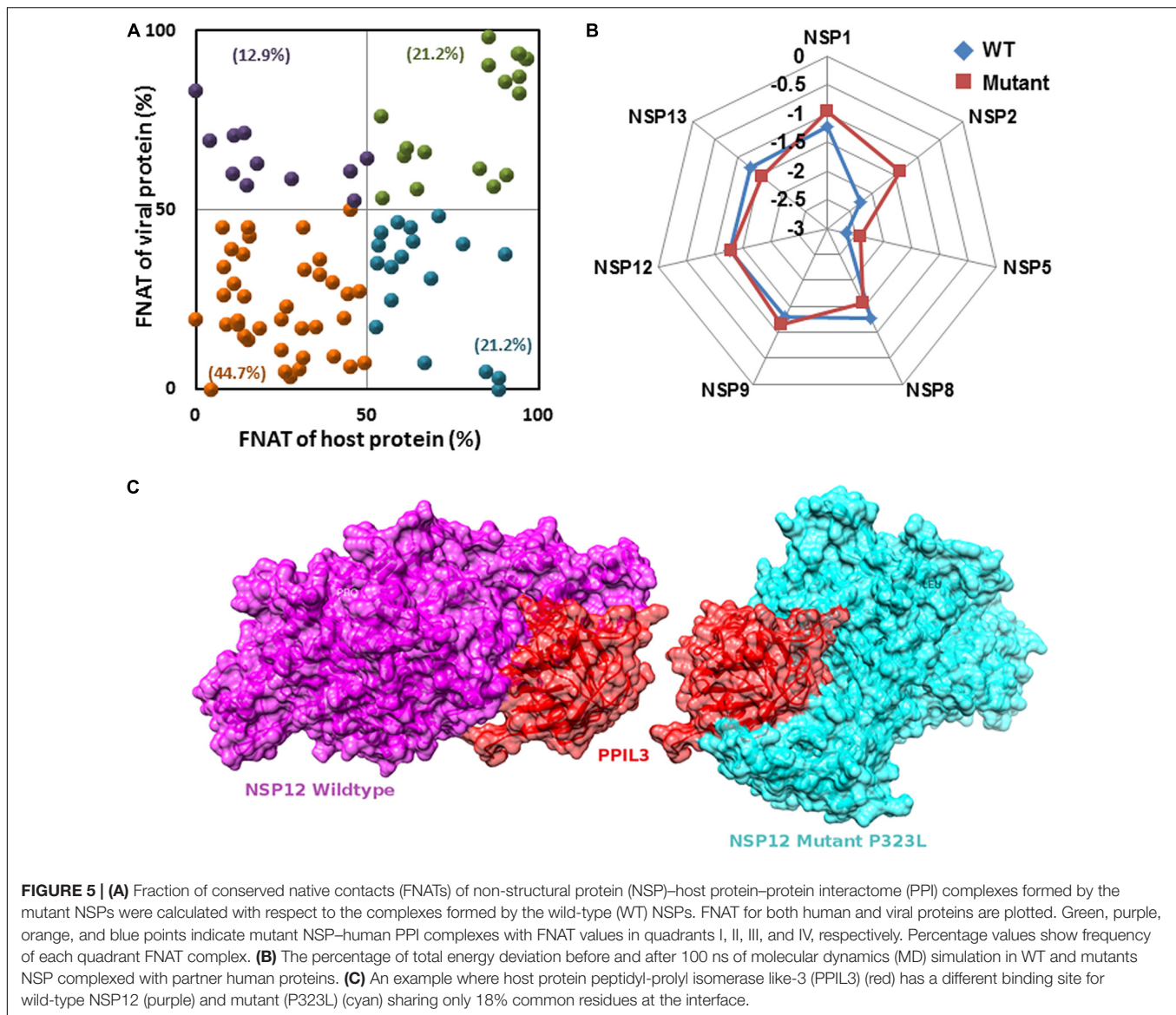
binding energy deviation ranges indicate the stability of the complexes (**Figure 8**).

## Molecular Dynamics Analysis

Structural flexibilities represented by RMSD and RMSF of the WT and mutant NSPs were calculated and compared to evaluate the probable structural and functional alterations that might be due to the mutations. The current version of DbNSP InC provides MD results of WT and mutated NSP1, NSP2, NSP5, NSP8, and NSP12. **Figure 9** shows the RMSD, RMSF, and energy profiles of selected mutants from NSP2 and NSP12 as examples to demonstrate marked variations with respect to their WT counterparts. For NSP2, a mutation at 496(Q→P) resulted in lower RMSD (**Figure 9A**) and higher energy (**Figure 9C**), whereas RMSF remains almost equally fluctuating compared with WT NSP2 (**Figure 9B**). For the most prevalent mutation in India, 323(P→L) of NSP12, RMSD has increased (**Figure 9D**), RMSF (**Figure 9E**) has reduced significantly, and energy has reduced (**Figure 9F**) compared with those in the WT variant. It indicates that 323(P→L) is likely to be a stable mutation for NSP12.

Similarly, viral–human protein complexes were also undertaken for MD simulations, and the energy profiles of the complexes during the simulation run were compared between selected mutants and their respective WT NSPs. The current version of DbNSP InC provides MD results of complexes of WT and mutant NSP1, NSP2, NSP4, NSP5, NSP9, NSP12, and NSP13. For each NSP, a complex with one human interactor protein was simulated. The interactor protein was selected based on their topological importance in the corresponding interactome network. **Figure 10** shows the representative data for NSP2 and NSP12. For WT and mutant 496(Q→P) complex of NSP2 with human protein EIF4E2, RMSD (**Figure 10A**), RMSF (**Figure 10B**), and energy variation (**Figure 10C**) are shown. EIF4E2 is known to be associated with interferon gamma signaling and innate immune system pathways (Stelzer et al., 2016). In the interactome of NSP2, EIF4E2 appears as an IIP,





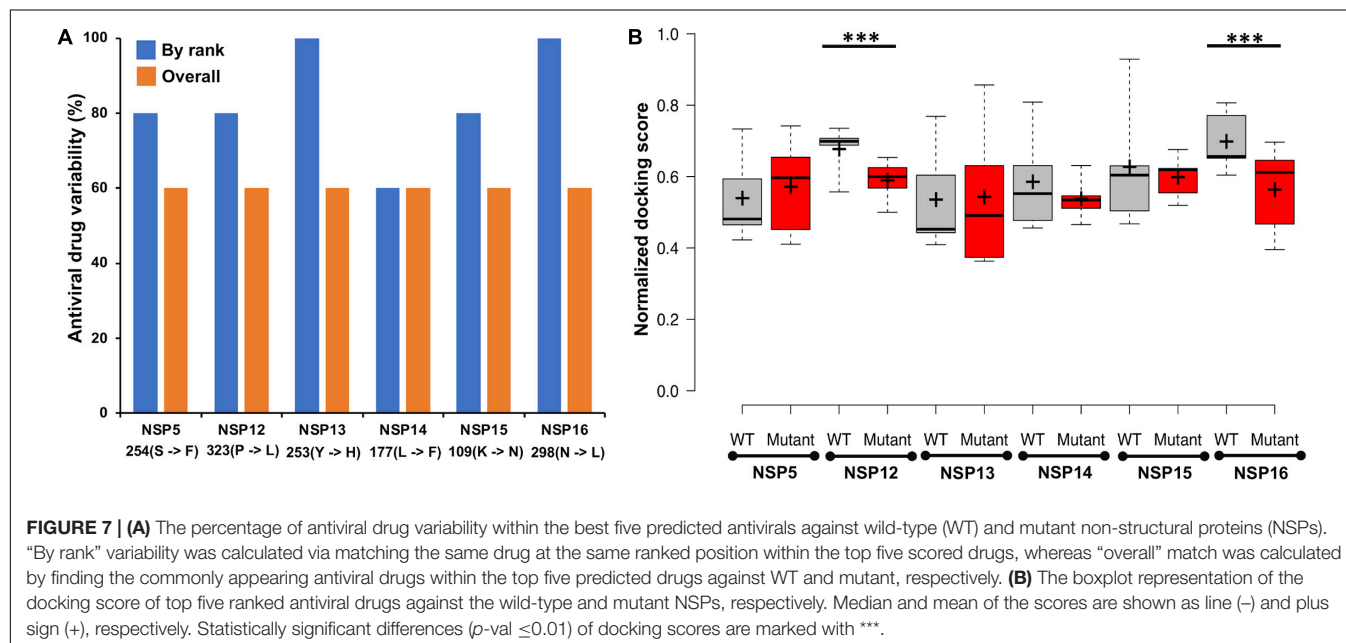
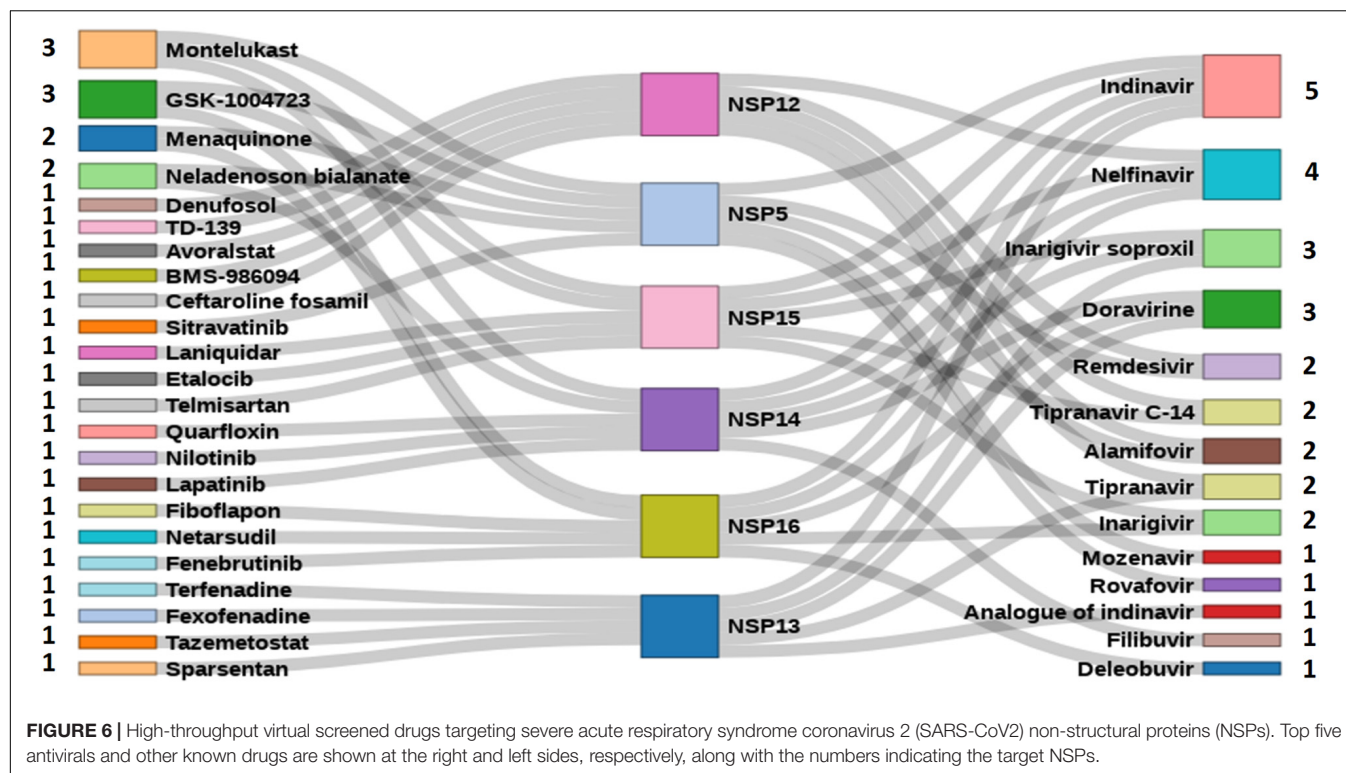
indicating its topological significance. The binding of EIF4E2 with NSP2 may disrupt the immune response of host. The EIF4E2–NSP2 complex is being targeted by zotatifin drug and is under clinical trial (Yoshimoto, 2020). **Supplementary Table S2** shows that EIF4E2 complex has a lower average docking score with mutant NSP2 [496(Q→P)] compared with WT complex. The RMSD profile (**Figure 10A**) shows that the mutant complex is less stable than the WT complex. Although 496(Q→P) mutation results in slightly lower energy (more stable from) (**Figure 9C**), binding of EIF4E2 makes the complex less energetically favorable (**Figure 10C**) than their respective WT counterparts.

**Figures 10D–F** show the outcomes of MD simulation for complex of WT NSP12 and mutant 323(P→L) with human protein PPIL3. PPIL3, a protein coding gene, helps in protein-folding events (Stelzer et al., 2016) and appears as an IIP in the interactome network of NSP12 (**Figure 4**). **Figure 10D** shows that

PPIL3 has a stable complex with the mutated structure of NSP12 compared with the WT structure. **Supplementary Table S2** also shows that the mutant complex has a higher average docking score compared with the WT structure, while RMSFs are quite less for most of the residues (**Figure 10E**). The favored association of PPIL3 with most prevalent mutant variation of NSP12 may disrupt the protein-folding mechanisms of host.

## DISCUSSION

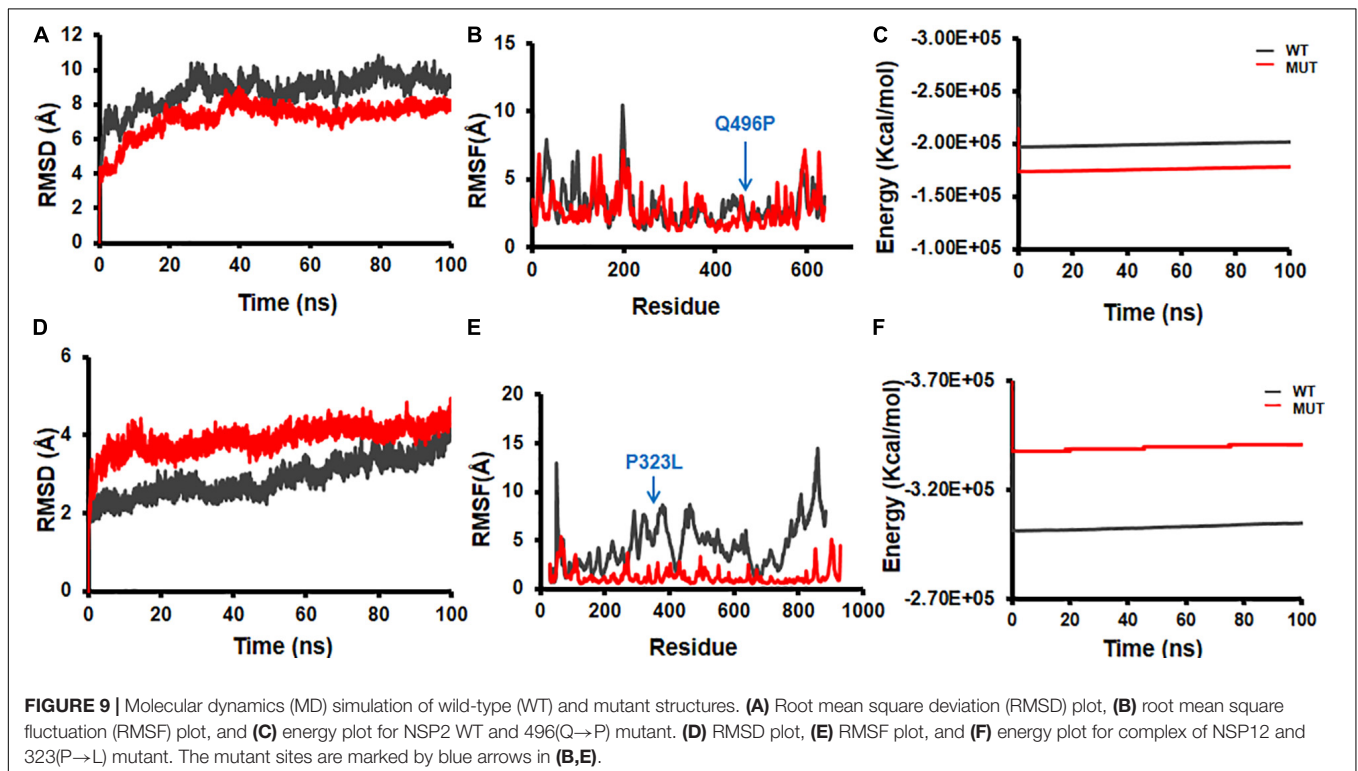
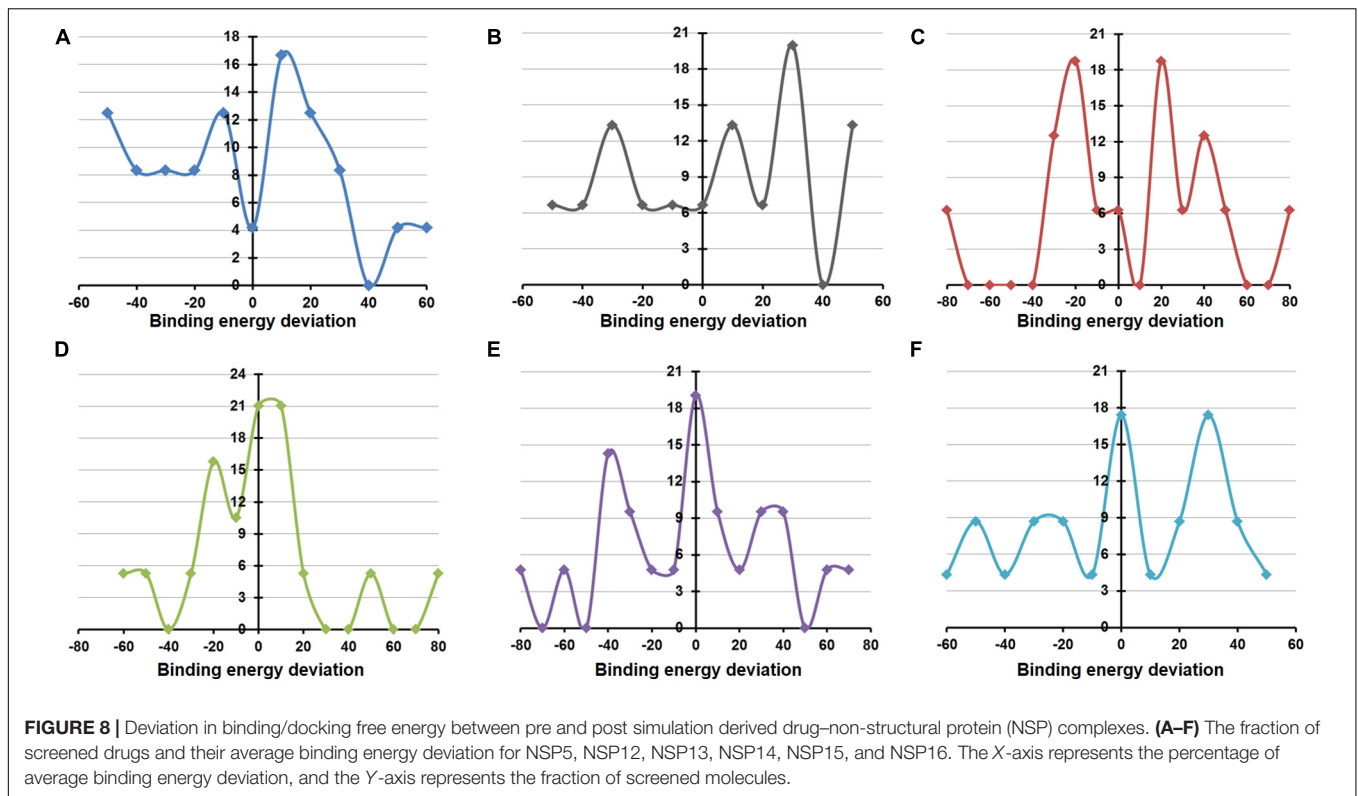
COVID-19 disease has caused an unprecedented pandemic, affecting millions around the globe in manifolds. A complete understanding of the underlying virus, SARS-CoV2, is an utmost necessity. Compared with the source samples from Wuhan, SARS-CoV2 has already demonstrated several mutations across the globe, and the mutations are often region specific



(Khan et al., 2020; Mercatelli and Giorgi, 2020). In this context, we concentrate on the Indian variants of SARS-CoV2 genomes. The major part of the SARS-CoV2 genome consists of a poly-protein, which comprises 16 NSPs. Our database, DbNSP InC, is dedicated to holistic studies of NSPs of SARS-CoV2 virus obtained from samples collected from different places of India. It showcases the mutational variations of SARS-CoV2 virus along with the impact of the mutations in different aspects

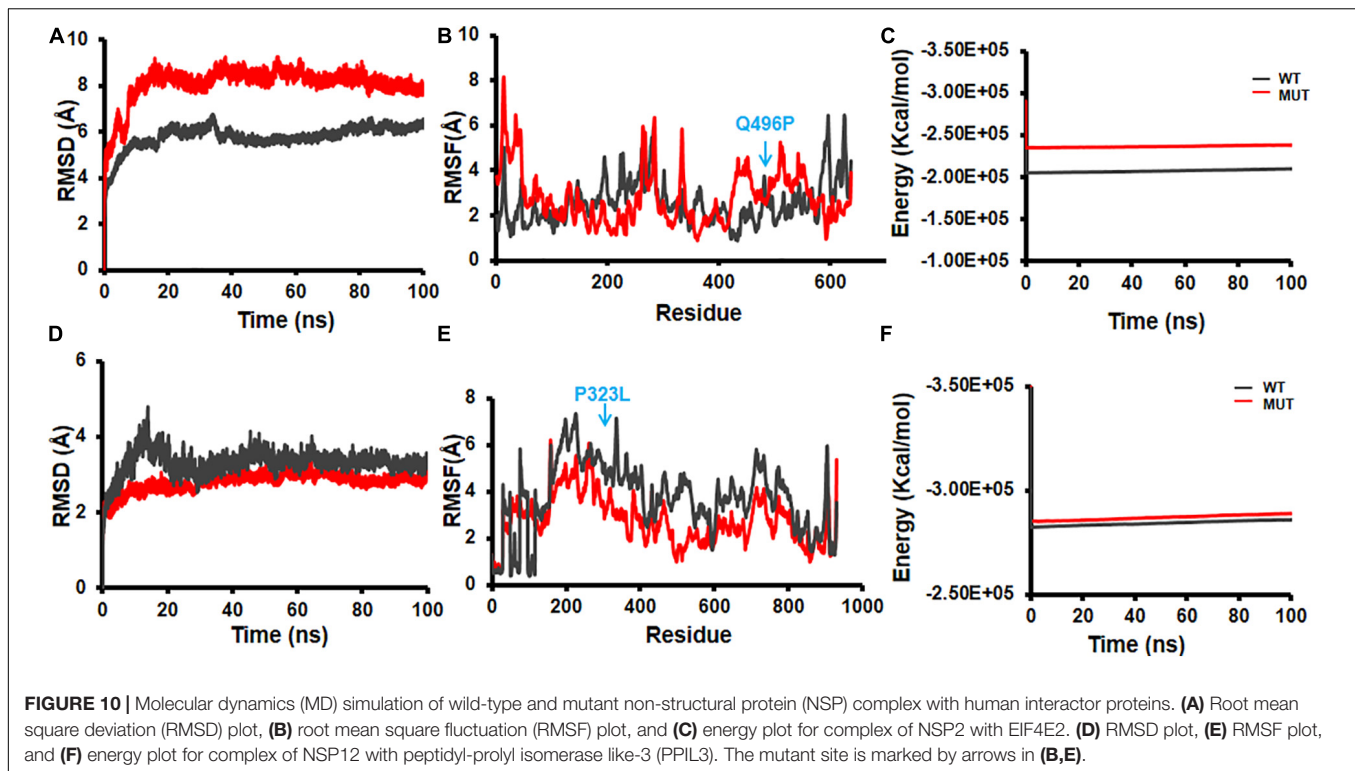
including disease severity and spread in different Indian states. This database provides a pool of combinatorial information regarding the probable impact of the mutations on structural and energetic stabilities of the viral NSPs and subsequently on host protein interaction. Moreover, it also provides critical and useful information about the probable antivirals and known drugs that could be testified for development of effective drugs against the novel coronavirus 2 (nCoV2) virus. We are hopeful that DbNSP





InC database will be a very useful repository to understand the nature of the nCoV2 variants that prevailed in India and their probable impact on the patho-physiology of the disease.

Over the last 1 year, numerous works have been performed to characterize the SARS-CoV2 proteins and the associated mutations. Several databases and online resources have been



developed to aid the fight against the deadly COVID-19 pandemic. Databases like EpiCoV<sup>TM</sup> platform from GISAID (GISAID, 2020), NCBI-SARS-CoV2 resources (NCBI-SARS-CoV2 Resources, 2020), COVID-19 data portal (EMBL-EBI, 2020), Virus Pathogen Database and Analysis Resource (ViPR) (Pickett et al., 2012), GESS (Fang et al., 2021), CovDB (Zhu et al., 2020), and ViruSurf (Canakoglu et al., 2021) systematically categorized thousands of nCoV2 genome sequences deposited from all over the world. Similarly, resources like Cov3D (Gowthaman et al., 2021), SWISS-MODEL SARS-CoV2 portal (Swiss-Model, 2020), and Zhang lab COVID-19 resource (Zhang Lab, 2020) developed 3D models for SARS-CoV2 proteins for structural characterizations, whereas exhaustive experimental characterization of host protein interactions was revealed by works from Gordon et al. (2020a,b). In addition, countless efforts have been put forward using *in silico* drug screening approaches to identify potential inhibitors of the SARS-CoV2 proteins. Some of the works from India also highlighted the genomic diversity and the phylogenetic profiles of the prevalent strains in the country (Banu et al., 2020; Thakur et al., 2020; INDICOV, 2021; Jain et al., 2021; Phylovis, 2021). However, most of these works are discrete in nature, and a combined unified effort characterizing a country- or region-specific mutational profile of the SARS-CoV2 proteins, especially for the NSPs, is warranted. DbNSP InC aims to encompass the country- and state-specific mutational profile of the prevalent SARS-CoV2 genomes and to further provide a comprehensive characterization of the frequently observed mutations in terms of the probable impacts on their structure, function, and interactions with host proteins and target small molecule inhibitors. To the best of our knowledge, this kind of

large-scale, multilevel characterization of country (India) specific SARS-CoV2 NSP mutational analysis followed by estimation of the probable impact of the mutant proteins has not been reported before.

The mutation analysis of the NSP sequences of SARS-CoV2 virus collected from Indian patients reveals several mutations that were not observed in the samples collected in Wuhan, China, from where the virus spread by human contact. Also, some mutations, which are frequently observed in the Wuhan samples, were not observed in the Indian samples. It seems that NSP12 (RdRp) is the most changing protein among the NSPs found in the Indian population. The mutation at site 323 of NSP12 is caused by change of amino acid from P to L. This mutation was observed in 78.44% samples. Moreover, this mutation was observed in 93.24% of samples where patients did not survive. It implies that 323(P→L) mutation of NSP12 is the most lethal mutation among all mutations of all NSPs. From the PAM250 substitution matrix, the score of P→L transition is -3, indicating strong dissimilarity between the mutated and reference sequences. However, 323(P→L) mutation of NSP12 is not unique to the Indian samples. Although not observed in the Wuhan samples, its occurrence is already reported as prevalent in European countries and also in North America (Kannan et al., 2020; Pachetti et al., 2020). This mutation also has a prevalence of a co-occurrence with other mutations (Pachetti et al., 2020). NSP12 creates the core polymerase complex with NSP7 and NSP8 (Hillen et al., 2020; Peng et al., 2020; Wang et al., 2020), and site 323 locates near the binding interface of NSP8 and NSP12 (Hillen et al., 2020). The proline (P) amino acid creates hydrogen bond with NSP8 (Mutlu et al., 2020). The P→L mutation is preferable

to NSP8–NSP12 binding and thus promotes viral replication (Kannan et al., 2020). Hence, the role of 323(P→L) mutation needs attention while designing antiviral drugs targeting the polymerase complex. Moreover, 323(P→L) of NSP12 has a strong co-occurrence with spike protein mutation at 614(D→G) worldwide (Kannan et al., 2020). **Supplementary Figure S3** illustrates the co-occurrence of 323(P→L) and 614(D→G) in the Indian samples also. 323(P→L) is also known to co-occur with 241(C→U) mutation of 5'-UTR of SARS-CoV2 (Kannan et al., 2020). These co-occurrences perhaps enhance the viral activity, making it lethal for human survival. The other mutation 97(A→V) of NSP12 appeared in Singapore, Malaysia, and Europe (GISAID, 2020). 1198(T→K) mutation of NSP3 is prevalent in Asian countries, such as Singapore Malaysia, and also in the United Kingdom (GISAID, 2020). 37(L→F) mutation of NSP6 is also observed in other countries including in samples from Wuhan, China (**Figure 1A** and **Supplementary Figure S1**). It reduces the stability of the protein structure (Benvenuto et al., 2020; Mercatelli and Giorgi, 2020). Hence, this mutation appears favorable to human beings, and also, it is not associated with deceased samples (**Figure 1C**). We also compared the frequencies of the most frequent mutations in India in the global scenario. **Supplementary Figure S4** compares the frequencies of the mutations shown in **Figure 1A** in different continents. Here, Asian data are considered, excluding India data. We observed 323(P→L) mutation of NSP12 across the globe. Mutation 37(L→F) of NSP6 is also observed in different continents but more frequently in India and Asia. Mutations 97(A→V) of NSP12 and 1198(T→K) of NSP3 appear specific to India and Asia. Mutation 994(A→D) of NSP3 emerges as specific to India.

Depending on the availability, the crystal structures and/or 3D models of WTs and mutated NSPs are listed in the DbNSP InC database. The crystal structures are available for WTs NSP5, NSP7, NSP9, NSP10, NSP12, NSP15, and NSP16. We have constructed 3D model structures of WTs NSP1, NSP6, NSP8, NSP13, and NSP14 by homology modeling, and we further validated them using multiple structure validation tools. 3D models retrieved from the Zhang Lab (2020) are also shared for comparison purposes. In general, validation scores of our models are comparable and/or better than those obtained from the Zhang lab models. We observed for NSP1 that QMEAN and Verify3D scores are better for our model than the corresponding scores from Zhang lab NSP1 model, whereas our model has a lower ERRAT score. For NSP6, our model obtained better scores for all the validation methods, whereas for NSP8, ProSA z-score and ERRAT quality factor are comparable with those of the Zhang lab. For NSP13, the QMEAN score is better, but Verify3D and ERRAT scores are not compared with that achieved from the Zhang lab-derived model. Verify3D and QMEAN scores are better for our NSP14 model. However, we have listed the WT model structures NSP2, NSP3, and NSP4 obtained from the Zhang lab in our DbNSP InC database. Based on the crystal and modeled structures of WT NSPs, 36 mutant model structures were generated. All these 3D models were evaluated using various structure validation tools such as PROCHECK (Zhang Lab, 2020), ERRAT (Laskowski

et al., 1993), Verify 3D (Colovos and Yeates, 1993), QMEAN (Eisenberg et al., 1997), and ProSA (Benkert et al., 2011). The validation scores of these mutant models are comparable and/or better than those of the WT counterparts. This advocates their comparable stability and utilization of these mutant structures in downstream analyses of protein–protein interaction as well as protein–drug interactions.

We further constructed interactome for each NSP with their human host proteins, along with their first layer of interactors. The virus–host protein interactome is necessary for understanding how the virus proteins interact with human immune systems and proteins involved in various biological pathways (Perrin-Cocon et al., 2020). We observed that NSP8 has the highest number of interactors, 232, followed by NSP7, which has 133 interactors (**Table 2**). NSP7 interactome produced the highest number of IIPs, 11, followed by NSP8, which has 8 IIPs. Overall, 59 IIPs were identified out of 802 human interactor proteins for 15 NSPs. A composite interactome involving all 15 NSPs and their 802 human interactors (first layer) were also created to examine the interconnectivity between them where only NSP10 and NSP6 interactomes were found to be disjointed (**Supplementary Figure S5**). Guided by the interactome analysis, we generated 113 complex structures using 48 (WT and mutant) NSP and 28 human proteins. Further, structural and chemical properties calculated from the predicted interfaces have shown significant alterations of the interface formed by the mutant NSPs with respect to the WT protein complex. These findings may provide mechanistic insight toward differential host interaction pattern of the variants NSPs, which could relate to varied host responses of the patients infected with the variant nCoV2 virus. However, these preliminary analyses need to be verified by in-depth experimental studies to establish altered interaction and its connections to the patho-physiology of the disease. Nevertheless, our findings on host protein interactions provide clues and direction to future in-depth analyses of specific viral–host protein interaction studies.

A total of 111 antiviral and 8,736 known drugs were screened against various enzymes (NSP5, NSP12, NSP13, NSP14, NSP15, and NSP16) of SARS-CoV2 using a rigorous HTVS procedure to identify the probable candidate that can act against SARS-CoV2 NSP enzymes. Several drug candidates have been identified that can act on multiple targets (**Figure 6**). The antiviral drug indinavir is targeting five SARS-CoV2 enzymes (NSP5, NSP13, NSP14, NSP15, and NSP16). Indinavir is a known HIV-1 protease inhibitor (Lv et al., 2015). Some of these antivirals (e.g., remdesivir, nelfinavir, and tipranavir) are part of ongoing clinical trials (ASHP, 2021), whereas drugs like nilotinib, lapatinib, indinavir, nelfinavir, tipranavir, montelukast, and telmisartan are also reported as potential inhibitors of NSPs (Ghahremanpour et al., 2020). Nelfinavir has also been identified as a SARS-CoV2 protease inhibitor by supervised MD simulation (Bolcato et al., 2020). It also appears as a drug effective in saving SARS-CoV2-affected cells from death (Ianevski et al., 2020; Musarrat et al., 2020). Similarly, other antiviral drugs like doravirine, alamifovir, inarigivir, and inarigivir soproxil were found to target multiple targets. Among the drug bank drugs, montelukast

targets three NSPs. Montelukast has anti-inflammatory effects, reduces oxidative stress, and appears as a potential treatment of COVID-19 (Fidan and Aydoğdu, 2020). It is currently being used in a clinical trial (Clinical Trials Gov, 2020). The other known drugs neladenoson bialanate and menaquinone were also found to act against multiple SARS-CoV2 enzymes. Menaquinone (vitamin K2) deficiency may lead to severity for SARS-CoV2-infected patients and appears as a supplementary in reducing COVID-19 mortality rate (Berenjian and Sarabadani, 2020). These multi-target drugs can be efficient drug candidates against SARS-CoV2. However, screening against the mutant forms of the NSPs yielded quite different antiviral drug populations, at least within the top five ranked antivirals selected based on the normalized composite docking scores (Figure 7). This finding is exciting and indicates a probable alteration of drug sensitivity of the NSPs due to the acquired mutations. However, further in-depth testing is required to confirm the likelihood of the effective alteration of drug sensitivity. Several studies have been reported in the past few months involving drug screening against SARS-CoV2 proteins. However, to the best of our knowledge, our study is one of the few (Swiss-Model, 2020; Gowthaman et al., 2021) to screen both antivirals and other known drugs against all six WT and mutant NSPs (NSP5, NSP12, NSP13, NSP14, NSP15, and NSP16) together. This composite HTVS provides a uniform perspective and platform for shortlisting drugs that could be further testified via in-depth cell free and cell-based assays. Drug repurposing with approved or investigational drugs is perhaps the most effective, rational, and timely strategy for identification of effective drugs against COVID-19. We believe that our findings, which have been made freely available through DbNSP InC, will help the community to attest to the effectiveness of some of the top-scoring drugs.

We have further complimented our molecular modeling and docking analyses with rigorous, atomistic, and solvent-implicit MD simulations. Atomic-level MD simulations offer a computational route toward characterizing both structural and energetic stabilities of protein–protein as well as protein–ligand complexes. In the absence of sufficient experimental information regarding the host protein and drug binding properties of the SARS-CoV2 NSPs, we utilized MD simulations to characterize and evaluate the predictive docking complexes formed by the WT and mutants. Findings from the MD simulation studies suggest acceptable structural and energetic stabilities of the 3D models as well as protein–protein complexes formed by them. Similarly, our MD simulations using the drug–NSP complexes retrieved from the molecular docking-based screening procedure provide additional screening and filtering criteria for selection of the most likely drug candidates. Drug–NSP complexes with progressive stabilized binding free energy profiles suggest better stability and hence can be used as a selection tool. Our MD analyses with drug–NSP complexes show that a higher fraction of the complexes remains stable ( $\pm 20\%$  deviation) or becomes more stable ( $> 20\%$  deviation) in terms of binding free energy throughout the duration of the simulation. This would definitely aid current and future drug discovery and re-purposing efforts against COVID-19.

## CONCLUSION

In conclusion, DbNSP InC emerges as a platform where researchers can get updated information on NSPs of SARS-CoV2 specific to Indian patients. Since many of the mutations, reported in our manuscript as well as provided in DbNSP InC, are observed globally, the corresponding analysis bears relevance even in the global context. In the future, we will enrich DbNSP InC by including more information obtained via structure analysis, host protein interaction, MD simulation, and drug screening. The database will also be updated regularly with the availability of newer sequencing and mutational data.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

NB performed all the sequence analyses and mutational and phylogenetic analyses. KK performed the drug screening and MD analysis. PM and SD performed the protein–protein interaction study. IK, SB, and AC performed all the modeling and partial MD analysis. NB and SC wrote the manuscript and conceptualized and coordinated the project. All authors contributed to the article and approved the submitted version.

## FUNDING

SC acknowledges financial support from MLP-132 grant. NB acknowledges the Systems Medicine Cluster (SyMeC) grant (GAP357), Department of Biotechnology (DBT), Government of India for fellowship. KK and IK acknowledge the Department of Biotechnology (DBT), Government of India for fellowships. SD and SB acknowledge the Council of Scientific and Industrial Research (CSIR), Government of India, for their fellowships. PM and AC acknowledge University Grants Commission, Government of India, for the fellowships. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## ACKNOWLEDGMENTS

We acknowledge CSIR-Indian Institute of Chemical Biology for infrastructural support.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.626642/full#supplementary-material>



## REFERENCES

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2, 19–25. doi: 10.1016/j.softx.2015.06.001
- Angeletti, S., Benvenuto, D., Bianchi, M., Giovanetti, M., Pascarella, S., and Ciccozzi, M. (2020). COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis. *J. Med. Virol.* 92, 584–588. doi: 10.1002/jmv.25719
- Angelini, M. M., Akhlaghpour, M., Neuman, B. W., and Buchmeier, M. J. (2013). Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *MBio* 4:e00524–13.
- Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40, W597–W603.
- ASHP (2021). Available online at: <https://www.ashp.org/COVID-19?loginreturnUrl=SSOCheckOnly> (accessed January 14, 2021).
- Banu, S., Jolly, B., Mukherjee, P., Singh, P., Khan, S., Zaveri, L., et al. (2020). A distinct phylogenetic cluster of indian severe acute respiratory syndrome coronavirus 2 isolates. *Open Forum Infect. Dis.* 7:ofaa434.
- Benkert, P., Biasini, M., and Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27, 343–350. doi: 10.1093/bioinformatics/btq662
- Benvenuto, D., Angeletti, S., Giovanetti, M., Bianchi, M., Pascarella, S., Cauda, R., et al. (2020). Evolutionary analysis of SARS-CoV-2: how mutation of non-structural protein 6 (NSP6) could affect viral autophagy. *J. Infect.* 81, e24–e27.
- Berenjian, A., and Sarabadani, Z. (2020). How menaquinone-7 deficiency influences mortality and morbidity among COVID-19 patients. *Biocatal. Agric. Biotechnol.* 29:101792. doi: 10.1016/j.cbac.2020.101792
- Bhattacharyya, M., and Chakrabarti, S. (2015). Identification of important interacting proteins (IIPs) in *Plasmodium falciparum* using large-scale interaction network analysis and in-silico knock-out studies. *Malar. J.* 14:70.
- Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M., and Moro, S. (2020). Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci. Rep.* 10:20927.
- Bowers, K. J., Chow, D. E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., et al. (2006). “Scalable algorithms for molecular dynamics simulations on commodity clusters,” in *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)*, (Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE)), 43–43.
- Buchan, D. W. A., and Jones, D. T. (2019). The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* 47, W402–W407.
- Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., et al. (2019). RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 47, D464–D474.
- Canakoglu, A., Pinoli, P., Bernasconi, A., Alfonsi, T., Melidis, D. P., and Ceri, S. (2021). ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Res.* 49, D817–D824.
- Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., et al. (2005). The amber biomolecular simulation programs. *J. Comput. Chem.* 26, 1668–1688. doi: 10.1002/jcc.20290
- Clinical Trials Gov (2020). *The COvid-19 Symptom MOnTelukast Trial*. Available online at: <https://clinicaltrials.gov/ct2/show/NCT04389411> (accessed January 14, 2021).
- Colovos, C., and Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* 2, 1511–1519. doi: 10.1002/pro.5560020916
- Dayhoff, M. O. (ed.). (1969). *Atlas of Protein Sequence and Structure [Internet]*. (Silver Spring, MD: The National Biomedical Research Foundation).
- DrugBank (2020). Available online at: <https://www.drugbank.com/> (accessed November 3, 2020).
- Eisenberg, D., Lüthy, R., and Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* 277, 396–404. doi: 10.1016/S0076-6879(97)70222-8
- EMBL-EBI (2020). *COVID-19 Data Portal 2020*. Available online at: <https://www.covid19dataportal.org/> (accessed November 2, 2020).
- Evans, D. J., and Holian, B. L. (1985). The Nose-Hoover thermostat. *J. Chem. Phys.* 83, 4069–4074.
- Fang, S., Li, K., Shen, J., Liu, S., Liu, J., Yang, L., et al. (2021). GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences. *Nucleic Acids Res.* 49, D706–D714.
- Fehr, A. R., and Perlman, S. (2015). “Coronaviruses: an overview of their replication and pathogenesis,” in *Coronaviruses: Methods and Protocols*, eds H. J. Maier, E. Bickerton, and P. Britton (Cham: Springer), 1–23. doi: 10.1007/978-1-0716-0900-2\_1
- Fidan, C., and Aydoğdu, A. (2020). As a potential treatment of COVID-19: montelukast. *Med. Hypotheses* 142:109828. doi: 10.1016/j.mehy.2020.109828
- Forni, D., Cagliani, R., Clerici, M., and Sironi, M. (2017). Molecular evolution of human coronavirus genomes. *Trends Microbiol.* 25, 35–48. doi: 10.1016/j.tim.2016.09.001
- Frieman, M., Ratia, K., Johnston, R. E., Mesecar, A. D., and Baric, R. S. (2009). Severe acute respiratory syndrome coronavirus papain-like protease ubiquitin-like domain and catalytic domain regulate antagonism of IRF3 and NF-kappaB signaling. *J. Virol.* 83, 6689–6705. doi: 10.1128/jvi.02220-08
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565
- Ghahremanzour, M. M., Tirado-rives, J., Deshmukh, M., Ippolito, J. A., Zhang, C., de Vaca, I. C., et al. (2020). Identification of 14 known drugs as inhibitors of the main protease of SARS-CoV-2. *ACS Med. Chem.* 11, 2526–2533.
- GISAID (2020). Available online at: <https://www.gisaid.org/>. (accessed October 8, 2020).
- Gopalan, H. S., and Misra, A. (2020). COVID-19 pandemic and challenges for socio-economic issues, healthcare and national health programs in India. *Diabetes Metab. Syndr. Clin. Res. Rev.* 14, 757–759. doi: 10.1016/j.dsx.2020.05.041
- Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten, C., Gulyaeva, A. A., et al. (2020). The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* 5, 536–544. doi: 10.1038/s41564-020-0695-z
- Gordon, D. E., Hiatt, J., Bouhaddou, M., Rezeli, V. V., Ulferts, S., Braberg, H., et al. (2020a). Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* 370:eabe9403.
- Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., et al. (2020b). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468.
- Gowthaman, R., Guest, J. D., Yin, R., Adolf-Bryfogle, J., Schief, W. R., and Pierce, B. G. (2021). CoV3D: a database of high resolution coronavirus protein structures. *Nucleic Acids Res.* 49, D282–D287.
- Hillen, H. S., Kokic, G., Farnung, L., Dienemann, C., Tegunov, D., and Cramer, P. (2020). Structure of replicating SARS-CoV-2 polymerase. *Nature* 584, 154–156. doi: 10.1038/s41586-020-2368-8
- Hoffman, J. I. E. (2019). *Basic Biostatistics for Medical and Biomedical Practitioners. Biostatistics for Medical and Biomedical Practitioners*. Amsterdam: Elsevier, 1–734.
- Ianevski, A., Yao, R., Fenstad, M. H., Biza, S., Zusinaite, E., Reisberg, T., et al. (2020). Potential antiviral options against SARS-CoV-2 infection. *Viruses* 12:642.
- INDICOV (2021). Available online at: <http://clingen.igib.res.in/indicov/>. (accessed January 14, 2021).
- Jain, A., Rophina, M., Mahajan, S., and Balaji, B. (2021). Analysis of the potential impact of genomic variants in global SARS-CoV-2 genomes on molecular diagnostic assays. *Int. J. Infect. Dis.* 102, 460–462. doi: 10.1016/j.ijid.2020.10.086
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267, 727–748. doi: 10.1006/jmbi.1996.0897
- Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* 105, 6474–6487. doi: 10.1021/jp003919d
- Kannan, S. R., Spratt, A. N., Quinn, T. P., Heng, X., Lorson, C. L., Sönnnerborg, A., et al. (2020). Infectivity of SARS-CoV-2: there is something more than D614G? *J. Neuroimmune Pharmacol.* 15, 574–577. doi: 10.1007/s11481-020-09954-3



- Khan, M. I., Khan, Z. A., Baig, M. H., Ahmad, I., Farouk, A.-E., Song, Y. G., et al. (2020). Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: an in silico insight. *PLoS One*. 15:e0238344. doi: 10.1371/journal.pone.0238344
- Krafchikova, P., Silhan, J., Nencka, R., and Boura, E. (2020). Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. *Nat. Commun.* 11:3717.
- Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* 372, 774–797. doi: 10.1016/j.jmb.2007.05.022
- Kumar, P., Gunalan, V., Liu, B., Chow, V. T. K., Druce, J., Birch, C., et al. (2007). The nonstructural protein 8 (nsp8) of the SARS coronavirus interacts with its ORF6 accessory protein. *Virology* 366, 293–303. doi: 10.1016/j.virol.2007.04.029
- Kumari, R., Kumar, R., and Lynn, A. (2014). G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.* 54, 1951–1962. doi: 10.1021/ci500020m
- Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26, 283–291. doi: 10.1107/s0021889892009944
- Lensink, M. F., Nadzirin, N., Velankar, S., and Wodak, S. J. (2020). Modeling protein–protein, protein–peptide, and protein–oligosaccharide complexes: CAPRI 7th edition. *Proteins Struct. Funct. Bioinform.* 88, 916–938. doi: 10.1002/prot.25870
- Lin, S., Chen, H., Ye, F., Chen, Z., Yang, F., Zheng, Y., et al. (2020). Crystal structure of SARS-CoV-2 nsp10 / nsp16 2' -O-methylase and its implication on antiviral drug design. *Signal Transduct. Target Ther.* 5:131.
- Lv, Z., Chu, Y., and Wang, Y. (2015). HIV protease inhibitors: a review of molecular selectivity and toxicity. *HIV/AIDS Res. Palliat. Care* 7, 95–104. doi: 10.2147/hiv.s79956
- Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641.
- Mark, P., and Nilsson, L. (2001). Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *J. Phys. Chem. A* 105, 9954–9960. doi: 10.1021/jp003020w
- Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* 101, 4177–4189. doi: 10.1063/1.467468
- Mercatelli, D., and Giorgi, F. M. (2020). Geographic and genomic distribution of SARS-CoV-2 mutations. *Front. Microbiol.* 11:1800. doi: 10.3389/fmicb.2020.01800
- Ministry of Health and Family Welfare Government of India (2020). Available online at: <https://www.mohfw.gov.in/> (accessed December 31, 2020).
- Muramatsu, T., Takemoto, C., Kim, Y., Wang, H., Nishii, W., and Terada, T. (2016). SARS-CoV 3CL protease cleaves its C-terminal autoproteolytic site by novel subsite cooperativity. *Proc. Natl. Acad. Sci. U.S.A.* 113, 12997–13002. doi: 10.1073/pnas.1601327113
- Musarrat, F., Chouljenko, V., Dahal, A., Nabi, R., Chouljenko, T., Jois, S. D., et al. (2020). The anti-HIV drug nelfinavir mesylate (Viracept) is a potent inhibitor of cell fusion caused by the SARS-CoV-2 spike (S) glycoprotein warranting further evaluation as an antiviral against COVID-19 infections. *J. Med. Virol.* 92, 2087–2095. doi: 10.1002/jmv.25985
- Mutlu, O., Ugurel, O. M., Sariyer, E., Ata, O., Inci, T. G., Ugurel, E., et al. (2020). Targeting SARS-CoV-2 Nsp12/Nsp8 interaction interface with approved and investigational drugs: an in silico structure-based approach. *J. Biomol. Struct. Dyn.* 1–13. doi: 10.1080/07391102.2020.1819882
- Narayanan, K., Ramirez, S. I., Lokugamage, K. G., and Makino, S. (2020). Coronavirus nonstructural protein 1: common and distinct functions in the regulation of host and viral gene expression. *Virus Res.* 202, 89–100. doi: 10.1016/j.virusres.2014.11.019
- NCBI-SARS-CoV2 Resources (2020). Available online at: <https://www.ncbi.nlm.nih.gov/sars-cov-2/> (accessed November 1, 2020).
- NC\_045512 (2020). Available online at: [https://www.ncbi.nlm.nih.gov/nucleotide/NC\\_045512.2](https://www.ncbi.nlm.nih.gov/nucleotide/NC_045512.2) (accessed October 10, 2020).
- Nocedal, J., and Wright, S. J. (2006). *Numerical Optimization*. New York, NY: Springer, 636.
- Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* 18:179.
- Peng, Q., Peng, R., Yuan, B., Qi, J., Gao, G. F., Shi, Y., et al. (2020). Structural and biochemical characterization of the nsp12-nsp7-nsp8 core polymerase complex from SARS-CoV-2. *Cell Rep.* 31:107774. doi: 10.1016/j.celrep.2020.107774
- Perrin-Cocon, L., Diaz, O., Jacquemin, C., Barthel, V., Ogire, E., Ramière, C., et al. (2020). The current landscape of coronavirus-host protein-protein interactions. *J. Transl. Med.* 18:319.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF chimera - a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084
- Phyllovis (2021). Available online at: <http://clingen.igib.res.in/genepi/phylovis/> (accessed January 14, 2021).
- Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., et al. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* 40, D593–D598.
- Romano, M., Ruggiero, A., Squeglia, F., Maga, G., and Berisio, R. (2020). A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. *Cells* 9:1267. doi: 10.3390/cells9051267
- Saputro, D. R. S., and Widyandingsih, P. (2017). “Limited memory broyden-fletcher-goldfarb-shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR),” in *Proceeding of the AIP Conference Proceedings*, (College Park, MA: American Institute of Physics Inc), 040009.
- Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33, W363–W367.
- Shank, S. D., Weaver, S., and Pond, S. L. K. (2018). Phylotree.js – a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinform.* 19:276. doi: 10.1186/s12859-018-2283-2
- Shannon, A., Le, N. T., Selisko, B., Eydyous, C., Alvarez, K., Guillemot, J., et al. (2020). Remdesivir and SARS-CoV-2: structural requirements at both nsp12 RdRp and nsp14 Exonuclease active-sites. *Antivir. Res.* 178:104793. doi: 10.1016/j.antiviral.2020.104793
- Snijder, E. J., Decroly, E., and Ziebuhr, J. (2016). “The nonstructural proteins directing coronavirus RNA synthesis and processing,” in *Advances in Virus Research: Coronaviruses*, 1st Edn, Vol. 96, ed. J. Ziebuhr (Amsterdam: Elsevier Inc), 59–126. doi: 10.1016/bs.aivir.2016.08.008
- Sousa Da Silva, A. W., and Vranken, W. F. (2012). ACPYPE – antechamber python parser interface. *BMC Res. Notes* 5:367. doi: 10.1186/1756-0500-5-367
- Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–D539.
- Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatic* 2016, 1.30.1–1.30.33.
- Straeter, T. A. (1971). *On the Extension of the Davidon-Broyden Class of Rank One, Quasi-Newton Minimization Methods to an Infinite Dimensional Hilbert Space With Applications to Optimal Control Problems*. Raleigh, NC: North Carolina State University.
- Swiss-Model (2020). Available online at: <https://swissmodel.expasy.org/repository/species/2697049> (accessed November 1, 2020).
- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-cepas, J., et al. (2019). STRING v11: protein – protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, 607–613.
- Thakur, M., Singh, A., Joshi, B. D., Ghosh, A., Singh, S. K., Singh, N., et al. (2020). Time-lapse sentinel surveillance of SARS-CoV-2 spread in India. *PLoS One* 15, e0241172. doi: 10.1371/journal.pone.0241172
- The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515.
- Thoms, M., Buschauer, R., Ameisemeier, M., Koepke, L., Denk, T., et al. (2020). Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science* 369, 1249–1255. doi: 10.1126/science.abc8665

- Tuckerman, M., Berne, B. J., and Martyna, G. J. (1992). Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* 97, 1990–2001. doi: 10.1063/1.463137
- Viswanathan, T., Arya, S., Chan, S., Qi, S., Dai, N., Misra, A., et al. (2020). Structural basis of RNA cap modification by SARS-CoV-2. *Nat. Commun.* 11:3718.
- Wang, Q., Wu, J., Wang, H., Guddat, L. W., Gong, P., and Rao, Z. (2020). Structural basis for RNA replication by the SARS-CoV-2 polymerase. *Cell* 182, 417–428.
- Webb, B., and Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinform.* 54, 5.6.1–5.6.37.
- Wiederstein, M., and Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* 35, 407–410.
- Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.
- Yoshimoto, F. K. (2020). The proteins of severe acute respiratory syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. *Protein J.* 39, 198–216.
- Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* 3:e59. doi: 10.1371/journal.pcbi.0030059
- Yuen, C., Lam, J., Wong, W., Mak, L., Chu, H., Cai, J., et al. (2020). SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists. *Emerg. Microb. Infectet.* 9, 1418–1428.
- Zhang, C., Chen, Y., Li, L., Yang, Y., He, J., Chen, C., et al. (2020). Structural basis for the multimerization of nonstructural protein nsp9 from SARS-CoV-2. *Mol Biomed.* 1:5.
- Zhang Lab (2020). *Genome-Wide Structure and Function Modeling of SARS-COV-2*. Available online at: <https://zhanglab.ccmb.med.umich.edu/COVID-19/> (accessed November 4, 2020).
- Zhu, Z., Meng, K., Liu, G., and Meng, G. (2020). A database resource and online analysis tools for coronaviruses on a historical and global scale. *Database* 2020:baaa070.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Biswas, Kumar, Mallick, Das, Kamal, Bose, Choudhury and Chakrabarti. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Structural Insights on the SARS-CoV-2 Variants of Concern Spike Glycoprotein: A Computational Study With Possible Clinical Implications

Marni E. Cueno\* and Kenichi Imai

Department of Microbiology, Nihon University School of Dentistry, Tokyo, Japan

## OPEN ACCESS

### Edited by:

Nimisha Ghosh,  
Siksha O Anusandhan University, India

### Reviewed by:

Neetika Nath,  
Universitätsmedizin Greifswald,  
Germany  
Kira Vyatkina,  
Saint Petersburg Academic University  
(RAS), Russia

### \*Correspondence:

Marni E. Cueno  
marni.cueno@nihon-u.ac.jp

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

Received: 10 September 2021

Accepted: 07 October 2021

Published: 22 October 2021

### Citation:

Cueno ME and Imai K (2021) Structural  
Insights on the SARS-CoV-2 Variants  
of Concern Spike Glycoprotein: A  
Computational Study With Possible  
Clinical Implications.  
Front. Genet. 12:773726.  
doi: 10.3389/fgene.2021.773726

Coronavirus disease 2019 (COVID-19) pandemic has been attributed to SARS-CoV-2 (SARS2) and, consequently, SARS2 has evolved into multiple SARS2 variants driving subsequent waves of infections. In particular, variants of concern (VOC) were identified to have both increased transmissibility and virulence ascribable to mutational changes occurring within the spike protein resulting to modifications in the protein structural orientation which in-turn may affect viral pathogenesis. However, this was never fully elucidated. Here, we generated spike models of endemic HCoV (HCoV 229E, HCoV OC43, HCoV NL63, HCoV HKU1, SARS CoV, MERS CoV), original SARS2, and VOC (alpha, beta, gamma, delta). Model quality check, structural superimposition, and structural comparison based on RMSD values, TM scores, and contact mapping were all performed. We found that: 1) structural comparison between the original SARS2 and VOC whole spike protein model have minor structural differences (TM > 0.98); 2) the whole VOC spike models putatively have higher structural similarity (TM > 0.70) to spike models from endemic HCoVs coming from the same phylogenetic cluster; 3) original SARS2 S1-CTD and S1-NTD models are structurally comparable to VOC S1-CTD (TM = 1.0) and S1-NTD (TM > 0.96); and 4) endemic HCoV S1-CTD and S1-NTD models are structurally comparable to VOC S1-CTD (TM > 0.70) and S1-NTD (TM > 0.70) models belonging to the same phylogenetic cluster. Overall, we propose that structural similarities (possibly ascribable to similar conformational epitopes) may help determine immune cross-reactivity, whereas, structural differences (possibly associated with varying conformational epitopes) may lead to viral infection (either reinfection or breakthrough infection).

**Keywords:** conformational epitopes, endemic HCoV, SARS-CoV-2, spike glycoprotein, variants of concern

## INTRODUCTION

Coronaviruses (CoV) are categorized as enveloped positive-stranded RNA viruses belonging to family Coronaviridae, order *Nidovirales*, and subfamily *Othocoronavirinae* comprising four genera (King et al., 2018). Currently, seven human-infecting CoVs have been identified as early as the 1960s, namely: human CoV (HCoV)-229E (1962), HCoV-OC43 (1967), severe acute respiratory syndrome (SARS)-CoV 1 (SARS1) (2002), HCoV-NL63 (2004), HCoV-HKU1 (2005), and Middle East respiratory syndrome (MERS)-CoV (2012) [all six are endemic to the human population] with

SARS-CoV 2 (SARS2) (2019) being the latest CoV capable of infecting humans (Hamre and Procknow, 1966; Kapikian et al., 1969; Ksiazek et al., 2003; Fouchier et al., 2004; Woo et al., 2005; Zaki et al., 2012; Zhu et al., 2020). Moreover, the spike (a common structural protein among the CoVs) is classified as a class I viral fusion protein involved in host tropism, viral entry and pathogenesis, and host immune response induction (Lu et al., 2015; Millet and Whittaker, 2015; Hulswit et al., 2016; Li, 2016). Additionally, the spike has three segments, namely: the large ectodomain which is divided into the S1 receptor-binding subunit (involved in viral attachment) and S2 membrane-fusion subunit (assists virus-cell fusion) (Hulswit et al., 2016; Li, 2016), single-pass transmembrane anchor, and short intracellular tail (Li, 2016).

Among the human-infecting CoVs, only SARS2 infection resulted to a pandemic causing the coronavirus disease 2019 (COVID-19) (Tay et al., 2020). Moreover, multiple SARS2 variants were produced ascribable to various mutations occurring within the spike and, among the SARS2 variants produced, variants of concern (VOC) were identified to have increased transmissibility and virulence while having decreased response to available therapeutic strategies (Koyama et al., 2020). Considering VOC are a product of mutational changes occurring within the spike and structural orientation modifications are a product of amino acid alterations which in-turn may affect viral pathogenesis (Chen and Bahar, 2004), we hypothesize that the VOC spike glycoprotein may have structural modifications that may affect both immune cross-reactivity and viral pathogenesis. However, this has likewise not been fully investigated. A better understanding of the possible structural differences and similarities occurring within the VOC spike proteins may give us a better understanding of the potential of cross-reactivity to occur and, likewise, could give a possible explanation for the occurrence of both SARS2 reinfection and breakthrough infections which in-turn may lead to novel therapeutic strategies.

## MATERIALS AND METHODS

### SARS2 VOC and HCoV Spike Modeling

Representative CoV spike amino acid sequences were collected from the National Center for Biological Information (NCBI) website. In order to obtain an accurately generated representative spike model, at least five sequence models were initially analyzed, whereby, spike models having similar Root Mean Square Deviation (RMSD) values and Template Modeling scores (TM-scores) based on superimposition done by TM-align (Zhang and Skolnick, 2005) were utilized for further downstream analyses. For generating SARS2 VOC spike models, the following representative amino acid sequences were used with Genebank accession number indicated: alpha (QTC11018), beta (QTJ24451), gamma (QRX39401), and delta (QUF59047). For generating the endemic HCoV spike models, the following representative amino acid sequences were used with Genebank accession number indicated: 229E (ABB90513), OC43 (AXX83297), NL63 (QED88040), HKU1 (ARB07617), SARS1 (AAR07625), MERS (AHX00731), and original SARS2 (YP\_009724390). Similarly, representative original SARS2 spike

S1 C-terminal domain (S1-CTD) and N-terminal domain (S1-NTD) models were generated based on UniProt reference number P0DTC2. All models generated were through the Phyre2 web server (Kelley and Sternberg, 2009) while Jmol applet (Herraez, 2006) was used for protein visualization.

### Spike Model Quality Assessment

All CoV spike models generated throughout the study were initially assessed for quality before further downstream analyses. In this regard, protein model:crystal structure superimposition and contact mapping were performed. Representative crystal structure used for model quality comparison was the 2021 strain (PDB ID: 7BNM) which already has the D614G mutation (Tomaszewski et al., 2020). Moreover, a monomeric 7BNM crystal model (based on the 7BNM crystal structure) was generated using Phyre2 and superimposed to the 7BNM crystal structure to likewise serve as an additional model quality check. Representative CoV spike models and crystal structure were superimposed using TM align (Zhang and Skolnick, 2005). For this study, we considered spike models as suitable for further downstream analyses if TM scores between superimposed sequence model:crystal structure, crystal model:crystal structure, and crystal model:sequence model are close to 1.0. Subsequently, CMView applet (Contact type: Cα; Distance cut-off: 8.0; Needleman-Wunsch alignment) was used to determine protein common contact among the superimpositions made (Vehlow et al., 2011). Briefly, higher common contact would indicate that there is more structural similarities between the superimposed models and crystal structure (Holm and Sander, 1996) which in-turn implies that the generated spike models are suitable for further downstream analyses.

### CoV Spike Model Comparison

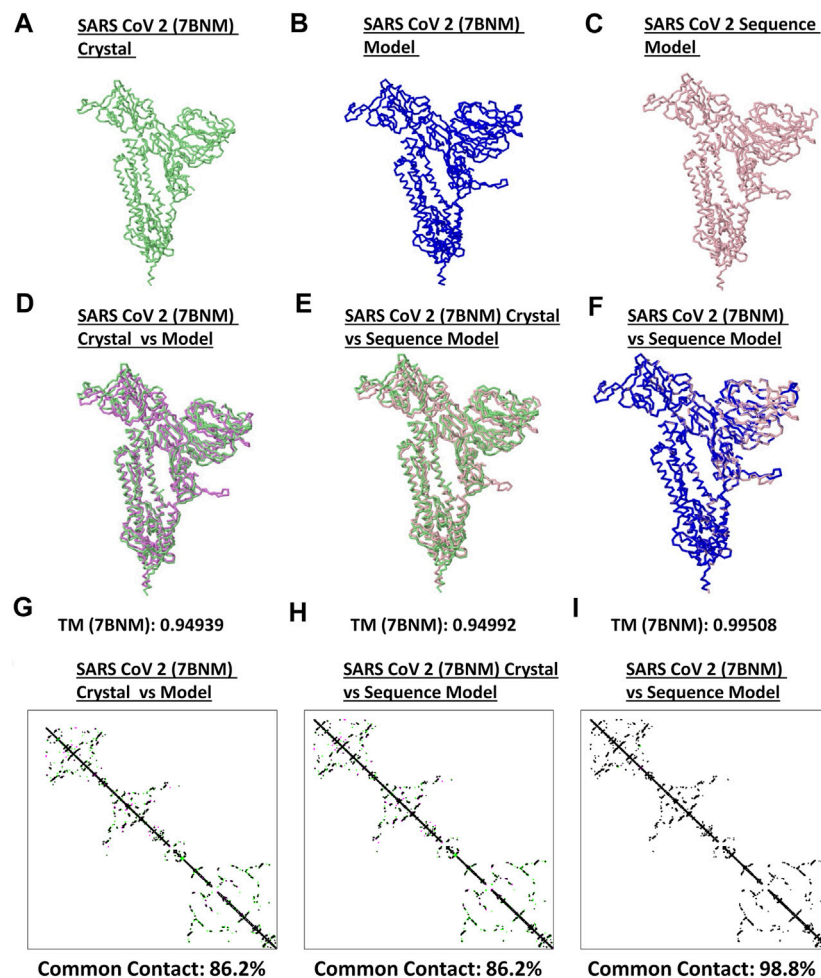
Three different sets of protein structural differentiation were performed: 1) whole protein structural comparison among VOC spike models, whereby, all generated models were compared (RMSD value, TM score, common contact) to the original SARS2 and among VOC spike models through superimposition and contact mapping; 2) whole protein structural comparison between VOC and endemic HCoVs spike models, whereby, generated VOC spike models were compared (RMSD value, TM score, common contact) to generated endemic HCoV spike models also through superimposition and contact mapping; and 3) spike domain structural comparison, whereby, generated S1-CTD and S1-NTD models derived from the VOC and endemic HCoV spike models were compared (TM score only) through original SARS2:VOC and VOC:endemic HCoV superimposition. RMSD value, Tm score, and protein common contact were established using TM align and CMView, respectively.

## RESULTS

### Generated Spike Models Are Fit for Downstream Analyses

Model quality assessment has been highly recommended before performing any downstream structural analyses using generated protein structures from either experimental (i.e. crystallized) or





**FIGURE 1** | Quality check of generated monomeric SARS2 spike protein models. Representative SARS2 (A) 7BNM crystal (B) 7BNM model, and (C) sequence model of monomeric spike proteins are presented. Superimposition between (D) 7BNM crystal and 7BNM model (E) 7BNM crystal and sequence model, and (F) 7BNM model and sequence models are shown. TM scores relative to the 7BNM crystal (when superimposed with either the 7BNM model or sequence model) and 7BNM model (when superimposed with the sequence model) of the superimposed protein structures are indicated below. SARS2 7BNM crystal (green), 7BNM model (blue), and sequence model (pink) are presented.

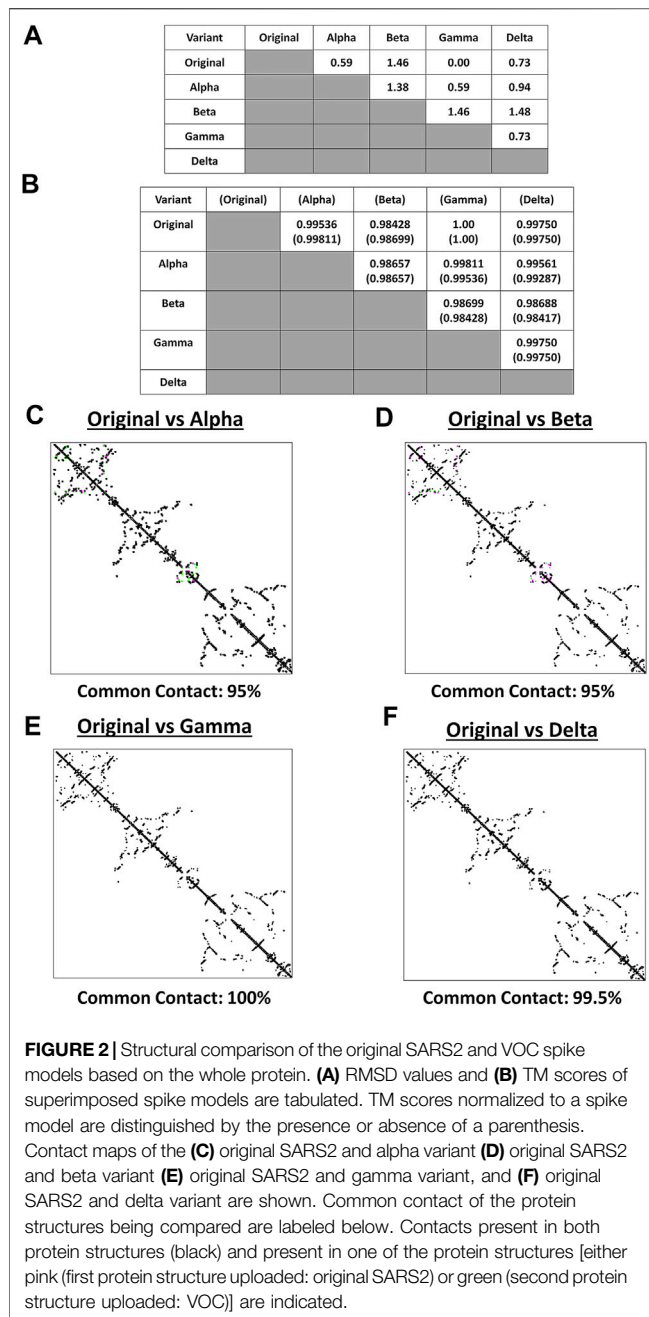
theoretical (i.e. computer-based) approaches (Berman et al., 2006). To determine the quality and correctness of all spike models generated, both protein structural superimpositions and contact mapping were done. Representative SARS2 crystal structure (Figure 1A), generated SARS2 crystal model (Figure 1B) and SARS2 sequence model (Figure 1C) were all utilized for superimposition. We found that TM scores between crystal structure:crystal model [TM (based on the crystal structure): 0.94939] (Figure 1D), crystal structure:sequence model [TM (based on the crystal structure): 0.94992] (Figure 1E), and crystal model:sequence model [TM (based on the crystal model): 0.99508] (Figure 1F) were TM > 0.90 which we considered adequate for further analyses (Hevener et al., 2009). Additionally, protein contact mapping between crystal structure:crystal model [common contact: 86.2%] (Figure 1G),

crystal structure: sequence model [common contact: 86.2%] (Figure 1H), and crystal model:sequence model [common contact: 98.8%] (Figure 1I) have high common contact (>85%), thereby, insinuating that there is high protein contact similarity between the structures. Taken together, these results would suggest that the generated spike models are fit for further downstream structural analyses.

### Original SARS2 and VOC Spike Models Putatively Have Minor Structural Differences

Both protein structure and conformation dynamics are associated to biological function (Chen and Bahar, 2004). To establish the possible spike structural variations among

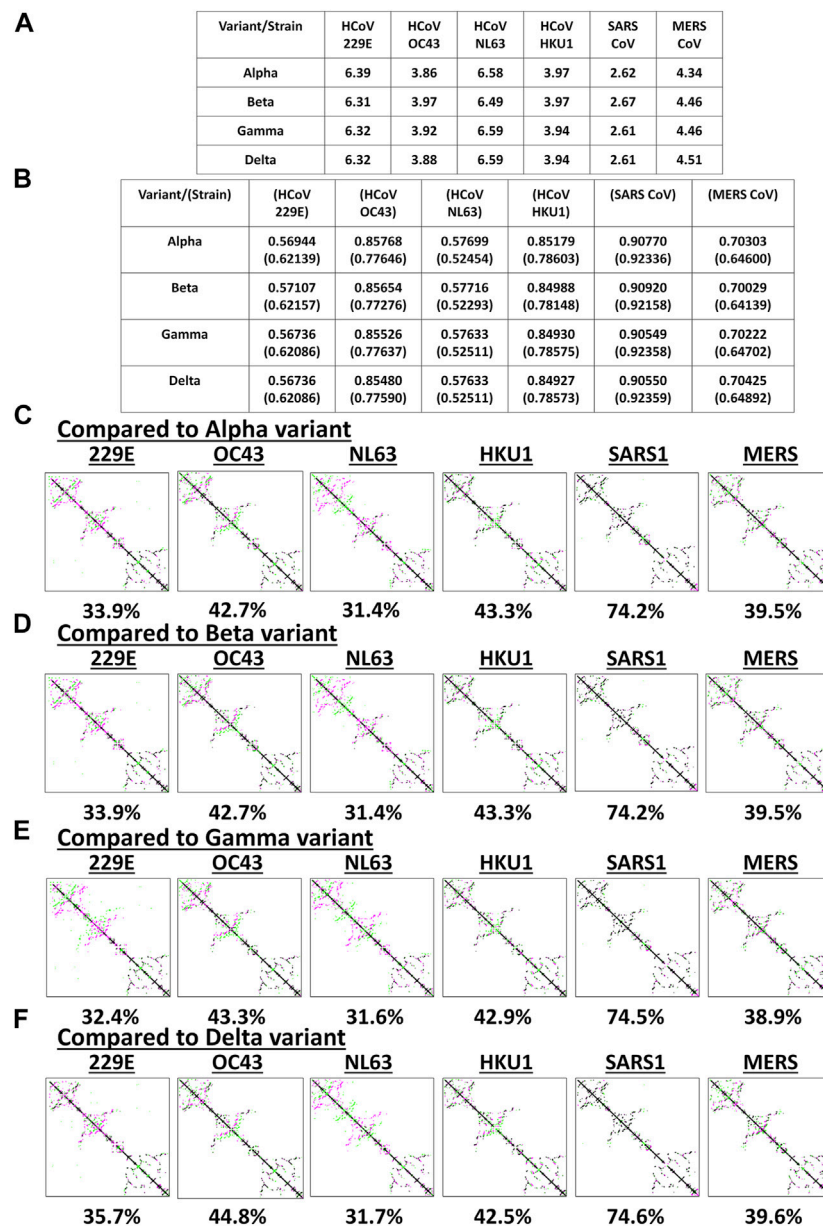




the VOC, spike models of each VOC (alpha, beta, gamma, delta) and the original SARS2 were superimposed and analyzed using RMSD values, TM scores, and contact map overlap (CMO) analyses. Measurements involving RMSD values focus on similarities between superimposed atomic coordinates (including amino acid residues), whereas, measurements involving TM scores focus on similarities between protein structures regardless of protein size (Zhang and Skolnick, 2005; Kufareva and Abagyan, 2012). Additionally, common contacts obtained through CMO analyses provide information related to pairwise spatial and functional relationship of residues within a protein

while unifying certain features related to protein folding and structure prediction (Wang and Xu, 2013; Bittrich et al., 2019). Original SARS2 and VOC spike models used were generated by Phyre2 (**Supplementary Figure S1**). As seen in **Figure 2A**, alpha, gamma, and delta variants are possibly similar with the original SARS2 (RMSD <1.00), whereas, the beta variant has a higher structural difference compared to the original SARS2 and other VOC (RMSD >1.00). These observations are likewise generally consistent with TM scores (**Figure 2B**). Moreover, CMO analyses between the original SARS2 and VOC showed similar common contact (95%) between the original and both alpha (**Figure 2C**) and beta (**Figure 2D**) variants while both gamma (**Figure 2E**) and delta (**Figure 2F**) variants had higher common contact at 100 and 99.5%, respectively. Taken together, we hypothesize that no major structural difference within the spike glycoprotein occurred among the original SARS2, alpha, gamma, and delta variants (RMSD <1.00; TM > 0.99), whereas, the beta variant putatively may have differed with regards to atomic coordinates when compared to the original SARS2 and VOC (RMSD >1.00). However, considering TM score, we likewise presume that no major structural difference occurred in the beta variant (TM > 0.98). Furthermore, similar common contact between the alpha and beta variants could suggest similar functional residues in both variants, whereas, the close to similar common contact (0.5% difference) between gamma and delta variants may likewise imply that functional residues are somewhat the same albeit with some minor difference. These results are consistent with SARS2 maintaining its genomic integrity across propagation (Mercatelli and Giorgi, 2020) and varying VOC transmissibility (Campbell et al., 2021). In this regard, we postulate that the overall spike model among VOC generally did not have a major deviation in terms of protein structural conformation from the original SARS2 spike model. Nevertheless, the minor structural deviation observed may contribute to each VOC having a unique biological characteristic especially in terms of viral transmissibility and immune evasion consistent with an earlier report (Campbell et al., 2021) showing that the effective reproduction numbers of the VOC differ among themselves, namely: alpha (4% compared to alpha), beta (4% compared to beta), gamma (10% compared to alpha; 17% compared to beta), and delta (55% compared to alpha; 60% compared to beta; 34% compared to gamma).

It is worth mentioning that the spike model of the gamma variant potentially has similar atomic coordinates (RMSD value), protein structure (TM score), and functional residues (CMO analyses) when compared to the original SARS2 spike model. Considering the gamma variant is more transmissible compared to the original SARS2 (Campbell et al., 2021), we hypothesize that the biological difference between the gamma variant and original SARS2 in terms of spike function is mainly associated with amino acid residue changes and not on protein structural variations. Additionally, it is also worth mentioning that individuals



**FIGURE 3 |** Structural comparison of the original SARS2 and endemic HCoV spike models based on the whole protein. **(A)** RMSD values and **(B)** TM scores of superimposed spike models are tabulated. TM scores normalized to a spike model is distinguished by the presence or absence of a parenthesis. Contact maps of the **(C)** alpha variant relative to other endemic HCoV **(D)** beta variant relative to other endemic HCoV **(E)** gamma variant relative to other endemic HCoV, and **(F)** delta variant relative to other endemic HCoV are shown. Common contact of the protein structures being compared are labeled below. Endemic HCoVs [HCoV 229E (229E), HCoV OC43 (OC43), HCoV NL63 (NL63), HCoV HKU1 (HKU1), SARS-CoV-1 (SARS1), and MERS CoV (MERS)] are indicated. Contacts present in both protein structures (black) and present in one of the protein structures [either pink (first protein structure uploaded: VOC) or green (second protein structure uploaded: endemic HCoV)] are presented.

infected with the beta variant have a higher chance of needing critical care and death occurrence compared to infections associated with alpha, gamma, and delta variants (Callaway, 2021) possibly due to high levels of immune evasion associated to the beta variant (Madhi et al., 2021). In this regard, we think that the difference in atomic coordinates of the beta variant (RMSD >1.00) compared to the other VOC (RMSD <1.00) is a contributing factor in COVID-19 infection severity.

Admittedly, additional work is needed to further explore these two points.

## VOC Spike Models May Have Varying Structural Similarity to Endemic HCoVs

Among the known endemic HCoVs, both 229E and NL63 strains are classified under the alpha-CoV phylogenetic cluster while the

other remaining strains are classified under the beta-CoV phylogenetic cluster which is further divided into lineages, specifically: OC43 and HKU1 belong to the A lineage; SARS1 and SARS2 belong to the B lineage; and MERS belong to the C lineage (Hamre and Procknow, 1966; Kapikian et al., 1969; Ksiazek et al., 2003; Chiu et al., 2005; Woo et al., 2005; Letko et al., 2020; Zhu et al., 2020). To determine the potential spike structural differences and similarities between VOC and endemic HCoVs, model superimposition and analyses (RMSD values, TM scores, and CMO analyses) were performed. All endemic HCoV spike models were generated by Phyre2 (**Supplementary Figure S2**). In terms of atomic coordinates (RMSD values), we found that VOC spike models differed (RMSD >2.6) from endemic HCoVs (**Figure 3A**). However, in terms of protein structure (TM scores), we observed that VOC spike models (**Figure 3B**) potentially have similar protein structural conformation (TM > 0.50) (Yang et al., 2015). Moreover, VOC spike models putatively have high structural similarity when compared to endemic HCoVs in the same phylogenetic cluster [SARS1 (TM > 0.90), OC43 (TM > 0.85), HKU1 (TM > 0.849), MERS (TM > 0.70)] while those in a different phylogenetic cluster have lower structural similarity [229E (TM > 0.569), NL63 (TM > 0.57)]. Interestingly, in terms of CMO analyses, we found that endemic HCoV spike models have the same common contact difference when compared to spike models from the alpha (**Figure 3C**) and beta (**Figure 3D**) variants which we suspect to be due to alpha and beta variants having putatively the same functional residues (common contact) consistent with our earlier results (**Figures 2C,D**) and reported biological characteristics wherein effective reproduction numbers between the two variants are the same (Campbell et al., 2021). In contrast, both gamma (**Figure 3E**) and delta (**Figure 3F**) variants have varying common contact when compared to the endemic HCoV spike models which we likewise believe to be attributable to the difference in functional residues between the two variants consistent with our earlier results (**Figures 2E,F**) and reported biological characteristics wherein the effective reproduction numbers of both gamma and delta variants differ between the two (Campbell et al., 2021). Noticeably, VOC spike models have high common contact (74.2–74.6%) with SARS1 which coincidentally belongs to the same lineage as that of SARS2. This would emphasize the close structural dynamics between SARS1 and VOC spike models which we attribute to high nucleotide similarity (Robson, 2020). Taken together, we postulate that the overall VOC spike models have varying atomic coordinates and functional residues while generally having the same protein structural conformation when compared to the endemic HCoV spike models.

Considering the results at this point (**Figures 2, 3**), we wish to highlight that data obtained from RMSD values, TM score, and CMO analyses were all based on superimposing full-length CoV spike protein models. However, since it is probable that the protein structural dynamics along a receptor binding site may be composed of different atomic coordinates (particularly, protein length and structure) while having a similar binding surface (Di Rienzo et al., 2017) [consistent with what we observed (TM > 0.98) (**Figure 2B**)], further structural comparison is

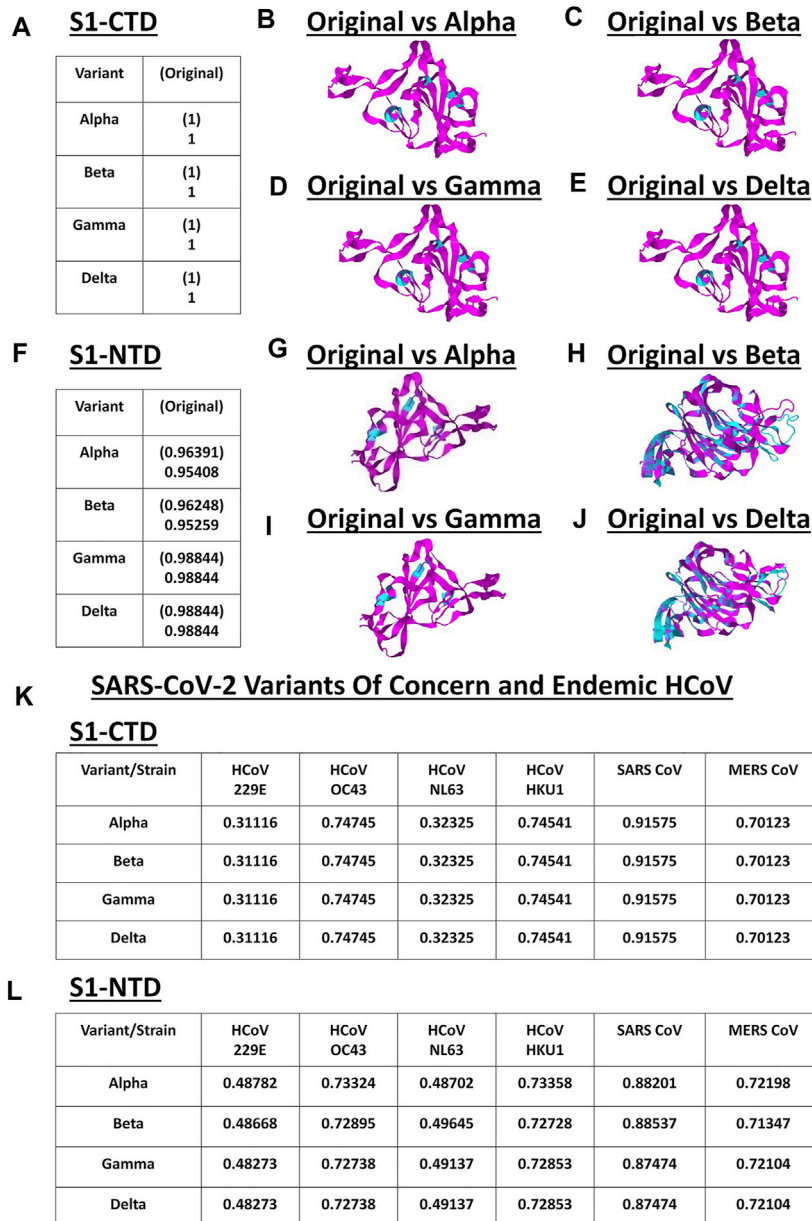
merited which would mainly focus on both S1-CTD and S1-NTD of the VOC spike models.

## VOC S1-CTD and S1-NTD Models Are Structurally Comparable to the Original SARS2 and Endemic HCoV

S1 subunit of CoV spike glycoproteins is made up of the C-terminal domain (S1-CTD) and N-terminal domain (S1-NTD) which in-turn have been associated to host cell binding (Hulswit et al., 2016; Li, 2016). To elucidate the structural similarities and differences within the SARS2 S1-CTD and S1-NTD, VOC S1-CTD and S1-NTD models were superimposed with models from the original SARS2 and endemic HCoV. Structural analyses were done using TM score measurements. Surprisingly, when comparing the original SARS2 and VOC S1-CTD models, we found that they are structurally similar (TM = 1.00) (**Figure 4A**). Moreover, ocular inspection of the model superimposition between the original SARS2 and VOC S1-CTD models showed no difference (**Figures 4B–E**). SARS2 pathogenesis and host tropism were linked to the SARS2 furin-like cleavage site (FLC) (Xing et al., 2020), however, protein structural analyses have shown that the SARS2 S1-CTD [alternatively known as the receptor binding domain (RBD)] is unaffected in the absence of the SARS2 FLC (Cueno et al., 2021; Papa et al., 2021). This emphasizes the structural importance of maintaining the structural conformation of the SARS2 S1-CTD with regards to viral pathogenesis and host tropism consistent with our results. In this regard, we postulate that regardless of successive SARS2 variants being generated, S1-CTD would most likely maintain its structural conformation. In contrast, we observed that the original SARS2 and VOC S1-NTD models had varying structural differences (TM > 0.95) (**Figure 4F**) which can be further seen upon ocular inspection of the model superimposition between the original SARS2 and VOC S1-NTD models (**Figures 4G–J**). Mutations along the S1-NTD have been linked to viral escape from humoral immune response (Graham et al., 2021; Kemp et al., 2021) and S1-NTD was shown to bind to heme metabolites (in particular to biliverdin and bilirubin) which has been proposed to have a role in immune evasion (Rosa et al., 2021). This could putatively mean that structural alterations within the S1-NTD may contribute to immune evasion. Admittedly, additional experimentation is needed to further prove this point.

Subsequently, when comparing VOC and endemic HCoV S1-CTD models, we noted a consistent structural difference (**Figure 4K**) which we ascribe to VOC S1-CTD models being structurally similar (**Figure 4A**). On the other hand, when VOC and endemic HCoV S1-NTD models were structurally compared (**Figure 4L**), we likewise observed varying structural differences consistent with our earlier results (**Figure 4F**). Noticeably, both S1-CTD and S1-NTD models belonging to the same phylogenetic cluster (SARS1, OC43, HKU1, MERS) possibly have the same structural conformation (TM > 0.50) (Yang et al., 2015) with the VOC S1-CTD and S1-NTD models, respectively. These results are consistent with our earlier work and further emphasizes the possibility of the receptor binding structural conformation (S1-

### Original SARS-CoV-2 and SARS-CoV-2 Variants Of Concern



**FIGURE 4 |** Structural comparison of the VOC spike models relative to the original SARS2 and endemic HCoV based on S1-CTD and S1-NTD models. (A–J) Original SARS2 and VOC. (A) TM scores of superimposed S1-CTD models. TM scores normalized to a spike model are distinguished by the presence (original SARS2) or absence (VOC) of a parenthesis. Structural superimposition of SARS2 S1-CTD models between (B) original SARS2 and alpha variant (C) original SARS2 and beta variant (D) original SARS2 and gamma variant, and (E) original SARS2 and delta variant are shown. (F) TM scores of superimposed S1-NTD models. TM scores normalized to a spike model are distinguished by the presence (original SARS2) or absence (VOC) of a parenthesis. Structural superimposition of SARS2 S1-NTD models between (G) original SARS2 and alpha variant (H) original SARS2 and beta variant (I) original SARS2 and gamma variant, and (J) original SARS2 and delta variant are shown. Original SARS2 is colored magenta while the VOC is colored cyan. (K–L) VOC and endemic HCoV. (K) TM scores of superimposed S1-CTD models. TM scores normalized to VOC models. (L) TM scores of superimposed S1-NTD models. TM scores normalized to VOC models. Endemic HCoVs [HCoV 229E (229E), HCoV OC43 (OC43), HCoV NL63 (NL63), HCoV HKU1 (HKU1), SARS-CoV-1 (SARS1), and MERS CoV (MERS)] are indicated.

CTD and S1-NTD) being somewhat conserved in the same phylogenetic cluster and lineage (Cueno and Imai, 2021).

It is worth mentioning that gamma and delta S1-NTD models have similar TM scores (Figure 4F) when compared to the

original SARS2 S1-NTD insinuating that both variants have similar S1-NTD structural conformation. Considering both S1-CTD and S1-NTD models are structurally similar between the gamma and delta variants while having varying viral



transmissibility (Campbell et al., 2021), we hypothesize that amino acid residue changes unique in each variant play a significant role in contributing to viral pathogenesis (Harvey et al., 2021). In a possible future work, it would be interesting to test this hypothesis.

## DISCUSSION

SARS2 genome has mutated consistently with genetic changes occurring almost every week (Day et al., 2020; Mercatelli and Giorgi, 2020). Similarly, nonsynonymous nucleotide changes occurred which in-turn causes amino acid changes (Day et al., 2020). Additionally, these mutations are either high-effect (contribute to viral adaptation and fitness) or low-effect mutations (deleterious and rapidly purged) (Frost et al., 2018; Harvey et al., 2021). Moreover, heavily mutated SARS2 lineages have emerged since the original SARS2 was detected in December 2019 giving rise to VOC (Harvey et al., 2021; Sanyaolu et al., 2021). Throughout this study, we attempted to show that VOC spike models have structural similarities and differences with the original SARS2 and endemic HCoV spike models.

Spike protein binding is the initial step in all CoV infections which is why it is the first CoV antigen targeted by the immune system (Hulswit et al., 2016; Li, 2016; Salvatori et al., 2020). In general, epitopes found along antigen regions are classified as either sequential (continuous or linear amino acid stretch) or conformational (discontinuous amino acid stretch) epitopes (Jerne, 1960; Benjamin et al., 1984; Gershoni et al., 2007). Moreover, antigen:antibody complexes formed are mainly composed of conformational epitopes (~90%) (Haste Andersen et al., 2006). Additionally, antibody paratopes found in the antibody variable region primarily identify and interact with antigen epitopes thereby forming epitope: paratope complementarity which goes beyond amino acid sequence recognition but instead protein structure dynamics (Vojtek et al., 2019). Furthermore, every antibody paratope could interact with multiple antigen epitopes which in-turn could induce a polyclonal immune response resulting to cross-reactivity (Sewell, 2012; Vojtek et al., 2019). These would highlight the potential significance of protein structure formation (particularly conformational epitopes) when considering SARS2 immune response induction. In fact, it was found that viral epitopes (such as Influenza and CMV) that lack sequence identity with SARS2 are able to stimulate an immune response (Mahajan et al., 2021) which we believe is attributable to similar protein structural formation. In this regard, we postulate that high VOC S1-CTD and S1-NTD structural similarity ( $TM > 0.70$ ) with either the original SARS2 or endemic HCoV could putatively have cross-reactivity with the original SARS2 and endemic HCoV spike models (Ladner et al., 2021) possibly ascribable to having multiple similar conformational epitopes that are considered valuable in neutralizing viral pathogenesis (Khare et al., 2021). This is consistent with previous work showing that T cell frequencies against the original SARS2 have likewise been correlated to VOC (Stankov et al., 2021)

which we suspect to be due to structural similarity (particularly S1-CTD). Moreover, VOC have been shown to partially escape humoral immune response, however, VOC are found to be unable to escape cellular immune response among convalescent donors and vaccinees (Geers et al., 2021). This would highlight the putative significance of cellular immune response [particularly Th1 and Tfh cells (Poland et al., 2020)] in providing lasting protection against VOC and, more importantly, the T cell-recognizing conformational epitopes that can counteract viral infectivity (Khare et al., 2021).

It is worth mentioning that VOC emergence is distinguished by having reduced susceptibility to polyclonal antibody responses which can potentially lead to increased reinfections or breakthrough infections (Geers et al., 2021). In this regard, we speculate that both reinfections and breakthrough infections are ascribable to T cell-recognizing conformational changes along the VOC spike glycoprotein [particularly S1-NTD (Graham et al., 2021; Kemp et al., 2021)]. Admittedly, these speculations would need both laboratory and clinically-derived data to prove.

In summary, we putatively showed that: 1) minor structural differences occur in the whole original SARS2 and VOC spike protein model; 2) the whole VOC spike models possibly have differing structural similarity to spike models from endemic HCoVs, wherein, those belonging in the same phylogenetic cluster have high structural similarities while those belonging in a different phylogenetic cluster have low structural similarities; 3) original SARS2 S1-CTD and S1-NTD models are structurally similar to VOC S1-CTD and S1-NTD models; and 4) endemic HCoV S1-CTD and S1-NTD models are structurally similar to VOC S1-CTD and S1-NTD models belonging to the same phylogenetic cluster. Overall, we propose that structural similarities (possibly ascribable to similar conformational epitopes) may help determine immune cross-reactivity, whereas, structural differences (possibly associated with varying conformational epitopes) may lead to viral infection (either reinfection or breakthrough infection)

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## FUNDING

This work was supported by the JSPS KAKENHI Grant Numbers 19K10078 and 19K10097, Uemura Fund provided by the Dental Research Center, Nihon University School of Dentistry, and Nihon University Multidisciplinary Research Grant for 2021.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.773726/full#supplementary-material>

## REFERENCES

- Benjamin, D. C., Berzofsky, J. A., East, I. J., Gurd, F. R. N., Hannum, C., Leach, S. J., et al. (1984). The Antigenic Structure of Proteins: a Reappraisal. *Annu. Rev. Immunol.* 2, 67–101. doi:10.1146/annurev.iy.02.040184.000435
- Berman, H. M., Burley, S. K., Chiu, W., Sali, A., Adzhubei, A., Bourne, P. E., et al. (2006). Outcome of a Workshop on Archiving Structural Models of Biological Macromolecules. *Structure* 14, 1211–1217. doi:10.1016/j.str.2006.06.005
- Bittrich, S., Schroeder, M., and Labudde, D. (2019). StructureDistiller: Structural Relevance Scoring Identifies the Most Informative Entries of a Contact Map. *Sci. Rep.* 9, 18517. doi:10.1038/s41598-019-55047-4
- Callaway, E. (2021). Remember Beta? New Data Reveal Variant's Deadly powers. *Nature*. doi:10.1038/d41586-021-02177-3
- Campbell, F., Archer, B., Laurenson-Schafer, H., Jinnai, Y., Konings, F., Batra, N., et al. (2021). Increased Transmissibility and Global Spread of SARS-CoV-2 Variants of Concern as at June 2021. *Euro Surveill.* 26, 2100509. doi:10.2807/1560-7917.ES.2021.26.24.2100509
- Chen, S. C., and Bahar, I. (2004). Mining Frequent Patterns in Protein Structures: a Study of Protease Families. *Bioinformatics* 20, i77–85. doi:10.1093/bioinformatics/bth912
- Chiu, S. S., Chan, K. H., Chu, K. W., Kwan, S. W., Guan, Y., Man Poon, L. L., et al. (2005). Human Coronavirus NL63 Infection and Other Coronavirus Infections in Children Hospitalized with Acute Respiratory Disease in Hong Kong, China. *Clin. Infect. Dis.* 40, 1721–1729. doi:10.1086/430301
- Cueno, M. E., and Imai, K. (2021). Structural Comparison of the SARS CoV 2 Spike Protein Relative to Other Human-Infecting Coronaviruses. *Front. Med.* 7, 594439. doi:10.3389/fmed.2020.594439
- Cueno, M. E., Ueno, M., Iguchi, R., Harada, T., Miki, Y., Yasumaru, K., et al. (2021). Insights on the Structural Variations of the Furin-like Cleavage Site Found Among the December 2019–July 2020 SARS-CoV-2 Spike Glycoprotein: A Computational Study Linking Viral Evolution and Infection. *Front. Med.* 8, 613412. doi:10.3389/fmed.2021.613412
- Day, T., Gandon, S., Lion, S., and Otto, S. P. (2020). On the Evolutionary Epidemiology of SARS-CoV-2. *Curr. Biol.* 30, R849–R857. doi:10.1016/j.cub.2020.06.031
- Di Rienzo, L., Milanetti, E., Lepore, R., Olimpieri, P. P., and Tramontano, A. (2017). Superposition-free Comparison and Clustering of Antibody Binding Sites: Implications for the Prediction of the Nature of Their Antigen. *Sci. Rep.* 7, 45053. doi:10.1038/srep45053
- Fouchier, R. A. M., Hartwig, N. G., Bestebroer, T. M., Niemeyer, B., De Jong, J. C., Simon, J. H., et al. (2004). A Previously Undescribed Coronavirus Associated with Respiratory Disease in Humans. *Proc. Natl. Acad. Sci.* 101, 6212–6216. doi:10.1073/pnas.0400762101
- Frost, S. D. W., Magalis, B. R., and Kosakovsky Pond, S. L. (2018). Neutral Theory and Rapidly Evolving Viral Pathogens. *Mol. Biol. Evol.* 35, 1348–1354. doi:10.1093/molbev/msy088
- Geers, D., Shamier, M. C., Bogers, S., Den Hartog, G., Gommers, L., Nieuwkoop, N., et al. (2021). SARS-CoV-2 Variants of Concern Partially Escape Humoral but Not T-Cell Responses in COVID-19 Convalescent Donors and Vaccinees. *Sci. Immunol.* 6, eabj1750. doi:10.1126/sciimmunol.abj1750
- Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Tarnovitski Freund, N., and Weiss, Y. (2007). Epitope Mapping: The First Step in Developing Epitope-Based Vaccines. *BioDrugs* 21, 145–156. doi:10.2165/00063030-200721030-00002
- Graham, C., Seow, J., Huettner, I., Khan, H., Kouphou, N., Acors, S., et al. (2021). Neutralization Potency of Monoclonal Antibodies Recognizing Dominant and Subdominant Epitopes on SARS-CoV-2 Spike Is Impacted by the B.1.1.7 Variant. *Immunity* 54, 1276–1289. doi:10.1016/j.immuni.2021.03.023
- Hamre, D., and Procknow, J. J. (1966). A New Virus Isolated from the Human Respiratory Tract. *Proc. Soc. Exp. Biol. Med.* 121, 190–193. doi:10.3181/00379727-121-30734
- Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., et al. (2021). SARS-CoV-2 Variants, Spike Mutations and Immune Escape. *Nat. Rev. Microbiol.* 19, 409–424. doi:10.1038/s41579-021-00573-0
- Haste Andersen, P., Nielsen, M., and Lund, O. (2006). Prediction of Residues in Discontinuous B-Cell Epitopes Using Protein 3D Structures. *Protein Sci.* 15, 2558–2567. doi:10.1110/ps.062405906
- Herráez, A. (2006). Biomolecules in the Computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.* 34, 255–261. doi:10.1002/bmb.2006.494034042644
- Hevener, K. E., Zhao, W., Ball, D. M., Babaoglu, K., Qi, J., White, S. W., et al. (2009). Validation of Molecular Docking Programs for Virtual Screening against Dihydropterolate Synthase. *J. Chem. Inf. Model.* 49, 444–460. doi:10.1021/ci800293n
- Holm, L., and Sander, C. (1996). Mapping the Protein Universe. *Science* 273, 595–602. doi:10.1126/science.273.5275.595
- Hulswit, R. J., De Haan, C. A., and Bosch, B. J. (2016). Coronavirus Spike Protein and Tropism Changes. *Adv. Virus. Res.* 96, 29–57. doi:10.1016/b.s.aivir.2016.08.004
- Jerne, N. K. (1960). Immunological Speculations. *Annu. Rev. Microbiol.* 14, 341–358. doi:10.1146/annurev.mi.14.100160.002013
- Kapikian, A. Z., James, H. D., Jr., Kelly, S. J., Dees, J. H., Turner, H. C., McIntosh, K., et al. (1969). Isolation from Man of “avian Infectious Bronchitis Virus-like” Viruses (Coronaviruses) Similar to 229E Virus, with Some Epidemiological Observations. *J. Infect. Dis.* 119, 282–290. doi:10.1093/infdis/119.3.282
- Kelley, L. A., and Sternberg, M. J. E. (2009). Protein Structure Prediction on the Web: a Case Study Using the Phyre Server. *Nat. Protoc.* 4, 363–371. doi:10.1038/nprot.2009.2
- Kemp, S. A., Collier, D. A., Datir, R. P., Ferreira, I., Gayed, S., Jahun, A., et al. (2021). SARS-CoV-2 Evolution during Treatment of Chronic Infection. *Nature* 592, 277–282. doi:10.1038/s41586-021-03291-y
- Khare, S., Azevedo, M., Parajuli, P., and Gokulan, K. (2021). Conformational Changes of the Receptor Binding Domain of SARS-CoV-2 Spike Protein and Prediction of a B-Cell Antigenic Epitope Using Structural Data. *Front. Artif. Intell.* 4, 630955. doi:10.3389/frai.2021.630955
- King, A. M. Q., Lefkowitz, E. J., Mushegian, A. R., Adams, M. J., Dutilh, B. E., Gorbalenya, A. E., et al. (2018). Changes to Taxonomy and the International Code of Virus Classification and Nomenclature Ratified by the International Committee on Taxonomy of Viruses (2018). *Arch. Virol.* 163, 2601–2631. doi:10.1007/s00705-018-3847-1
- Koyama, T., Platt, D., and Parida, L. (2020). Variant Analysis of SARS-CoV-2 Genomes. *Bull. World Health Organ.* 98, 495–504. doi:10.2471/blt.20.253591
- Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., et al. (2003). A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *N. Engl. J. Med.* 348, 1953–1966. doi:10.1056/nejmoa030781
- Kufareva, I., and Abagyan, R. (2012). Methods of Protein Structure Comparison. *Methods Mol. Biol.* 857, 231–257. doi:10.1007/978-1-61779-588-6\_10
- Ladner, J. T., Henson, S. N., Boyle, A. S., Engelbrektson, A. L., Fink, Z. W., Rahee, F., et al. (2021). Epitope-resolved Profiling of the SARS-CoV-2 Antibody Response Identifies Cross-Reactivity with Endemic Human Coronaviruses. *Cell Rep. Med.* 2, 100189. doi:10.1016/j.xcrm.2020.100189
- Letko, M., Marzi, A., and Munster, V. (2020). Functional Assessment of Cell Entry and Receptor Usage for SARS-CoV-2 and Other Lineage B Betacoronaviruses. *Nat. Microbiol.* 5, 562–569. doi:10.1038/s41564-020-0688-y
- Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu. Rev. Virol.* 3, 237–261. doi:10.1146/annurev-virology-110615-042301

- Lu, G., Wang, Q., and Gao, G. F. (2015). Bat-to-human: Spike Features Determining 'host Jump' of Coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends Microbiol.* 23, 468–478. doi:10.1016/j.tim.2015.06.003
- Madhi, S. A., Baillie, V., Cutland, C. L., Voysey, M., Koen, A. L., Fairlie, L., et al. (2021). Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* 384, 1885–1898. doi:10.1056/nejmoa2102214
- Mahajan, S., Kode, V., Bhojak, K., Karunakaran, C., Lee, K., Manoharan, M., Ramesh, A., et al. (2021). Immunodominant T-cell Epitopes From the SARS-CoV-2 Spike Antigen Reveal Robust Pre-Existing T-cell Immunity in Unexposed Individuals. *Sci. Rep.* 11, 13164.
- Mercatelli, D., and Giorgi, F. M. (2020). Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Front. Microbiol.* 11, 1800. doi:10.3389/fmicb.2020.01800
- Millet, J. K., and Whittaker, G. R. (2015). Host Cell Proteases: Critical Determinants of Coronavirus Tropism and Pathogenesis. *Virus. Res.* 202, 120–134. doi:10.1016/j.virusres.2014.11.021
- Papa, G., Mallery, D. L., Albecka, A., Welch, L. G., Cattin-Ortolá, J., Luptak, J., et al. (2021). Furin Cleavage of SARS-CoV-2 Spike Promotes but Is Not Essential for Infection and Cell-Cell Fusion. *Plos Pathog.* 17, e1009246. doi:10.1371/journal.ppat.1009246
- Poland, G. A., Ovsyannikova, I. G., and Kennedy, R. B. (2020). SARS-CoV-2 Immunity: Review and Applications to Phase 3 Vaccine Candidates. *Lancet* 396, 1595–1606. doi:10.1016/s0140-6736(20)32137-1
- Robson, B. (2020). Computers and Viral Diseases. Preliminary Bioinformatics Studies on the Design of a Synthetic Vaccine and a Preventative Peptidomimetic Antagonist Against the SARS-CoV-2 (2019-nCoV, COVID-19) Coronavirus. *Comput. Biol. Med.* 119, 103670. doi:10.1016/j.combiomed.2020.103670
- Rosa, A., Pye, V. E., Graham, C., Muir, L., Seow, J., Ng, K. W., et al. (2021). SARS-CoV-2 Can Recruit a Heme Metabolite to Evade Antibody Immunity. *Sci. Adv.* 7, eabg7607. doi:10.1126/sciadv.abg7607
- Salvatori, G., Luberto, L., Maffei, M., Aurisicchio, L., Roscilli, G., Palombo, F., et al. (2020). SARS-CoV-2 SPIKE PROTEIN: an Optimal Immunological Target for Vaccines. *J. Transl. Med.* 18, 222. doi:10.1186/s12967-020-02392-y
- Sanyaolu, A., Okorie, C., Marinkovic, A., Haider, N., Abbasi, A. F., Jaferi, U., et al. (2021). The Emerging SARS-CoV-2 Variants of Concern. *Ther. Adv. Infect. Dis.* 8, 20499361211024372. doi:10.1177/20499361211024372
- Sewell, A. K. (2012). Why Must T Cells Be Cross-Reactive. *Nat. Rev. Immunol.* 12, 669–677. doi:10.1038/nri3279
- Stankov, M. V., Cossmann, A., Bonifacius, A., Dopfer-Jablonka, A., Ramos, G. M., Godecke, N., et al. (2021). Humoral and Cellular Immune Responses against SARS-CoV-2 Variants and Human Coronaviruses after Single BNT162b2 Vaccination. *Clin. Infect. Dis.* doi:10.1093/cid/ciab555
- Tay, M. Z., Poh, C. M., Rénia, L., Macary, P. A., and Ng, L. F. P. (2020). The trinity of COVID-19: Immunity, Inflammation and Intervention. *Nat. Rev. Immunol.* 20, 363–374. doi:10.1038/s41577-020-0311-8
- Tomaszewski, T., Devries, R. S., Dong, M., Bhatia, G., Norsworthy, M. D., Zheng, X., et al. (2020). New Pathways of Mutational Change in SARS-CoV-2 Proteomes Involve Regions of Intrinsic Disorder Important for Virus Replication and Release. *Evol. Bioinform. Online* 16, 1176934320965149. doi:10.1177/1176934320965149
- Vehlow, C., Stehr, H., Winkelmann, M., Duarte, J. M., Petzold, L., Dinse, J., et al. (2011). CMView: Interactive Contact Map Visualization and Analysis. *Bioinformatics* 27, 1573–1574. doi:10.1093/bioinformatics/btr163
- Vojtek, I., Buchy, P., Doherty, T. M., and Hoet, B. (2019). Would Immunization Be the Same without Cross-Reactivity. *Vaccine* 37, 539–549. doi:10.1016/j.vaccine.2018.12.005
- Wang, Z., and Xu, J. (2013). Predicting Protein Contact Map Using Evolutionary and Physical Constraints by Integer Programming. *Bioinformatics* 29, i266–i273. doi:10.1093/bioinformatics/btt211
- Woo, P. C., Lau, S. K., Chu, C. M., Chan, K. H., Tsoi, H. W., Huang, Y., et al. (2005). Characterization and Complete Genome Sequence of a Novel Coronavirus, Coronavirus HKU1, from Patients with Pneumonia. *J. Virol.* 79, 884–895. doi:10.1128/jvi.79.2.884-895.2005
- Xing, Y., Li, X., Gao, X., and Dong, Q. (2020). Natural Polymorphisms Are Present in the Furin Cleavage Site of the SARS-CoV-2 Spike Glycoprotein. *Front. Genet.* 11, 783. doi:10.3389/fgene.2020.00783
- Yang, J., Zhang, W., He, B., Walker, S. E., Zhang, H., Govindarajoo, B., et al. (2015). Template-based Protein Structure Prediction in CASP11 and Retrospect of I-TASSER in the Last Decade. *Proteins* 84, 233–246. doi:10.1002/prot.24918
- Zaki, A. M., Van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E., and Fouchier, R. A. M. (2012). Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *N. Engl. J. Med.* 367, 1814–1820. doi:10.1056/nejmoa1211721
- Zhang, Y., and Skolnick, J. (2005). TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic. Acids. Res.* 33, 2302–2309.
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi:10.1056/nejmoa2001017

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Cueno and Imai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Perspectives About Modulating Host Immune System in Targeting SARS-CoV-2 in India

## OPEN ACCESS

### Edited by:

Indrajit Saha,  
National Institute of Technical  
Teachers' Training and Research,  
India

### Reviewed by:

Michael Poidinger,  
Murdoch Childrens Research  
Institute, Royal Children's Hospital,  
Australia  
Balaji Banoth,  
St. Jude Children's Research  
Hospital, United States  
Ashok Sharma,  
Augusta University, United States

### \*Correspondence:

Sudipto Saha  
ssaha4@jcbosc.ac.in;  
ssaha4@gmail.com

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 03 December 2020

**Accepted:** 19 January 2021

**Published:** 16 February 2021

### Citation:

Majumdar S, Verma R, Saha A,  
Bhattacharyya P, Maji P, Surjit M,  
Kundu M, Basu J and Saha S (2021)  
Perspectives About Modulating Host  
Immune System in Targeting  
SARS-CoV-2 in India.  
Front. Genet. 12:637362.  
doi: 10.3389/fgene.2021.637362

Sreyashi Majumdar<sup>1</sup>, Rohit Verma<sup>2</sup>, Avishek Saha<sup>3</sup>, Parthasarathi Bhattacharyya<sup>4</sup>,  
Pradipta Maji<sup>5</sup>, Milan Surjit<sup>2</sup>, Manikuntala Kundu<sup>6</sup>, Joyoti Basu<sup>6</sup> and Sudipto Saha<sup>1\*</sup>

<sup>1</sup> Division of Bioinformatics, Bose Institute, Kolkata, India, <sup>2</sup> Virology Laboratory, Vaccine and Infectious Disease Research Centre, Translational Health Science and Technology Institute, NCR Biotech Science Cluster, Faridabad, India, <sup>3</sup> Ubiquitous Analytical Techniques, CSIR-Central Scientific Instruments Organisation, Chandigarh, India, <sup>4</sup> Department of Respiratory Medicine, Institute of Pulmocare and Research, Kolkata, India, <sup>5</sup> Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India, <sup>6</sup> Department of Chemistry, Bose Institute, Kolkata, India

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), the causative agent of coronavirus induced disease-2019 (COVID-19), is a type of common cold virus responsible for a global pandemic which requires immediate measures for its containment. India has the world's largest population aged between 10 and 40 years. At the same time, India has a large number of individuals with diabetes, hypertension and kidney diseases, who are at a high risk of developing COVID-19. A vaccine against the SARS-CoV-2, may offer immediate protection from the causative agent of COVID-19, however, the protective memory may be short-lived. Even if vaccination is broadly successful in the world, India has a large and diverse population with over one-third being below the poverty line. Therefore, the success of a vaccine, even when one becomes available, is uncertain, making it necessary to focus on alternate approaches of tackling the disease. In this review, we discuss the differences in COVID-19 death/infection ratio between urban and rural India; and the probable role of the immune system, co-morbidities and associated nutritional status in dictating the death rate of COVID-19 patients in rural and urban India. Also, we focus on strategies for developing masks, vaccines, diagnostics and the role of drugs targeting host-virus protein-protein interactions in enhancing host immunity. We also discuss India's strengths including the resources of medicinal plants, good food habits and the role of information technology in combating COVID-19. We focus on the Government of India's measures and strategies for creating awareness in the containment of COVID-19 infection across the country.

**Keywords:** SARS-CoV-2, genetic variations, host immuno-modulation, repurposed drugs, vaccines, medicinal plants, CT scans, artificial intelligence



## INTRODUCTION

Coronavirus disease 2019 (COVID-19) outbreak, caused by the novel coronavirus (SARS-CoV-2) has emerged as a global epidemic and posed serious worldwide public health concerns owing to the contagious nature of the virus and high death rate. Transmission through droplets facilitated its rapid spread and caused panic across the globe. There were 80,776,890 confirmed cases worldwide till December 30, 2020<sup>1</sup>. India has also been largely affected by instances of COVID-19. SARS-CoV-2 viral protein interacts with various host proteins to mediate viral entry and replication in the human host (Khorsand et al., 2020). Targeting virus and host protein-protein interactions or downstream signaling cascades using novel or repositioned drugs, serves as one of the strategies for COVID-19 therapy. Several drugs such as remdesivir, dexamethasone, hydroxychloroquine, ivermectin, azithromycin, tocilizumab, famotidine, thalidomide have been evaluated in different countries for their efficacy in treating COVID-19 (Omolo et al., 2020). Convalescent plasma therapy has been recommended by FDA as an alternative therapeutic strategy for severe forms of COVID-19 infection<sup>2</sup>. Vaccination has been considered as the major option for containing the COVID-19 pandemic. Presently, 172 vaccine candidates are in developmental stage, while 63 have entered clinical trials<sup>3</sup>. The Oxford COVID-19 group have clinically proven the safety of the ChAdOx1 nCoV-19 vaccine in triggering humoral and cellular immune response against SARS-CoV-2. The vaccine is presently under the phase 3 trial program across the world (Folegatti et al., 2020). The phase 3 trials of Covishield, the Oxford vaccine in India have been conducted under the supervision of Serum Institute of India, Pune and the vaccine has been approved for emergency supply and use in India<sup>4,5</sup>. COVAXIN has been developed as India's first indigenous vaccine by Bharat Biotech in association with Indian Council of Medical Research (ICMR). COVAXIN has currently gone into Phase III clinical trial after successful completion of Phase I and II clinical trials started by Bharat Biotech from July, 2020 onwards<sup>6</sup>. Recently, the Drug Controller General of India (DCGI) has granted emergency approval to COVAXIN in India<sup>7</sup>. The Ministry of Ayush under the Govt. of India has emphasized the importance of exploiting medicinal herbs in the context of COVID-19. Indian indigenous medicinal plants with immune regulating properties have often served to boost immunity and render protection against

viral infections (Akram et al., 2018; Mohanraj et al., 2018). Besides these, the Govt. of India has adopted certain strategies such as social distancing and extensive lockdown for effective containment of COVID-19 and has launched the artificial intelligence (AI) based mobile application Aarogya Setu to create public awareness.

Computational bioinformatics and AI have been exploited for better management of COVID-19. Machine learning techniques (MLTs) have been employed for taxonomic and hierarchical classification of SARS-CoV-2 strains (Randhawa et al., 2020). Computational approaches used in CRISPR based detection systems and neural network for COVID-19 detection have increased diagnostic accuracy (Alimadadi et al., 2020; Li et al., 2020). Deep learning technology based on pulmonary CT scan images has successfully allowed differentiation of COVID-19 from other respiratory diseases such as community acquired pneumonia (Li et al., 2020). Novel text mining based collection of COVID-19 related big data, followed by subsequent analyses using advanced machine learning techniques have enabled real time surveillance of viral epidemiology and live tracking of COVID-19 cases. Access to these digital big data through mobile applications allows potential risk assessment and rapid information dissemination in public for creating social awareness and better mitigation of COVID-19 (Alimadadi et al., 2020; Bragazzi et al., 2020; Srinivasa Rao and Vazquez, 2020; Ting et al., 2020). Besides, bioinformatics tools and AI are of prime importance in drug discovery and vaccine development for SARS-CoV-2. Repurposing of existing drugs and computation based drug target identification have been extensively performed to accelerate the therapy of COVID-19. In silico docking and deep learning based drug designing have been employed to develop novel drugs against SARS-CoV-2 (Alimadadi et al., 2020; Bragazzi et al., 2020; Senior et al., 2020). A deep learning system Alphafold was designed by Google DeepMind for identification of protein structures linked with COVID-19 that might be valuable for vaccine formulation (Senior et al., 2020). Vaxign reverse vaccinology tool amalgamated with machine learning has also been used to predict vaccine candidates for COVID-19 (Ong et al., 2020). B-cell epitopes and MHC Class II epitopes can also be predicted using bioinformatics tools for peptide based vaccine development (Jabbari and Rezaei, 2019). Potential vaccine adjuvants can also be screened using the AI based program named Search Algorithm for Ligands (SAM) (Ahuja et al., 2020).

In a nutshell, this review highlights the current scenario of COVID-19 across India, with special emphasis on death/infection ratio in urban and rural India and disease association with co-morbidities. This review also deals with strains of SARS-CoV-2 circulating in India and the immunomodulatory action of viral proteins. It discusses the various diagnostic kits, masks and disinfection techniques in use for diagnosing and combating COVID-19. This review further focuses on various approaches that may be followed to tackle the problem of SARS-CoV-2 infection (summarized in **Figure 1**), including immuno-regulating drugs, drugs targeting host-viral protein-protein interactions, vaccines

<sup>1</sup> Covid19.Who.Int/

<sup>2</sup> <https://www.fda.gov/Vaccines-Blood-Biologics/Investigational-New-Drug-Ind-or-Device-Exemption-Ide-Process-Cber/Recommendations-Investigational-Covid-19-Convalescent-Plasma>

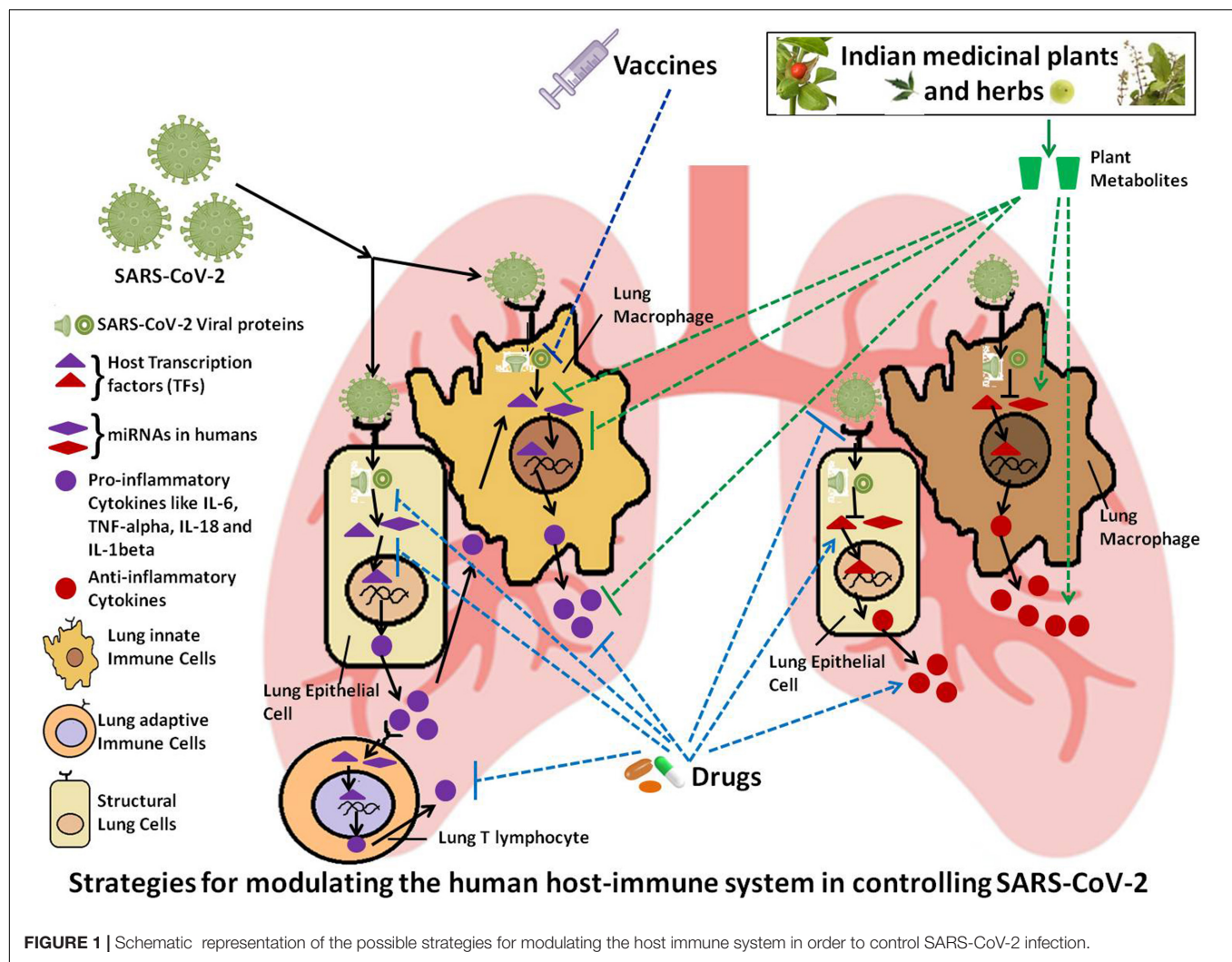
<sup>3</sup> <https://www.who.int/Publications/M/Item/Draft-Landscape-of-Covid-19-Candidate-Vaccines>

<sup>4</sup> <https://Timesofindia.Indiatimes.Com/India/Oxford-Covid-Vaccine-Set-to-Become-First-to-Get-Approval-in-India-Report/Articleshow/80059745.Cms>

<sup>5</sup> <https://Vaccine.Icmr.Org.In/Covid-19-Vaccine>

<sup>6</sup> <https://www.bharatbiotech.com/covaxin.html>

<sup>7</sup> <https://www.expresspharma.in/covid19-updates/dcgi-approves-covishield-and-covaxin-for-restricted-emergency-use-in-india/>



and Indian herbs and plants with medicinal and immuno-modulating properties. Lastly, it deals with role of AI and various Government strategies adopted in India for addressing the COVID-19 pandemic.

## GENERAL SCENARIO IN INDIA

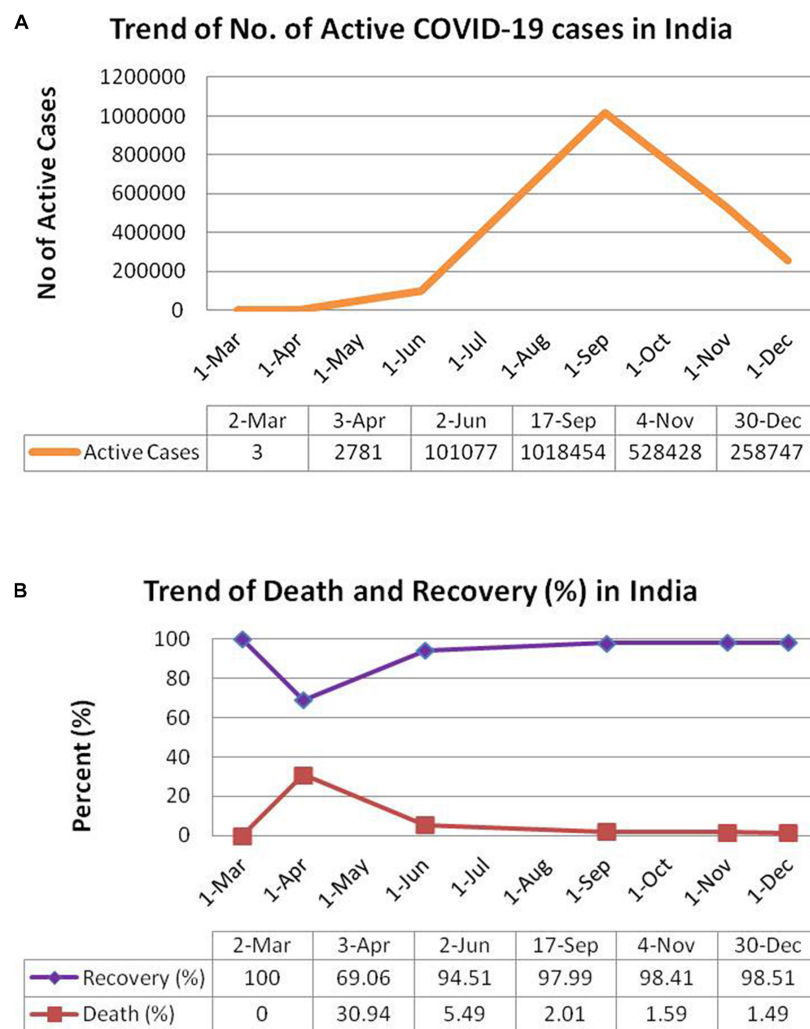
### COVID-19 Infection and Death Rate in States and Union territories (UTs) of India: Association With Lifestyle Habits, Proximity to Airport and Urbanization

The first case of COVID-19 in India was diagnosed in March, 2020. From that time onwards, there has been rise in the incidence of COVID-19 in India with 2,871 and 101,077 active cases on April 3, 2020 and June 2, 2020 respectively. The number of active cases reached a peak of 1,018,454 on September 17, 2020 followed by a decrease to 528,428 active cases on November 4, 2020 and 258,747 active cases on December 30, 2020 and (as evident from **Figure 2A**). Till December 30, 2020, there

has been a total of 10,267,283 confirmed cases in India with a total of 148,774 deaths<sup>8</sup>. In terms of total number of cases, India occupies the second position after United States and is followed by developed nations such as Brazil, Russia, France, and the United Kingdom (as shown in **Supplementary Table 1**). The numbers of daily new cases have reduced considerably in December, 2020 as compared to that in September, 2020. There has been a consistent increase in percent recovery from April to December with a minimum recovery rate of 69.06% in April 3, 2020 to a recovery rate of 98.51% on December 30, 2020 (as shown in **Figure 2B**). Likewise, the death percentage has declined to 1.49% on December 30, 2020 after a surge of 30.94% in April 3, 2020 (see text footnote 71). The decrease in COVID-19 deaths (in terms of death/total confirmed case ratio) across different states of India from June, 2020 to December, 2020 has been tabulated in **Supplementary Table 2**.

States and cities of India harboring busy international airports (such as Kolkata in West Bengal; Ahmedabad, Surat in Gujarat; Mumbai in Maharashtra; New Delhi in Delhi and Chennai

<sup>8</sup><https://www.worldometers.info/coronavirus/country/india/>



**FIGURE 2 |** Schematic representation of trend of COVID-19 pandemic in India. **(A)** Total no. of active cases in India. **(B)** Percentage death and recovery in India. Panels **(A)** and **(B)** have been adopted from <https://www.worldometers.info/coronavirus/country/india/>.

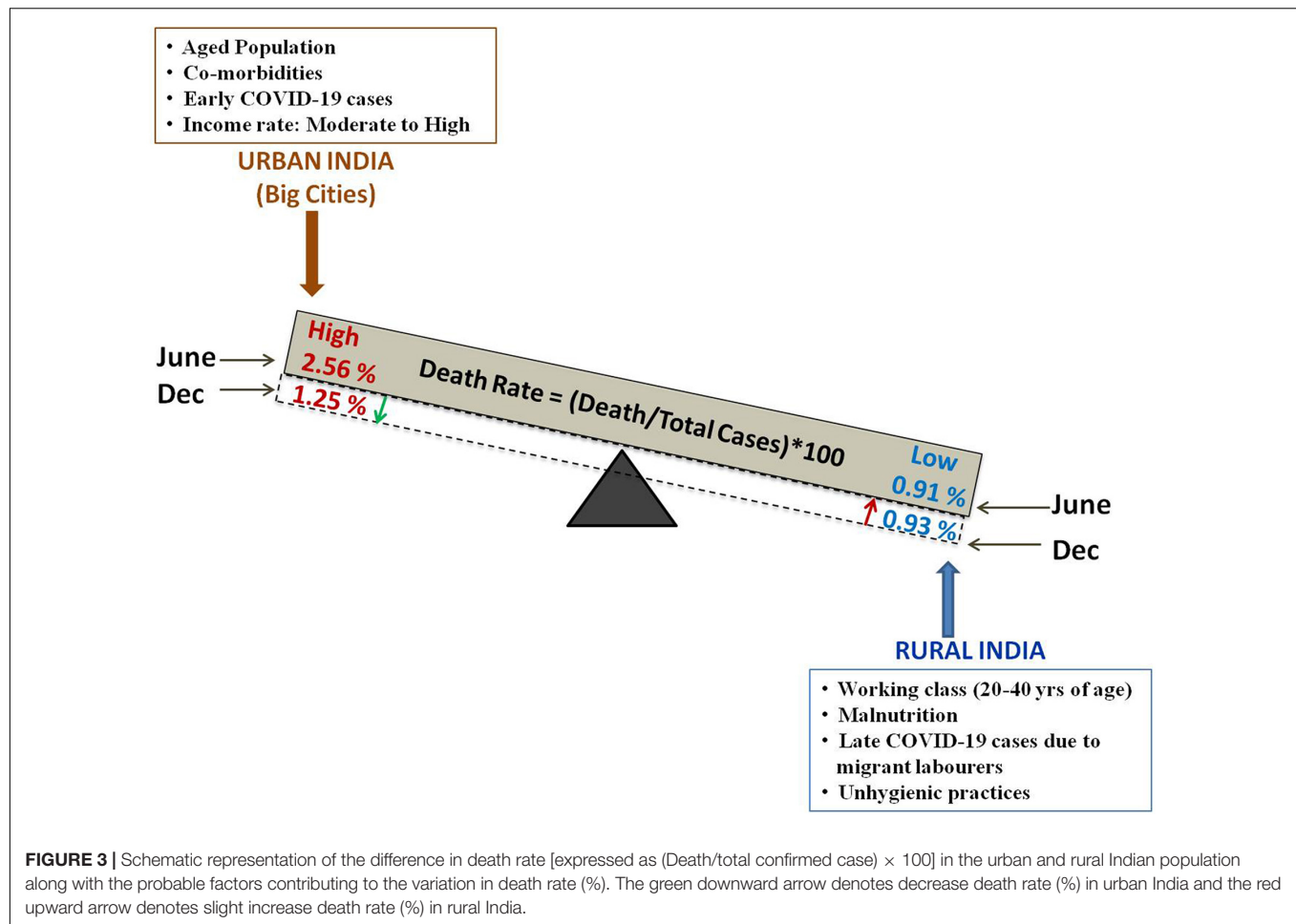
in Tamil Nadu) have shown high COVID-19 death rates. The total number of confirmed cases and death rates in Indian cities with important international airports has been tabulated in **Supplementary Table 3**.

The incidences of COVID-19, death/total cases ratio and death rate have been higher in urban India than in rural areas. The death rate in urban India showed a decline in November, 2020 but, there is no significant change in the rural COVID-19 death rate (as evident from **Figure 3**, **Table 1** and **Supplementary Table 4**). Early COVID-19 cases in India were primarily diagnosed in cities. Subsequent to the movement of migrant laborers from urban to rural areas and easing of transportation between rural and urban areas, there has been an increase in COVID-19 cases in rural India. High population density, greater economic activity, infrastructure development and movement of people contribute to constraints in social distancing in urban areas. Urban food habits (fast food, alcohol consumption), lifestyle patterns (improper sleep pattern, lack of

physical exercise, stress) and high levels of pollution result in non-communicable lifestyle diseases (such as obesity, diabetes, and hypertension), which create additional complications in COVID-19 patients. Instances of such disorders are lower among rural population. Besides, rural lifestyle practices such as consumption of hot food, prolonged periods of sun exposure due to agricultural field work, lesser crowding, limited instances of handshaking may prove to be advantageous in conferring protection from COVID-19 (Mishra S. et al., 2020). Correlation analyses carried out in rural and semi rural areas indicate very weak positive correlation of COVID Fatality Rate (CFR) and hypertension; mild negative correlation of CFR with diabetes, implying that CFR is not necessarily related to co-morbidities such as hypertension and diabetes in rural areas<sup>9</sup>. Although,

<sup>9</sup><https://www.thehindu.com/Data/Data-Lower-Covid-19-Fatality-Rate-in-Rural-Areas-Not-Necessarily-Due-to-Lower-Share-of-Co-Morbidities-among-Rural-Population/Article32620632.Ece>





reduced to some extent by the Swachh Bharat Mission, a large proportion of rural households avail open defecation and public toilet facilities. Also, many rural households travel long distances to carry drinking water from community source. Social distancing becomes a difficult proposition in such situations (Mishra S. V. et al., 2020). Further, many rural households do not have exclusive rooms for individuals, thus making self isolation difficult. So, careful monitoring of urban-rural movement and augmentation of rural healthcare facilities, wherever necessary, is required to control rise in rural COVID-19 cases and death rates.

### Association of COVID-19 With Other Co-morbidities in India

Globally, common co-morbidities such as hypertension, diabetes, asthma, cardiovascular disease (CVD), chronic obstructive pulmonary disease (COPD), obesity, chronic kidney disease (CKD), cerebro-vascular accident (CVA), malignancy, and inflammatory conditions have been noted to worsen health status in COVID-19 positive patients (Renu et al., 2020). Medications used in these conditions often lead to upregulation of Angiotensin-converting enzyme 2 (ACE-2) receptor; thereby enhancing the possibility of ACE2 mediated viral entry

and susceptibility to SARS-CoV-2 infection (Shahid et al., 2020). Communicable diseases such as tuberculosis and HIV-AIDS (Human immunodeficiency virus – Acquired Immuno Deficiency Syndrome) have also been associated with escalated severity and death rate in COVID-19 patients across the world. Sporadic studies from different Indian states/cities such as West Bengal and Jaipur revealed association of one or more co-morbid conditions with deaths in COVID-19 patients. Computational analysis based on Boolean search highlighted diabetes as the most prevalent co-morbidity in Indian COVID-19 patients, followed by hypertension (Singh and Misra, 2020). Co-morbidities in COVID-19 patients result in increased medical complications, incidence of hospitalization and high mortality rate. In order to deal with medical complications arising from COVID-19, it is vital to have knowledge regarding the SARS-CoV-2 strains and the viral mode of action within the host system.

### SARS-CoV-2 Strains Available in India and Their Evolution

Phylogenetic studies denote the causative agent of COVID-19 as belonging to the family *Coronaviridae*. Viruses belonging to this family have a single-stranded, (+) sense RNA genome of



**TABLE 1** | Percentage death rate in urban and rural population across some states of India.

Sl.No	States	Urban or rural	Districts considered for study	Death rate (in %) (as on 10.06.2020)	Death rate (in %) (as on 04.11.2020)	Death rate (in %) (as on 30.12.2020)
1.	West Bengal	Urban	Kolkata, North 24 Parganas, South 24 Parganas, Howrah, Hooghly	4.45	2.24	2.15
		Rural	Malda, Paschim Medinipur, Purba Medinipur, Nadia, Puruliya	0.91	1.08	1.15
2.	Odisha	Urban	Khordha, Nayagarh, Cuttack, Puri, Malkangiri	0.44	0.55	0.64
		Rural	Ganjam, Balangir, Debagarh, Mayurbhanj, Jagatsinghpur	0.12	0.51	0.62
3.	Bihar	Urban	Patna, Gaya, Nalanda, Bhagalpur, Bengusarai	0.56	0.68	0.70
		Rural	Samastipur, Banka, Madhubani, Kaimur, Madhepura	0.31	0.47	0.50
4.	Uttar Pradesh	Urban	Lucknow, Ghaziabad, Agra, Meerut, Kanpur Nagar	4.14	1.75	1.60
		Rural	Allahabad, Azamgarh, Jaunpur, Sitapur, Gorakhpur	2.22	1.50	1.53
5.	Jharkhand	Urban	Ranchi, Dhanbad, Bokaro, Purbi Singhbhum, Ramgarh	1.95	1.08	1.12
		Rural	Giridih, Palamu, Hazaribagh, Pashchimi Singhbhum, Simdega	0.38	0.52	0.57
6.	Madhya Pradesh	Urban	Bhopal, Indore, Gwalior, Jabalpur, Ujjain	4.21	1.91	1.61
		Rural	Dindori, Jhabua, Bhind, Morena, Rewa	2.74	0.75	0.73
7.	Haryana	Urban	Gurgaon, Panipat, Faridabad, Rohtak, Sonipat	2.12	0.87	0.92
		Rural	Palwal, Rewari, Mewat, Jhajjar, Fatehabad	0	1.19	1.41

Death rate = (Death/total confirmed case) × 100; Death rate computed with data obtained on 10.06.2020, 04.11.2020 and 30.12.2020 <https://Bing.Com/Covid/Local/India>.

~30 kb (Yadav et al., 2020). During the 18th to 19th centuries, viruses from these families were known to cause infections only in animals (Cui et al., 2019). The first time it was discovered in humans was in mid 1965. This strain was referred as HCoV 229E in the United States. This was followed by an outbreak of coronavirus in France caused by another member of the same family, HCoV OC43 that led to 501 confirmed cases in 2000–2001. Till date seven different coronaviruses have been identified in this family that cause infection in humans. There have been five subsequent outbreaks in two decades prior to the recent pandemic caused by SARS-CoV-2 in December 2019 that originated from Wuhan city, China. Bioinformatics based analyses on SARS-CoV-2 genomes isolated from different countries shows its close relation with two bat origin SARS-CoV (bat-SL-CoVZC45 and bat-SL-CoVZXC21). Further, in-depth analysis of SARS-CoV-2 sequence exhibits 96.3% genome similarity with Bat CoV RaTG13 (Yadav et al., 2020). Upon comparison of SARS-CoV-2 with SARS-CoV, six different mutations were identified in ORF1a/b, S, ORF7b, and ORF8 genes. Moreover, similarity between RdRp and 3CLPro proteins has been reported. ORF8 and ORF10 show no homology with that of SARS-CoV strain (Kaur et al., 2020). Till now, no confirmed animal reservoir has been identified, although pangolins are claimed to be natural reservoirs due to the high similarity of the spike region between human SARS-CoV-2 and pangolin SARS-CoV (Andersen et al., 2020). Viruses belonging to this family have an anomalous feature of rapid mutation in their genome that causes variability in

the strain. Studies are being conducted to understand the genetic diversity and evolution to establish a reference sequence for SARS-CoV-2 through mathematical modeling and Single Nucleotide Polymorphism (SNP) analysis of all the available sequences (Wang et al., 2020b). In the context of therapeutic drug and vaccine development, it is essential to monitor and track local and global genetic variations in the genome (Yin, 2020). A study of 3636 SARS-CoV-2 RNA sequences from 55 different countries revealed a remarkable mutation in the S protein at D614 amino acid position (D614G) among all the high-frequency mutations and was classified as A2a subtype. These high-frequency mutations in the SARS CoV-2 genome have resulted in 11 different clusters of related sequences. Among these, O type is an ancestral type that arose from China. SARS-CoV-2 genotypes A, B, C have been described previously. These have been further divided into subtypes B, B1, B2, and B4 on the basis of mutations in the ORF8 region of SARS CoV-2. Genotype A possesses a mutation that is carried by all the B2 subtypes. A1a, a subtype of A, possesses a mutation similar to type C that may merge all these previously reported genetic variations in one cluster. There is inadequate information about the A2a subtype of SARS-CoV-2 that had spread widely in March. The A2a genotype of SARS CoV-2 consists of a non-synonymous mutation located near the S1-S2 junction similar to the A2 subtype. This non-synonymous mutation could possibly impact viral entry into the host cell (Biswas and Majumder, 2020). Thus, A2a variants could be important genetic variants for the development of effective

vaccines and drugs against this virus. Further, sub-genotypes A3, A7, A1a, A2, and A6 have evolved from genotype A due to variation at the ORF1a, ORF3a, S, and nucleotide T514C respectively (Samaddar et al., 2020). Another group in India has examined 591 different novel coronaviruses and grouped them in five different clades. A total of 43% synonymous and 57% non-synonymous nucleotide substitutions were observed. The maximum number of non-synonymous substitutions was observed in the S protein (Saha et al., 2020). The presence of four SNPs at genomic positions 241, 3037, 144410, 23405 among 50–60% of the novel coronavirus population was deciphered by combining different bioinformatics (Tiwari and Mishra, 2020). However, epidemiological studies undertaken from time to time and surveillance of genetic variants among humans as well as in animals could be a major aid in the management of such outbreaks.

## **VIRAL MODE OF ACTION: IMMUNO-MODULATORY ACTION OF VIRAL PROTEINS**

Binding of SARS-CoV-2 spike (S) protein with host cell angiotensin-converting enzyme 2 (ACE2) aided by TMPRSS2 mediates viral entry. SARS-CoV-2 viral proteins (enlisted in **Table 2** and **Supplementary Table 5**) modulate host immune system and antagonize IFN response. COVID-19 pathophysiology is associated with aggressive pro-inflammatory responses (including IL-6, IL-1 $\beta$ , IP-10, macrophage inflammatory protein 1 $\alpha$  (MIP1 $\alpha$ ), MIP1 $\beta$  and MCP1) and airway damage. Disease severity depends on viral load and the host immune response. Severe COVID-19 patients exhibit high level of pro-inflammatory macrophages, neutrophils and monocytes, which contribute to the cytokine storm with very high plasma levels of TNF, IL-12, IL-6, IL-10, IL-7, G-CSF, IP-10, MIP1 $\alpha$ , and MCP1 (Chen et al., 2020; Liao et al., 2020; Zhou et al., 2020). Vigorous pro-inflammatory response leads to airway epithelial and endothelial cell apoptosis, respiratory microvascular damage, vascular leakage and edema, thereby causing hypoxia and compromising blood gas exchange, resulting in acute respiratory distress syndrome (ARDS) (Ye et al., 2020; Zhang B. et al., 2020). Activation of complement pathways has been associated with microvascular injury and thrombosis in severe COVID infection (Magro et al., 2020).

## **DIAGNOSTIC METHODS, THERAPEUTIC STRATEGIES AND GOVERNMENT INITIATIVES TO COMBAT COVID-19 IN INDIA**

### **COVID-19 Diagnosis in India**

Similarity in signs and symptoms with other respiratory infectious diseases (fever, chills, cough, and shortness of breath) put an extra burden on specialized COVID-19 diagnosis (Kaushik

et al., 2020). Clinical manifestation of COVID 19 patients vary day to day and asymptomatic carriers of SARS-CoV-2 pose a challenge to our diagnostic approaches. ICMR and WHO have categorized COVID-19 as mild, moderate, and severe<sup>10</sup> (Sivasankarapillai et al., 2020). Accurate and rapid diagnosis is needed to minimize substantial morbidity and mortality. Virus isolation, electron microscopy, genomic sequencing—the standard procedures for coronavirus diagnosis are time-consuming and costly. Thus, to examine a large number of patients, serological and laboratory-based methods such as CBC, AST, ALT, creatinine, LDH, ferritin examination, and molecular-based assays are being used on priority (Balachandar et al., 2020). India has set up several diagnostic and labs all over the country to test COVID-19 patients on the basis of qRT-PCR (Kaushik et al., 2020; Lamba, 2020). Diagnosis depends on several SARS-CoV-2 proteins, namely, spike (S), M, envelope (E), N, RdRp and ORF-1b-nsp14 (Alagarasu et al., 2020; Mourya et al., 2020). Initially, in India the first two SARS-COV-2 viruses were identified and confirmed by screening for viral genes (E, RdRp, and N protein of SARS-CoV-2) in 881 suspected cases by RT-PCR and next-generation sequencing (Yadav et al., 2020). The limited supply of positive controls has been overcome by the introduction of *in vitro* transcribed RNA from the National Institute of Virology (NIV) (Choudhary et al., 2020). SOPs for types of specimen collection and transportation were initially documented by ICMR-NIV (Gupta et al., 2020). To enhance the speed of detection, various rapid detection kits, CT scan and X-ray based techniques have been introduced from time to time. However, lack of accuracy of these techniques has prevented them from being used as standard procedures (Iyer et al., 2020). The production of IgG and IgM against COVID-19 takes 10–15 days from infection. This is a limitation for any antigen and antibody-based rapid detection kit (Hou et al., 2020). Recently, a CRISPR based fast and highly accurate diagnostic approach for COVID-19 has been introduced which employs nucleic acid readout of SARS-COV-2 (Lotfi and Rezaei, 2020). However, its implementation is highly challenging. CSIR-Institute of Genomics and Integrative Biology (CSIR-IGIB), India has also developed an efficient and accurate detection tool named Feluda based on CRISPR-Cas9 technology, as an alternative to current gold standard RT-PCR technique. Feluda has received approval from the DCGI for commercial launch<sup>11</sup>. In continuing efforts to discover a fast and rapid detection technique for SARS-CoV-2, an aptamer based assay has been developed at Translational Health Science and Technology Institute (THSTI). In this assay, nasal swab is used as the specimen for detection<sup>12</sup>. Gargle lavage sample collected from COVID-19 patients was identified as an easy, alternative showing comparable efficiency as nasopharyngeal and oropharyngeal swab samples (Mittal et al., 2020). Monitoring

<sup>10</sup><https://www.expresshealthcare.in/covid19-updates/icmr-revises-treatment-protocol-for-covid-19-patients/421792/>

<sup>11</sup><https://Science.TheWire.In/the-Sciences/Explained-Feluda-Covid-19-Test-India-Crispr-Technology/>

<sup>12</sup><https://Journalsdiary.Com/2020/07/12/Aptamer-Based-Assay-Developed-for-Coronavirus-Detection/>

**TABLE 2 |** Comparative list of SARS-CoV and SARS-CoV-2 viral proteins involved in modulating the host anti-viral immune response.

Sl.No.	Viral protein	Category	Host immuno modulating function of viral proteins	
			SARS-CoV	SARS-CoV-2
1.	M protein	Structural Protein (Important component of viral envelope)	Increased M protein expression is linked with RIG-I, TBK1, IKK $\epsilon$ , and TRAF3 and hence, prevention of IRF3 and IRF7 activation. This results in significant decrease in induction of the interferon- $\beta$ promoter by dsRNA (Weiss and Leibowitz, 2011).	–
2.	N protein	Structural Protein (Encodes for Nucleocapsid protein)	Overexpression is associated with decreased IFN response via inhibition of IRF3 and NF- $\kappa$ B responsive promoter mediated activation (Weiss and Leibowitz, 2011).	–
3.	Nsp1	Non-Structural Protein	Suppresses the activation of IRF3, c-Jun and NF- $\kappa$ B, thereby blocking the interferon response and subsequent activation of interferon-dependent anti-viral proteins (such as ISG15 and ISG56) (Weiss and Leibowitz, 2011).	Modulates and suppresses the host anti-viral immune response (Gordon et al., 2020).
4.	Nsp3	Non-Structural Protein	Serves as papain like protease with de-ubiquitinating activity; could act as Type I Interferon antagonist (Weiss and Leibowitz, 2011).	–
5.	Nsp13	Non-Structural Protein	–	Targets innate immune pathways such as Interferon (IFN) and NF- $\kappa$ B pathways (Azkur et al., 2020).
6.	Nsp15	Non-Structural Protein	–	Targets the Interferon (IFN) pathway (Azkur et al., 2020).
7.	ORF3a	Accessory Protein	Raises level of fibrinogen in lungs.  Activates the NLRP3 inflammasome (Chen et al., 2019).  Activates NF- $\kappa$ B and JNK which in turn leads to upregulated expression of pro-inflammatory cytokines (such as IL8 and RANTES) (Narayanan et al., 2008).  Induces increased apoptosis via caspase 8 and caspase 9-mediated pathways. Bax, p53 and p38 MAP kinase are also involved in ORF3a mediated apoptosis (McBride and Fielding, 2012; Chen et al., 2019).	Activates the NLRP3 inflammasome (Gordon et al., 2020).  Activates caspase-1.  Mediates IL-1 $\beta$ and IL-18 secretion (Chen et al., 2019; Gordon et al., 2020).
8.	ORF3b	Accessory Protein	Enhances the production of cytokines and chemokines by regulating the transcriptional activity of RUNX1b. Inhibits Type I interferon (IFN) production and signaling (Narayanan et al., 2008; McBride and Fielding, 2012).	IFN antagonist; regulates IRF3 activity (Gordon et al., 2020).
9.	ORF6	Accessory Protein	Promotes DNA synthesis.  Hampers Type I IFN production and signaling (Narayanan et al., 2008; McBride and Fielding, 2012).	Serves as a Type I Interferon (IFN) antagonist (Gordon et al., 2020).
10.	ORF7a	Accessory Protein	Triggers inflammatory response through activation of NF- $\kappa$ B and IL8 promoter region (Narayanan et al., 2008). Promotes pro-inflammatory cytokines (such as IL8 and RANTES) production (McBride and Fielding, 2012).	Mediates virus induced apoptosis (Gordon et al., 2020).
11.	ORF8b	Accessory Protein	Blocks the IFN- $\beta$ signaling pathway by ubiquitin-proteasome mediated degradation of IRF3 (Wong et al., 2018).	–
12.	Orf9b	Accessory Protein	–	Serves as a Type I Interferon (IFN) antagonist (Azkur et al., 2020).
13.	Orf9c	Accessory Protein	–	Targets the NF- $\kappa$ B pathway and hence the anti-viral innate immune response (Azkur et al., 2020).

of patients before and after recovery through epidemiological and immunological assays in the large cohort would help understanding the prognosis and pathogenesis of COVID-19 and shall also aid in preventing community transmission and post recovery complications.

## Detection Equipment

Standard diagnosis for infection requires real-time thermal cyclers which are used to perform RT-PCR, a robust and reliable detection technology (Corman et al., 2020). Technology centres under MSMEs began manufacturing components of Real Time

Quantitative Micro PCR System in order to assemble the devices at a manufacturing unit in Visakhapatnam to ramp up the testing procedure<sup>13</sup>. Apart from RT-PCR based testing, other approaches have also been demonstrated which involve two-step detection methods involving more affordable thermal cyclers (conventional PCR) and fluorescence spectrometers<sup>14</sup>.

## The Treatment of SARS-CoV-2 Infection and COVID-19: The Present Scenario

The SARS-CoV-2 infection and the COVID-19 pandemic have posed an unprecedented challenge to the medical fraternity. The treatment is restricted to the best supportive care and experimental medications. Targeting the viral entry, interaction of the virus with its host and the downstream signaling pathways using novel or repurposed drugs, is one of the strategies for the management of COVID-19. Several agents (enlisted in **Table 3** and **Supplementary Table 6**) have been tried based on their role in similar viral infections, or their prospective action on the novel corona virus.

Indian Pharmaceuticals Cadila has tested the immunomodulator drug named Sepsivac (containing heat-killed *Mycobacterium W* (Mw)), on COVID-19 patients at PGIMER, Chandigarh in partnership with the Council of Scientific and Industrial Research (CSIR) to reduce the mortality of critically ill COVID-19 patients and have obtained promising results<sup>15</sup>.

Apart from these drugs, the US FDA has approved use of convalescent plasma for severe life-threatening COVID infection as an investigational new drug (Duan et al., 2020). Its use has been documented in a series of cases (Huang et al., 2020; Zeng et al., 2020)<sup>16</sup>. One small trial with five ventilated patients showed success. Its role is still not clear and US FDA is facilitating the use of hyperimmune globulin for COVID treatment (Mehta et al., 2020). US FDA recommended the use of convalescent plasma for emerging infections including COVID-19 on May 1, 2020 (see text footnote 2). The Indian Council of Medical Research (ICMR) began clinical trials with convalescent plasma in India to evaluate its safety and efficiency in controlling COVID-19 symptoms<sup>17,18</sup>. ICMR has recommended use of convalescent plasma for COVID-19 therapy. A plasma bank has been established in Delhi and Project PLATINA has been established in Maharashtra for treatment cum trial with plasma therapy<sup>19</sup>.

Another approach for developing drugs targeting host immunity has been to express SARS-CoV-2 proteins in human cell lines and identify their human protein interacting partners. Of 332 interactions, 66 human proteins were found as druggable candidates that could be targeted by 29 FDA approved drugs, 12 compounds in clinical trials and 28 compounds in preclinical stage (Gordon et al., 2020). Further screening has helped in the identification of two pharmacological candidates that inhibit mRNA translation and are predicted to regulate Sigma1 and Sigma2 receptors. Besides, inhibitors targeting endocytosis have shown activity in vitro against other coronaviruses such as SARS CoV and MERS-CoV. These include chlorpromazine, ouabain and bufalin (de Wilde et al., 2014; Burkard et al., 2015). Their efficacy against SARS CoV-2 is yet to be tested. However, very high EC<sub>50</sub>/C<sub>max</sub> (half-maximal effective concentration value/peak serum concentration level) ratio at the typical dosages used is limiting their possible clinical use.

Natural killer cells play a role in the clearance of SARS-CoV. NK cell based products are in various stages of trial as anti-COVID-19 agents. The US-based Company Celularity has developed placenta derived NK cells CYNK-001 (Tu et al., 2020). Recombinant Interferon Type I exhibits broad spectrum activity against coronaviruses (Cinatl et al., 2003b; Sheahan et al., 2020). Clinical trials are currently in motion for the treatment of COVID-19 pneumonia (NCT04293887). Trials are also ongoing to test the efficacy of mesenchymal stem cells from the umbilical cord and dental pulp to attenuate the inflammatory response of COVID-19 (NCT04293692, NCT04269525, NCT04288102, NCT04302519). The World Health Organization's (WHO) Solidarity trial including randomized and controlled clinical trials are set to test several protocols against COVID-19.

## COVID-19 Vaccine Developments – Present Indian Scenario

In the global fight against COVID-19, scientists from different countries are trying to decipher a potential therapeutic drug, vaccine, and early diagnostic tools. The SARS-CoV-2 'S' protein interacts with the ACE2 receptor and is a glycosylated protein, making this protein a good candidate for vaccine development (Othman et al., 2020). Globally several vaccine generation methods are being used against COVID-19, including a live attenuated vaccine, inactivated vaccine, replicating viral vector, non-replicating viral vector, DNA vaccine, peptide-based vaccine, recombinant protein, virus-like particle (VLP) and mRNA-based vaccine (Le et al., 2020). According to the WHO, there are currently 63 COVID-19 vaccines in clinical development and 172 vaccine candidates in pre clinical developmental stage as on 6th January, 2021 (see text footnote 3). Out of these 63 vaccines, about 20 vaccine candidates are in Phase III clinical trial (as enlisted in **Table 4**). Among these 20 vaccines, the efficacy report is available for five vaccines that include "BNT162 (Pfizer), mRNA 1273 (Moderna), chAdOX1nCOV19 (University of Oxford and AstraZeneca), BBIBP-CorV (Sinopharm) and Sputnik-V (Gamaleya Research Institute)<sup>20</sup>. However, only

<sup>13</sup><https://Pib.Gov.In/Pressreleasepage.aspx?Prid=1623027>

<sup>14</sup><https://Www.Hindustantimes.Com/India-News/Iisc-Comes-up-with-an-Affordable-Two-Step-Method-to-Scale-up-Rt-Pcr-Testing/Story-Xbztjylgeldprouohp6o.html>

<sup>15</sup><https://Www.Hindustantimes.Com/Health/Indian-Trials-on-Multiple-Covid-19-Drugs-Make-Progress-Have-Atmanirbhar-Bharat-Tilt/Story-Nk0owrrrsyragqhvrk2a9i.html>

<sup>16</sup><https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-coordinates-national-effort-develop-blood-related-therapies-covid-19>

<sup>17</sup><https://Www.Indialegallive.Com/Special-Story/Convalescent-Plasma-Therapy-a-Treatment-for-Covid-19/>

<sup>18</sup><https://Www.Medrxiv.Org/Content/10.1101/2020.09.03.20187252v2>

<sup>19</sup><https://Swachhindia.Ndtv.Com/Coronavirus-Outbreak-Explained-What-Is-Convalescent-Plasma-Therapy-and-How-Effective-Is-It-in-Treating-Covid-19-Patients-46568/>

<sup>20</sup>[https://Www.TheLancet.Com/Journals/Lanmic/Article/Piis2666-5247\(20\)30226-3/Fulltext](https://Www.TheLancet.Com/Journals/Lanmic/Article/Piis2666-5247(20)30226-3/Fulltext)



**TABLE 3 |** List of immuno-modulating drugs which are tried for COVID-19.

Sl.No.	Name of drugs	Type of anti viral or immune boosting action
1.	Chloroquine and Hydroxychloroquine (HCQS)	<p>May keep the virus out of host cells by blocking host receptor glycosylation or by breaking down viral protein production.</p> <p>May lead to suppression of pH-dependent steps of viral replication (Choudhary and Sharma, 2020).</p> <p>May exert immune-modulatory effects by inhibiting TNF-<math>\alpha</math> and IL6 production and may serve as a potent autophagy inhibitor.</p> <p>Active against SARS-CoV-2 in vitro (Yao et al., 2020).</p> <p>Leads to fast symptomatic improvement (fever, cough and chest imaging) (Cortegiani et al., 2020)</p> <p>HCQS and azithromycin combination leads to early viral clearance compared to HCQS alone (Gautret et al., 2020).</p> <p>US-FDA have cautioned against the use of HCQS for COVID-19 outside hospital settings <a href="https://MedicalDialogues.In/Medicine/News/Fda-Cautions-against-Use-of-Chloroquine-or-Hcqs-in-Covid-19-65165">https://MedicalDialogues.In/Medicine/News/Fda-Cautions-against-Use-of-Chloroquine-or-Hcqs-in-Covid-19-65165</a> (accessed on May 5, 2020)..</p>
2.	Corticosteroids	<p>Exert immune-modulatory effects by inhibiting expression of genes encoding inflammatory molecules (Shaffer, 2020).</p> <p>Dexamethasone proven to be a life saving drug for severe COVID-19.</p>
3.	Tocilizumab, Sarilumab and Situximab	<p>Monoclonal antibody (MAB) antagonists of the IL6 receptor.</p> <p>Drugs commonly used for treatment of rheumatoid arthritis.</p> <p>Severe forms of COVID-19 are associated with elevated levels of IL6, causing <u>acute respiratory distress syndrome (ARDS)</u> even upon reduction of viral load.</p> <p>These MABs may play a vital role in reducing IL6 level and reduce instances of ARDS in COVID-19 patients (Chan et al., 2013; Luo et al., 2020; Michot et al., 2020; Shaffer, 2020).</p>
4.	Fluvoxamine	This serotonin re-uptake inhibitor may serve as an immune modulatory agent and shut down the inflammatory cascade from the endoplasmic reticulum by binding to the sigma-1 receptor (Shaffer, 2020).
5.	Remdesivir	<p>Antiviral pro drug.</p> <p>The active analog of the pro drug inhibits the viral RNA dependent RNA polymerase (RdRp) and preventing viral replication.</p> <p>Remdesivir also evades the proofreading mechanism (exoribonuclease) of coronavirus (Ferner and Aronson, 2020; Wang M. et al., 2020) <a href="https://www.fda.gov/media/137564/download">https://www.fda.gov/media/137564/download</a> (downloaded on May 5, 2020)..</p>
6.	Azithromycin	<p>Broad spectrum macrolide antibiotic.</p> <p>Used mainly for treatment of pulmonary, enteric and genitourinary tract infections. Acts as an acidotropic lipophilic weak base which modifies the pH of the endosome and trans-Golgi network and affects viral replication.</p> <p>Interferes with viral entry by binding to viral spike (S) protein and human receptor protein ACE2 (angiotensin converting enzyme-2).</p> <p>May exert interferon mediated anti viral immune response (Choudhary and Sharma, 2020; Damle et al., 2020).</p>
7.	Baricitinib, Fedratinib, and Ruxolitinib	<p>Potent JAK inhibitors selectively inhibiting JAK-STAT signaling <a href="https://www.clinicaltrials.gov/ct2/show/study?term=baricitinib&amp;rank=1">https://www.clinicaltrials.gov/ct2/show/study?term=baricitinib&amp;rank=1</a> (Spinelli et al., 2020; Stebbing et al., 2020).</p> <p>Exerts anti-inflammatory effects.</p> <p>Might be effective in controlling the cytokine storm in COVID-19.</p> <p>Baricitinib is also predicted to hamper ACE2 mediated endocytosis (Richardson et al., 2020).</p>
8.	Gimsilumab, Lenzilumab, Namilumab	<p>These are anti-granulocyte-macrophage colony-stimulating factor (GM-CSF) antibodies. Blocks the inflammatory pathway in its early steps.</p> <p>Being clinically tested for efficacy in COVID-19 (Tay et al., 2020).</p>
9.	Thalidomide	<p>Synthetic glutamic acid derivative.</p> <p>Possess anti-inflammatory, anti-fibrotic, anti-angiogenesis, and immuno-modulatory effects.</p> <p>Inhibits and downregulates COX2, PGE2, TNF-<math>\alpha</math>, IL6 and IL1.</p> <p>Used to treat severe H1N1 influenza-associated lung injury.</p> <p>Being tested for its efficacy in treating cytokine storm and reducing lung injury and respiratory complications in COVID-19 <a href="https://clinicaltrials.gov/ct2/show/study?term=thalidomide&amp;rank=1">https://clinicaltrials.gov/ct2/show/study?term=thalidomide&amp;rank=1</a> (Khalil et al., 2020).</p>
10.	Nafamostat and Camostat	Serine protease inhibitors which prevent SARS-CoV-2 entry by acting as antagonists to the serine protease TMPRSS2 (Yamamoto et al., 2016; Hoffmann et al., 2020; Zhang H. et al., 2020).
11.	Famotidine	H2 receptor antagonist; may bind to SARS-CoV-2 encoded papain like protease and impair entry of SARS-CoV-2 (Shaffer, 2020).
12.	Ivermectin	Broad spectrum anti-parasitic macrolide drug.
13.	Favipiravir	<p>Functions by binding and impairing the cell transport proteins that are vital for entry into the nucleus (Choudhary and Sharma, 2020).</p> <p>Inhibits virus replication by binding and blocking the RdRp enzyme (Furuta et al., 2013).</p> <p>Its incorporation in RNA also terminates viral protein synthesis (Jin et al., 2013).</p> <p>Classically used against influenza virus.</p> <p>Also acts on SARS-CoV-2 replication; used for mild and moderate COVID cases (Agrawal et al., 2020).</p>

(Continued)

TABLE 3 | Continued

Sl.No.	Name of drugs	Type of anti viral or immune boosting action
14.	Lopinavir/ Ritonavir	Antiretroviral protease inhibitors, successfully used in HIV infection (Huang et al., 2015). Combination of lopinavir/ritonavir used successfully for treatment of SARS with significantly fewer adverse clinical outcomes (Chu et al., 2004). Lopinavir/ Ritonavir with Interferon 1b found promising in the marmoset model (Chan et al., 2015).
15.	APN01	Soluble form of ACE2 delivered in high concentrations. Could potentially block SARS-CoV-2 entry into target cells. Under clinical trial <a href="https://Pipelinereview.Com/Index.Php/2020022673884/Proteins-and-Peptides/Apeirons-Respiratory-Drug-Product-to-Start-Pilot-Clinical-Trial-to-Treat-Coronavirus-Disease-Covid-19-in-China.html">https://Pipelinereview.Com/Index.Php/2020022673884/Proteins-and-Peptides/Apeirons-Respiratory-Drug-Product-to-Start-Pilot-Clinical-Trial-to-Treat-Coronavirus-Disease-Covid-19-in-China.html</a> .

three vaccines are available with the data published in peer reviewed journals till 5th January, 2021 namely, mRNA1273, BNT162, and chAdOX1nCOV19 (Baden et al., 2020; Polack et al., 2020; Voysey et al., 2020). Supporting data to answer such important question such as duration of herd immunity upon vaccination, requirement of booster doses for long term immunity and whether vaccine could help in the prevention of transmission is available for only chAdOX1nCOV19 vaccine till to date. Further, safety of the above mentioned vaccines needs to be evaluated in the populations that have not been included in the trials such as pregnant women (see text footnote 21).

Apart from this, global mass immunization also encountered several challenges including financial, logistic and vaccine storage-related issues. The upper middle income countries started the vaccination in 2020. However, successful global vaccination or complete eradication of virus is possible only when low income and middle income countries get immunized in parallel. To overcome this situation, an International initiative termed COVAX facility has been set up to ensure equitable access to vaccine doses in Low income and middle income countries (LCMICs). COVAX aims at fixed vaccination for 20% of population belonging to the LCMICs by 2021. The vaccine will be provided by the AstraZeneca<sup>21</sup>.

In India, COVAXIN, an indigenous inactivated COVID-19 vaccine, stable at 2–8°C, manufactured by Bharat Biotech (Hyderabad, India) has currently entered the Phase III Human clinical trial and has recently been given emergency approval in India by the DCGI (see text footnote 7)<sup>22</sup>. A plasmid DNA based vaccine, ZyCoV-D has been developed by Ahmedabad-based pharma company Cadila Healthcare (Zydus Cadila). It has been claimed that this vaccine is stable for 3 months at a temperature of 30°C and longer at 2–8°C. This thermostability could be beneficial for nationwide vaccination program due to minimalistic cold storage requirements. Phase III human clinical trials are being initiated for this vaccine. Besides these indigenous vaccines, several non indigenous vaccines of foreign origin are presently in various stages of clinical trial in India. Covishield, the vaccine developed by Oxford University and AstraZeneca has entered phase III trials in collaboration with

the Serum Institute, Pune, India. Serum Institute has applied to DCGI for emergency regulatory authorization for Covishield use in India and has submitted additional requisite vaccine datasheet in this regard. This vaccine has been approved for emergency use in United Kingdom and has become the first COVID-19 vaccine candidate to have obtained emergency approval in India (see text footnote 4)<sup>23</sup>. Covishield has the advantage of storage at 2–8°C<sup>24</sup>. Besides this, Dr. Reddy's Laboratories Limited and Sputnik LLC (Russia) have been jointly conducting the clinical trial of Sputnik-V, the world's first registered vaccine in India. This vaccine, ranking among world's top 10 vaccine candidates is presently in Phase II Human Clinical Trial in India<sup>25</sup>. The Biological E's novel Covid-19 vaccine is also in the Phase I/II Human Clinical Trial in India (see text footnote 5). Ecological studies have highlighted lower number of infections and reduced COVID-19 mortality in countries, where BCG vaccination is made mandatory (Urashima et al., 2020). Randomized controlled trials of BCG-Danish have been conducted in Netherlands and Australia (NCT04327206, NCT04328441). Serum Institute, Pune, India has conducted phase III trial of BCG vaccine VPM1002 to evaluate cross-protection to COVID-19<sup>26</sup>. BCG vaccine could serve as a booster of innate immunity against COVID-19 via metabolic and epigenetic changes in a process called trained immunity (Netea et al., 2020).

Another important vaccine, BNT162 from Pfizer, which has already been rolled out in United Kingdom, United States and received emergency use approval in more than 10 countries, has extreme cold chain and storage requirement at –75°C to keep its potency intact<sup>27</sup>. Similarly, Moderna vaccine also has stringent storage requirement at –20°C<sup>28</sup>. Such stringent refrigeration needs may be difficult to achieve in developing countries and may render mass vaccination in India extremely challenging. Although India has cold storage facilities, they

<sup>21</sup>[https://Www.TheLancet.Com/Pdfs/Journals/Lanmic/Piis2666-5247\(20\)30226-3.Pdf](https://Www.TheLancet.Com/Pdfs/Journals/Lanmic/Piis2666-5247(20)30226-3.Pdf)

<sup>22</sup><https://Www.Indiatoday.In/News-Analysis/Story/Why-Covaxin-Covishield-Best-Option-for-India-against-Covid19-Pandemic-1755517-2021-01-03>

<sup>23</sup><https://Www.Thehindu.Com/Sci-Tech/Health/Uk-Becomes-First-Country-to-Approve-Astrazeneca-Oxford-Vaccine-for-Covid-19/Article33451081.Ece>

<sup>24</sup><https://Www.Businesstoday.In/Coronavirus/Oxford-Serum-Institute-Vaccine-Stored-Fridge-Temperature-2-8-Degrees/Story/422741.html>

<sup>25</sup><https://sputnikvaccine.com/about-vaccine/>

<sup>26</sup><https://health.economictimes.indiatimes.com/news/pharma/serum-institute-conducting-phase-iii-clinical-trial-of-tuberculosis-vaccine-dbt/77210270>

<sup>27</sup>[https://www.pfizer.com/news/hot-topics/covid\\_19\\_vaccine\\_u\\_s\\_distribution\\_fact\\_sheet](https://www.pfizer.com/news/hot-topics/covid_19_vaccine_u_s_distribution_fact_sheet)

<sup>28</sup><https://indianexpress.com/article/explained/covid-19-vaccine-storage-optimal-temperature-cold-chain-india-explained-quixplained-7063369/>

**TABLE 4 |** List of vaccines in Phase III trial across the world.

Sl.No.	Name of vaccine	Nature of vaccine	Clinical Trial Phase	Country of origin
1	Ad5-nCoV	Recombinant vaccine (adenovirus type 5 vector)	Phase III	CanSino Biologics Inc/ Beijing Institute of Biotechnology (China)
2	Covishield (Code name: AZD1222)	Replication-deficient viral vector vaccine (adenovirus from chimpanzees)	Phase III (received approval for emergency use in United Kingdom and India)	The University of Oxford; AstraZeneca; IQVIA; Serum Institute of India (Multinational) (see text footnote 4) <a href="https://www.bbc.com/news/health-55280671">https://www.bbc.com/news/health-55280671</a>
3	CoronaVac	Inactivated viral vaccine (formalin with alum adjuvant)	Phase III	Sinovac (China)
4	COVAXIN	Inactivated viral vaccine	Phase III (approved for emergency use in India)	Bharat Biotech; National Institute of Virology (India) (see text footnote 7)
5	JNJ-78436735 (formerly Ad26.COV2-S)	Non-replicating viral vector	Phase III	Johnson & Johnson (Janssen Pharmaceutical) (United States)
6	mRNA-1273	mRNA-based vaccine	Phase III (received approval and presently in use in United States)	Moderna; National Institute of Allergy and Infectious Diseases (NIAID) (United States) <a href="https://www.nature.com/articles/D41586-020-03593-7">https://www.nature.com/articles/D41586-020-03593-7</a>
7	New Crown COVID-19 Vaccine	Inactivated vaccine	Phase III	Wuhan Institute of Biological Products; China National Pharmaceutical Group (Sinopharm, China) <a href="https://www.precisionvaccinations.com/vaccines/new-crown-covid-19-vaccine">https://www.precisionvaccinations.com/vaccines/new-crown-covid-19-vaccine</a>
8	NVX-CoV2373	Protein based vaccine (Full length recombinant SARS CoV-2 spike protein nanoparticle vaccine adjuvanted with Matrix M)	Phase III	Novavax (Maryland); Serum Institute of India <a href="https://lr.novavax.com/news-releases/news-release-details/novavax-announces-covid-19-vaccine-clinical-development-progress">https://lr.novavax.com/news-releases/news-release-details/novavax-announces-covid-19-vaccine-clinical-development-progress</a> <a href="https://www.verywellhealth.com/novavax-covid-19-vaccine-5093292">https://www.verywellhealth.com/novavax-covid-19-vaccine-5093292</a>
9	BNT162 (3 LNP-mRNAs)	mRNA-based vaccine	Phase II/III (Already in use in United Kingdom and United States)	Pfizer; BioNTech; Fosun Pharma; Jiangsu Provincial Center for Disease Prevention and Control (Multinational) <a href="https://www.thehindu.com/news/international/uk-approves-pfizer-biontech-covid-19-vaccine-for-use/article33228634.ece">https://www.thehindu.com/news/international/uk-approves-pfizer-biontech-covid-19-vaccine-for-use/article33228634.ece</a> ( <a href="https://www.nature.com/articles/d41586-020-03593-7">https://www.nature.com/articles/d41586-020-03593-7</a> ) Gamaleya Research Institute; Health Ministry of the Russian Federation (Russia)
10	Sputnik-V Vaccine (rAd26-S+rAd5-S)	Adeno viral vector based technology	Phase III	
11	BBIBP-CorV	Inactivated viral vaccine	Phase III	Sinopharm + Beijing Institute of Biological Products (China) (Xia et al., 2021)
12		Recombinant SARS-CoV-2 vaccine	Phase III	Anhui Zhifei Longcom Biopharmaceutical; Institute of Microbiology, Chinese Academy of Sciences (China)
13	INO-4800	DNA based vaccine	Phase II/III	Inovio Pharmaceuticals and International Vaccine Institute (South Korea)
14	CoVLP	Coronavirus-Like Particle based vaccine	Phase II/III	Medicago Inc. (Canada)
15	CVnCoV	RNA based vaccine	Phase II/III	CureVac AG (Germany)
16	UB-612	Multitope peptide based S1-RBD-protein based vaccine	Phase II/III	COVAXX; United Biomedical Inc
17	ZyCoV-D nCov vaccine	DNA based vaccine	Phase III	Cadila Healthcare Ltd. Zydus Cadila, (India) <a href="https://economictimes.indiatimes.com/markets/stocks/news/cadila-healthcare-gains-3-as-dcgi-plays-phase-iii-trials-of-covid-vaccine/articleshow/80091363.cms">https://economictimes.indiatimes.com/markets/stocks/news/cadila-healthcare-gains-3-as-dcgi-plays-phase-iii-trials-of-covid-vaccine/articleshow/80091363.cms</a>
18	QazCovid-in	Inactivated viral vaccine	Phase III	Research Institute for Biological Safety Problems (Rep of Kazakhstan)
19	SARS-CoV-2 vaccine (vero cell)	Inactivated viral vaccine	Phase III	Institute of Medical Biology; Chinese Academy of Medical Sciences (China)
20	AG0301-COVID19	DNA based vaccine	Phase II/III	AnGes + Takara Bio + Osaka University (Japan)

Vaccine Information obtained from World Health Organization (WHO) as on 06.01.2021 (see text footnote 3).

are limited and there are constraints especially in the case of handling large numbers of doses in a country as densely populated as India. Ramping up of cold chain and restructuring of cold storage facilities with synergistic aid from food storage and supply cold chain, may aid in overcoming vaccine storage issues to some extent<sup>29</sup>. Another important factor involved in mass vaccination is the economic burden. The COVAX facility (led by WHO, CEPI, and GAVI) have emerged to financially support and enable equitable distribution of COVID-19 vaccines across the world<sup>30</sup>. The Government of India (GOI) has also taken initiatives to bear the entire cost of vaccination and ensure mass immunization at nominal price<sup>31</sup>. The Global Alliance for Vaccines and Immunizations (GAVI) has estimated an expenditure of \$1.4 billion to \$1.8 billion on part of India (the second most populous country after China) for the first phase vaccination, even after support from the COVAX facility<sup>32</sup>. Moderna vaccine, apart from its cold requirement, is highly priced (at \$10–450 per dose), which might be difficult to cater to the Indian population. However, aid from COVAX alliance and Government may help Moderna reduce its cost for India. Covishield and COVAXIN are reasonably priced with respect to the Indian scenario. Covishield have been priced at \$3 per dose for government and approximately, \$10 for private entities. Because of their local origin and normal refrigeration temperatures, these two vaccines will be easy for handling and supply chain distribution in India<sup>33, 34</sup>.

COVID-19 mass vaccination drive in India shall soon be initiated with 30 crore people receiving the vaccines in the first phase. Healthcare workers, frontline workers and individuals aged above 50 will be vaccinated first according to the recommendations of the National Expert Group on Vaccine Administration for COVID-19 (NEGVAC). In this regard, the COVID Vaccine Intelligence Network (Co-WIN) system has been developed as a digital platform for registration and real time monitoring of vaccination to pre registered individuals in India<sup>35</sup>.

## Neutralizing Antibodies: Another Approach

Neutralization of the virus by antibodies is an important strategy for containing SARS-CoV-2. In SARS-CoV, the RBD122

(amino acids 318 to 510) of the S protein is primarily being targeted by neutralizing antibodies (Wong et al., 2004). The RBD of SARS-CoV and SARS CoV-2 are poorly conserved, so the majority of the monoclonal antibodies to SARS-CoV do not bind with or neutralize SARS CoV-2 (Wang et al., 2020a). Therapeutic monoclonal antibodies to SARS CoV-2 are being developed with the aid of phage library display, cloning of human B cell sequences from recovering patients and mouse immunization and hybridoma isolation. Anaïve semi synthetic library has been used to identify the anti-SARS-CoV-2 RBD human monoclonal antibody. This approach holds promise since the entire RBD remains conserved as of now (Parrray et al., 2020). However, caution must be exercised, since animal studies of SARS CoV infection show that neutralizing antibodies to S protein may increase lung injury by aggravating inflammatory responses (Liu et al., 2019). Anti-S-IgG mediated proinflammatory responses occur due to binding of virus-anti-S-IgG complex with the Fc receptors (FcR) present on monocytes and macrophages (Liu et al., 2019). In addition, virus-anti-S-IgG complex may trigger the classical complement pathway leading to cellular damage.

## Indian Government Initiatives and Strategies to Combat COVID-19 Personal Protective Equipment

Personal Protective Equipment (PPE) including face piece respirators, gloves, shoe covers and face shields are necessary for the protection of health workers from infection<sup>36</sup>. N95 respirators, surgical masks or cloth masks are recommended to prevent respiratory transmission. Cloth masks may possibly be cost-effective in preventing community transmission in densely populated Asian countries (Sra et al., 2020). Unlike N95 respirators, simple cloth and surgical masks are non-disposable and can be potentially decontaminated routinely using alcohol/detergent washing, and moist heat treatment (Viscusi et al., 2009). To prevent contact transmission, disposable gloves are recommended for patient examination. Government of India is funding enterprises and enabling transfer of advanced technology for increased PPE production<sup>37</sup>. However, supply of raw materials may be dependent on import and could be a bottle neck for large scale production in India (Feinmann, 2020).

## Disinfection Instruments

COVID-19 may potentially remain transmissible on inanimate surfaces up to several days. Effective disinfection could be achieved using biocidal chemicals such as 70% ethanol, 0.1% aqueous sodium hypochlorite and 0.5% hydrogen peroxide solutions (Kampf, 2020; Kampf et al., 2020). 60–70% ethanol is recommended for sterilizing high-end biomedical equipment, while 0.1% aqueous sodium hypochlorite could be a viable solution for decontamination of large areas such as mass transit systems, hospital outdoors etc. Scientists from the

<sup>29</sup><https://Health.Economictimes.Indiatimes.Com/News/Pharma/Ramping-up-Cold-Storage-Facilities-Critical-as-India-Preps-for-Covid-19-Vaccine-Experts/78550153>

<sup>30</sup><https://Www.Who.Int/Emergencies/Diseases/Novel-Coronavirus-2019/Covid-19-Vaccines>

<sup>31</sup><https://Www.Livemint.Com/Budget/News/Govt-Will-Bear-the-Entire-Cost-of-Covid-19-Vaccination-Gulera-11608563339012.html>

<sup>32</sup><https://Science.Thewire.In/Health/India-Covid-19-Vaccine-First-Phase-1-8-Billion/>

<sup>33</sup><https://indianexpress.com/article/explained/coronavirus-vaccines-india-covishield-bharat-biotech-covaxin-7131057/>

<sup>34</sup><https://Timesofindia.Indiatimes.Com/Life-Style/Health-Fitness/Health-News/Coronavirus-Vaccine-Can-India-Get-Its-Hands-on-Modernas-Covid-19-Vaccine-3-Challenges-We-Have/Photostory/79411299.Cms?Picid=79411390>

<sup>35</sup><https://Www.Livemint.Com/News/India/India-S-Covid-19-Vaccination-Drive-to-Start-Soon-Registrations-Details-to-Guidelines-All-You-Need-to-Know-11608363252706.html>

<sup>36</sup><https://www.health.state.mn.us/facilities/patientsafety/infectioncontrol/ppe/index.html>

<sup>37</sup><https://www.investindia.gov.in/siru/personal-protective-equipment-india-INR-7000-cr-industry-in-the-making>



Council of Scientific and Industrial Research (CSIR), India have claimed to develop a spraying procedure by using induction charged electrostatic spraying apparatus involving lower amounts of chemicals, charge based disinfection and large coverage in comparison with conventional high-volume sprayers<sup>38,39</sup> (Lyons et al., 2011). In line with other countries, drone-based disinfection methods have been proposed by Indian enterprises<sup>40</sup>. Concern about the potential hazards of inhaling the aerosolized disinfectants still poses a challenge for large area disinfection (Kim et al., 2020).

### Biomedical Equipment

Various medical equipment such as ventilators, sensor equipments including pulse-oximeter, infrared thermometer, multi parametric photo plethysmography (PPG) sensor, portable X-ray machine, fiberoptic bronchoscopes, video laryngoscopes, are required in monitoring and treatment of COVID-19 patients (Wax and Christian, 2020). The ventilator is a crucial equipment for critically ill patients with respiratory problems (Iyengar et al., 2020). Ventilators are costly (~\$30,000) and there is a world-wide shortage of ventilators during the pandemic. India alone has a requirement for tens of thousands of ventilators<sup>41</sup>. There is global endeavor to enhance production, lower cost and find alternatives. Engineers from Rail Coach Factory have claimed production of a low-cost prototype ventilator<sup>42</sup>. Scientists at the CSIR laboratories are also developing 3D printed automatic ventilators and mechanical ventilators<sup>43</sup> (Iyengar et al., 2020). An alternative for the ventilator, “Artificial Manual Breathing Unit (AMBU)” has been designed by researchers from the Postgraduate Institute of Medical Education and Research, Chandigarh (Iyengar et al., 2020). Recently, an Indian manufacturer has reported production of state-of-the-art ventilator costing less than \$2000 (Agrawal, 2020).

### Indigenous Medicinal Plants for Combating COVID-19

Antiviral herbal therapy has made enormous progress in the past decade (Dhama et al., 2018). Various medicinal plants and bioactive phyto-metabolites have been widely explored for effective control of several viral diseases such as influenza, hepatitis, human immunodeficiency virus (HIV), herpes simplex virus (HSV) and coxsackievirus infections (Akram et al., 2018). India harbors a diverse variety of medicinal plants and herbs with therapeutic potential (Mohanraj et al., 2018). The major indigenous medicinal plants with immuno-modulatory properties, which can potentially be explored for their role

in boosting immunity and rendering protection from SARS-CoV-2 infection, have been summarized in **Table 5** and **Supplementary Table 7**.

The Ministry of Ayush under the Govt. of India has recommended use of indigenous herbal plants and spices namely, tulsi, cinnamon, dry ginger, black pepper, turmeric, coriander, cumin and garlic for enhancing immunity<sup>44</sup>. Besides, the Ministry of Ayush has formulated a collection of four ayurvedic herbs namely, ashwagandha, guduchi, yasthimadhu, peepli; and a drug named Ayush 64 for combating COVID-19. The Ministry of Ayush along with the CSIR have initiated the process of validating the efficacy of these formulations against COVID-19 in the month of May, 2020 and the outcomes of these trials are expected to be available soon<sup>45, 46, 47</sup>.

### Artificial Intelligence in Combating COVID-19

The worldwide outbreak of SARS-CoV-2 has resulted in a tremendous dearth of clinical equipment. In order to contain the pandemic effectively, large scale testing and diagnosis are required. This is evident from the successful containment of SARS-CoV-2 virus in countries that have been able to perform mass testing of possibly infected people and contact tracing. RT-PCR serves as the gold standard test for validating SARS-CoV-2 infection. Inadequate testing capability in most countries, along with the high dependency of the RT-PCR test on the swab technique, has spurred the need to search for alternative methods that allow COVID-19 diagnosis.

#### CT scan in COVID-19

The chest X-ray and thoracic computed tomography (CT) are examples of easily accessible medical imaging equipment, which assists clinicians in diagnosis. CT images may serve as a visual indicator of coronavirus infection for radiologists (Duncan and Ayache, 2000). While RT-PCR may take up to 24 h and needs multiple tests for conclusive results, chest CT combined with certain health symptoms can be used as an effective diagnostic tool in clinical practice for rapid screening of COVID-19 patients. There is a high chance that COVID-19 patients can be diagnosed accurately by using chest radiography images (van Ginneken et al., 2001; Sluimer et al., 2006). However, manual examination of CT scans for COVID-19 diagnosis is a labor-intensive and time-taking process. Besides, clinical presentation of COVID-19 in CT images is similar to other forms of viral pneumonia, which makes diagnosis even more difficult. A dependable computer-aided diagnostic system for COVID-19

<sup>38</sup><https://www.igi-global.com/Chapter/Fundamentals-of-Electrostatic-Spraying/232957>

<sup>39</sup><https://www.tribuneindia.com/News/Nation/Csio-Develops-Electrostatic-Disinfection-Technology-to-Combat-Covid-78098>

<sup>40</sup><https://pib.gov.in/PressReleasePage.aspx?Prid=1620351>

<sup>41</sup><https://www.medrxiv.org/content/10.1101/2020.03.26.20044511v1.Full.Pdf>

<sup>42</sup><https://www.tribuneindia.com/News/Nation/Rail-Coach-Factory-Kapurthala-Develops-Ventilator-66118>

<sup>43</sup><https://www.csir.res.in/Csir-Labs-Initiatives-against-Covid-19>

<sup>44</sup><https://www.ayush.gov.in/>

<sup>45</sup><https://newsonair.nic.in/News?Title=Ministry-of-Ayush%2c-Csir-Working-Together-on-Validating-Four-Ayush-Formulations-against-Covid-19&Id=388575> (accessed on June 22, 2020).

<sup>46</sup><https://timesofindia.indiatimes.com/life-style/health-fitness/home-remedies/covid-19-ministry-of-ayush-starts-clinical-trials-for-ashwagandha-and-4-other-ayurvedic-herbs-here-is-what-you-need-to-know/photostory/75692669.cms>

<sup>47</sup><https://www.hindustantimes.com/india-news/trials-for-4-ayush-formulations-against-covid-19-to-start-within-a-week-says-minister/story-XU9RsNDC3vLrFukt3gOApK.html>

**TABLE 5 |** List of medicinal plants with major immune modulating properties.

Sl.No.	NAME OF PLANTS	Type of anti viral or immune targeting effects exerted
1.	Turmeric	Curcumin in turmeric is an immune-modulatory agent. Has an anti-viral, anti-microbial, anti-inflammatory and anti-oxidant activity. Reduces pro-inflammatory cytokines like TNF- $\alpha$ , IFN- $\gamma$ , IL-1 and IL-8 via interaction with signal transducers such as NF- $\kappa$ B, JAKs/STATs, MAPKs and $\beta$ -catenin (Lelli et al., 2017; Kahkhaie et al., 2019).
2.	Ashwagandha	Activates immune response. Triggers Th1 cytokines and interferon expression. Increases expression of co-stimulatory molecules and integrins (Khan et al., 2009).
3.	Cinnamon	Inhibits allergen specific immune responses. Protects from systemic inflammation and lung injury by attenuating NLRP3 inflammasome activation (Sharma et al., 2016; Xu et al., 2019; Ose et al., 2020).
4.	Cardamom	Has anti-microbial activity (Agnihotri and Wakode, 2010). Exerts anti-inflammatory effect by inhibiting mediators such as COX2, TNF- $\alpha$ and IL-6 (Majdalawieh and Carr, 2010; Kandikattu et al., 2017).
5.	Holy Basil	Exerts anti-inflammatory effects by modulating cellular and humoral immunity. Elevates IFN- $\gamma$ and IL-4. Increases percentages of T-helper cells and NK-cells (Mondal et al., 2011; Kamyab and Eshraghian, 2013).
6.	Cumin	Thymoquinone in cumin has immuno-modulatory and anti-inflammatory properties. Suppresses inflammation by downregulation of COX2, IL-6, TNF- $\alpha$ and NO production, and enhancement of IL-10 production. Modulates cellular and humoral immunity and regulates Th1/Th2 immune response. Enhances NK cell mediated cytotoxicity (Majdalawieh and Fayyad, 2015; Gholamnezhad et al., 2016, 2019).
7.	Neem	Has anti-inflammatory, antibacterial and antioxidant effects. Attenuates release of pro-inflammatory cytokines such as TNF- $\alpha$ and IL-6, thus modulating immune response; inhibits MCP-1 (monocyte chemoattractant protein-1) expression and recruitment of inflammatory cells (Hao et al., 2014; Lee et al., 2017).
8.	Saffron	Has anti-inflammatory, radical scavenging and immuno-modulatory properties (Bolhassani et al., 2014; Moshiri et al., 2015).
9.	Amlaki	Has anti-inflammatory and immune-regulating activities. Promotes NK cell function and Antibody-dependent cellular cytotoxicity (ADCC) (Yang and Liu, 2014).
10.	Brahmi	Has immunomodulatory effects <a href="http://Nopr.Niscar.Res.In/Handle/123456789/41986">Http://Nopr.Niscar.Res.In/Handle/123456789/41986</a> . Mediates anti inflammatory effects by preventing the release of pro-inflammatory cytokines such as IL6 and TNF- $\alpha$ from microglial cells and the immune cells of the brain (Nemetchek et al., 2017).
11.	Moringa	Activates CD8 <sup>+</sup> T cells, promotes IL-10, IL-2, IL-6 and TNF- $\alpha$ production (Coriolano et al., 2018).
12.	Liquorice Root (Yashtimadhu)	Glycyrrhizin, the active compound of the liquorice root, inhibits SARS-associated coronavirus replication (Cinatl et al., 2003a). Reduces virus uptake by host cells (especially in case of influenza virus) (Mousa, 2017). Glycyrrhizin also stimulates IFN- $\gamma$ production by T cells. Exerts anti-inflammatory effects by inhibiting iNOS, COX2, IL-1 $\beta$ , TNF- $\alpha$ , IL-5 and IL-6 or by blocking trans-activation of NF- $\kappa$ B (Kuang et al., 2018; Fouladi et al., 2019).
13.	Shatavari	Modulates the Th1/Th2 balance; promotes IgG secretion and IL-12 production; and inhibits IL-6 production (Pise et al., 2015).
14.	Coriander	Has anti-inflammatory activity and boosts immunity (Li et al., 2016).
15.	Kapikacchu (Velvet Beans)	Modulates immune mediators such as NF- $\kappa$ B, IL-6, IFN- $\lambda$ , TNF- $\alpha$ , IL-1 $\beta$ , iNOS and IL-2 in the central nervous system (Rai et al., 2017). Boosts the innate immune response (Saiyad Musthafa et al., 2018).
16.	Ajwain	Acts as an anti-inflammatory agent and exerts bronchodilatory effect (Boskabady et al., 2005, 2007; Bairwa et al., 2012).
17.	Manjishtha	Serves as potential anti-inflammatory agent and immune modulator. Increases functions of the lymphatic system <a href="http://www.ljtsrd.Com/Papers/ljtsrd9616.Pdf">Http://www.ljtsrd.Com/Papers/ljtsrd9616.Pdf</a> (Shen et al., 2018).
18.	Bibhitaki	Boosts immunity <a href="https://www.Netmeds.Com/Health-Library/Post/Bibhitaki-5-Ways-This-Traditional-Fruit-Boosts-Your-Immunity">https://www.Netmeds.Com/Health-Library/Post/Bibhitaki-5-Ways-This-Traditional-Fruit-Boosts-Your-Immunity</a> .
19.	Guduchi, Giloy (Tinospora)	Serves as anti-oxidant and anti-inflammatory agent. Regulates NF- $\kappa$ B signaling and production of pro-inflammatory mediators (Dhama et al., 2017; Haque et al., 2017).
20.	Haritaki	Possesses anti-inflammatory and wound healing properties (Ratha and Joshi, 2013).
21.	Cinchona Bark	Source of chloroquine, a common anti-malarial drug; exerts an effect on SAR CoV-2 by immune modulation and blockage of viral entry (Lentini et al., 2020).
22.	Shatapushpa (Fennel)	Suppresses the immune response (Darzi et al., 2018). Regulates Th17 and Treg immune response (Zhang et al., 2018).
23.	Triphala	Exerts an anti-inflammatory effect via decreased expression of inflammatory mediators such as IL-17, COX-2, iNOS, TNF- $\alpha$ , IL-1 $\beta$ , VEGF, IL-6 and RANKL by preventing NF- $\kappa$ B activation (Peterson et al., 2017).

(Continued)

TABLE 5 | Continued

Sl.No.	NAME OF PLANTS	Type of anti viral or immune targeting effects exerted
24.	Jatiphala (Nutmeg)	Has immuno-modulatory functions. Macelignan in nutmeg has anti-inflammatory property and inhibits Th2 cytokines such as IL-4 (Shin et al., 2013).
25.	Jatamansi	Has anti-inflammatory, immuno-modulatory and wound-healing properties (Pandey et al., 2013; Han et al., 2017).
26.	Vidanga	Ameliorates pro-inflammatory cytokines and suppresses TNF- $\alpha$ production (Shirole et al., 2015).
27.	Gokshura (Tribulus)	Can reduce inflammation and fibrosis in the lungs by lowering the expression of IL-6, IL-8, TNF- $\alpha$ and TGF $\beta$ 1 (Qiu et al., 2019).
28.	Bhringaraj (Eclipta)	Exerts anti-inflammatory effects by regulating the NF- $\kappa$ B pathway and the production of pro-inflammatory mediators (Feng et al., 2019).
29.	Punarnava (Boerhavia)	Has anti-inflammatory properties (Mishra et al., 2014). Punarnavine, an alkaloid in Boerhavia exerts immuno-modulatory activities by reducing TNF- $\alpha$ , IL-1 $\beta$ , IL-6 production, and by increasing the titer of circulating antibody (Manu and Kuttan, 2009).
30.	Bhunimba (Andrographis)	Has anti-inflammatory and immuno-modulatory effects (Islam et al., 2018).
31.	Shankha pushpi (Dwarf morning glory)	Possesses anti-bacterial and immuno-modulating activity (Nguyen et al., 2016).
32.	Vidari (Indian Kudzu)	Serves as an immune booster and an anti-inflammatory agent by inhibiting inflammatory mediators such as CRP, NF- $\kappa$ B, COX-2, iNOS, TNF- $\alpha$ , IL-1 $\beta$ and IL-6 (Maji et al., 2014).

may have huge implication in clinical practice for improving the detection efficiency while alleviating the radiologist's workload (Dong et al., 2020; Shi et al., 2020). COVID-19 lesions in CT scans have a wide range of presentation in terms of appearance, size, and location in lungs, so, developing a system using either classical image processing approaches or conventional machine learning techniques relying on handcrafted features, is a challenging task. Recently, artificial intelligence (AI) has shown promise. It warrants better safety, higher accuracy and efficacy in imaging compared to the traditional, laborious imaging workflows. Alongside pioneering the basic clinical research, AI have enormous application in recent COVID-19 scenario which include provision for well allocated imaging platform, segmentation of infected and unaffected regions of lungs, clinical evaluation and diagnosis (Wang J. et al., 2020; Wang X. et al., 2020).

### Role of deep learning

Deep learning technology which lies central to current concept of Artificial Intelligence has been effective in automated detection of lung diseases with high diagnostic accuracy. However, there are challenges when developing AI-empowered deep learning technologies for COVID-19 screening (Oh et al., 2020; Roy et al., 2020). Most of the deep learning based methods require annotating the lesions in CT volumes for effective disease detection. Annotating lesions and labeling of annotations are laborious and time consuming, and hence, are not desirable in times of rapid COVID-19 outbreak and simultaneous shortage of radiologists. Therefore, the major challenge of AI-empowered solutions is to determine the potential of a deep learning model based on patients' chest CT volumes for automated and accurate COVID-19 diagnosis. It should require nominal expert annotation and should be easily trained, which will be extremely advantageous in developing AI solution rapidly for COVID-19 diagnosis. Due to the constraints of hardware resources, a major challenge is to educate a deep learning model using volumetric CT scans. Another problem is the inter-class

similarity and variation across pneumonia lesions. Finally, the lung CT scan images from patients with pneumonia harbor large portions of non-lesion regions, which exhibit wide range of complex tissue level variations. These non-lesion regions often exert a negative impact on the overall performance of AI-based solutions.

### Mobile Applications and Social Distancing Strategies

Aarogya Setu has been developed as a digital mobile COVID-19 tracking application, by the National Informatics Centre under the initiative of the Ministry of Electronics and Information Technology, Govt. of India, for effective awareness, management and mitigation of COVID-19 (Kodali et al., 2020)<sup>48,49</sup>. The Delhi Government has also launched the Delhi Corona app to create public awareness regarding availability of hospital facilities for COVID-19 treatment and also for complaint redressal regarding refusal to admit COVID-19 patients by hospitals with available facilities<sup>50,51,52</sup>. Apart from these mobile applications, the Govt. of India has promoted strict social distancing to contain spread of COVID-19 amongst the Indian population (Paital et al., 2020).

## DISCUSSION

The COVID-19 health crisis has created a stir in the whole world including in India. There has been a global endeavor in terms of disease diagnosis, drug repurposing and vaccine development

<sup>48</sup>[https://en.wikipedia.org/wiki/Aarogya\\_Setu](https://en.wikipedia.org/wiki/Aarogya_Setu)

<sup>49</sup>[https://static.mygov.in/rest/s3fs-public/mygov\\_159056978751307401.pdf](https://static.mygov.in/rest/s3fs-public/mygov_159056978751307401.pdf) (retrieved on June 23, 2020).

<sup>50</sup><https://www.ndtv.com/india-news/arvind-kejriwal-launches-delhi-corona-app-for-information-on-availability-of-hospital-beds-in-delhi-2239276>

<sup>51</sup><https://indianexpress.com/article/explained/delhi-corona-mobile-application-covid-19-6438796/>

<sup>52</sup><https://www.thehindu.com/News/Cities/Delhi/Kejriwal-Launches-Delhi-Corona-App-for-Real-Time-Information-on-Availability-of-Hospital-Beds/Article31729239.Ece> (retrieved on June 23, 2020).

to combat this pandemic. In addition to actively participating in these efforts to improve therapeutics and vaccine development against SARS-CoV-2; the Government of India has taken several initiatives and measures to further contain the disease. The total numbers of active cases in India reached a peak in the month of September, 2020 and have reduced subsequently. Although there have been 148,774 deaths in India till December 30, 2020; the recovery rate of COVID-19 patients in India has increased to about 98.51% as on December 30, 2020 (see text footnote 8). COVID-19 infection may exert detrimental long term effects on organs such as lungs, liver, kidney, brain. and heart (Heneka et al., 2020). These may even last after recovery from COVID-19 and lead to life-threatening health issues<sup>53</sup>. Several clinical parameters such as blood levels of inflammatory mediators, neutrophil to lymphocyte ratio (NLR) and CT scan severity score have been evaluated for highlighting disease progression and the risk for development of post recovery complications such as pulmonary fibrosis, ARDS, neurological disorder or even multi-organ failure (Feng et al., 2020). Identification of blood borne easily detectable biomarkers could potentially stratify COVID-19 based on its severity and enable early prediction of progression to post recovery complications, thereby leading to better post COVID care and effective control of deaths due to such complications.

<sup>53</sup> <https://www.thehindu.com/sci-tech/health/the-hindu-explains-what-are-the-long-term-effects-of-covid-19/article32651206.ece>

## REFERENCES

- Agnihotri, S., and Wakode, S. (2010). Antimicrobial activity of essential oil and various extracts of fruits of greater cardamom. *Indian J. Pharm. Sci.* 72, 657–659. doi: 10.4103/0250-474x.78542
- Agrawal, D. (2020). Ventilator politics—the big picture. *Indian J. Neurotrauma* 17, 1–2.
- Agrawal, U., Raju, R., and Udwadia, Z. F. (2020). Favipiravir: a new and emerging antiviral option in COVID-19. *Med. J. Armed Forces India* 76, 370–376. doi: 10.1016/j.mjafi.2020.08.004
- Ahuja, A. S., Reddy, V. P., and Marques, O. (2020). Artificial intelligence and COVID-19: a multidisciplinary approach. *Integr. Med. Res.* 9:100434. doi: 10.1016/j.imr.2020.100434
- Akram, M., Tahir, I. M., Shah, S. M. A., Mahmood, Z., Altaf, A., Ahmad, K., et al. (2018). Antiviral potential of medicinal plants against HIV, HSV, influenza, hepatitis, and coxsackievirus: a systematic review. *Phytother. Res.* 32, 811–822. doi: 10.1002/ptr.6024
- Alagarasu, K., Choudhary, M. L., Lole, K. S., Abraham, P., Potdar, V., and Team, N. I. C. (2020). Evaluation of RdRp & ORF-1b-nsp14-based real-time RT-PCR assays for confirmation of SARS-CoV-2 infection: an observational study. *Indian J. Med. Res.* 151, 483–485.
- Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., and Cheng, X. (2020). Artificial intelligence and machine learning to fight COVID-19. *Physiol. Genomics* 52, 200–202. doi: 10.1152/physiolgenomics.00029.2020
- Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452. doi: 10.1038/s41591-020-0820-9
- Azkar, A. K., Akdis, M., Azkar, D., Sokolowska, M., Van De Veen, W., Bruggen, M. C., et al. (2020). Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. *Allergy* 75, 1564–1581. doi: 10.1111/all.14364
- Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., et al. (2020). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* doi: 10.1056/NEJMoa2035389 [Epub ahead of print].

## AUTHOR CONTRIBUTIONS

All the authors reviewed the literature, wrote the review, edited, and approved the manuscript.

## FUNDING

Research support and publication charges are funded by Bose Institute Intramural Fund.

## ACKNOWLEDGMENTS

SM was grateful to the Department of Science and Technology, Govt. of India for the DST-Inspire Fellowship. MK was supported by the Council for Scientific and Industrial Research Emeritus Scientist Scheme (21(1088)/19/EMR-II). JB was supported by the J.C. Bose National Fellowship (SB/S2/JCB-049/2016).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.637362/full#supplementary-material>

- Bairwa, R., Sodha, R. S., and Rajawat, B. S. (2012). *Trachyspermum ammi*. *Pharmacogn. Rev.* 6, 56–60. doi: 10.4103/0973-7847.95871
- Balachandhar, V., Mahalaxmi, I., Devi, S. M., Kaavya, J., Kumar, N. S., Laldinmawii, G., et al. (2020). Follow-up studies in COVID-19 recovered patients - is it mandatory? *Sci. Total Environ.* 729:139021. doi: 10.1016/j.scitotenv.2020.139021
- Biswas, N. K., and Majumder, P. P. (2020). Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type. *Indian J. Med. Res.* 151, 450–458. doi: 10.4103/ijmr.ijmr\_1125\_20
- Bolhassani, A., Khavari, A., and Bathaie, S. Z. (2014). Saffron and natural carotenoids: biochemical activities and anti-tumor effects. *Biochim. Biophys. Acta* 1845, 20–30. doi: 10.1016/j.bbcan.2013.11.001
- Boskabady, M. H., Alizadeh, M., and Jahanbin, B. (2007). Bronchodilatory effect of Carum copticum in airways of asthmatic patients. *Therapie* 62, 23–29. doi: 10.2515/therapie.2007007
- Boskabady, M. H., Jandaghi, P., Kiani, S., and Hasanzadeh, L. (2005). Antitussive effect of Carum copticum in guinea pigs. *J. Ethnopharmacol.* 97, 79–82. doi: 10.1016/j.jep.2004.10.016
- Bragazzi, N. L., Dai, H., Damiani, G., Behzadifar, M., Martini, M., and Wu, J. (2020). How Big Data and Artificial intelligence can help better manage the COVID-19 pandemic. *Int. J. Environ. Res. Public Health* 17:3176. doi: 10.3390/ijerph17093176
- Burkard, C., Verheije, M. H., Haagmans, B. L., Van Kuppeveld, F. J., Rottier, P. J., Bosch, B. J., et al. (2015). ATP1A1-mediated Src signaling inhibits coronavirus entry into host cells. *J. Virol.* 89, 4434–4448. doi: 10.1128/jvi.03274-14
- Chan, J. F., Chan, K. H., Kao, R. Y., To, K. K., Zheng, B. J., Li, C. P., et al. (2013). Broad-spectrum antivirals for the emerging Middle East respiratory syndrome coronavirus. *J. Infect.* 67, 606–616. doi: 10.1016/j.jinf.2013.09.029
- Chan, J. F., Yao, Y., Yeung, M. L., Deng, W., Bao, L., Jia, L., et al. (2015). Treatment with lopinavir/ritonavir or interferon-beta1b improves outcome of MERS-CoV infection in a nonhuman primate model of common marmoset. *J. Infect. Dis.* 212, 1904–1913. doi: 10.1093/infdis/jiv392



- Chen, G., Wu, D., Guo, W., Cao, Y., Huang, D., Wang, H., et al. (2020). Clinical and immunological features of severe and moderate coronavirus disease 2019. *J. Clin. Invest.* 130, 2620–2629. doi: 10.1172/jci137244
- Chen, I. Y., Moriyama, M., Chang, M. F., and Ichinohe, T. (2019). Severe acute respiratory syndrome coronavirus viroporin 3a activates the NLRP3 inflammasome. *Front. Microbiol.* 10:50. doi: 10.3389/fmicb.2019.00050
- Choudhary, M. L., Vipat, V., Jadhav, S., Basu, A., Cherian, S., Abraham, P., et al. (2020). Development of in vitro transcribed RNA as positive control for laboratory diagnosis of SARS-CoV-2 in India. *Indian J. Med. Res.* 151, 251–254.
- Choudhary, R., and Sharma, A. K. (2020). Potential use of hydroxychloroquine, ivermectin and azithromycin drugs in fighting COVID-19: trends, scope and relevance. *New Microbes New Infect.* 35, 100684. doi: 10.1016/j.nmni.2020.100684
- Chu, C. M., Cheng, V. C., Hung, I. F., Wong, M. M., Chan, K. H., Chan, K. S., et al. (2004). Role of lopinavir/ritonavir in the treatment of SARS: initial virological and clinical findings. *Thorax* 59, 252–256. doi: 10.1136/thorax.2003.012658
- Cinatl, J., Morgenstern, B., Bauer, G., Chandra, P., Rabenau, H., and Doerr, H. W. (2003a). Glycyrrhizin, an active component of liquorice roots, and replication of SARS-associated coronavirus. *Lancet* 361, 2045–2046. doi: 10.1016/s0140-6736(03)13615-x
- Cinatl, J., Morgenstern, B., Bauer, G., Chandra, P., Rabenau, H., and Doerr, H. W. (2003b). Treatment of SARS with human interferons. *Lancet* 362, 293–294. doi: 10.1016/s0140-6736(03)13973-6
- Coriolano, M. C., De Santana Brito, J., De Siqueira Patriota, L. L., De Araujo Soares, A. K., De Lorena, V. M. B., Paiva, P. M. G., et al. (2018). Immunomodulatory effects of the water-soluble lectin from *Moringa oleifera* seeds (WSMoL) on human peripheral blood mononuclear cells (PBMC). *Protein Pept. Lett.* 25, 295–301. doi: 10.2174/0929866525666180130141736
- Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., et al. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eur. Surveill.* 25:2000045.
- Cortegiani, A., Ingoglia, G., Ippolito, M., Giarratano, A., and Einav, S. (2020). A systematic review on the efficacy and safety of chloroquine for the treatment of COVID-19. *J. Crit. Care* 57, 279–283.
- Cui, J., Li, F., and Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi: 10.1038/s41579-018-0118-9
- Damle, B., Vourvahis, M., Wang, E., Leaney, J., and Corrigan, B. (2020). Clinical pharmacology perspectives on the antiviral activity of azithromycin and use in COVID-19. *Clin. Pharmacol. Ther.* 108, 201–211. doi: 10.1002/cpt.1857
- Darzi, S. E., Khazraei, S. P., and Amirghofran, Z. (2018). The immunoinhibitory and apoptosis-inducing activities of *Foeniculum vulgare* on human peripheral blood lymphocytes. *Res. Pharm. Sci.* 13, 103–110. doi: 10.4103/1735-5362.223792
- de Wilde, A. H., Jochmans, D., Posthuma, C. C., Zevenhoven-Dobbe, J. C., Van Nieuwkoop, S., Bestebroer, T. M., et al. (2014). Screening of an FDA-approved compound library identifies four small-molecule inhibitors of Middle East respiratory syndrome coronavirus replication in cell culture. *Antimicrob. Agents Chemother.* 58, 4875–4884. doi: 10.1128/aac.03011-14
- Dhama, K., Karthik, K., Khandia, R., Munjal, A., Tiwari, R., Rana, R., et al. (2018). Medicinal and therapeutic potential of herbs and plant metabolites / extracts countering viral pathogens - current knowledge and future prospects. *Curr. Drug Metab.* 19, 236–263. doi: 10.2174/1389200219666180129145252
- Dhama, K., Sachan, S., Khandia, R., Munjal, A., Iqbal, H. M. N., Latheef, S. K., et al. (2017). Medicinal and beneficial health applications of *Tinospora cordifolia* (Guduchi): a miraculous herb countering various diseases/disorders and its immunomodulatory effects. *Recent Pat. Endocr. Metab. Immune Drug Discov.* 10, 96–111. doi: 10.2174/1872214811666170301105101
- Dong, D., Tang, Z., Wang, S., Hui, H., Gong, L., Lu, Y., et al. (2020). The role of imaging in the detection and management of COVID-19: a review. *IEEE Rev. Biomed. Eng.* doi: 10.1109/RBME.2020.2990959 [Epub ahead of print].
- Duan, K., Liu, B., Li, C., Zhang, H., Yu, T., Qu, J., et al. (2020). Effectiveness of convalescent plasma therapy in severe COVID-19 patients. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9490–9496.
- Duncan, J. S., and Ayache, N. (2000). Medical image analysis: progress over two decades and the challenges ahead. *IEEE Trans. Pattern Anal. Machine Intell.* 22, 85–106. doi: 10.1109/34.824822
- Feinmann, J. (2020). PPE: what now for the global supply chain? *BMJ* 369:m1910. doi: 10.1136/bmj.m1910
- Feng, L., Zhai, Y. Y., Xu, J., Yao, W. F., Cao, Y. D., Cheng, F. F., et al. (2019). A review on traditional uses, phytochemistry and pharmacology of *Eclipta prostrata* (L.) L. *J. Ethnopharmacol.* 245:112109. doi: 10.1016/j.jep.2019.112109
- Feng, Z., Yu, Q., Yao, S., Luo, L., Zhou, W., Mao, X., et al. (2020). Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics. *Nat. Commun.* 11:4968.
- Ferner, R. E., and Aronson, J. K. (2020). Remdesivir in covid-19. *BMJ* 369:m1610. doi: 10.1136/bmj.m1610
- Folegatti, P. M., Ewer, K. J., Aley, P. K., Angus, B., Becker, S., Belij-Rammerstorfer, S., et al. (2020). Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet* 396, 467–478.
- Fouladi, S., Masjedi, M., Ganjalikhani Hakemi, M., and Eskandari, N. (2019). The review of in vitro and in vivo studies over the glycyrrhizic acid as natural remedy option for treatment of allergic asthma. *Iran J Allergy Asthma Immunol.* 18, 1–11.
- Furuta, Y., Gowen, B. B., Takahashi, K., Shiraki, K., Smeets, D. F., and Barnard, D. L. (2013). Favipiravir (T-705), a novel viral RNA polymerase inhibitor. *Antiviral Res.* 100, 446–454. doi: 10.1016/j.antiviral.2013.09.015
- Gautret, P., Lagier, J. C., Parola, P., Hoang, V. T., Meddeb, L., Mailhe, M., et al. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *Int. J. Antimicrob. Agents* 56:105949.
- Gholamnezhad, Z., Havakhah, S., and Boskabady, M. H. (2016). Preclinical and clinical effects of *Nigella sativa* and its constituent, thymoquinone: a review. *J. Ethnopharmacol.* 190, 372–386. doi: 10.1016/j.jep.2016.06.061
- Gholamnezhad, Z., Shakeri, F., Saadat, S., Ghorani, V., and Boskabady, M. H. (2019). Clinical and experimental effects of *Nigella sativa* and its constituents on respiratory and allergic disorders. *Avicenna J. Phytomed.* 9, 195–212.
- Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468.
- Gupta, N., Potdar, V., Praharaj, I., Giri, S., Sapkal, G., Yadav, P., et al. (2020). Laboratory preparedness for SARS-CoV-2 testing in India: harnessing a network of virus research & diagnostic laboratories. *Indian J. Med. Res.* 151, 216–225.
- Han, X., Beaumont, C., and Stevens, N. (2017). Chemical composition analysis and in vitro biological activities of ten essential oils in human skin cells. *Biochim. Open* 5, 1–7. doi: 10.1016/j.biopen.2017.04.001
- Hao, F., Kumar, S., Yadav, N., and Chandra, D. (2014). Neem components as potential agents for cancer prevention and treatment. *Biochim. Biophys. Acta* 1846, 247–257. doi: 10.1016/j.bbcan.2014.07.002
- Haque, M. A., Jantan, I., and Abbas Bukhari, S. N. (2017). *Tinospora* species: an overview of their modulating effects on the immune system. *J. Ethnopharmacol.* 207, 67–85. doi: 10.1016/j.jep.2017.06.013
- Heneka, M. T., Golenbock, D., Latz, E., Morgan, D., and Brown, R. (2020). Immediate and long-term consequences of COVID-19 infections for the development of neurological disease. *Alzheimers Res. Ther.* 12:69.
- Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 181, 271–280.e278.
- Hou, H., Wang, T., Zhang, B., Luo, Y., Mao, L., Wang, F., et al. (2020). Detection of IgM and IgG antibodies in patients with coronavirus disease 2019. *Clin. Transl. Immunol.* 9:e01136.
- Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506.
- Huang, X., Xu, Y., Yang, Q., Chen, J., Zhang, T., Li, Z., et al. (2015). Efficacy and biological safety of lopinavir/ritonavir based anti-retroviral therapy in HIV-1-infected patients: a meta-analysis of randomized controlled trials. *Sci. Rep.* 5:8528.
- Islam, M. T., Ali, E. S., Uddin, S. J., Islam, M. A., Shaw, S., Khan, I. N., et al. (2018). Andrographolide, a diterpene lactone from *Andrographis paniculata* and its therapeutic promises in cancer. *Cancer Lett.* 420, 129–145. doi: 10.1016/j.canlet.2018.01.074

- Iyengar, K., Bahl, S., Vaishya, R., and Vaish, A. (2020). Challenges and solutions in meeting up the urgent requirement of ventilators for COVID-19 patients. *Diabetes Metab. Syndr. Clin. Res. Rev.* 14, 499–501. doi: 10.1016/j.dsx.2020.04.048
- Iyer, M., Jayaramayya, K., Subramaniam, M. D., Lee, S. B., Dayem, A. A., Cho, S. G., et al. (2020). COVID-19: an update on diagnostic and therapeutic approaches. *BMB Rep.* 53, 191–205. doi: 10.5483/bmbrep.2020.53.4.080
- Jabbari, P., and Rezaei, N. (2019). Artificial intelligence and immunotherapy. *Expert Rev. Clin. Immunol.* 15, 689–691. doi: 10.1080/1744666x.2019.1623670
- Jin, Z., Smith, L. K., Rajwanshi, V. K., Kim, B., and Deval, J. (2013). The ambiguous base-pairing and high substrate efficiency of T-705 (Favipiravir) Ribofuranosyl 5'-triphosphate towards influenza A virus polymerase. *PLoS One* 8:e68347. doi: 10.1371/journal.pone.0068347
- Kahkhaie, K. R., Mirhosseini, A., Aliabadi, A., Mohammadi, A., Mousavi, M. J., Haftcheshmeh, S. M., et al. (2019). Curcumin: a modulator of inflammatory signaling pathways in the immune system. *Inflammopharmacology* 27, 885–900. doi: 10.1007/s10787-019-00607-3
- Kampf, G. (2020). Potential role of inanimate surfaces for the spread of coronaviruses and their inactivation with disinfectant agents. *Infect. Prevention Pract.* 2:100044. doi: 10.1016/j.infpip.2020.100044
- Kampf, G., Todt, D., Pfaender, S., and Steinmann, E. (2020). Persistence of coronaviruses on inanimate surfaces and their inactivation with biocidal agents. *J. Hosp. Infect.* 104, 246–251. doi: 10.1016/j.jhin.2020.01.022
- Kamyab, A. A., and Eshraghian, A. (2013). Anti-Inflammatory, gastrointestinal and hepatoprotective effects of *Ocimum sanctum* Linn: an ancient remedy with new application. *Inflamm. Allergy Drug Targets* 12, 378–384. doi: 10.2174/1871528112666131125110017
- Kandikattu, H. K., Rachitha, P., Jayashree, G. V., Krupashree, K., Sukhith, M., Majid, A., et al. (2017). Anti-inflammatory and anti-oxidant effects of Cardamom (*Elettaria repens* (Sonn.) Baill) and its phytochemical analysis by 4D GCXGC TOF-MS. *Biomed. Pharmacother.* 91, 191–201. doi: 10.1016/j.biopha.2017.04.049
- Kaur, N., Singh, R., Dar, Z., Bijarnia, R. K., Dhingra, N., and Kaur, T. (2020). Genetic comparison among various coronavirus strains for the identification of potential vaccine targets of SARS-CoV2. *Infect. Genet. Evol.* doi: 10.1016/j.meegid.2020.104490 [Epub ahead of print].
- Kaushik, S., Kaushik, S., Sharma, Y., Kumar, R., and Yadav, J. P. (2020). The Indian perspective of COVID-19 outbreak. *Virusdisease* 31, 1–8.
- Khalil, A., Kamar, A., and Nemer, G. (2020). Thalidomide-revisited: are COVID-19 patients going to be the latest victims of yet another theoretical drug-repurposing? *Front. Immunol.* 11:1248. doi: 10.3389/fimmu.2020.01248
- Khan, S., Malik, F., Suri, K. A., and Singh, J. (2009). Molecular insight into the immune up-regulatory properties of the leaf extract of Ashwagandha and identification of Th1 immunostimulatory chemical entity. *Vaccine* 27, 6080–6087. doi: 10.1016/j.vaccine.2009.07.011
- Khorsand, B., Savadi, A., and Naghibzadeh, M. (2020). SARS-CoV-2-human protein-protein interaction network. *Inform. Med. Unlocked* 20:100413. doi: 10.1016/j.imu.2020.100413
- Kim, K. M., Han, S. H., Yoo, S. Y., and Yoo, J. H. (2020). Potential hazards of concern in the walk-through screening system for the corona virus disease 2019 from the perspective of infection preventionists. *J. Korean Med. Sci.* 35:e156.
- Kodali, P. B., Hense, S., Kopparty, S., Kalapala, G. R., and Haloi, B. (2020). How Indians responded to the Arogya Setu app? *Indian J. Public Health* 64, S228–S230.
- Kuang, Y., Li, B., Fan, J., Qiao, X., and Ye, M. (2018). Antitussive and expectorant activities of licorice and its major compounds. *Bioorg. Med. Chem.* 26, 278–284. doi: 10.1016/j.bmc.2017.11.046
- Lamba, I. (2020). Why India needs to extend the nationwide lockdown. *Am. J. Emerg. Med.* 38, 1528–1529. doi: 10.1016/j.ajem.2020.04.026
- Le, T. T., Andreadakis, Z., Kumar, A., Gómez Román, R., Tollefsen, S., Saville, M., et al. (2020). The COVID-19 vaccine development landscape. *Nat. Rev. Drug Discov.* 19, 305–306. doi: 10.1038/d41573-020-00073-5
- Lee, J. W., Ryu, H. W., Park, S. Y., Park, H. A., Kwon, O. K., Yuk, H. J., et al. (2017). Protective effects of neem (*Azadirachta indica* A. Juss.) leaf extract against cigarette smoke- and lipopolysaccharide-induced pulmonary inflammation. *Int. J. Mol. Med.* 40, 1932–1940.
- Lelli, D., Sahebkar, A., Johnston, T. P., and Pedone, C. (2017). Curcumin use in pulmonary diseases: state of the art and future perspectives. *Pharmacol. Res.* 115, 133–148. doi: 10.1016/j.phrs.2016.11.017
- Lentini, G., Cavalluzzi, M. M., and Habtemariam, S. (2020). COVID-19, chloroquine repurposing, and cardiac safety concern: chirality might help. *Molecules* 25:1834. doi: 10.3390/molecules25081834
- Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., et al. (2020). Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology* 296, E65–E71.
- Li, Y., Yao, J., Han, C., Yang, J., Chaudhry, M. T., Wang, S., et al. (2016). Quercetin, inflammation and immunity. *Nutrients* 8:167. doi: 10.3390/nu8030167
- Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* 26, 842–844. doi: 10.1038/s41591-020-0901-9
- Liu, L., Wei, Q., Lin, Q., Fang, J., Wang, H., Kwok, H., et al. (2019). Anti-spike IgG causes severe acute lung injury by skewing macrophage responses during acute SARS-CoV infection. *JCI Insight* 4:e123158.
- Lotfi, M., and Rezaei, N. (2020). CRISPR/Cas13: a potential therapeutic option of COVID-19. *Biomed. Pharmacother.* 131:110738. doi: 10.1016/j.biopha.2020.110738
- Luo, P., Liu, Y., Qiu, L., Liu, X., Liu, D., and Li, J. (2020). Tocilizumab treatment in COVID-19: a single center experience. *J. Med. Virol.* 92, 814–818. doi: 10.1002/jmv.25801
- Lyons, S. M., Harrison, M. A., and Law, S. E. (2011). Electrostatic application of antimicrobial sprays to sanitize food handling and processing surfaces for enhanced food safety. *J. Phys.* 301:012014. doi: 10.1088/1742-6596/301/1/012014
- Magro, C., Mulvey, J. J., Berlin, D., Nuovo, G., Salvatore, S., Harp, J., et al. (2020). Complement associated microvascular injury and thrombosis in the pathogenesis of severe COVID-19 infection: a report of five cases. *Transl. Res.* 220, 1–13. doi: 10.1016/j.trsl.2020.04.007
- Majdalawieh, A. F., and Carr, R. I. (2010). In vitro investigation of the potential immunomodulatory and anti-cancer activities of black pepper (*Piper nigrum*) and cardamom (*Elettaria cardamomum*). *J. Med. Food* 13, 371–381. doi: 10.1089/jmf.2009.1131
- Majdalawieh, A. F., and Fayyad, M. W. (2015). Immunomodulatory and anti-inflammatory action of *Nigella sativa* and thymoquinone: a comprehensive review. *Int. Immunopharmacol.* 28, 295–304. doi: 10.1016/j.intimp.2015.06.023
- Maji, A. K., Pandit, S., Banerji, P., and Banerjee, D. (2014). *Pueraria tuberosa*: a review on its phytochemical and therapeutic potential. *Nat. Prod. Res.* 28, 2111–2127. doi: 10.1080/14786419.2014.928291
- Manu, K. A., and Kuttan, G. (2009). Immunomodulatory activities of Punarnavine, an alkaloid from *Boerhaavia diffusa*. *Immunopharmacol. Immunotoxicol.* 31, 377–387. doi: 10.1080/08923970802702036
- McBride, R., and Fielding, B. C. (2012). The role of severe acute respiratory syndrome (SARS)-coronavirus accessory proteins in virus pathogenesis. *Viruses* 4, 2902–2923. doi: 10.3390/v4112902
- Mehta, P., McAuley, D. F., Brown, M., Sanchez, E., Tattersall, R. S., Manson, J. J., et al. (2020). COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* 395, 1033–1034. doi: 10.1016/s0140-6736(20)30628-0
- Michot, J. M., Albiges, L., Chaput, N., Saada, V., Pommeret, F., Griscelli, F., et al. (2020). Tocilizumab, an anti-IL-6 receptor antibody, to treat COVID-19-related respiratory failure: a case report. *Ann. Oncol.* 31, 961–964. doi: 10.1016/j.annonc.2020.03.300
- Mishra, S., Aeri, V., Gaur, P. K., and Jachak, S. M. (2014). Phytochemical, therapeutic, and ethnopharmacological overview for a traditionally important herb: *Boerhaavia diffusa* Linn. *Biomed. Res. Int.* 2014:808302.
- Mishra, S., Mohapatra, A., Kumar, R., Singh, A., Bhadoria, A. S., and Kant, R. (2020). Restricting rural-urban connect to combat infectious disease epidemic as India fights COVID-19. *J. Family Med. Prim. Care* 9, 1792–1794. doi: 10.4103/jfmpc.jfmpc\_451\_20
- Mishra, S. V., Haque, S. M., and Gayen, A. (2020). COVID-19 in India transmits from the urban to the rural. *Int. J. Health Plann. Manage.* 35, 1623–1625. doi: 10.1002/hpm.3047

- Mittal, A., Gupta, A., Kumar, S., Surjit, M., Singh, B., Soneja, M., et al. (2020). Gargle lavage as a viable alternative to swab for detection of SARS-CoV-2. *Indian J. Med. Res.* 152, 77–81. doi: 10.4103/ijmr.ijmr\_2987\_20
- Mohanraj, K., Karthikeyan, B. S., Vivek-Ananth, R. P., Chand, R. P. B., Aparna, S. R., Mangalapandi, P., et al. (2018). IMPPAT: a curated database of Indian medicinal plants, phytochemistry and therapeutics. *Sci. Rep.* 8:4329.
- Mondal, S., Varma, S., Bamola, V. D., Naik, S. N., Mirdha, B. R., Padhi, M. M., et al. (2011). Double-blinded randomized controlled trial for immunomodulatory effects of Tulsi (*Ocimum sanctum* Linn.) leaf extract on healthy volunteers. *J. Ethnopharmacol.* 136, 452–456. doi: 10.1016/j.jep.2011.05.012
- Moshiri, M., Vahabzadeh, M., and Hosseinzadeh, H. (2015). Clinical applications of saffron (*Crocus sativus*) and its constituents: a review. *Drug Res. (Stuttg)* 65, 287–295. doi: 10.1055/s-0034-1375681
- Mourya, D. T., Sapkal, G., Yadav, P. D., Sk, M. B., Shete, A., and Gupta, N. (2020). Biorisk assessment for infrastructure & biosafety requirements for the laboratories providing coronavirus SARS-CoV-2/(COVID-19) diagnosis. *Indian J. Med. Res.* 151, 172–176.
- Mousa, H. A. (2017). Prevention and Treatment of influenza, influenza-like illness, and common cold by herbal, complementary, and natural therapies. *J. Evid. Based Complementary Altern. Med.* 22, 166–174. doi: 10.1177/21565872166641831
- Narayanan, K., Huang, C., and Makino, S. (2008). SARS coronavirus accessory proteins. *Virus Res.* 133, 113–121. doi: 10.1016/j.virusres.2007.10.009
- Nemetcheck, M. D., Stierle, A. A., Stierle, D. B., and Lurie, D. I. (2017). The Ayurvedic plant *Bacopa monnieri* inhibits inflammatory pathways in the brain. *J. Ethnopharmacol.* 197, 92–100. doi: 10.1016/j.jep.2016.07.073
- Netea, M. G., Dominguez-Andres, J., Barreiro, L. B., Chavakis, T., Divangahi, M., Fuchs, E., et al. (2020). Defining trained immunity and its role in health and disease. *Nat. Rev. Immunol.* 20, 375–388. doi: 10.1038/s41577-020-0285-6
- Nguyen, K. N., Nguyen, G. K., Nguyen, P. Q., Ang, K. H., Dedon, P. C., and Tam, J. P. (2016). Immunostimulating and Gram-negative-specific antibacterial cyclotides from the butterfly pea (*Clitoria ternatea*). *FEBS J.* 283, 2067–2090. doi: 10.1111/febs.13720
- Oh, Y., Park, S., and Ye, J. C. (2020). Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans. Med. Imaging* 39, 2688–2700. doi: 10.1109/tmi.2020.2993291
- Omolo, C. A., Soni, N., Fasiku, V. O., Mackraj, I., and Govender, T. (2020). Update on therapeutic approaches and emerging therapies for SARS-CoV-2 virus. *Eur. J. Pharmacol.* 883:173348. doi: 10.1016/j.ejphar.2020.173348
- Ong, E., Wong, M. U., Huffman, A., and He, Y. (2020). COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *Front. Immunol.* 11:1581. doi: 10.3389/fimmu.2020.01581
- Ose, R., Tu, J., Schink, A., Maxeiner, J., Schuster, P., Lucas, K., et al. (2020). Cinnamon extract inhibits allergen-specific immune responses in human and murine allergy models. *Clin. Exp. Allergy* 50, 41–50. doi: 10.1111/cea.13507
- Othman, H., Bouslama, Z., Brandenburg, J. T., da Rocha, J., Hamdi, Y., Ghedira, K., et al. (2020). Interaction of the spike protein RBD from SARS-CoV-2 with ACE2: similarity with SARS-CoV, hot-spot analysis and effect of the receptor polymorphism. *Biochem. Biophys. Res. Commun.* 527, 702–708. doi: 10.1016/j.bbrc.2020.05.028
- Patil, B., Das, K., and Parida, S. K. (2020). Inter nation social lockdown versus medical care against COVID-19, a mild environmental insight with special reference to India. *Sci. Total Environ.* 728:138914. doi: 10.1016/j.scitotenv.2020.138914
- Pandey, M. M., Katara, A., Pandey, G., Rastogi, S., and Rawat, A. K. (2013). An important Indian traditional drug of ayurveda jatamansi and its substitute bhootkeshi: chemical profiling and antioxidant activity. *Evid. Based Complement Alternat. Med.* 2013:142517.
- Parray, H. A., Chiranjivi, A. K., Asthana, S., Yadav, N., Shrivastava, T., Mani, S., et al. (2020). Identification of an anti-SARS-CoV-2 receptor-binding domain-directed human monoclonal antibody from a naive semisynthetic library. *J. Biol. Chem.* 295, 12814–12821. doi: 10.1074/jbc.ac120.014918
- Peterson, C. T., Denniston, K., and Chopra, D. (2017). Therapeutic uses of triphala in ayurvedic medicine. *J. Altern. Complement Med.* 23, 607–614. doi: 10.1089/acm.2017.0083
- Pise, M. V., Rudra, J. A., and Upadhyay, A. (2015). Immunomodulatory potential of shatavaris produced from *Asparagus racemosus* tissue cultures. *J. Nat. Sci. Biol. Med.* 6, 415–420. doi: 10.4103/0976-9668.160025
- Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., et al. (2020). Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N. Engl. J. Med.* 383, 2603–2615.
- Qiu, M., An, M., Bian, M., Yu, S., Liu, C., and Liu, Q. (2019). Terrestrosin D from *Tribulus terrestris* attenuates bleomycin-induced inflammation and suppresses fibrotic changes in the lungs of mice. *Pharm. Biol.* 57, 694–700. doi: 10.1080/13880209.2019.1672754
- Rai, S. N., Birla, H., Zahra, W., Singh, S. S., and Singh, S. P. (2017). Immunomodulation of Parkinson's disease using *Mucuna pruriens* (Mp). *J. Chem. Neuroanat.* 85, 27–35. doi: 10.1016/j.jchemneu.2017.06.005
- Randhawa, G. S., Soltysiak, M. P. M., El Roz, H., De Souza, C. P. E., Hill, K. A., and Kari, L. (2020). Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS One* 15:e0232391. doi: 10.1371/journal.pone.0232391
- Ratha, K. K., and Joshi, G. C. (2013). Haritaki (Chebulic myrobalan) and its varieties. *Ayu* 34, 331–334. doi: 10.4103/0974-8520.123139
- Renu, K., Prasanna, P. L., and Valsala Gopalakrishnan, A. (2020). Coronaviruses pathogenesis, comorbidities and multi-organ damage - a review. *Life Sci.* 255:117839. doi: 10.1016/j.lfs.2020.117839
- Richardson, P., Griffin, I., Tucker, C., Smith, D., Oechsle, O., Phelan, A., et al. (2020). Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *Lancet* 395, e30–e31.
- Roy, S., Menapace, W., Oei, S., Luijten, B., Fini, E., Saltori, C., et al. (2020). Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE Trans. Med. Imaging* 39, 2676–2687. doi: 10.1109/tmi.2020.2994459
- Saha, I., Ghosh, N., Maity, D., Sharma, N., and Mitra, K. (2020). Inferring the genetic variability in Indian SARS-CoV-2 genomes using consensus of multiple sequence alignment techniques. *Infect. Genet. Evol.* 85:104522. doi: 10.1016/j.meegid.2020.104522
- Saiyad Musthafa, M., Asgari, S. M., Kurian, A., Elumalai, P., Jawahar Ali, A. R., Paray, B. A., et al. (2018). Protective efficacy of *Mucuna pruriens* (L.) seed meal enriched diet on growth performance, innate immunity, and disease resistance in *Oreochromis mossambicus* against *Aeromonas hydrophila*. *Fish Shellfish Immunol.* 75, 374–380. doi: 10.1016/j.fsi.2018.02.031
- Samaddar, A., Gadeballi, R., Nag, V. L., and Misra, S. (2020). The enigma of low COVID-19 fatality rate in India. *Front. Genet.* 11:854. doi: 10.3389/fgene.2020.00854
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710.
- Shaffer, L. (2020). 15 drugs being tested to treat COVID-19 and how they would work. *Nat. Med.* doi: 10.1038/d41591-020-00019-9 [Epub ahead of print].
- Shahid, Z., Kalayanamitra, R., McClafferty, B., Kepko, D., Ramgobin, D., Patel, R., et al. (2020). COVID-19 and older adults: what we know. *J. Am. Geriatr. Soc.* 68, 926–929.
- Sharma, U. K., Sharma, A. K., and Pandey, A. K. (2016). Medicinal attributes of major phenylpropanoids present in cinnamon. *BMC Complement Altern. Med.* 16:156. doi: 10.1186/s12906-016-1147-4
- Sheahan, T. P., Sims, A. C., Leist, S. R., Schafer, A., Won, J., Brown, A. J., et al. (2020). Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV. *Nat. Commun.* 11:222.
- Shen, C. H., Liu, C. T., Song, X. J., Zeng, W. Y., Lu, X. Y., Zheng, Z. L., et al. (2018). Evaluation of analgesic and anti-inflammatory activities of *Rubia cordifolia* L. by spectrum-effect relationships. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 1090, 73–80. doi: 10.1016/j.jchromb.2018.05.021
- Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., et al. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* doi: 10.1109/RBME.2020.2987975 [Epub ahead of print].
- Shin, K., Chung, H. C., Kim, D. U., Hwang, J. K., and Lee, S. H. (2013). Macelignan attenuated allergic lung inflammation and airway hyper-responsiveness in murine experimental asthma. *Life Sci.* 92, 1093–1099. doi: 10.1016/j.lfs.2013.04.010
- Shirole, R. L., Shirole, N. L., and Saraf, M. N. (2015). Embelia ribes ameliorates lipopolysaccharide-induced acute respiratory distress syndrome. *J. Ethnopharmacol.* 168, 356–363. doi: 10.1016/j.jep.2015.03.009



- Singh, A. K., and Misra, A. (2020). Impact of COVID-19 and comorbidities on health and economics: focus on developing countries and India. *Diabetes Metab. Syndr.* 14, 1625–1630. doi: 10.1016/j.dsx.2020.08.032
- Sivasankarapillai, V. S., Pillai, A. M., Rahdar, A., Sobha, A. P., Das, S. S., Mitropoulos, A. C., et al. (2020). On facing the SARS-CoV-2 (COVID-19) with combination of nanomaterials and medicine: possible strategies and first challenges. *Nanomaterials (Basel)* 10:852. doi: 10.3390/nano10050852
- Sluimer, I., Schilham, A., Prokop, M., and Van Ginneken, B. (2006). Computer analysis of computed tomography scans of the lung: a survey. *IEEE Trans. Med. Imaging* 25, 385–405. doi: 10.1109/tmi.2005.862753
- Spinelli, F. R., Conti, F., and Gadina, M. (2020). Hijacking SARS-CoV-2? The potential role of JAK inhibitors in the management of COVID-19. *Sci. Immunol.* 5:eabc5367. doi: 10.1126/sciimmunol.abc5367
- Sra, H. K., Sandhu, A., and Singh, M. (2020). Use of Face Masks in COVID-19. *Indian J. Pediatr.* 87:553.
- Srinivasa Rao, A. S. R., and Vazquez, J. A. (2020). Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine. *Infect. Control Hosp. Epidemiol.* 47, 826–830. doi: 10.1017/ice.2020.61
- Stebbing, J., Phelan, A., Griffin, I., Tucker, C., Oechsle, O., Smith, D., et al. (2020). COVID-19: combining antiviral and anti-inflammatory treatments. *Lancet Infect. Dis.* 20, 400–402. doi: 10.1016/s1473-3099(20)30132-8
- Tay, M. Z., Poh, C. M., Renia, L., Macary, P. A., and Ng, L. F. P. (2020). The trinity of COVID-19: immunity, inflammation and intervention. *Nat. Rev. Immunol.* 20, 363–374. doi: 10.1038/s41577-020-0311-8
- Ting, D. S. W., Carin, L., Dzau, V., and Wong, T. Y. (2020). Digital technology and COVID-19. *Nat. Med.* 26, 459–461.
- Tiwari, M., and Mishra, D. (2020). Investigating the genomic landscape of novel coronavirus (2019-nCoV) to identify non-synonymous mutations for use in diagnosis and drug design. *J. Clin. Virol.* 128:104441. doi: 10.1016/j.jcv.2020.104441
- Tu, Y. F., Chien, C. S., Yarmishyn, A. A., Lin, Y. Y., Luo, Y. H., Lin, Y. T., et al. (2020). A review of SARS-CoV-2 and the ongoing clinical trials. *Int. J. Mol. Sci.* 21:2657. doi: 10.3390/ijms21072657
- Urahima, M., Otani, K., Hasegawa, Y., and Akutsu, T. (2020). BCG vaccination and mortality of COVID-19 across 173 countries: an ecological study. *Int. J. Environ. Res. Public Health* 17:5589. doi: 10.3390/ijerph17155589
- van Ginneken, B., Ter Haar Romeny, B. M., and Viergever, M. A. (2001). Computer-aided diagnosis in chest radiography: a survey. *IEEE Trans. Med. Imaging* 20, 1228–1241. doi: 10.1109/42.974918
- Viscusi, D. J., Bergman, M. S., Eimer, B. C., and Shaffer, R. E. (2009). Evaluation of five decontamination methods for filtering facepiece respirators. *Ann. Occup. Hyg.* 53, 815–827.
- Voysey, M., Clemens, S. A. C., Madhi, S. A., Weckx, L. Y., Folegatti, P. M., Aley, P. K., et al. (2020). Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* 397, 99–111.
- Wang, C., Li, W., Drabek, D., Okba, N. M. A., Van Haperen, R., Osterhaus, A., et al. (2020a). A human monoclonal antibody blocking SARS-CoV-2 infection. *Nat. Commun.* 11:2251.
- Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., et al. (2020b). The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J. Med. Virol.* 92, 667–674. doi: 10.1002/jmv.25762
- Wang, J., Bao, Y., Wen, Y., Lu, H., Luo, H., Xiang, Y., et al. (2020). Prior-attention residual learning for more discriminative COVID-19 screening in CT images. *IEEE Trans. Med. Imaging* 39, 2572–2583. doi: 10.1109/tmi.2020.2994908
- Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., et al. (2020). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* 30, 269–271. doi: 10.1038/s41422-020-0282-0
- Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., et al. (2020). A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Trans. Med. Imaging* 39, 2615–2625. doi: 10.1109/tmi.2020.2995965
- Wax, R. S., and Christian, M. D. (2020). Practical recommendations for critical care and anesthesiology teams caring for novel coronavirus (2019-nCoV) patients. *Can. J. Anesth.* 67, 568–576. doi: 10.1007/s12630-020-01591-x
- Weiss, S. R., and Leibowitz, J. L. (2011). Coronavirus pathogenesis. *Adv. Virus Res.* 81, 85–164. doi: 10.1016/b978-0-12-385885-6.00009-2
- Wong, H. H., Fung, T. S., Fang, S., Huang, M., Le, M. T., and Liu, D. X. (2018). Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3. *Virology* 515, 165–175. doi: 10.1016/j.virol.2017.12.028
- Wong, S. K., Li, W., Moore, M. J., Choe, H., and Farzan, M. (2004). A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2. *J. Biol. Chem.* 279, 3197–3201. doi: 10.1074/jbc.c300520200
- Xia, S., Zhang, Y., Wang, Y., Wang, H., Yang, Y., Gao, G. F., et al. (2021). Safety and immunogenicity of an inactivated SARS-CoV-2 vaccine, BBIBP-CorV: a randomised, double-blind, placebo-controlled, phase 1/2 trial. *Lancet Infect. Dis.* 21, 39–51. doi: 10.1016/s1473-3099(20)30831-8
- Xu, F., Sang, W., Li, L., He, X., Wang, F., Wen, T., et al. (2019). Protective effects of ethyl acetate extracts of *Rimulus cinnamom* on systemic inflammation and lung injury in endotoxin-poisoned mice. *Drug Chem. Toxicol.* 42, 309–316. doi: 10.1080/01480545.2018.1509987
- Yadav, P. D., Potdar, V. A., Choudhary, M. L., Nyayanit, D. A., Agrawal, M., Jadhav, S. M., et al. (2020). Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J. Med. Res.* 151, 200–209.
- Yamamoto, M., Matsuyama, S., Li, X., Takeda, M., Kawaguchi, Y., Inoue, J. I., et al. (2016). Identification of nafamostat as a potent inhibitor of middle east respiratory syndrome coronavirus s protein-mediated membrane fusion using the split-protein-based cell-cell fusion assay. *Antimicrob. Agents Chemother.* 60, 6532–6539. doi: 10.1128/aac.01043-16
- Yang, B., and Liu, P. (2014). Composition and biological activities of hydrolyzable tannins of fruits of *Phyllanthus emblica*. *J. Agric. Food Chem.* 62, 529–541. doi: 10.1021/jf404703k
- Yao, X., Ye, F., Zhang, M., Cui, C., Huang, B., Niu, P., et al. (2020). In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Clin. Infect. Dis.* 71, 732–739. doi: 10.1093/cid/ciaa237
- Ye, Q., Wang, B., and Mao, J. (2020). The pathogenesis and treatment of the 'Cytokine Storm' in COVID-19. *J. Infect.* 80, 607–613. doi: 10.1016/j.jinf.2020.03.037
- Yin, C. (2020). Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics* 112, 3588–3596. doi: 10.1016/j.ygeno.2020.04.016
- Zeng, Q. L., Yu, Z. J., Gou, J. J., Li, G. M., Ma, S. H., Zhang, G. F., et al. (2020). Effect of convalescent plasma therapy on viral shedding and survival in patients with coronavirus disease 2019. *J. Infect. Dis.* 222, 38–43. doi: 10.1093/infdis/jiaa228
- Zhang, B., Zhou, X., Qiu, Y., Song, Y., Feng, F., Feng, J., et al. (2020). Clinical characteristics of 82 cases of death from COVID-19. *PLoS One* 15:e0235458. doi: 10.1371/journal.pone.0235458
- Zhang, H., Penninger, J. M., Li, Y., Zhong, N., and Slutsky, A. S. (2020). Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Med.* 46, 586–590. doi: 10.1007/s00134-020-05985-9
- Zhang, S., Chen, X., Devshilt, I., Yun, Q., Huang, C., An, L., et al. (2018). Fennel main constituent, transanethole treatment against LPS-induced acute lung injury by regulation of Th17/Treg function. *Mol. Med. Rep.* 18, 1369–1376.
- Zhou, Y. F. B., Zheng, X., Wang, D., Zhao, C., Qi, Y., Sun, R., et al. (2020). Pathogenic T cells and inflammatory monocytes incite inflammatory storm in severe COVID-19 patients. *Natl. Sci. Rev.* 7, 998–1002. doi: 10.1093/nsr/nwaa041

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Majumdar, Verma, Saha, Bhattacharyya, Maji, Surjit, Kundu, Basu and Saha. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.





# AI Aided Design of Epitope-Based Vaccine for the Induction of Cellular Immune Responses Against SARS-CoV-2

Giovanni Mazzocco<sup>1\*</sup>, Iga Niemiec<sup>1</sup>, Alexander Myronov<sup>1,2</sup>, Piotr Skoczylas<sup>1</sup>, Jan Kaczmarczyk<sup>1</sup>, Anna Sanecka-Duin<sup>1</sup>, Katarzyna Gruba<sup>1,2</sup>, Paulina Król<sup>1</sup>, Michał Drwał<sup>1</sup>, Marian Szczepanik<sup>3</sup>, Krzysztof Pyrc<sup>4</sup> and Piotr Stępnia<sup>1</sup>

<sup>1</sup> Ardigen, Krakow, Poland, <sup>2</sup> Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, <sup>3</sup> Department of Medical Biology, Faculty of Health Sciences, Jagiellonian University Medical College, Krakow, Poland, <sup>4</sup> Virogenetics Laboratory of Virology, Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

## OPEN ACCESS

### Edited by:

Nimisha Ghosh,  
Siksha 'O' Anusandhan University,  
India

### Reviewed by:

Michael Poidinger,  
Murdoch Childrens Research  
Institute, Royal Children's Hospital,  
Australia  
Gustavo Fioravanti Vieira,  
Universidade La Salle Canoas, Brazil

### \*Correspondence:

Giovanni Mazzocco  
giovanni.mazzocco@ardigen.com

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 02 September 2020

**Accepted:** 28 January 2021

**Published:** 25 March 2021

### Citation:

Mazzocco G, Niemiec I,  
Myronov A, Skoczylas P,  
Kaczmarczyk J, Sanecka-Duin A,  
Gruba K, Król P, Drwał M,  
Szczepanik M, Pyrc K and Stępnia P  
(2021) AI Aided Design  
of Epitope-Based Vaccine  
for the Induction of Cellular Immune  
Responses Against SARS-CoV-2.  
Front. Genet. 12:602196.  
doi: 10.3389/fgene.2021.602196

The heavy burden imposed by the COVID-19 pandemic on our society triggered the race toward the development of therapies or preventive strategies. Among these, antibodies and vaccines are particularly attractive because of their high specificity, low probability of drug-drug interaction, and potentially long-standing protective effects. While the threat at hand justifies the pace of research, the implementation of therapeutic strategies cannot be exempted from safety considerations. There are several potential adverse events reported after the vaccination or antibody therapy, but two are of utmost importance: antibody-dependent enhancement (ADE) and cytokine storm syndrome (CSS). On the other hand, the depletion or exhaustion of T-cells has been reported to be associated with worse prognosis in COVID-19 patients. This observation suggests a potential role of vaccines eliciting cellular immunity, which might simultaneously limit the risk of ADE and CSS. Such risk was proposed to be associated with FcR-induced activation of proinflammatory macrophages (M1) by Fu et al. (2020) and Iwasaki and Yang (2020). All aspects of the newly developed vaccine (including the route of administration, delivery system, and adjuvant selection) may affect its effectiveness and safety. In this work we use a novel *in silico* approach (based on AI and bioinformatics methods) developed to support the design of epitope-based vaccines. We evaluated the capabilities of our method for predicting the immunogenicity of epitopes. Next, the results of our approach were compared with other vaccine-design strategies reported in the literature. The risk of immuno-toxicity was also assessed. The analysis of epitope conservation among other *Coronaviridae* was carried out in order to facilitate the selection of peptides shared across different SARS-CoV-2 strains and which might be conserved in emerging zoonotic coronavirus strains. Finally, the potential applicability of the selected epitopes for the development of a vaccine eliciting cellular immunity for COVID-19 was discussed, highlighting the benefits and challenges of such an approach.

**Keywords:** SARS-CoV-2, COVID-19, *Coronaviridae*, vaccines, cellular immunity, epitopes, CD8+, CTL

## INTRODUCTION

As of August 6, 2020, more than 19 million cases of COVID-19 were reported worldwide, leading to more than 700 thousands deaths<sup>1</sup>. The disease was first recorded on December 26, 2019, when a 41-year-old patient with no history of hepatitis, tuberculosis, or diabetes was hospitalized at the Central Hospital of Wuhan due to respiratory problems (Wu F. et al., 2020). The metagenomic RNA sequencing of bronchoalveolar lavage (BAL) fluid sample obtained from that patient led to the identification of the seventh coronavirus (CoV) strain known to infect humans.

Coronaviruses are well known human respiratory pathogens associated with the common cold. Until the 21st century they were neglected by the medical world, but the emergence and subsequent spread of the SARS-CoV in the 2002/2003 season raised interest in this virus family and increased awareness of the potential threat. At present, there are four seasonal coronaviruses infecting humans and they cluster within alphacoronaviruses (HCoV-NL63, HCoV-229E) and betacoronaviruses (HCoV-OC43, HCoV-HKU1) genera. Further, three zoonotic strains were reported – severe acute respiratory syndrome coronavirus (SARS-CoV; 2002–2003), the Middle East respiratory syndrome coronavirus (MERS-CoV; 2012–), and SARS-CoV-2 (2019–), all of which belong to the betacoronavirus genus (Wu A. et al., 2020). The highly pathogenic species cluster in two subgenera – sarbecoviruses (SARS-CoVs) and merbecoviruses (MERS-CoVs) (Hu et al., 2018; Wu F. et al., 2020; Zhou et al., 2020).

While generally, viruses infect one host, some have broader specificity or can cross the interspecies borders, causing outbreaks, epidemics, and pandemics. In this context, it is worth mentioning viruses like the Ebola virus, dengue fever virus, Nipah virus, rabies virus, or Hendra virus. However, these are well known and long studied animal viruses that only sometimes enter the human population. Coronaviruses are slightly different, as among the myriads of viral species and subspecies found in animals, it is unlikely to predict the place, the time, and the genotype of the coronavirus that will emerge. The classic transmission route of these viruses encompasses the spillover of the bat species to wild or domesticated animals, rapid evolution in this intermediate host, and subsequent transmission to humans. Coronaviruses emerge at different sites of the globe where the interaction between humans and animals is broad, such as the Asian wet markets and the dromedary camel farms in the Arabian peninsula. While these high-risk regions were identified, the next spillover may occur in Europe or the Americas, as sarbecoviruses are prevalent around the globe (Andersen et al., 2020).

The coronavirus genome is a single-stranded RNA of positive polarity, which ranges in size from 26,000 up to 32,000 bases. Two-thirds of the genome on the 5' end are occupied by two large open reading frames (ORFs) that may be read along due to the ribosomal slippery site. The resulting polyprotein undergoes subsequent autoproteolysis, and the matured proteins form the complete replicatory machinery and re-shape the microenvironment of the infection. Downstream of the 1ab ORFs, a number of ORFs are found that encode structural

and accessory proteins (Cui et al., 2019; Song et al., 2019). Four major structural proteins are: spike surface glycoprotein (S), envelope protein (E), membrane glycoprotein (M), and nucleocapsid phosphoprotein (N). Of them the S protein is the primary determinant of the species and cell tropism, interacting with the receptors and co-receptors on the host cells (Li, 2016; Zhu et al., 2020).

Evolutionary studies indicate that CoV genomes display high plasticity in terms of gene content and recombination (Forni et al., 2016). The long CoV genome expands the sequence space available for adaptive mutations, and the spike glycoprotein used by the virus to engage target cells can adapt with relative ease to exploit homologs of cellular receptors in different species. While coronaviruses are rapidly evolving, their mutation rate is lower than expected for an RNA virus. The large genomes require proofreading machinery to maintain their functions, and proteins required for such activity are among the 1a/1ab proteins.

While sarbecoviruses and merbecoviruses are associated with severe, potentially lethal diseases and are known for their epidemic potential in humans and animals, several years of research did not allow for the development of effective and safe vaccines. In addition to the high variability and ability to elude immune recognition, there are several aspects to be considered. First, the antibody-dependent enhancement (ADE) of the infection was previously reported for some coronaviruses, including sarbecoviruses. ADE is based on the fact that the virus exploits non-neutralizing antibodies to enter the host's cells utilizing the Fc receptor (FcR). The ADE phenomenon was originally observed for antibodies specific to certain dengue virus serotypes developed after a primary infection. During subsequent dengue infections, caused by a different virus serotype, these antibodies were able to recognize the virus but were not capable of neutralizing it. Instead, antibodies bridged the dengue virus and the Fc receptors of the immune cells, such as macrophages and B-cells, mediating viral entry into these cells and transforming the disease from a relatively mild illness to a life-threatening infection. A similar mechanism was later observed for HIV and Ebola infections (Takada et al., 2003, 2001; Guzman et al., 2007; Whitehead et al., 2007; Beck et al., 2008; Dejnirattisai et al., 2010; Willey et al., 2011; Katzelnick et al., 2017). Importantly, ADE has also been reported for some coronaviruses. The best-documented ADE cases are associated with feline infectious peritonitis virus. It was shown that immunization of cats with feline coronavirus spike protein leads to increased severity during future infections due to the induction of infection-enhancing antibodies (Corapi et al., 1992; Hohdatsu et al., 1998). Further, some studies show that antibodies induced by the SARS-CoV spike protein enhance viral entry into FcR-expressing cells (Kam et al., 2007; Jaume et al., 2011; Wang et al., 2014). It was confirmed that this Abs-dependent SARS-CoV entry was independent of the classical ACE2 receptor-mediated entry (Jaume et al., 2011). A recent study investigated the molecular mechanism behind antibody-dependent and receptor-dependent viral entry of MARS-CoV and SARS-CoV pseudoviruses *in vitro* (Wan et al., 2019). The authors demonstrated that MERS-CoV and SARS-CoV neutralizing monoclonal antibodies (mAbs) binding to the receptor-binding domain region of the respective spike protein

<sup>1</sup><https://coronavirus.jhu.edu/map.html>

were capable of mediating viral entry into FcR-expressing human cells, confirming the possibility of coronavirus-mediated ADE. Given the critical role of antibodies in host immunity, ADE causes serious concerns in epidemiology, vaccine design, and antibody-based drug therapy.

The consequences of ADE may be dramatic, as it may cause lymphopenia and induce or increase the frequency of the cytokine storm syndrome (CSS). This may result directly from the active infection of immune cells, which in response produce large amounts of the inflammatory markers or indirectly, when virus-antibody complex binds to FcR and activates pro-inflammatory signaling, skewing macrophages responses to the accumulation of pro-inflammatory M1 macrophages in lungs. The macrophages secrete inflammatory cytokines, such as MCP-1 and IL-8, which lead to worsened lung injury (Fu et al., 2020). In both animal models and patients who eventually died from SARS, extensive lung damage was associated with high initial viral loads, increased accumulation of inflammatory monocytes/macrophages in the lungs, and elevated levels of serum pro-inflammatory cytokines and chemokines (IL-1, IL-6, IL-8, CXCL-10, and MCP1) (Channappanavar et al., 2016). Moreover, during the SARS-CoV outbreak in Hong Kong (2003–2004), 80% of the patients developed acute respiratory distress syndrome after 12 days from the diagnosis, coinciding with IgG seroconversion (Peiris et al., 2003). Another study by Huang et al. (2020) highlighted an increased release of IL-1 $\beta$ , IL-4, IL-10, IFN $\gamma$ , MCP-1, and IP-10 in COVID-19 patients. Interestingly, compared with non-severe cases, severe patients in the intensive care unit showed higher plasma concentrations of TNF $\alpha$ , IL-2, IL-7, IL-10, MIP-1A, MCP-1, and G-CSF, supporting the hypothesis of a possible correlation between CSS and severity of the disease. An extensive study done by Liu et al. (2019) demonstrated that anti-spike IgGs enhanced the induction of pro-inflammatory cytokines (i.e., IL-6, IL-8, and MPC-1) in Chinese rhesus monkeys through the stimulation of alternatively activated monocyte-derived macrophages (MDM) upon SARS-CoV rechallenge. The presence of high MDM infiltrations was shown by histochemical staining of the lung tissue from 3 deceased SARS patients. The blockade of Fc-receptors for IgG (Fc $\gamma$ Rs) reduced proinflammatory cytokine production, suggesting a potential role of Fc $\gamma$ Rs for the reprogramming of alternatively activated macrophages. Putting these results in the context of other works in literature (Pahl et al., 2014), one has to consider that anti-S IgG may promote pro-inflammatory cytokine production and, consequently, CSS development.

Taking into account the risk associated with the improper humoral response and high variability of sites targeted by the neutralizing antibodies, together with the low effectiveness of IgG-mediated immunity during mucosal infection, it is of importance to consider the anticoronaviral vaccine in a broader perspective. This may include alternative delivery systems/routes based on, e.g., virus-like particles and intranasal delivery for the IgA mediated response, but it is also important to consider combining the humoral response with the cell-mediated response. Ideally, such an approach might allow for the design of a vaccine carrying carefully selected epitopes to induce only the neutralizing antibodies and epitopes targeted for induction

of the cellular response. While neutralizing antibodies impair the virus entry, activated CD8+ cytotoxic T-cells can identify and eliminate infected cells. Moreover, CD4+ helper T-cells are required to stimulate the production of antibodies. Antibody response was found to be short-lived in convalescent SARS-CoV patients (Tang et al., 2011) in contrast to T-cell responses, which have been shown to provide long-term protection (Peng et al., 2006; Fan et al., 2009; Tang et al., 2011), up to 11 years post-infection (Ng et al., 2016). The activation of CD8+ cytotoxic T-cells capable of recognizing and destroying infected cells represents a crucial second line of defense against the virus that should be considered. The importance of both CD8+ and CD4+ T-cell activation has been reported in several SARS-CoV studies for both animal models and humans (Channappanavar et al., 2014). Moreover, several recent studies indicate a strong correlation between the reduction of lymphocyte counts (CD4+ and CD8+) and the severity of COVID-19 cases (Chen et al., 2020; Liao et al., 2020; Wan et al., 2020).

The selection of epitopes capable of eliciting either B-cell or T-cell responses is a critical step for the development of subunit vaccines. Most of the efforts in this area are directed toward the stimulation of neutralizing antibodies, whereas the cellular immune response is less explored. Considering the importance of T-cell activation for vaccine efficacy, the focus of the work here presented is on the latter. Despite the apparent similarity between SARS-CoV and SARS-CoV-2, there is still a considerable genetic variation between these two. Thus, it is not trivial to assess if epitopes eliciting an immune response against previous coronaviruses are likely to be effective against SARS-CoV-2, with the exception of identical peptides shared among subgenera. A restricted list of SARS-CoV epitopes identical to those present in SARS-CoV-2 and resulting positive in immunoassays, has been recently reported (Ahmed et al., 2020). Nonetheless, the 29 T-cell epitopes described therein are mostly limited to S, N, and M antigens and encompass an exiguous number of Class I Human Leukocyte Antigen (HLA) alleles. In order to extend the search area to other epitopes, computational predictive models might be applied. Methods for the selection of vaccine peptides are typically based on the predicted binding affinity (or probability of presentation on the cell surface) of peptide-HLA (pHLA) complexes or defined by the physicochemical properties of the peptides (Baruah and Bose, 2020; Grifoni et al., 2020; Lee and Koohy, 2020). These methods take into account only restricted parts of processes contributing to the final immunogenicity of an epitope, and thus their prediction capabilities are limited. In addition to pHLA binding, proteasome cleavage, pHLA loading, and presentation, as well as direct activation of CD8+ T-cell to the pHLA complex should be taken into account.

Here, we use a machine learning model for the prediction of epitope immunogenicity. The model is trained on data including the experimental T-cell immunogenicity data of viral epitopes. We validate our model on publicly available immunogenicity data of epitopes from the *Coronaviridae* virus family (held out from training). Assessment of the risk of immuno-toxicity and the analysis of epitope conservation among different strains are also performed.

## MATERIALS AND METHODS

### Presentation Data

A curated dataset containing peptides presented by class I HLAs on the surface of host cells was extracted from publicly available databases (Abelin et al., 2017; Di Marco et al., 2017; Sarkizova et al., 2020). The presentation of each peptide within the dataset was experimentally confirmed by mass-spectroscopy experiments. All peptides were of human origin and were presented on the surfaces of monoallelic human cell lines (see **Figure 1** and **Table 1**). Synthetic negative data (non-presented peptides) were also prepared based on human proteome (GRCh38, release 98).

### Immunogenicity Data

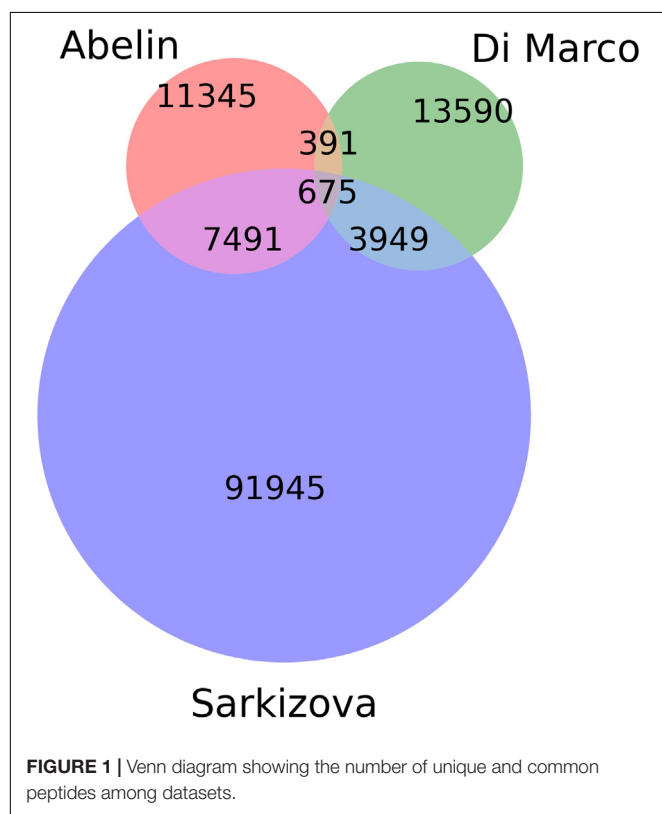
All peptides collected from the IEDB database (Vita et al., 2019) were of viral origin and were confirmed in experimental immunoassays. Similar data were extracted from selected publications (Wang et al., 2004; Chen et al., 2005; Tsao et al., 2006; Liu et al., 2010; Zhang, 2013; Ogishi and Yotsuyanagi, 2019). The number of pHLAs (per immunoassay category) used

for training is given in **Table 2**. Most of the peptides were obtained from human hosts, with a minority obtained from transgenic mice. Only peptides containing 8–11 amino acids were included in the analysis. In some cases, multiple experimental settings and protocols were used to validate immunogenicity for a given pHLA, occasionally leading to non-consensual results. Each pHLA was considered immunogenic if at least one experiment conducted on human cells positively confirmed that immunological event. If no experiments conducted on human cells were available, the pHLA was considered immunogenic, if at least one such confirming experiment was conducted in transgenic mice. The remaining pHLAs were used as negative examples. From this dataset we held out the *Coronaviridae* family as a separate test set.

### Predictive Model Design

Our computational methods are based on machine learning and predict (1) the probability of pHLAs to be presented on the host's cell surface and (2) the immunogenicity of such complexes. The model for pHLA presentation is based on artificial neural networks and has been trained on a curated collection of peptide presentation data (Abelin et al., 2017; Di Marco et al., 2017; Sarkizova et al., 2020). Both peptide sequence and HLA type were taken into consideration as separate inputs. We use bootstrapping and select 80% of positive examples during training with the remaining ones used for early stopping. We then ensemble the results of a collection of 27 such neural networks. Our model is pan-specific and can be used to generate predictions for any peptide and any canonical class I HLA (i.e., A, B, and C). Note, that the accuracy of our method depends on the considered HLA type, as in the case of other machine learning methods for predicting pHLA properties.

The model mentioned above was also used as a starting point for training the immunogenicity model. The latter was fine-tuned using the viral peptide immunogenicity data collected from IEDB (Vita et al., 2019) and Ogishi and Yotsuyanagi (2019). The immunogenicity model was validated using a Leave One Group Out (LOGO) cross-validation scheme with groups defined by viral families. The final model is an ensemble of 11 models – one per each LOGO split. An additional group “others” was defined by aggregating data from viruses that belong to several families, having a small number of observations. Such an approach provides data splits according to the virus families and leads to a better measure of performance on virus families not seen in training (e.g., *Coronaviridae*). Moreover, it reveals the differences in model performance on various virus families.



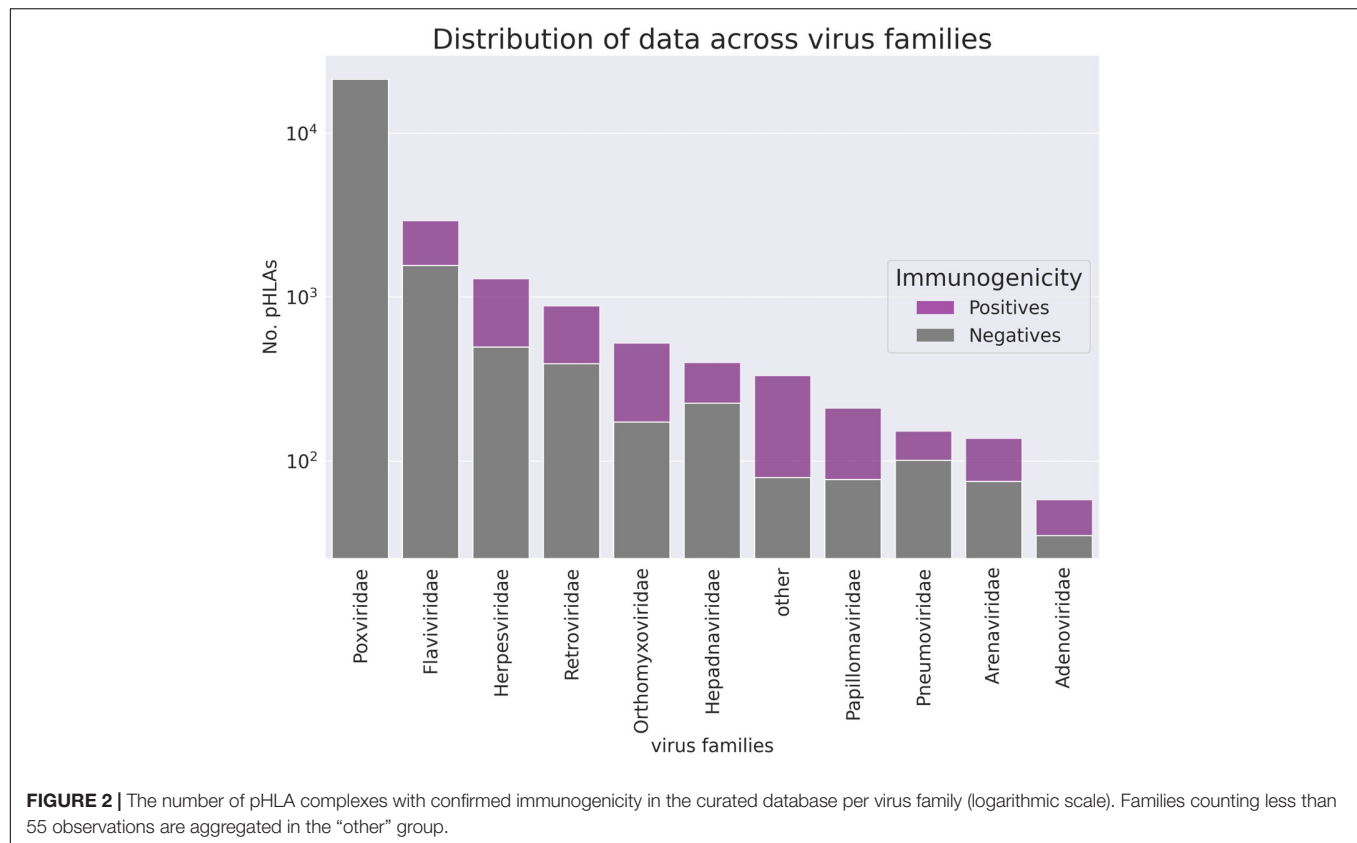
**TABLE 1** | The total number of pHLAs included in our model from each dataset.

Source publication	No. pHLAs
Abelin et al. (2017)	22,999
Di Marco et al. (2017)	22,889
Sarkizova et al. (2020)	146,739

**TABLE 2** | The number of pHLA complexes used for training per immunogenic assay group.

Source publication	Negative	Positive
IFN( $\gamma$ )	23,249	2,598
Cytotoxicity	218	524
Proliferation	7	34
cytokines/chemokines	0	13
TNF( $\alpha$ )	1	8





The final predictions of our model (called ArdImmune Rank) are obtained by combining the predictions of both models (i.e., the pHLA presentation and the immunogenicity model).

Both models were implemented in Python 3.7 using the keras 2.4.3 package, which is a high-level API of TensorFlow. For our usage TensorFlow with GPU support was deployed, i.e., tensorflow-gpu 2.2.0. For GPU-based computations we used cudnn 7.6.5 and cudatoolkit 10.2.89 and a machine equipped with NVIDIA Tesla V100 GPU card with CUDA® 7.0 architecture, 640 Tensor Cores, 5,120 CUDA® Cores and 32 GB HBM2 GPU Memory. Additionally, scikit-learn, pandas, and numpy were used to perform standard machine learning tasks while images were produced using matplotlib and seaborn.

## Validation Scheme

In order to validate the ArdImmune Rank model over different virus families not seen during the training procedure, a LOGO strategy was applied. The peptides associated with coronaviruses were held out from the dataset and left for testing purposes only. At each LOGO iteration, the dataset was split into training and validation sets, and the model was tested accordingly. Peptides within the training set highly similar to the ones in the validation set were removed from the training set. The similarity of peptides was assessed using a clustering algorithm classifying their sequences into groups of peptides sharing a common root (differing only by short prefixes or suffixes of lengths of at most three amino acids). The number of pre-processed peptides in

each group is given in **Figure 2**. Finally, the immunogenicity model (an ensemble of 11 models from the LOGO scheme) was validated on the held-out *Coronaviridae* dataset.

## SARS-CoV-2 Data Analysis

### Selection of HLA Alleles

Class I HLA types were chosen based on their frequency of occurrence in the United States and Europe. HLA-allele frequency data were downloaded from<sup>2</sup>, accounting for all the populations within the regions of choice and all ethnicities. The overall frequency for each allele was computed as the weighted average with weights corresponding to the size of each population, separately for the United States and Europe, encompassing all ethnic populations. All HLA-alleles with frequency  $\geq 0.01$  were chosen for the study.

### Toxicity/Tolerance Evaluation

In order to evaluate the risk for a given pHLA to be cross-reactive or tolerogenic with respect to self-epitopes within the human proteome, a procedure for the evaluation of potential toxicity/tolerance was implemented. Initially, each SARS-CoV-2 peptide was queried against the reference human proteome (GRCh38, release 100) using the BLASTp algorithm and a BLOSUM45 substitution matrix. All matches with *E*-values less than or equal to four were included in the analysis. The selected peptides are available in **Supplementary Data 1**.

<sup>2</sup><http://www.allelefrequencies.net/>

## Selection of Peptides

The dataset consisting of SARS-CoV-2 peptides was generated according to the following procedure: (1) all the reference sequences of the virus proteins were collected from the NCBI database<sup>3</sup>, (2) from each protein, all possible peptides of length 8–11 amino acids were selected. In addition, for proteins encoded by the ORF1a and ORF1ab genes (i.e., pp1a, pp1ab, respectively), the peptides within the cleavage sites were excluded. Finally, all the peptide duplicates were removed from the dataset. A total of 47,612 peptide sequences were collected.

## Estimation of SARS-CoV-2 Genome Diversity

The analysis of conservation of SARS-CoV-2 genomic sequences was performed using 8,639 complete genomic sequences obtained from the GISAID database<sup>4</sup> and GenBank<sup>5</sup>. All sequences were aligned to the SARS-CoV-2 reference genome (NCBI Reference Sequence: NC\_045512.2). The R DECIPHER package (Wright, 2015) v2.14.0 was used to perform the multiple sequence alignment (MSA) of long SARS-CoV-2 whole genome sequences. The following parameters were applied: AlignSeqs(sequences, iterations = 2, refinements = 1, gapOpening = c(-18, -16), gapExtension = c(-2, -1), FUN = AdjustAlignment, processors = 18). In order to align short sequences with partial fragments of the SARS-CoV-2 genome, the R Biostrings v2.54.0 package was used, adopting the following parameters: Biostrings:pairwiseAlignment(pattern = sequences, subject = reference\_genome, type = "local," and scoreOnly = F). Next, all the nucleotides within the coding cDNA sequence (CDS) regions of the reference genome were translated into amino acids using the *translate* function available in the R Biostring package v2.45.0 (Pagès and DebRoy, 2020) with the following parameters: Biostrings:translate[DNAStringSet(sequences), if.fuzzy.codon = "solve"]. The Standard Genetic Code provided by default was used for the encoding. All the fuzzy codons were marked as unknown amino acids by setting the *if.fuzzy.codon* = "solve" parameter. For each protein, all sequences containing indels or being inconveniently aligned were removed. Inconvenient sequences include those having short reading frameshifts, marked as transcription artifacts. Mutation frequencies for both long and short genomics fragments were computed for each amino acid in the SARS-CoV-2 proteome. The mutation frequency of each amino acid was defined as the ratio between the number of translated protein sequences containing the mutation and the number of sequences containing a valid nucleotide (sequences containing unknown nucleotides in this position were excluded). The maximum mutation frequency score for each peptide was computed as the maximum value of the mutation frequency scores among all amino acid positions of the peptide. Mutation frequency values for all positions within SARS-CoV-2 proteome are available in **Supplementary Data 2**.

<sup>3</sup><https://www.ncbi.nlm.nih.gov/search/all/?term=SARS-CoV-2>

<sup>4</sup>[https://bigd.big.ac.cn/ncov/release\\_genome?lang=en](https://bigd.big.ac.cn/ncov/release_genome?lang=en)

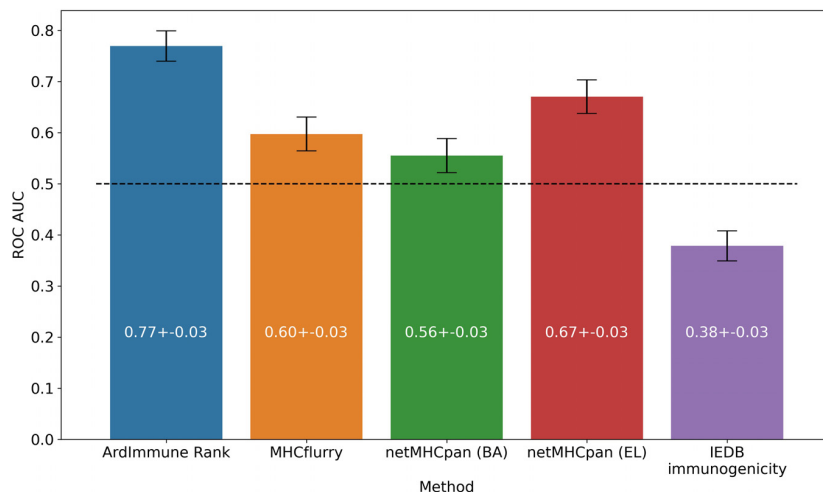
<sup>5</sup><https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/>

## Datasets for External Comparison

In order to highlight similarities and differences of our approach with respect to other methods, we compare the scores of our model with scores relative to the same pHLAs reported in a list of selected studies. A peptide missing from the reference proteome ("QSADAQSFLNR") was removed. Only peptides between 8 and 11 amino acids were considered. The peptides arising from the cleavage sites of the ORF1a/ab polyprotein were also removed from the datasets. These sites are defined as nucleic acids within the NCBI reference sequence: NC\_045512.2 but outside the range of the ORF1a/ab coding sequences.

The ORF1a and ORF1ab cleavage sites were corrected for reading frameshift which occurs for ORF1ab (as opposed to ORF1a), when independently translating RNA polymerase and nsp11, respectively.

- Baruah and Bose (2020): Five epitopes from the surface glycoprotein of SARS-CoV-2 and their corresponding HLA class I supertype representative were reported by the authors (Table 1 in the reference publication). Bioinformatics protocols, machine learning methods, and structural analysis were applied in the original paper for the selection of these pHLAs.
- Lee and Koohy (2020): 19 A\*02:01 restricted epitopes were selected applying TCR-specific Position Weight Matrices (PWM) previously published by the authors. The geometric mean of the three scores was used as an estimator for immunogenicity (Tables 4, 5 in the reference publication).
- Grifoni et al. (2020):
  - 1st dataset: 386 SARS-CoV-2 CD8+ predicted epitopes were collected (Supplementary Table 6 in the reference publication) and 41 peptides were excluded as a result of our filtering procedure.
  - 2nd dataset: 28 SARS-CoV-2 CD8+ epitopes mapped to immunodominant SARS-CoV epitopes were selected (Table 5 in the reference publication). One peptide was excluded as a result of our filtering procedure.
- Gupta et al. (2020): 10 HLA-A\*11:01 restricted peptides from the surface glycoprotein of SARS-CoV-2 were selected by the authors (Table 4 in the reference publication). Bioinformatics protocols, machine learning methods, and structural analysis were used for the selection of those pHLAs. A candidate with an optimal docking score is reported.
- Prachar et al. (2020): 138 peptides with pHLA complex stability measurements performed using Immunotrack's NeoScreens<sup>®</sup> assay were made available by the authors. A peptide absent in our dataset was excluded from the comparison.
- Rammensee et al. (2020): 5 HLA class I peptides were used by the authors for the experimental vaccination of self-experimenting healthy volunteers. IFN $\gamma$  ELISPOT assays for the measurement of CD8+ activation were negative for all these peptides.



**FIGURE 3 |** Predictive performance of the selected models on the *Coronaviridae* dataset. ArdImmune Rank, blue bars; MHCflurry, orange bars. netMHCpan, green and red bars for the predicted binding affinity (BA) and ligand likelihood (EL); IEDB immunogenicity, purple bars.

- Smith et al. (2020): Predictions for ~615 k peptides were extracted from the Supplementary Table 1 of the reference publication. Approximately 7,600 peptides were excluded as a result of our filtering procedure.

The ArdImmune Rank percentile rank for the pHLAs described in the above datasets was computed for groups of peptides according to their HLA allele. Only pHLAs with a binding affinity percentile rank score < 0.02 (computed using NetMHCpan 4.0) were considered. The predictions were calculated separately for peptides of structural and non-structural origin.

## RESULTS

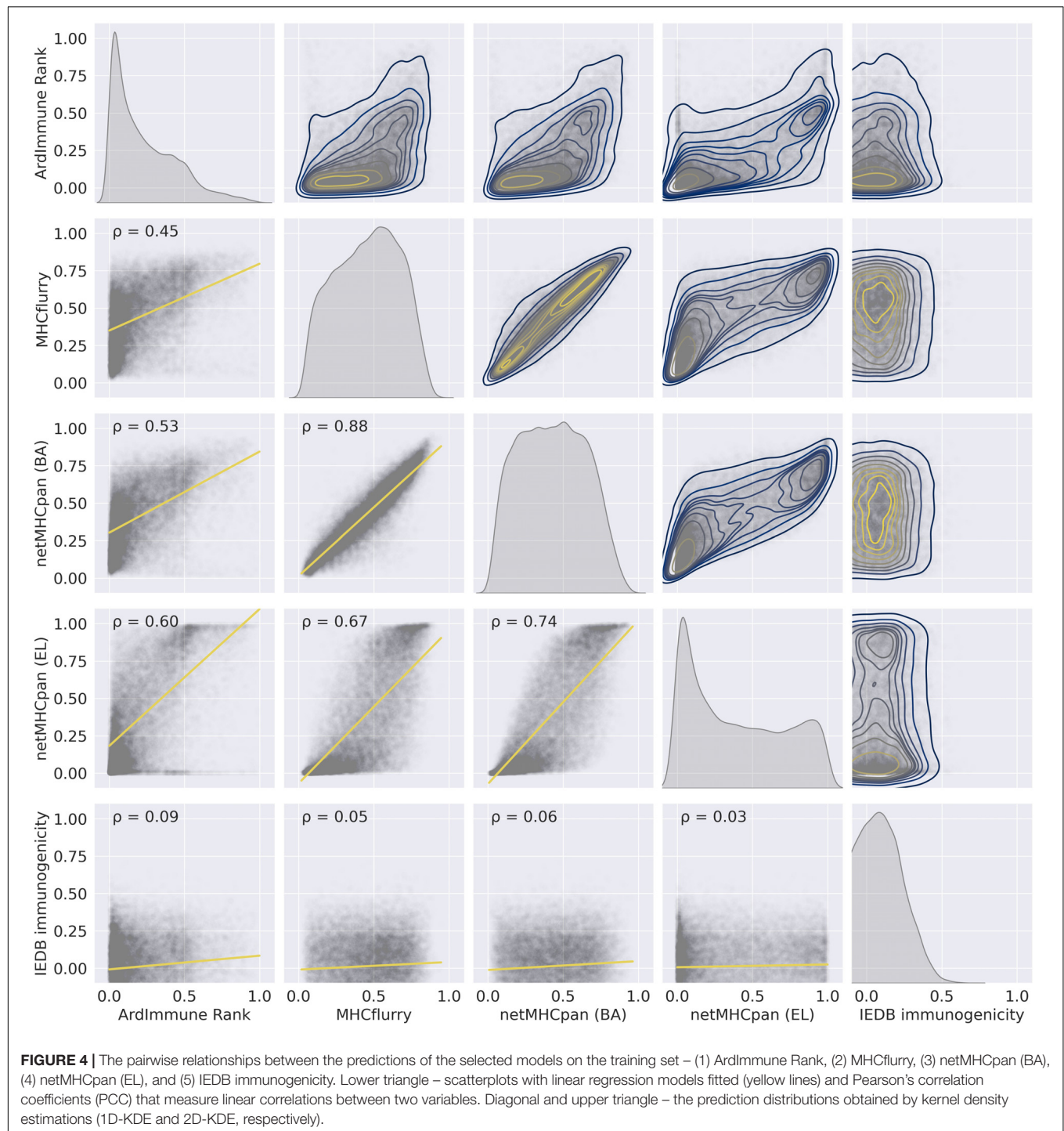
### Model Performance

The performance of our method on the test set encompassing *Coronaviridae* epitopes (excl. SARS-CoV-2 epitopes) is shown in **Figure 3**. In addition, the results of our approach are compared to those obtained by other commonly used pHLA binding affinity and pHLA presentation probability predictors, namely netMHCpan 4.0 (Jurtz et al., 2017) and MHCflurry (O'Donnell et al., 2018), as well as the IEDB immunogenicity predictor, version 3.0 (Calis et al., 2013). For both binding affinity tools [MHCflurry and netMHCpan (BA)], the binding affinity predictions in nanomoles (nM) are converted into (0, 1) range with a widely used logarithmic transformation [i.e., first the predictions are bounded from above by 50,000 nM and from below by 1 nM and then transformed with  $\left(1 - \frac{\log_{10}x}{\log_{10}50,000}\right)$ ]. The difference in the predictive performance (measured with ROC AUC) of our model with respect to the other methods is statistically significant (and ranges from 0.10 to 0.39). Moreover (as verified on our training dataset across virus families), the high Pearson correlation between the results produced by the binding predictors (corr. coeff.  $\rho = 0.88$ ) and the low correlation of such results with the predictions of our model ( $\rho = 0.45$  and  $\rho = 0.53$ )

demarcate substantial differences between our approach and the approaches based on those methods for predicting immunogenic epitopes (see **Figure 4**).

We apply the LOGO cross-validation scheme according to the procedure described in the Materials and methods section. While we observe a significant variation in ROC AUC scores depending on the tested groups (i.e., virus families), the performance of each method is not correlated with the number of observations within each group. The *Pneumoviridae* family might be an outlier in our dataset as the predictive performance of all the considered models are substantially different for this family than those observed for the other families. Although some groups display a noticeable correlation between pHLA immunogenicity and pHLA binding affinity predictions (e.g., *Pneumoviridae* and *Orthomyxoviridae*), this trend is not confirmed across all groups. The performance (median ROC AUC across virus families) of our method is comparable to those obtained for binding affinity and ligand likelihood predictors, usually with a smaller variance of prediction performance (see **Figures 5, 6**).

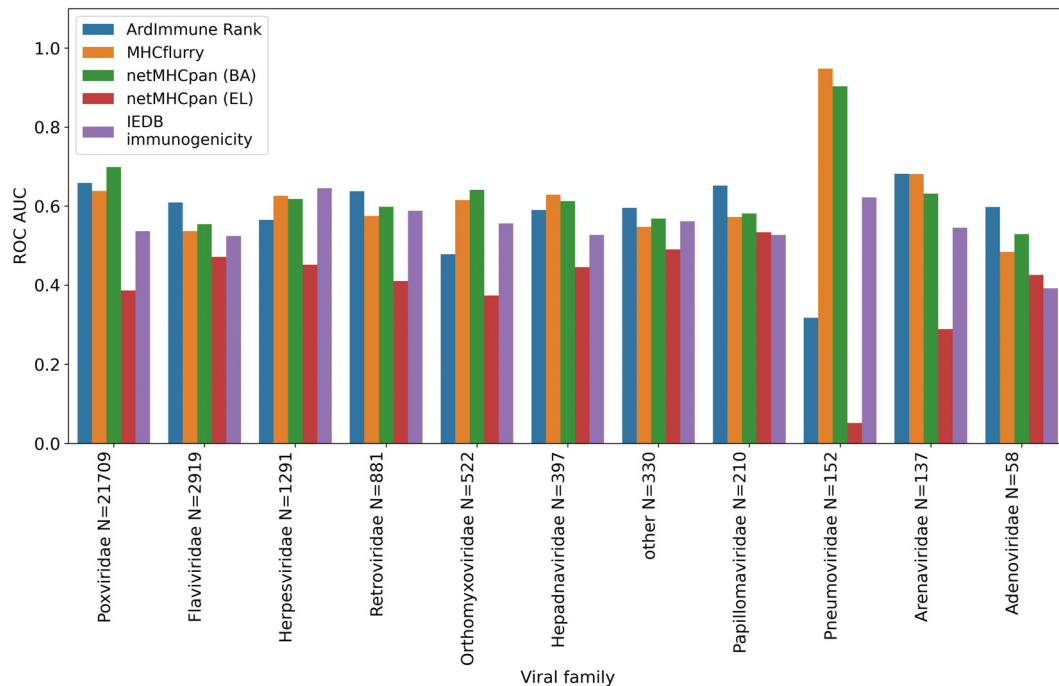
Note that the *Coronaviridae* dataset (**Figure 3**) is the most relevant dataset to the problem at hand, but it also is a very small dataset containing 67 epitopes. Hence, the variation of performance of the selected methods is expected to be high and their performance on the training set (**Figures 5, 6**) might be different (note also that in the LOGO validation – in **Figures 5, 6** – we use a single immunogenicity model instead of 11 models, as in **Figure 3**). On the other hand, evaluation on the *Coronaviridae* dataset might still reflect performance of the selected methods on the epitopes from the SARS-CoV-2 genome. The dataset encompassing all other virus families used in our LOGO cross-validation procedure (training dataset) is much larger, but is also very heterogeneous. For example the *Poxviridae* family contains predominantly *Vaccinia* virus, which is a model organism with mostly non-immunogenic epitopes reported in IEDB. Namely, there are 1.6% immunogenic observations in the *Poxviridae* family, whereas for *Herpesviridae* 62% of the



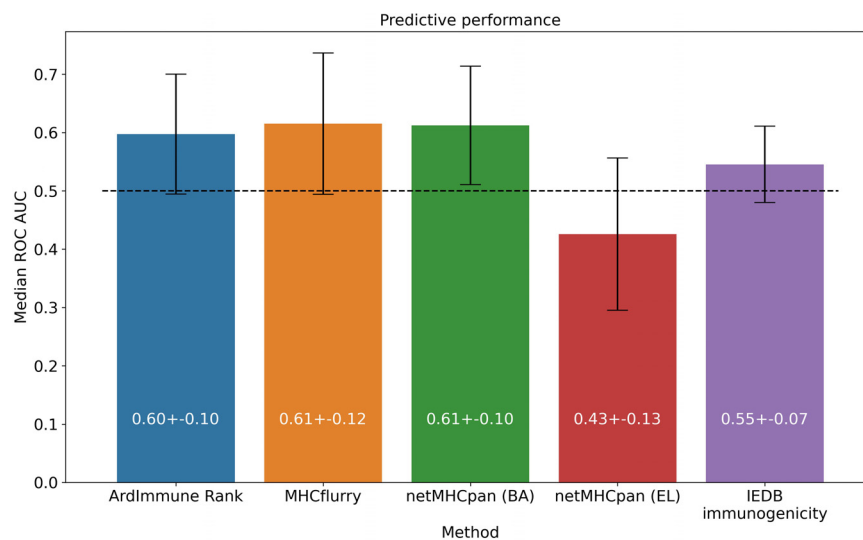
observations are immunogenic. Moreover, IEDB observations are very small in size within some families (e.g., Adenoviridae with  $N = 58$ ) and much larger in others (e.g., Poxviridae with  $N = 21709$ ). In such a situation, the large variance of performance of predictive methods when evaluated on different viral families is expected and originates from both the underlying biological and experimental factors, as well as from the small number of observations for some virus families.

The model was then used to predict the immunogenicity of peptides from the SARS-CoV-2 proteome. Target peptides and HLA types considered for the analysis were selected according to the procedure described in the “Selection of peptides” and “Selection of HLA alleles” sections, respectively. A considerable number of peptides with high scores are observed in both structural and non-structural proteins, encompassing different HLA alleles. Structural epitopes are dominated by the





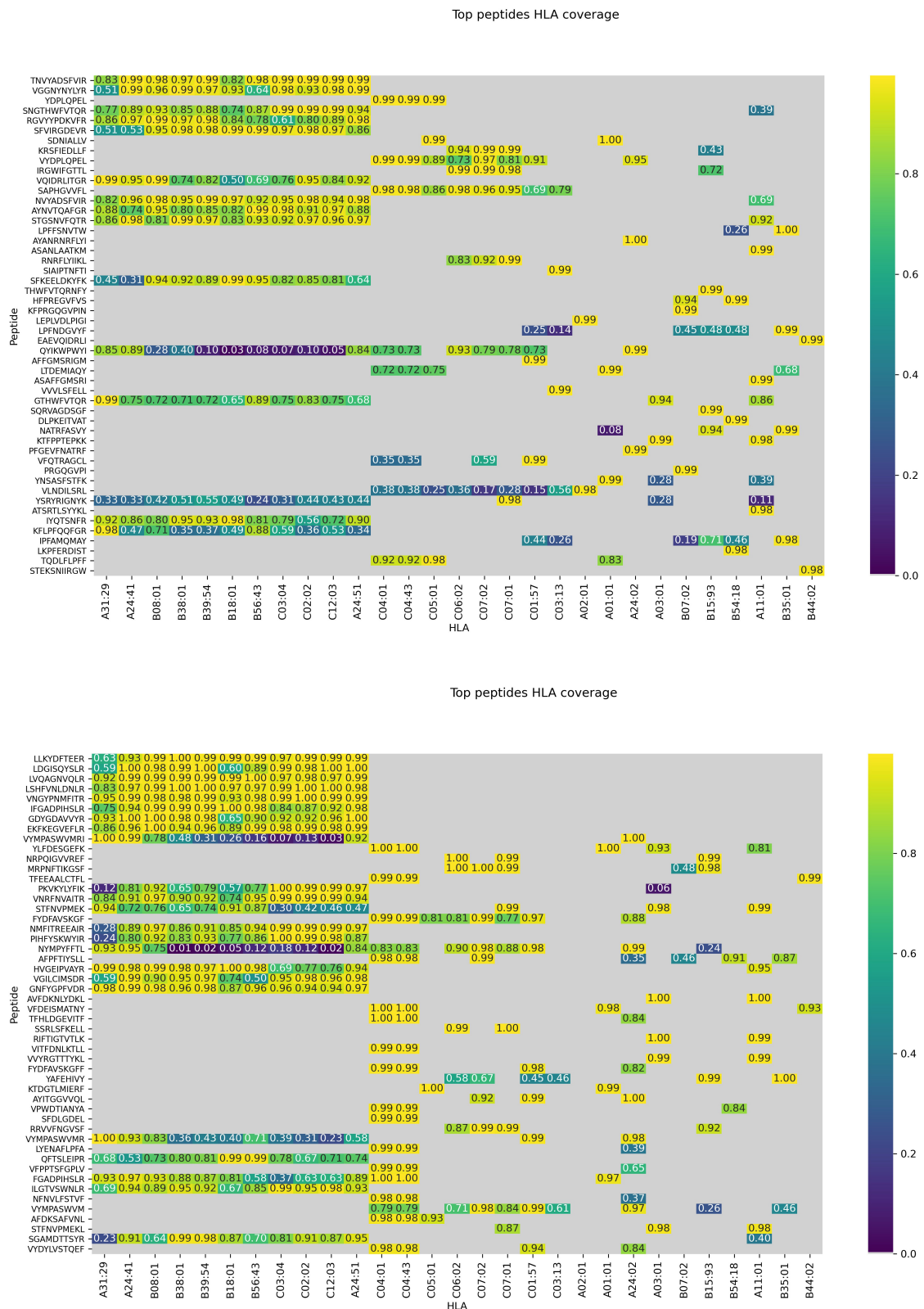
**FIGURE 5 |** Predictive performance of the selected models obtained in a LOGO cross validation and measured with ROC AUC. ArdImmune Rank, blue bars; MHCflurry, orange bars; NetMHCpan (BA), green bars; NetMHCpan (EL), red bars; IEDB immunogenicity, purple bars.



**FIGURE 6 |** Predictive performance of the selected models, averaged across virus groups in the training dataset.

Spike protein, whereas the non-structural ones mostly originate from the ORF1a/ORF1ab-encoded polyproteins. Peptides with percentile rank  $\leq 2$  presented across the selected HLAs, were considered for both structural (Table 3) and non-structural (Table 4) viral proteins. We noticed that some HLA alleles exhibit a large number of highly-ranked peptides, in particular A\*02:01, A\*11:01, A\*24:41 and C\*12:03. Interestingly, the presence of some of these alleles was earlier reported to be

statistically correlated with the immune protection in SARS cases. Namely, A\*02:01 was found to present immunogenic peptides (Ahmed et al., 2020; Lee and Koohy, 2020) whereas A\*11:01-restricted epitopes were proposed to be included in a SARS-CoV vaccine by Sylvester-Hvid et al. (2004). Groups of peptides predicted to be associated with multiple HLAs are shown in Figure 7. These epitopes originate from both structural and non-structural antigens.



**FIGURE 7 |** Peptides presented across multiple HLAs. Immunogenicity scores are reported for epitopes from both structural (**top**) and non-structural (**bottom**) proteins. Peptide-HLA combinations marked in gray are predicted non-binders (netMHCpan 4.0 percentile rank > 2). For the remaining pHLAs, the color relates to the percentile rank of our predictions for a given HLA type (0.95 means that the prediction is among top 5% of the predictions for that particular HLA allele).

**TABLE 3 |** Peptides with ArdImmune Rank percentile rank  $\leq 2$  obtained from SARS-CoV-2 structural proteins, sorted by (1) the number of HLA types capable of binding and presenting given peptide and (2) the median rank across different HLA types.

No.	Peptide	Prot. start	Prot. end	Protein	HLA% rank $\leq 2$	Median HLA%_rank	Max mut. freq
1	TNVYADSFVIR	393	403	S	0.994	A24:41  A24:51  B39:54  C02:02  C03:04  C12:03	0.00012
2	VGGNYNYLYR	445	454	S	0.989	A24:41  A24:51  B38:01  C12:03	0.00013
3	YDPLQPEL	1,138	1,145	S	0.995	C04:01  C04:43  C05:01	0.00049
4	SNGTHWFTQQR	1,097	1,107	S	0.989	C02:02  C03:04  C12:03	0.00012
5	RGVYYPDKVFR	34	44	S	0.983	A24:51  B08:01  B39:54	0.00012
6	SFVIRGDEVIR	399	408	S	0.989	B18:01  B56:43  C02:02	0.00012
7	SDNIALLV	214	221	M	0.995	A01:01  C05:01	0.00047
8	KRSFIEDLLF	814	823	S	0.99	C07:01  C07:02	0.00024
9	VYDPLQPEL	1,137	1,145	S	0.989	C04:01  C04:43	0.00049
10	IRGWIFGTTL	101	110	S	0.992	C06:02  C07:02	0.00012
11	VQIDRLITGR	991	1,000	S	0.992	A31:29  B08:01	0.00000
12	SAPHGWFL	1,055	1,063	S	0.984	C04:01  C04:43	0.00024
13	NVYADSFVIR	394	403	S	0.986	B08:01  B39:54	0.00012
14	AYNVTAFAFR	267	276	N	0.989	B56:43  C03:04	0.00035
15	STGSNVFQTR	637	646	S	0.986	A24:41  B38:01	0.00000
16	LPFFSNVTW	56	64	S	0.996	B35:01	0.00024
17	AYANRNRFLYI	38	48	M	0.995	A24:02	0.00058
18	ASANLAATKM	1,020	1,029	S	0.995	A11:01	0.00024
19	RNRFLYIIKL	42	51	M	0.995	C07:01	0.00023
20	SIAPTNFTI	711	720	S	0.995	C03:13	0.00024
21	SFKEELDKYFK	1,147	1,157	S	0.994	B18:01	0.00049
22	THWFTQQRNFY	1,100	1,110	S	0.994	B15:93	0.00012
23	HFPREGVFS	1,088	1,097	S	0.994	B54:18	0.00012
24	KFPRGQGVPI	65	75	N	0.993	B07:02	0.00035
25	LEPLVDLPIGI	223	233	S	0.992	A02:01	0.00000
26	LPFNDGVYF	84	92	S	0.991	B35:01	0.00049
27	EAEVQIDRLI	988	997	S	0.991	B44:02	0.00000
28	QYIKWPWYI	1,208	1,216	S	0.991	A24:02	0.00024
29	AFFGMSRIGM	313	322	N	0.991	C01:57	0.00071
30	LTDEMAQY	865	873	S	0.99	A01:01	0.00024
31	ASAFFGMSRI	311	320	N	0.99	A11:01	0.00012
32	VVLSFELL	510	518	S	0.989	C03:13	0.00013
33	GTHWFTQQR	1,099	1,107	S	0.989	A31:29	0.00012
34	SQRVAGDSGF	184	193	M	0.989	B15:93	0.00000
35	DLPKEITVAT	163	172	M	0.988	B54:18	0.00012
36	NATRFASVY	343	351	S	0.987	B35:01	0.00024
37	KTFPPTPEPKK	361	370	N	0.993	A03:01	0.00036
38	PFGEVFNATRF	337	347	S	0.986	A24:02	0.00024
39	VFQTRAGCL	642	650	S	0.986	C01:57	0.00012
40	PRGQGVPI	67	74	N	0.986	B07:02	0.00035
41	YNSASFSTFK	369	378	S	0.986	A01:01	0.00025
42	VLNDILSRL	976	984	S	0.984	A02:01	0.00012
43	YSRYRIGNYK	196	205	M	0.984	C07:01	0.00012
44	ATSRTLSEYK	171	181	M	0.984	A11:01	0.02876
45	IYQTSNFR	312	319	S	0.983	B18:01	0.00014
46	KFLPFQFGR	558	567	S	0.983	A31:29	0.00036
47	IPFAMQMAY	896	904	S	0.982	B35:01	0.00000
48	LKPFERDIST	461	470	S	0.982	B54:18	0.00025
49	TQDLFLPFF	51	59	S	0.982	C05:01	0.00292
50	STEKSNIIRGW	94	104	S	0.982	B44:02	0.00073

Peptides marked in red are considered as highly variable (HV) due to maximum mutation frequency score  $\geq 0.05$ .

**TABLE 4 |** Peptides with model percentile rank  $\leq 2$  obtained from SARS-CoV-2 non-structural proteins, sorted by (1) the number of HLA types capable of binding and presenting given peptide and (2) the median rank across different HLA types.

No.	Peptide	Prot. start	Prot. end	Protein	HLA% rank $\leq 2$	Median HLA%_rank	Max mut. freq
1	LLKYDFTEER	4,662	4,671	ORF1ab	0.991	A24:51  B08:01  B18:01  B38:01  B39:54  B56:43  C02:02  C12:03	0.00012
2	LDGISQYSLR	570	579	ORF1a	0.997	A24:41  A24:51  B08:01  B38:01  B39:54  C03:04  C12:03	0.00372
3	LVQAGNVQLR	3,330	3,339	ORF1a	0.993	A24:41  A24:51  B08:01  B18:01  B38:01  B39:54  B56:43	0.00565
4	LSHFVNLDNLR	2,518	2,528	ORF1a	0.997	A24:51  B08:01  B38:01  B39:54  C02:02  C03:04  C12:03	0.00414
5	VNGYPNMFITR	5,991	6,001	ORF1ab	0.995	A24:41  A24:51  B39:54  C02:02  C03:04  C12:03	0.00036
6	IFGADPIHSLR	1,153	1,163	ORF1a	0.993	B08:01  B18:01  B38:01  B39:54  B56:43	0.00332
7	GDYGDVAVYR	5,527	5,536	ORF1ab	0.997	A24:41  A24:51  B08:01  B38:01  B39:54	0.00084
8	EKFKEGVFLR	633	643	ORF1a	0.986	A24:51  B08:01  B56:43  C02:02  C03:04	0.00371
9	VYMPASWVMRI	3,653	3,663	ORF1a	0.998	A24:02  A24:41  A31:29	0.00412
10	YLFDESGEFK	906	915	ORF1a	0.995	A01:01  C04:01  C04:43	0.00413
11	NRPQIGVREF	5,813	5,823	ORF1ab	0.993	B15:93  C06:02  C07:01	0.00024
12	MRPNFTIKGSF	3,393	3,403	ORF1a	0.997	C06:02  C07:01  C07:02	0.00425
13	TFEEAALCTFL	3,174	3,184	ORF1a	0.992	B44:02  C04:01  C04:43	0.00399
14	PKVKYLYFIK	4,223	4,232	ORF1a	0.993	C02:02  C03:04  C12:03	0.00398
15	VNRFNVAITR	5,882	5,891	ORF1ab	0.991	C02:02  C03:04  C12:03	0.00000
16	STFNVPMEK	2,600	2,608	ORF1a	0.989	A03:01  A11:01  C07:01	0.00550
17	FYDFAVSKGF	4,811	4,820	ORF1ab	0.988	C04:01  C04:43  C07:02	0.00048
18	NMFITREEAIR	5,996	6,006	ORF1ab	0.99	C02:02  C03:04  C12:03	0.00060
19	PIHFYSKWYIR	38	48	ORF8	0.988	C02:02  C03:04  C12:03	0.00023
20	NYMPYFFTL	2,167	2,175	ORF1a	0.981	A24:02  C01:57  C07:02	0.00415
21	AFPFTIYSL	8	17	ORF10	0.98	C04:01  C04:43  C07:02	0.00168
22	HVGEIPVAYR	110	119	ORF1a	0.991	A31:29  B08:01  B18:01	0.00206
23	VGILCIMSDR	5,894	5,903	ORF1ab	0.983	A24:41  A24:51  C02:02	0.00132
24	GNFYGPFVDR	3,442	3,451	ORF1a	0.983	A24:41  A31:29  B08:01	0.00467
25	AVFDKNLYDKL	1,176	1,186	ORF1a	0.998	A03:01  A11:01	0.00386
26	VFDEISMATNY	5,696	5,706	ORF1ab	0.998	C04:01  C04:43	0.00024
27	TFHLDGEVITF	1,543	1,553	ORF1a	0.997	C04:01  C04:43	0.00440
28	SSRLSFKELL	4,755	4,764	ORF1ab	0.996	C06:02  C07:01	0.00012
29	RIFTIGTVTLK	6	16	ORF3a	0.995	A03:01  A11:01	0.01995
30	VITFDNLKTLL	1,550	1,560	ORF1a	0.994	C04:01  C04:43	0.00385
31	VYRGTTTYKL	5,533	5,543	ORF1ab	0.993	A03:01  A11:01	0.00024
32	FYDFAVSKGFF	4,811	4,821	ORF1ab	0.993	C04:01  C04:43	0.00048
33	YAFEHIVY	6,682	6,689	ORF1ab	0.993	B15:93  B35:01	0.00024
34	KTDGTLMIERF	5,241	5,251	ORF1ab	0.992	A01:01  C05:01	0.00000
35	AYITGGVQQL	599	608	ORF1a	0.991	A24:02  C01:57	0.00427
36	VPWDTIANYA	2,133	2,142	ORF1a	0.991	C04:01  C04:43	0.00401
37	SFDLGDEL	142	149	ORF1a	0.99	C04:01  C04:43	0.00014
38	RRVVFNGVSF	3,163	3,172	ORF1a	0.989	C07:01  C07:02	0.00399
39	VYMPASWVMR	3,653	3,662	ORF1a	0.992	A31:29  C01:57	0.00412
40	LYENAFLPFA	3,606	3,615	ORF1a	0.987	C04:01  C04:43	0.17819
41	QFTSLEIPR	5,910	5,918	ORF1ab	0.987	B18:01  B56:43	0.00060
42	VFPPTSFGPLV	4,712	4,722	ORF1ab	0.986	C04:01  C04:43	0.55016
43	FGADPIHSLR	1,154	1,163	ORF1a	0.999	C04:01  C04:43	0.00332
44	ILGTVSWNLR	1,367	1,376	ORF1a	0.985	C03:04  C12:03	0.00398
45	NFNVLFTSTV	4,704	4,713	ORF1ab	0.985	C04:01  C04:43	0.00012
46	VYMPASWVM	3,653	3,661	ORF1a	0.985	C01:57  C07:02	0.00412
47	AFDKSAFVNL	6,355	6,364	ORF1ab	0.984	C04:01  C04:43	0.00029
48	STFNVPMEKL	2,600	2,609	ORF1a	0.983	A03:01  A11:01	0.00550
49	SGAMDTTSYR	3,218	3,227	ORF1a	0.984	B38:01  B39:54	0.00508
50	VYDYLVTQEF	3,810	3,820	ORF1a	0.983	C04:01  C04:43	0.00412

Peptides marked in red are considered as Highly Variable (HV) due to maximum mutation frequency score  $\geq 0.05$ .



**TABLE 5 |** The most frequently mutated positions within the SARS-CoV-2 proteome.

No.	Protein	Protein position	Mutation frequency
1	ORF1ab	4,715	0.5502
2	S	614	0.5478
3	ORF3a	57	0.1789
4	ORF1a	3,606	0.1781
5	N	203	0.1770
6	N	204	0.1765
7	ORF1a	265	0.1646
8	ORF3a	251	0.1439
9	ORF8	84	0.1384
10	ORF1ab	5,865	0.0926
11	ORF1ab	5,828	0.0924
12	ORF1a	765	0.0668
13	ORF1a	739	0.0590

## SARS-CoV-2 Genome Diversity Analysis

In order to enable the exclusion of peptides originating from genetically highly variable areas, the mutation frequency of each amino acid within the SARS-CoV-2 genome was computed (see section “Materials and Methods” for details). The genes that those peptides originate from are likely to mutate, hence the inclusion of such peptides might lower the vaccine efficacy over time. From the analysis of 8,639 complete genome sequences, obtained from different SARS-CoV-2 isolates, which then were translated into protein sequences, the mutation frequency at each amino acid position was computed.

For each peptide in the SARS-CoV-2 proteome, the maximum mutation frequency was calculated (see section “Materials and Methods”), and peptides with the resulting score  $\geq 0.05$  (marked in color in **Tables 3, 4**) are considered to be highly variable (HV) and should be disregarded as vaccine components. 13 amino acid positions were observed to contain mutations in at least 5% of the selected sequences. Among these, as many as nine amino acid positions were mutated in more than 10% of the selected sequences, while two positions showed mutations in fully half of the samples (more than 50%). In **Table 5** we present the most frequently mutated positions within the SARS-CoV-2 proteome. Mutation frequency values for all positions are available in the **Supplementary Data 2**. Figures presenting distribution of mutation frequency are available in the **Supplementary Data 3**.

Within the top-50 immunogenic peptides originating from the SARS-CoV-2 structural and non-structural proteins (NSPs), 1 and 3 HV peptides were found, respectively.

## Toxicity/Tolerance Results

Each peptide derived from the SARS-CoV-2 proteome was studied to ascertain the lack of similarity with peptides present in the reference human proteome. When administered in a vaccine, epitopes highly similar to peptides presented by the host's healthy tissues could either trigger an unwanted immune self-reaction or be tolerated by the immune system.

In both cases, these peptides should be eliminated from the vaccine composition. A total of 11 SARS-CoV-2-derived peptides with moderate similarity to human proteins were found ( $E$ -value  $\leq 4$ ). Of these, four were significantly similar ( $E$ -value  $\leq 1$ ) and thus should be avoided (see **Supplementary Data 1**). None of these peptides were found within the top-100 ranked peptides.

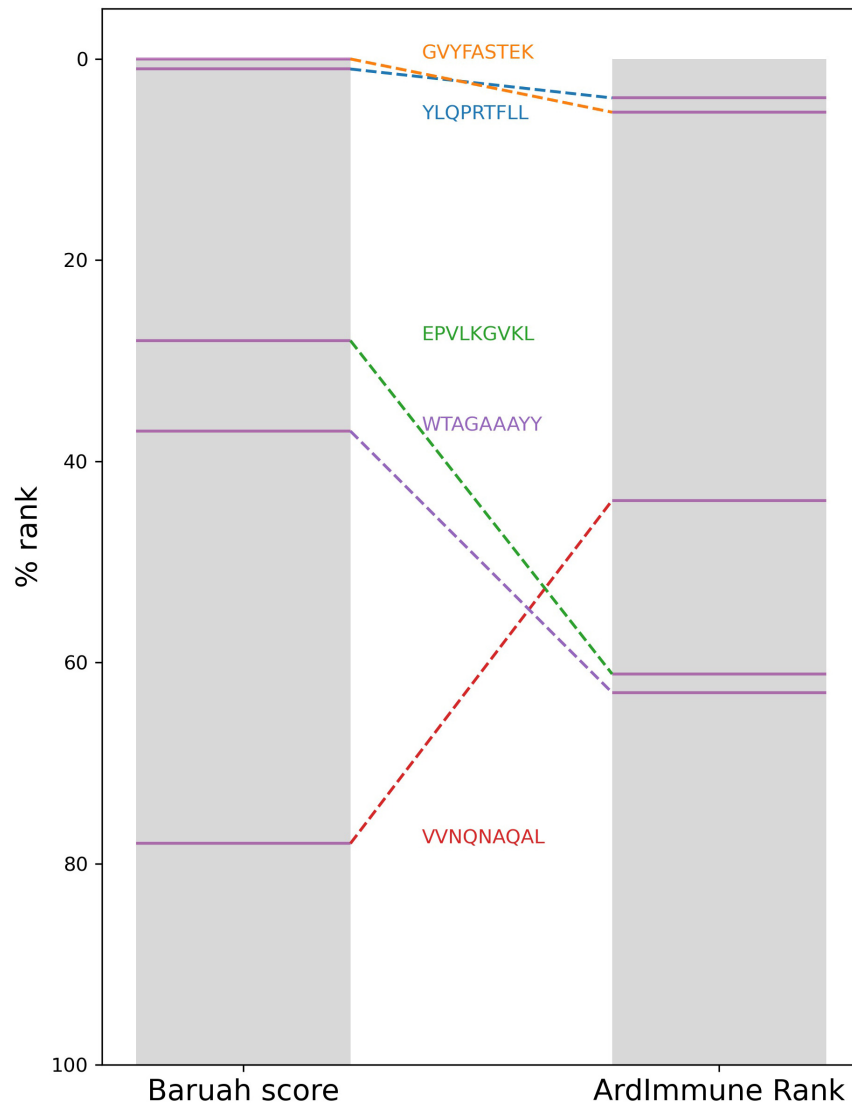
## Comparison With Other Methods

Results from a list of selected publications were compared with percentile ranks computed by our method for the same pHLAs. We did not find any significant correlation with the *in silico* predictions from Grifoni et al. (2020), Lee and Koohy (2020), and Gupta et al. (2020) highlighting a clear distinction between our methodology and the procedures used in these studies. Although the best candidate selected by Gupta et al. is not among our best candidates for HLA-A\*11:01, it is scored by the model as the top candidate among those proposed by the authors. A moderate negative correlation ( $\rho \cong -0.45$ ) was observed between the percentile rank scores of our method and the scores presented by Smith et al. (2020). Although our top peptide candidates associated with the HLAs proposed by Baruah and Bose (2020) do not include any of the five peptides proposed by the authors, we noticed a consensus between the HLA percentile rank of the pHLAs selected by the authors, and our percentile rank scores (**Figure 8**).

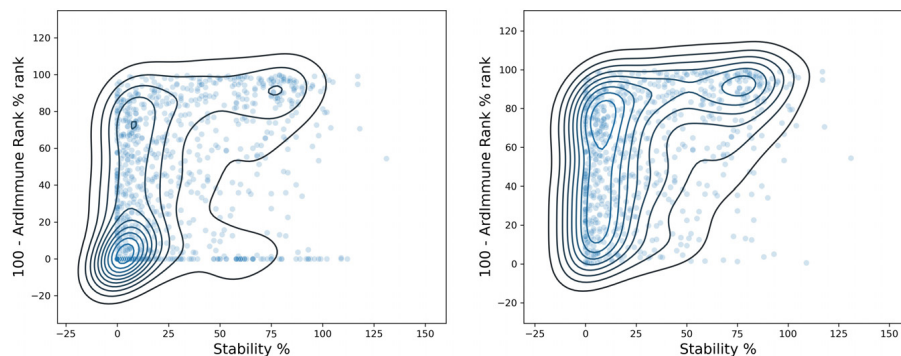
The immunogenicity scores predicted by our model were then compared with the experimental measurement of pHLA binding stability done by Prachar et al. (2020). Peptide candidates with low immunogenicity ranks are enriched in regions with a low stability percentage. The results are shown in **Figure 9**, on the left. The immunogenicity score is expressed as the complement to 100 of the immunogenicity percentile rank. The stability percentage is defined relative to reference peptides (see Prachar et al., 2020 for details). The concordance between high immunogenicity (or low immunogenicity rank) and high stability percentage is more noticeable after the exclusion of peptides with low predicted binding affinity (**Figure 9**, right). The Spearman correlation between pHLA stability percentage and the predicted immunogenicity ( $\rho = 0.392$ ) is higher than the correlation between the stability percentage and the predicted binding affinity ( $\rho = 0.313$ ). The binding affinity was computed using NetMHCpan 4.0 (Jurtz et al., 2017).

A noticeable difference in the distributions of experimentally measured pHLA stability percentage was obtained by ranking using binding affinity predictors and our immunogenicity predictions. A clear distinction between stable and unstable pHLAs was obtained through the selection of the top-10% and the bottom-10% scores predicted by the immunogenicity model, whereas the use of filters relying on standard binding affinity thresholds (e.g., 100 nM) leads to a less defined separation (**Figure 10**).

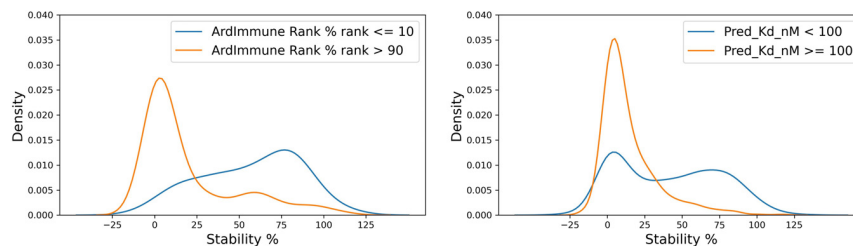
Finally, we report low scores for all the five class I pHLAs which were experimentally confirmed to be non-immunogenic by Rammensee et al. (2020). None of these peptides were recommended by ArdImmune Rank as a candidate to be included in a vaccine formulation against SARS-CoV-2.



**FIGURE 8 |** The HLA percentile ranks of the five peptides selected by Baruah et al. as computed from Baruah score and ArdImmune Rank.



**FIGURE 9 |** Comparison between ArdImmune Rank percentile ranks for pHLA immunogenicity and pHLA stability data measured by Prachar et al. (2020). Scatter plots and kernel density estimations are shown with (right) and without (left) the exclusion of pHLA predicted non-binders ( $K_d$  percentile rank  $\geq 2$ ). The complement of the ArdImmune Rank percentile rank is shown on the y-axis (higher value = lower rank), while the stability percentage as reported by Prachar et al. (2020) is shown on the x-axis.



**FIGURE 10 |** Distribution of stability percentage for different filtering procedures. The respective pHLA stability score densities of the 10% top ranked and the 10% lowest ranked peptides in terms of predicted immunogenicity is shown on the **left**. The pHLA stability score densities computed according to the binding affinity ranges reported by Prachar et al. (2020). ( $K_d \geq 100$  nM,  $K_d < 100$  nM, based on predicted binding affinity) is shown on the **right**.

## DISCUSSION

The high selective pressure exerted upon coronaviruses, caused by the need of a viable host for survival, together with their high genetic variability, facilitates their rapid evolution and the prompt generation of escape mutants. Despite the vigorous effort of the industry, vaccine design, clinical trials, and production require at least several months and most likely several years. Many investigations aimed at developing vaccines protecting humans and animals from coronaviruses were initiated in the last few decades, setting the basis for the recent scientific advancement in COVID-19 treatment. Nonetheless, a limiting aspect associated with the approval and commercialization of a vaccine is that the demand for a vaccine is limited to the outbreak period, and its market value is proportional to the number of people affected. This represented a major issue for the development of vaccines for SARS and MERS (Du et al., 2009; Dhama et al., 2020). In addition, the majority of coronavirus biotherapeutics (i.e., antibodies and vaccines) are designed to leverage neutralizing antibodies directed against the S protein. Safety issues such as those associated with the ADE and CSS events, make the development of vaccine and antibody-based therapies even more problematic.

In combination with the stimulation of humoral immune response, which is aimed at the direct neutralization of the virus, the targeted elimination of infected cells is a crucial element of the immune response against viruses. This might be induced either by the administration of a vaccine eliciting protective CD8+ Cytotoxic T Lymphocyte (CTL) or by transferring CD8+ cells engineered to recognize viral antigens specifically. Previous studies have confirmed a strong correlation between the depletion and exhaustion of T-cells and worse prognosis in critical coronavirus patients (Diao et al., 2020) highlighting the potential of vaccines inducing T-cell responses for COVID-19 prevention. This strategy has beneficial features such as a lower risk of stimulating ADE and CSS with respect to antibody-based strategies (Jaume et al., 2011; Channappanavar et al., 2016) and the stimulation of the immune response against intracellular epitopes not reachable by the antibodies but potentially highly immunogenic. In both cases, the selection of effective immunogenic epitopes is of paramount importance.

The aim of this study was to identify SARS-CoV-2 epitopes for the development of a vaccine composition focused

on T-cell activation. We investigated several aspects pre-determining whether viral epitopes may induce an effective T-cell response, including the MHC-I peptide presentation and immunogenicity potential, SARS-CoV-2 genome variability, and possible toxicity/immune tolerance of the peptides considered.

In contrast to the majority of works on this topic either relying of pHLA binding and presentation events or modeling single pHLA structural interactions, the model applied herein was designed to leverage simultaneously information about the propensity of a peptide to be presented by its cognated HLA and the probability that such pHLA is immunogenic, inferred from similar experimental data. As we show in **Figure 3** when evaluated on the experimentally-validated *Coronaviridae* immunogenicity data, our approach has higher performance than the widely-used predictors assessing pHLA binding affinity, presentation or immunogenicity (i.e., IEDB).

By applying our method, a considerable amount of highly scored T-cell epitopes was found across the SARS-CoV-2 proteome, encompassing the structural proteins and NSPs, as shown in **Tables 3, 4**. The majority of selected epitopes were conserved across different SARS-CoV-2 isolates. Only 16 epitopes were excluded because of their significant mutability (see **Table 5**). The availability of epitopes from NSPs allows for the design of vaccine components dedicated to T-cell responses, and might be further integrated with other components focused on B-cell responses. The adoption of such a compartmentalized strategy might help to lower the risk of non-neutralizing antibody production, which constituted a reason of concern during the development of a vaccine formulation for SARS. Moreover, during the early stages of viral infection, the expression of non-structural proteins is significantly higher than the expression of structural ones. The targeted stimulation of the immune response toward epitopes originating from non-structural proteins might be useful to induce an immune response at the early phase of the disease. Some highly ranked peptides were found to be presented across multiple HLAs and could be used to increase population coverage while decreasing the number of epitopes needed to be included in the vaccine formulation. This aspect could be particularly relevant for solutions relying on delivery systems of limited capacity.

The risk of eliciting potentially harmful and sometimes deadly (Linette et al., 2013) cross-reactivities is an issue to be carefully

addressed in vaccine design. On the other hand, epitopes shared with proteins from the host could also be tolerated by the host's immune system, being not useful for vaccine purposes. Considering the importance of such an aspect, the analysis of potential toxicity and tolerance was addressed in this study, leading to the identification of four highly ranked epitopes having a certain degree of similarity with proteins within the human proteome. Such peptides were removed for safety and efficacy reasons.

The substantial difference between the selection of pHLA candidates performed by our methodology with respect to those presented by Grifoni et al. (2020), Lee and Koohy (2020), and Gupta et al. (2020) highlights a clear distinction between these approaches. Nonetheless, our method supported the selection of top candidates in small datasets obtained by applying hand-crafted filtering stages (Baruah and Bose, 2020; Gupta et al., 2020). The mild correlation with the results from Smith et al. (2020) might indicate the usage of equivalent components during some steps of the selection process. A relative concordance between the pHLA stability scores from Prachar et al. (2020) and the associated immunogenic scores computed by our method was observed (Figure 9). Moreover, we show that the peptide ranks produced by our immunogenicity model have a higher correlation with the experimentally measured pHLA stability than the ranks obtained by methods relying solely on binding affinity or ligand likelihood predictions. This observation is consistent with works reported in the literature (Harndahl et al., 2012). We also obtained low immunogenicity scores for all five peptides which have been experimentally confirmed by Rammensee to be unable to activate CD8+ lymphocytes.

## CONCLUSION

In this paper we suggested a SARS-CoV-2 vaccine composition in the form of the list of epitopes optimized for their (predicted) immunogenicity and HLA population coverage. Our motivation is that cellular immune response is fundamental for an effective SARS-CoV-2 vaccine and it also mitigates the risks of ADE and CSS which are typically associated with modalities relying on the activation of humoral immune response. We showed that the predictive model, on which our methodology is based outperforms, on *Coronaviridae* data, other methods used to date for the design of epitope-based vaccines against SARS-CoV-2. Our approach differs from other existing methods and shows a higher correlation with the measured pHLA stability in comparison with methods based solely on binding affinity predictions. The limitations of our method have the same roots as those found in other *in silico* approaches based on predicting various pHLA properties, i.e., the accuracy of these predictive methods. We expect that with the increasing amount of experimentally validated data and with further algorithmic enhancements in the field of artificial intelligence, the accuracy of such models and the effectiveness of vaccine design will continue to improve. Computational methods have proven to be of considerable

support in optimizing the vaccine design process on several occasions. Moreover, a notable improvement in the predicting skills of such methods has been recorded in recent years, admittedly due to the increasing advancements in machine learning coupled with a surge in the availability of powerful computational resources. However, it is important to mention that such tools do not represent a substitute for the laboratory experiments necessary to verify and optimize the safety and efficacy of vaccines. Their role is to support the design of such experiments in order to reduce their number, the time needed and cost.

## DATA AVAILABILITY STATEMENT

The lists containing the predicted immunogenic peptides with percentile rank  $\leq 2$  are included in this study (Tables 3 and 4). The lists of all the predicted immunogenic peptides generated during this study are available from the corresponding author upon reasonable request.

## AUTHOR CONTRIBUTIONS

GM wrote the article with contributions from IN, PSk, and JK. AM, PSk, IN, and KG performed the analyses and generated figures and tables included in the article. GM, IN, AM, PSk, JK, AS-D, KG, PK, and MD developed the applied methodology. PSt conceived the idea for the project and coordinated the work. AS-D, MS, and KP gave essential contributions to the interpretation of immunological and virological aspects of the study. All the authors reviewed, edited, contributed to the article and approved the submitted version.

## FUNDING

The study was sponsored by Ardigen. The applied methodology was in part developed prior to this study with support from the regional Polish grant RPMP.01.02.01-12-0301/17 (European Funds, Regional Program) approved by the Małopolska Centre for Entrepreneurship.

## ACKNOWLEDGMENTS

Ardigen and COVID-19 Vaccine Corporation (CVC) announced that they entered a research collaboration aimed at the development of SARS-CoV-2 vaccine.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.602196/full#supplementary-material>



## REFERENCES

- Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., et al. (2017). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46, 315–326. doi: 10.1016/j.immuni.2017.02.007
- Ahmed, S. F., Quadeer, A. A., and McKay, M. R. (2020). Preliminary identification of potential vaccine targets for the COVID-19 Coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses* 12:254. doi: 10.3390/v12030254
- Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452. doi: 10.1038/s41591-020-0820-9
- Baruah, V., and Bose, S. (2020). Immunoinformatics-aided identification of T Cell and B Cell Epitopes in the surface glycoprotein of 2019-nCoV. *J. Med. Virol.* 92, 495–500. doi: 10.1002/jmv.25698
- Beck, Z., Prohászka, Z., and Füst, G. (2008). Traitors of the immune system—enhancing antibodies in HIV infection: their possible implication in HIV vaccine development. *Vaccine* 26, 3078–3085. doi: 10.1016/j.vaccine.2007.12.028
- Calis, J. J. A., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., et al. (2013). Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* 9:e1003266. doi: 10.1371/journal.pcbi.1003266
- Channappanavar, R., Fehr, A. R., Vijay, R., Mack, M., Zhao, J., Meyerholz, D. K., et al. (2016). Dysregulated Type I interferon and inflammatory monocyte-macrophage responses cause lethal pneumonia in SARS-CoV-infected mice. *Cell Host Microb.* 19, 181–193. doi: 10.1016/j.chom.2016.01.007
- Channappanavar, R., Fett, C., Zhao, J., Meyerholz, D. K., and Perlman, S. (2014). Virus-specific memory CD8 T cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. *J. Virol.* 88, 11034–11044. doi: 10.1128/JVI.01505-14
- Chen, H., Hou, J., Jiang, X., Ma, S., Meng, M., Wang, B., et al. (2005). Response of Memory CD8<sup>+</sup> T cells to severe acute respiratory syndrome (SARS) Coronavirus in recovered SARS patients and healthy individuals. *J. Immunol.* 175, 591–598. doi: 10.4049/jimmunol.175.1.591
- Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel Coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 395, 507–513. doi: 10.1016/S0140-6736(20)30211-7
- Corapi, W. V., Olsen, C. W., and Scott, F. W. (1992). Monoclonal antibody analysis of neutralization and antibody-dependent enhancement of feline infectious peritonitis virus. *J. Virol.* 66, 6695–6705. doi: 10.1128/JVI.66.11.6695-6705.1992
- Cui, J., Li, F., and Shi, Z. (2019). Origin and evolution of pathogenic Coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi: 10.1038/s41579-018-0118-9
- Dejnirattisai, W., Jumnainsong, A., Onsirakul, N., Fitton, P., Vasanawathana, S., Limpitkul, W., et al. (2010). Cross-reacting antibodies enhance dengue virus infection in humans. *Science* 328, 745–748. doi: 10.1126/science.1185181
- Dhama, K., Sharun, K., Tiwari, R., Dadar, M., Malik, Y. S., Singh, K. P., et al. (2020). COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics. *Hum. Vacc. Immunotherap.* 16, 1232–1238. doi: 10.1080/21645515.2020.1735227
- Di Marco, M., Schuster, H., Backert, L., Ghosh, M., Rammensee, H. G., and Stevanović, S. (2017). Unveiling the peptide motifs of HLA-C and HLA-G from naturally presented peptides and generation of binding prediction matrices. *J. Immunol.* 199, 2639–2651. doi: 10.4049/jimmunol.1700938
- Diao, B., Wang, C., Tan, Y., Chen, X., Liu, Y., Ning, L., et al. (2020). Reduction and functional exhaustion of T cells in patients with Coronavirus disease 2019 (COVID-19). *Front. Immunol.* 11:827. doi: 10.3389/fimmu.2020.00827
- Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B. J., and Jiang, S. (2009). The spike protein of SARS-CoV — a target for vaccine and therapeutic development. *Nat. Rev. Microbiol.* 7, 226–236. doi: 10.1038/nrmicro2090
- Fan, Y. Y., Huang, Z. T., Li, L., Wu, M. H., Yu, T., Koup, R. A., et al. (2009). Characterization of SARS-CoV-specific memory T cells from recovered individuals 4 years after infection. *Archiv. Virol.* 154, 1093–1099. doi: 10.1007/s00705-009-0409-6
- Forni, D., Cagliani, R., Mozzi, A., Pozzoli, U., Al-Daghri, N., Clerici, M., et al. (2016). Extensive positive selection drives the evolution of nonstructural proteins in Lineage C Betacoronaviruses. *J. Virol.* 90, 3627–3639. doi: 10.1128/JVI.02988-15
- Fu, Y., Cheng, Y., and Wu, Y. (2020). Understanding SARS-CoV-2-mediated inflammatory responses: from mechanisms to potential therapeutic tools. *Virol. Sin.* 35, 266–271. doi: 10.1007/s12250-020-00207-4
- Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R. H., Peters, B., and Sette, A. (2020). A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microb.* 27, 671–680.e2. doi: 10.1016/j.chom.2020.03.002
- Gupta, E., Mishra, R. K., and Niraj, R. R. K. (2020). Identification of potential vaccine candidates against SARS-CoV-2, a step forward to fight novel Coronavirus 2019-NCoV: a reverse vaccinology approach. *bioRxiv* [Preprint], doi: 10.1101/2020.04.13.039198
- Guzman, M. G., Alvarez, M., Rodriguez-Roche, R., Bernardo, L., Montes, T., Vazquez, S., et al. (2007). Neutralizing antibodies after infection with dengue 1 virus. *Emerg. Infect. Dis.* 13, 282–286. doi: 10.3201/eid1302.060539
- Harndahl, M., Rasmussen, M., Roder, G., Dalgaard Pedersen, I., Sørensen, M., Nielsen, M., et al. (2012). Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity: antigen processing. *Eur. J. Immunol.* 42, 1405–1416. doi: 10.1002/eji.201141774
- Hohdatsu, T., Yamada, M., Tominaga, R., Makino, K., Kida, K., and Koyama, H. (1998). Antibody-dependent enhancement of feline infectious peritonitis virus infection in feline alveolar macrophages and human monocyte cell line U937 by Serum of cats experimentally or naturally infected with feline Coronavirus. *J. Veter. Med. Sci.* 60, 49–55. doi: 10.1292/jvms.60.49
- Hu, D., Zhu, C., Ai, L., He, T., Wang, Y., Ye, F., et al. (2018). Genomic characterization and infectivity of a novel SARS-like Coronavirus in Chinese Bats. *Emerg. Micro. Infect.* 7, 1–10. doi: 10.1038/s41426-018-0155-5
- Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel Coronavirus in Wuhan, China. *Lancet* 395, 497–506. doi: 10.1016/S0140-6736(20)30183-5
- Iwasaki, A., and Yang, Y. (2020). The potential danger of suboptimal antibody responses in COVID-19. *Nat. Rev. Immunol.* 20, 339–341. doi: 10.1038/s41577-020-0321-6
- Jaume, M., Yip, M. S., Cheung, C. Y., Leung, H. L., Li, P. H., Kien, F., et al. (2011). Anti-Severe acute respiratory syndrome Coronavirus spike antibodies trigger infection of human immune cells via a PH- and cysteine protease-independent Fc R pathway. *J. Virol.* 85, 10582–10597. doi: 10.1128/JVI.00671-11
- Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted Ligand and peptide binding affinity data. *J. Immunol.* 199, 3360–3368. doi: 10.4049/jimmunol.1700893
- Kam, Y. W., Kien, F., Roberts, A., Cheung, Y. C., Lamirande, E. W., Vogel, L., et al. (2007). Antibodies against Trimeric S glycoprotein protect hamsters against SARS-CoV challenge despite their capacity to mediate FcγRII-dependent entry into B Cells in vitro. *Vaccine* 25, 729–740. doi: 10.1016/j.vaccine.2006.08.011
- Katzelnick, L. C., Gresh, L., Halloran, M. E., Mercado, J. C., Kuan, G., Gordon, A., et al. (2017). Antibody-dependent enhancement of severe dengue disease in humans. *Science* 358, 929–932. doi: 10.1126/science.aan6836
- Lee, C. H., and Koohy, H. (2020). In silico identification of vaccine targets for 2019-NCoV. *F1000Research* 9:145. doi: 10.12688/f1000research.22507.2
- Li, F. (2016). Structure, function, and evolution of Coronavirus spike proteins. *Ann. Rev. Virol.* 3, 237–261. doi: 10.1146/annurev-virology-110615-042301
- Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* 26, 842–844. doi: 10.1038/s41591-020-0901-9
- Linette, G. P., Stadtmauer, E. A., Maus, M. V., Rapoport, A. P., Levine, B. L., Emery, L., et al. (2013). Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in Myeloma and melanoma. *Blood* 122, 863–871. doi: 10.1182/blood-2013-03-490565
- Liu, J., Sun, Y., Qi, J., Chu, F., Wu, H., Gao, F., et al. (2010). The membrane protein of severe acute respiratory syndrome Coronavirus acts as a dominant immunogen revealed by a clustering region of novel functionally and

- structurally defined cytotoxic T-lymphocyte epitopes. *J. Infect. Dis.* 202, 1171–1180. doi: 10.1086/656315
- Liu, L., Wei, Q., Lin, Q., Fang, J., Wang, H., Kwok, H., et al. (2019). Anti-Spike IgG causes severe acute lung injury by skewing macrophage responses during acute SARS-CoV infection. *JCI Insight* 4:e123158. doi: 10.1172/jci.insight.123158
- Ng, O. W., Chia, A., Tan, A. T., Jodi, R. S., Leong, H. N., Bertoletti, A., et al. (2016). Memory T cell responses targeting the SARS Coronavirus persist up to 11 years post-infection. *Vaccine* 34, 2008–2014. doi: 10.1016/j.vaccine.2016.02.063
- O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7, 129–132.e4. doi: 10.1016/j.cels.2018.05.014
- Ogishi, M., and Yotsuyanagi, H. (2019). Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. *Front. Immunol.* 10:827. doi: 10.3389/fimmu.2019.00827
- Page, A., and DeRoy, G. (2020). *Biostrings: Efficient Manipulation of Biological Strings. R Package Version 2.56.0*.
- Pahl, J. H. W., Kwappenberg, K. M. C., Varypataki, E. M., Santos, S. J., Kuijter, M. L., Mohamed, S., et al. (2014). Macrophages inhibit human osteosarcoma cell growth after activation with the bacterial cell wall derivative Liposomal Muramyl tripeptide in combination with Interferon- $\gamma$ . *J. Exper. Clin. Cancer Res.* 33:27. doi: 10.1186/1756-9966-33-27
- Peiris, J. S. M., Chu, C. M., Cheng, V. C. C., Chan, K. S., Hung, I. F. N., Poon, L. L. M., et al. (2003). Clinical progression and viral load in a community outbreak of Coronavirus-associated SARS pneumonia: a prospective study. *Lancet* 361, 1767–1772. doi: 10.1016/S0140-6736(03)13412-5
- Peng, H., Yang, L. T., Wang, L. Y., Li, J., Huang, J., Lu, Z. Q., et al. (2006). Long-lived memory T lymphocyte responses against SARS Coronavirus nucleocapsid Protein in SARS-recovered patients. *Virology* 351, 466–475. doi: 10.1016/j.virol.2006.03.036
- Prachar, M., Justesen, S., Steen-Jensen, D. B., Thorgrimsen, S., Jurgons, E., Winther, O., et al. (2020). COVID-19 vaccine candidates: prediction and validation of 174 SARS-CoV-2 epitopes. *bioRxiv* [Preprint], doi: 10.1101/2020.03.20.00794
- Rammensee, H. S., Stevanovic, S., Gouttefangeas, C., Heidt, S., Klein, R., Preuß, B., et al. (2020). Designing a therapeutic SARS-CoV-2 T-cell-inducing vaccine for high-risk patient groups. *bioRxiv* [Preprint], doi: 10.21203/rs.3.rs-27316/v1
- Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* 38, 199–209. doi: 10.1038/s41587-019-0322-9
- Smith, C. C., Entwistle, S., Willis, C., Vensko, S., Beck, W., Garness, J., et al. (2020). Landscape and selection of vaccine epitopes in SARS-CoV-2. *bioRxiv* [Preprint], doi: 10.1101/2020.06.04.135004
- Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., et al. (2019). From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses* 11:59. doi: 10.3390/v11010059
- Sylvester-Hvid, C., Nielsen, M., Lamberth, K., Roder, G., Justesen, S., Lundegaard, C., et al. (2004). SARS CTL vaccine candidates; HLA Supertype-, genome-wide scanning and biochemical validation. *Tissue Antig.* 63, 395–400. doi: 10.1111/j.0001-2815.2004.00221.x
- Takada, A., Feldmann, H., Ksiazek, T. G., and Kawaoka, Y. (2003). Antibody-dependent enhancement of ebola virus infection. *J. Virol.* 77, 7539–7544. doi: 10.1128/JVI.77.13.7539-7544.2003
- Takada, A., Watanabe, S., Okazaki, K., Kida, H., and Kawaoka, Y. (2001). Infectivity-enhancing antibodies to ebola virus glycoprotein. *J. Virol.* 75, 2324–2330. doi: 10.1128/JVI.75.5.2324-2330.2001
- Tang, F., Quan, Y., Xin, Z. T., Wrammert, J., Ma, M. J., Lv, H., et al. (2011). Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: a six-year follow-up study. *J. Immunol.* 186, 7264–7268. doi: 10.4049/jimmunol.0903490
- Tsao, Y. P., Lin, J. Y., Jan, J. T., Leng, C. H., Chu, C. C., Yang, Y. C., et al. (2006). HLA-A\*0201 T-cell epitopes in severe acute respiratory syndrome (SARS) Coronavirus nucleocapsid and spike proteins. *Biochem. Biophys. Res. Commun.* 344, 63–71. doi: 10.1016/j.bbrc.2006.03.152
- Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2019). The Immune Epitope Database (IEDB), 2018 update. *Nucleic Acids Res.* 47, D339–D343. doi: 10.1093/nar/gky1006
- Wan, S., Xiang, Y., Fang, W., Zheng, Y., Li, B., Hu, Y., et al. (2020). Clinical features and treatment of COVID-19 patients in northeast Chongqing. *J. Med. Virol.* 92, 797–806. doi: 10.1002/jmv.25783
- Wan, Y., Shang, J., Sun, S., Tai, W., Chen, J., Geng, Q., et al. (2019). Molecular mechanism for antibody-dependent enhancement of coronavirus entry. edited by tom Gallagher. *J. Virol.* 94:e02015-19.
- Wang, S. F., Tseng, S. P., Yen, C. H., Yang, J. Y., Tsao, C. H., Shen, C. W., et al. (2014). Antibody-dependent SARS Coronavirus infection is mediated by antibodies against spike proteins. *Biochem. Biophys. Res. Commun.* 451, 208–214. doi: 10.1016/j.bbrc.2014.07.090
- Wang, Y. D., Sin, W. Y. F., Xu, G. B., Yang, H. H., Wong, T. Y., Pang, X. W., et al. (2004). T-Cell Epitopes in severe acute respiratory syndrome (SARS) Coronavirus spike protein elicit a specific T-Cell immune response in patients who recover from SARS. *J. Virol.* 78, 5612–5618. doi: 10.1128/JVI.78.11.5612-5618.2004
- Whitehead, S. S., Blaney, J. E., Durbin, A. P., and Murphy, B. R. (2007). Prospects for a dengue virus vaccine. *Nat. Rev. Microbiol.* 5, 518–528. doi: 10.1038/nrmicro1690
- Wiley, S., Aasa-Chapman, M. M. I., O'Farrell, S., Pellegrino, P., Williams, I., Weiss, R. A., et al. (2011). Extensive complement-dependent enhancement of HIV-1 by autologous non-neutralising antibodies at early stages of infection. *Retrovirology* 8:16. doi: 10.1186/1742-4690-8-16
- Wright, E. S. (2015). DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinform.* 16:322. doi: 10.1186/s12859-015-0749-z
- Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., et al. (2020). Genome composition and divergence of the novel Coronavirus (2019-nCoV) originating in China. *Cell Host Microb.* 27, 325–328. doi: 10.1016/j.chom.2020.02.001
- Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new Coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi: 10.1038/s41586-020-2008-3
- Zhang, X. W. (2013). A Combination of Epitope prediction and molecular docking allows for good identification of MHC class I restricted T-Cell epitopes. *Comput. Biol. Chem.* 45, 30–35. doi: 10.1016/j.compbiolchem.2013.03.003
- Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new Coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel Coronavirus from patients with pneumonia in China, 2019. *New Engl. J. Med.* 382, 727–733. doi: 10.1056/NEJMoa2001017

**Conflict of Interest:** GM, IN, AM, PSK, JK, AS-D, KG, PK, MD, and PSt are employees at Ardigen or were in the past.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Mazzocco, Niemiec, Myronov, Skoczylas, Kaczmarczyk, Sanecka-Duin, Gruba, Król, Drwal, Szczepanik, Pyrc and Stepniak. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# A Peptide Vaccine Candidate Tailored to Individuals' Genetics Mimics the Multi-Targeted T Cell Immunity of COVID-19 Convalescent Subjects

Eszter Somogyi<sup>1,2</sup>, Zsolt Csiszovszki<sup>1,2</sup>, Levente Molnár<sup>1,2</sup>, Orsolya Lőrincz<sup>1,2</sup>, József Tóth<sup>1,2</sup>, Sofie Pattijn<sup>3</sup>, Jana Schockaert<sup>3</sup>, Aurélie Mazy<sup>3</sup>, István Miklós<sup>1,2,4</sup>, Katalin Pántya<sup>1,2</sup>, Péter Páles<sup>1,2</sup> and Enikő R. Tóke<sup>1,2\*</sup>

<sup>1</sup> Treos Bio Ltd., London, United Kingdom, <sup>2</sup> Treos Bio Zrt, Veszprém, Hungary, <sup>3</sup> ImmunXperts Société Anonyme, A Nexelis Group Company, Gosselies, Belgium, <sup>4</sup> Alfréd Rényi Institute of Mathematics, Eötvös Loránd Research Network, Budapest, Hungary

## OPEN ACCESS

### Edited by:

Nimisha Ghosh,  
Siksha O Anusandhan University, India

### Reviewed by:

Biju Issac,  
Leidos Biomedical Research, Inc.,  
United States  
Sandra Paulina Smieszek,  
Vanda Pharmaceuticals Inc.,  
United States

### \*Correspondence:

Enikő R. Tóke  
eniko.toke@treosbio.com

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

Received: 22 March 2021

Accepted: 24 May 2021

Published: 23 June 2021

### Citation:

Somogyi E, Csiszovszki Z, Molnár L, Lőrincz O, Tóth J, Pattijn S, Schockaert J, Mazy A, Miklós I, Pántya K, Páles P and Tóke ER (2021) A Peptide Vaccine Candidate Tailored to Individuals' Genetics Mimics the Multi-Targeted T Cell Immunity of COVID-19 Convalescent Subjects. *Front. Genet.* 12:684152. doi: 10.3389/fgene.2021.684152

Long-term immunity to coronaviruses likely stems from T cell activity. We present here a novel approach for the selection of immunoprevalent SARS-CoV-2-derived T cell epitopes using an *in silico* cohort of HLA-genotyped individuals with different ethnicities. Nine 30-mer peptides derived from the four major structural proteins of SARS-CoV-2 were selected and included in a peptide vaccine candidate to recapitulate the broad virus-specific T cell responses observed in natural infection. PolyPEPI-SCoV-2-specific, polyfunctional CD8<sup>+</sup> and CD4<sup>+</sup> T cells were detected in each of the 17 asymptomatic/mild COVID-19 convalescents' blood against on average seven different vaccine peptides. Furthermore, convalescents' complete HLA-genotype predicted their T cell responses to SARS-CoV-2-derived peptides with 84% accuracy. Computational extrapolation of this relationship to a cohort of 16,000 HLA-genotyped individuals with 16 different ethnicities suggest that PolyPEPI-SCoV-2 vaccination will likely elicit multi-antigenic T cell responses in 98% of individuals, independent of ethnicity. PolyPEPI-SCoV-2 administered with Montanide ISA 51 VG generated robust, Th1-biased CD8<sup>+</sup>, and CD4<sup>+</sup> T cell responses against all represented proteins, as well as binding antibodies upon subcutaneous injection into BALB/c and hCD34<sup>+</sup> transgenic mice modeling human immune system. These results have implications for the development of global, highly immunogenic, T cell-focused vaccines against various pathogens and diseases.

**Keywords:** global vaccine, HLA-genotype, ethnic diversity, SARS-CoV-2 immunity, *in silico* clinical trial

## INTRODUCTION

The pandemic caused by the novel coronavirus SARS-CoV-2 is still evolving after its outbreak in late 2019, reaching second/third peak in a single year. After demonstration of high protective efficacy against symptomatic COVID-19 in large phase III studies, the first vaccines are rapidly being approved for emergency use (Forni et al., 2021). Both the approved vaccines and the numerous vaccine candidates under clinical development are predominantly designed to generate neutralizing antibodies against the viral Spike (S) protein (WHO, 2020). But lessons learned from the SARS and MERS epidemic as well as COVID-19 pandemic indicate potential challenges (Altmann and Boyton, 2020; Green, 2020; Hellerstein, 2020; Peiris and Leung, 2020). Due to waning

antibody responses and continuously arising mutations in the S protein, long-term durability of protection remains unknown (Callaway, 2021; Williams and Burgers, 2021). However, T cell responses against coronavirus proteins can last for over a decade, as demonstrated for MERS and SARS, and data collected till today for SARS-CoV-2 seem to support this expectation (Channappanavar et al., 2014b; Le Bert et al., 2020; Schwarzkopf et al., 2021).

Importantly, virtually all subjects with a history of SARS-CoV-2 infection mount T cell responses against the virus, including seronegatives and subjects with severe disease (Grifoni et al., 2020; Hellerstein, 2020; Peng et al., 2020; Zou et al., 2020). Moreover, correlation between defective T cell responses and COVID-19 severity was observed (Diao et al., 2020). Higher CD8<sup>+</sup> T cell counts were also associated with improved overall survival in cancer patients hospitalized for COVID-19 (Huang et al., 2021).

T cell responses are diverse, recognizing 30–40 SARS-CoV-2 epitopes in each person (Tarke et al., 2021). They are directed against the whole antigenic repertoire of the virus, and less dominated by the S-protein (Nelde et al., 2020; Sekine et al., 2020; Tarke et al., 2021). This diversity is associated with asymptomatic/mild disease and likely confers protection against viral escape by mutations.

The first COVID-19 vaccines, while engender robust humoral responses, have mixed potential for inducing CD8<sup>+</sup> T cell responses (Anderson et al., 2020; Ewer et al., 2020; Jackson et al., 2020; Sahin et al., 2020; Zhang et al., 2021). Multi-epitope CD8<sup>+</sup> T cell responses against the S protein, as revealed for two studies, were obtained for only a fraction of subjects (24–60%) (Ewer et al., 2020; Sahin et al., 2020). Multi-epitope responses against multiple viral antigens could be theoretically elicited by vaccines using whole virus material, but the assessment of cellular immune reactions was not included in their studies (Zhang et al., 2021).

Therefore, strategies to better mimic the heterogeneity of multi-specific T cell immunity caused by the natural infection would be required to leverage the vital role of both CD8<sup>+</sup> and CD4<sup>+</sup> T cell responses in reducing the impact of COVID-19 and potentially provoking long-term immune responses (Dan et al., 2021).

The core problem that afflicts T cell-epitope selection, however, is that each human has a unique immune response profile to pathogens. Indeed, for SARS-CoV-2, the infection or the disease course varies according to the genetic diversity represented by different ethnicities and human leukocyte antigen (HLA) alleles, however, the reason is not yet well-understood (Aldridge et al., 2020; Nguyen et al., 2020; Pan et al., 2020; Poland, 2020; Mohammadpour et al., 2021). HLA alleles are the molecular determinants of antigen-specific T cell activation, to kill infected cells. Each human has six major HLA class I and eight major HLA class II alleles, therefore larger populations have hundreds of different alleles and their numerous combinations in each HLA-genotype. As a result, each person's T cells recognize 30–40 epitopes derived from SARS-CoV-2 and only a fraction of them are shared between convalescents, as recently reported by Tarke et al. in a very comprehensive study (Tarke et al., 2021). To capture this heterogeneity during a global SARS-CoV-2 T

cell focused vaccine design effort, epitope mapping based on limited number of frequent HLA alleles has been used widely (Ferretti et al., 2020; Nelde et al., 2020). However, in reality, these epitope mapping studies have a low yield (cca. 10%) in terms of confirmed T cell response in HLA-matched subjects (Nelde et al., 2020; Tarke et al., 2021). Therefore, actionable strategies to target not alleles but individuals and ethnic populations are required. We hypothesize that all HLA alleles (HLA genotype) of a subject regulate immune responses capable of killing infected cells, therefore we propose epitope mapping that involves real-subjects with complete HLA-genotype instead of single HLA alleles split from the complexity of allele combinations.

We present here a novel, computer-aided approach for the selection of immunogenic peptides using an ethnically diverse *in silico* human cohort of individuals with complete HLA genotypes. We selected multiple, so called Personal Epitopes (PEPIs, restricted to multiple HLA alleles of a person) shared among high proportion of subjects in each ethnic group of this model population. PolyPEPI-SCoV-2 contains 9 peptides and targets all four major structural proteins of SARS-CoV-2. We demonstrated, that T cells against each selected epitopes were present in majority of COVID-19 convalescent subjects tested, and the frequency was in good agreement with the frequency determined for the *in silico* cohort. More importantly we found, that subjects' complete HLA-genotype influenced their peptide-specific anti-SARS-CoV-2 immune responses, as hypothesized. Immunogenicity and safety of the designed candidate vaccine were confirmed in two mouse models, resulting in the induction of robust CD8<sup>+</sup> and CD4<sup>+</sup> T cell responses, against all four targeted SARS-CoV-2 proteins. Our novel approach enables, for the first-time, computational determination of the epitopes that immune systems of individuals in large cohorts can respond to, likely an indispensable tool for both the design of a global vaccine and for post-vaccination surveillance.

## MATERIALS AND METHODS

### Donors

Donors were recruited based on their clinical history of SARS-CoV-2 infection. Blood samples were collected from convalescent individuals ( $n = 15$ ) at an independent medical research center in The Netherlands under an approved protocol (NL57912.075.16.) or collected by PepTC Vaccines Ltd ( $n = 2$ ). Sera and PBMC samples from non-exposed individuals ( $n = 10$ ) were collected before 2018 and were provided by Nexelis-IMXP (Belgium). All donors provided written informed consent. The study was conducted in accordance with the Declaration of Helsinki. Blood samples from COVID-19 convalescent patients ( $n = 17$ ; 16 with asymptomatic/mild disease and one with severe disease) were obtained 17–148 days after symptom onset. Surprisingly, one positive IgM antibody response was found among the healthy donors, which was excluded from further analysis. Demographic and baseline information of the subjects are provided in **Supplementary Table 1**. HLA genotyping of the convalescent donor patients from The Netherlands was done by IMGm laboratories GmbH (Martinsried, Germany) using next-generation sequencing. This cohort uses a total of 46 different



HLA class I alleles (15 HLA-A\*, 18 HLA-B\*, and 13 HLA-C\*) and 35 different HLA class II alleles (14 DRB1, 12 DQB1, and 9 DPB1). HLA-genotype data of the subjects is provided in **Supplementary Table 2**.

### ***In silico* Human Cohorts**

#### ***Model Population (n = 433)***

The Model Population is a cohort of 433 individuals, representing several ethnic groups worldwide, for whom complete HLA class I genotypes were available (2 × HLA-A, 2 × HLA-B, 2 × HLA-C). The Model Population was assembled from 90 Yoruban African (YRI), 90 European (CEU), 45 Chinese (CHB), 45 Japanese (JPT), 67 subjects with mixed ethnicity (US, Canada, Australia, New Zealand), and 96 subjects from an HIV database (MIX). HLA genotypes were determined using PCR techniques, Affymetrix 6.0 and Illumina 1.0 Million SNP mass arrays, and high-resolution HLA typing of the six HLA genes by Reference Strand-mediated Conformational Analysis (RSCA) or sequencing-based typing (SBT). This cohort uses a total of 152 different HLA class I alleles (49 HLA-A\*, 71 HLA-B\* and 32 HLA-C\*) representative for 97.4% of the alleles documented in the current global Common, Intermediate and Well-Documented (CIWD) database, well-representing also major ethnicities (database 3.0 released 2020) (**Supplementary Table 3**) (Hurley et al., 2020). The frequency of the A\*, B\*, and C\* alleles of the Model population correlates with the frequency documented for >8 million HLA-genotyped subjects of the CIWD database ( $R = 0.943, 0.869, 0.942$ , respectively,  $p < 0.00001$ ) (**Supplementary Figure 1**).

#### ***HLA Class II Cohort (n = 356)***

A second cohort of 356 individuals with characterized HLA class II genotypes (2 × HLA-DRB, 2 × HLA-DP, and 2 × HLA-DQ) at four-digit allele resolution was obtained from the dbMHC database, an online available repository operated by the National Center for Biotechnology Information (NCBI) (Helmberg et al., 2004). HLA genotyping was performed by SBT. This cohort uses a total of 150 different HLA class II alleles (41 DRB1, 66 DQB1, and 43 DPB1).

#### ***Large, US Cohort (n = 16,000)***

The database comprising anonymized HLA genotype data from 16,000 individuals was created by obtaining 1,000 donors from each of 16 ethnic groups (500 male and 500 female) from the National Marrow Donor Program (NMDP) (Gragert et al., 2013). The 16 ethnic groups were: African, African American, Asian Pacific Islander, Filipino, Black Caribbean, Caucasian, Chinese, Hispanic, Japanese, Korean, Native American Indian, South Asian, Vietnamese, US, Mideast/North coast of Africa, Hawaiian, and other Pacific Islander. The ethnic groups represented in this large US cohort covers the composition of the global population but they were not weighted for their global representativeness (we intentionally used  $n = 1,000$  subjects for each ethnicity).<sup>1</sup> HLA genotyping was performed by NMDP recruitment labs using sequence-specific oligonucleotide (SSO) and sequence

specific primer (SSP) methods with an average “typing resolution score” >0.7. This cohort uses a total of 497 different HLA class I alleles (136 HLA-A\* 240 HLA-B\* and 121 HLA-C\*) representative for 99.8% of the alleles documented in the current global Common, Intermediate and Well-Documented (CIWD) database (database 3.0 released in 2020) (Hurley et al., 2020) and 140 HLA class II alleles (105 DRB1 and 35 DQB1, DPB1 was not available). HLA-alleles covered by this cohort are provided in **Supplementary Table 4**.

### **Animals**

#### **CD34<sup>+</sup> Transgenic Humanized Mouse (Hu-mouse)**

*Female* NOD/Shi-scid/IL-2R $\gamma$  null immunodeficient mice (Charles River Laboratories, France) were humanized using hematopoietic stem cells (CD34<sup>+</sup>) isolated from human cord blood. Only mice with a humanization rate (hCD45/total CD45) >50% were used during the study. Experiments were carried out with 20–23-week-old female mice.

#### **BALB/c Mouse**

Experiments were carried out with 6–8 week old female BALB/c mice (Janvier, France).

### **Vaccine Design**

#### **Tailoring PolyPEPISCOV-2 to SARS-CoV-2 Genetics**

SARS-CoV-2 structural proteins (S, N, M, E) were screened and nine different 30-mer peptides were selected during a multi-step process. First, sequence diversity analysis was performed (as of 28 March 2020 in the NCBI database).<sup>2</sup> The accession IDs were as follows: **NC\_045512.2**, MN938384.1, MN975262.1, MN985325.1, MN988713.1, MN994467.1, MN994468.1, MN997409.1, MN988668.1, MN988669.1, MN996527.1, MN996528.1, MN996529.1, MN996530.1, MN996531.1, MT135041.1, MT135043.1, MT027063.1, and MT027062.1. The bolded ID represents the GenBank reference sequence. Then, the translated coding sequences of the four structural protein sequences were aligned and compared using a multiple sequence alignment (Clustal Omega, EMBL-EBI, United Kingdom). Of the 19 sequences, 15 were identical; however, single AA changes occurred in four N protein sequences: MN988713.1, N 194 S->X; MT135043.1, N 343 D->V; MT027063.1, N 194 S->L; MT027062.1, N 194 S->L. The resulting AA substitutions affected only two positions of N protein sequence (AA 194 and 343), neither of which occurred in epitopes that have been selected as targets for vaccine development.

Recent report (Feb.2021) established four different lineages by analyzing 45,494 complete SARS-CoV-2 genome sequences in the world. Most frequent circulating mutations from this report identified 11 missense amino acid mutations, one in S protein (D614G), three located in N protein (R203K with two different DNA substitutions and G204R), and further seven mutations in NSP2, NSP12, NSP13, ORF3a, and ORF8 (Wang et al., 2021). None of these amino acid positions were included in the nine 30-mers, supporting the proper selection of the conservative regions

<sup>1</sup>Demographics of the world <https://www.quora.com/What-are-all-the-races-and-their-world-population-demographics-the-entire-world>.

<sup>2</sup>U.S. National Library of Medicine Severe acute respiratory syndrome coronavirus 2 <https://www.ncbi.nlm.nih.gov/genome/browse#!/viruses/86693/>.

and intention to identify universal vaccine candidate peptides. Additionally, none of PolyPEPI-SCoV-2 peptides is affected by the presently, emerging mutant SARS-CoV-2 strains, except one single amino acid substitution: B.1.1.7 (UK, 17 mutations: delH69, V70, and Y144, substitutions in S: N501Y, A570D, D614G, P681H, T716I, S982A, D1118H; in N: D3L, S235F; and five mutant positions in ORF1ab), B.1.351 (South Africa, 10 mutations: amino acid substitutions in S: L18F, D80A, D215G, R246I, K417N, E484K, N501Y, A701V; in N: T205I, a single P71L change that affected one amino acid position in our peptide E1, and one non-affecting mutation in ORF1ab) or B.1.1.28.1 (Brazil, 16 mutations in S: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I; in N: P80R, and five mutations in ORF1ab), or B.1.617 (India, “double mutant” with S protein substitutions L452R, E484Q, D614G), B.1.618 (India, “triple mutant” with S protein delH145-146, and substitutions L452R, E484Q, D614G) either<sup>3</sup> (Rambaut et al., 2020; Thomson et al., 2021; O’Toole et al., 2021a,b; Tada et al., 2021). Further details on peptide selection are provided in the Results section and the resulting composition of the nine selected 30-mer peptides is shown in **Table 1**.

### Cross-Reactivity With Human Coronavirus Strains

The sequence of PolyPEPI-SCoV-2 vaccine was compared with that of SARS-CoV, MERS-CoV and common (seasonal) human coronavirus strains to reveal possible cross-reactive regions. According to Centers for Disease Control and Prevention (CDC), common coronaviral infections in the human population are caused by four coronavirus groups: alpha coronavirus 229E and NL63, and beta coronavirus OC43 and HKU1. Pairwise alignment of the structural proteins was also performed using UniProt database with the following reference sequence IDs: 229E: P15423 (S), P15130 (N), P19741 (E), P15422 (M); NL63: Q6Q1S2 (S), Q6Q1R8 (N), Q6Q1S0 (E), Q6Q1R9 (M); OC43: P36334 (S), P33469 (N), Q04854 (E), Q01455 (M); HKU1 (Isolate N1): Q5MQD0 (S), Q5MQC6 (N), Q5MQC8 (E), Q5MQC7 (M) (Consortium, The UniProt, 2018). In addition, the coronavirus strains were aligned with the nine 30-mer peptides comprising the PolyPEPI-SCoV-2 vaccine. For the minimum requirement of an epitope, eight AA-long regions were screened (sliding window) as regions responsible for potential cross-reactivity. In addition, shorter (and longer) length matching consecutive peptide fragments were recorded and reported during the analysis.

### No Cross-Reactivity With Human Protein Sequences

The selected immunogenic peptide candidates of PolyPEPI-SCoV-2 were analyzed by Basic Local Alignment Search (BLAST) analysis to identify any unwanted immunogenic regions in the vaccine that overlap with any proteins or peptides of the human proteome, available at blast.ncbi.nlm.nih.gov. All nine 30-mer peptide sequences were evaluated for homology with human proteins by comparing the sequences against the human protein database (taxid:9606). No cross-reactivity defined by at least eight consecutive amino acid match has been found between the

PolyPEPI-SCoV-2 peptides and proteins in the human proteome, consequently, no related autoimmune reactions are expected due to sequence similarities.

### Peptides and PolyPEPI-SCoV-2 Vaccine Preparation

The 9-mer (s2, s5, s9, n1, n2, n3, n4, e1, m1) and 30-mer (S2, S5, S7, N1, N2, N3, N4, E1, M1) peptides were manufactured by Intavis Peptide Services GmbH&Co. KG (Tübingen, Germany) and PEPScan (Lelystad, The Netherlands) using solid-phase peptide synthesis. Amino acid sequences are provided in **Table 1** for both 9-mer test peptides (**Table 1**, bold) and the 30-mer vaccine peptides. Research grade PolyPEPI-SCoV-2 vaccine for the animal study was prepared by dissolving equal masses of the nine 30-mer peptides in DMSO (Sigma, Hungary) to achieve at a concentration of 1 mg/mL and then diluted with purified water to a final concentration of 0.2 mg/mL and stored frozen until use. Ready-to-inject vaccine preparations were prepared by emulsifying equal volumes of thawed peptide mix solution and Montanide ISA 51 VG adjuvant (Seppic, France) following the standard two-syringe protocol provided by the manufacturer.

### Epitope Prediction and Analysis

Prediction of  $\geq 3$ HLA class I allele binding epitopes (PEPIs) for each individual was performed using an Immune Epitope Database (IEDB)-based epitope prediction method. The antigens were scanned with overlapping 9-mer to identify peptides that bind to the subject’s HLA class I alleles. Selection parameters were validated with an in-house set of 427 HLA-epitope pairs that had been characterized experimentally by using direct binding assays (327 binding and 100 non-binding HLA-epitope pairs). Both specificity and sensitivity resulted in 93% for the prediction of true HLA allele-epitope pairs. HLA class II epitope predictions were performed by NetMHCpan (2.4) prediction algorithm for overlapping 15-mer peptides.

### Preclinical Animal Study Design

Thirty-six Hu-mice and 36 BALB/c mice received PolyPEPI-SCoV-2 vaccine (0.66 mg/kg/peptide in 200  $\mu$ L solution;  $n = 18$ ) or 20% DMSO/water (Sigma, Hungary and MilliQ purified water) emulsified in Montanide ISA 51 VG (Seppic, France) adjuvant (200  $\mu$ L vehicle;  $n = 18$ ) administered subcutaneously on days 0 and 14; the follow up period ended on day 28. Samples from days 14, 21, and 28 were analyzed ( $n = 6$  per cohort). The studies were performed at the Transcure Bioservices facility (Archamps, France). The mice were monitored daily for unexpected signs of distress. Complete clinical scoring was performed weekly by monitoring coat (score 0–2), movement (0–3), activity (0–3), paleness (0–2), and bodyweight (0–3); a normal condition was scored 0.

All procedures described in this study have been reviewed and approved by the local ethic committee (CELEAG) and validated by the French Ministry of Research. Vaccination-induced T cell responses were assessed by *ex vivo* ELISpot and intracellular cytokine staining (ICS) assays of mice splenocytes (detailed below). Antibody responses were investigated by the measurement of total IgG in plasma samples (detailed below).

<sup>3</sup>WHO. Retrieved 6 May 2021. “COVID-19 Weekly Epidemiological Update.”

**TABLE 1** | PolyPEPI-SCoV-2 peptides and comprising PEPI frequencies within the *in silico* human cohort.

SARS-CoV-2 fragment	ID	Peptide (30-mer)	Class I PEPI	Class II PEPI	B cell epitope in SARS (ref)
S (35–64)	S2	GVYYPDKVFRSSVLH <b>STQDLFLPFF</b> SNVTW	71%	94%	N/A
S (253–282)	S5	DSSSGWTAGAAAYYVGY <b>LQPRTFLL</b> KYNEN	84%	97%	N/A
S (893–922) <sup>†</sup>	S9	ALQIP <b>FAMQMAYRFN</b> GIGVTQNVLYENQKL	93%	99%	IgM, 50% ( <i>n</i> = 4) (Guo et al., 2004)
N (36–65) <sup>†</sup>	N1	RSKQRRPQGLPN <b>NTASWFTALT</b> QHKGEDLK	36%	36%	IgG, 62% ( <i>n</i> = 42) (He et al., 2004; Liu et al., 2006)
N (255–284)	N2	SKKPRQKRTAT <b>KAYNVTQAF</b> GRRGPEQTQG	48%	22%	N/A
N (290–319) <sup>†</sup>	N3	ELIRQGTDYKHWPQIA <b>QFAPSASAFF</b> GMSR	54%	50%	IgG, 34% ( <i>n</i> = 42) (He et al., 2004) IgG, IgM, 50% ( <i>n</i> = 4) (Guo et al., 2004)
N (384–413) <sup>†</sup>	N4	QRQKKQQTVTLLPAADLDD <b>FSKQLQQSMSS</b>	23%	36%	IgG, IgM, 95% ( <i>n</i> = 42) (He et al., 2004) IgG, IgM, 75% ( <i>n</i> = 4) (Guo et al., 2004)
M (93–122)	M1	LSYFIASF <b>RLFARTRSM</b> WSFNPETNILLNV	90%	100%	N/A
E (45–74)	E1	NIVNVSLVK <b>PSFYVYSRVK</b> NLNSSRVPDLL	46%	100%	N/A
<b>Combined frequency of PolyPEPI-SCoV-2 PEPIs</b>					
At least one peptide			100%	100%	N/A
At least two peptides			100%	100%	
At least three peptides			97%	100%	

**Bold:** 9-mer HLA class I PEPI sequences; **underlined:** 15-mer HLA class II PEPI sequences within the PolyPEPI-SCoV-2 comprising 30-mer peptides. Percentages show the proportion of individuals from the model population with at least one HLA class I (CD8<sup>+</sup> T cell specific) PEPI or at least one HLA class II (CD4<sup>+</sup> T cell specific) PEPI. Peptides labeled <sup>†</sup> contain experimentally confirmed B cell epitopes with antibody (Ig) responses found in convalescent patients with SARS. N/A, data not available.

## ELISpot/FluoroSpot Assays

*Ex vivo* ELISpot assays for animal studies were performed as follows. IFN- $\gamma$ -producing T cells were identified using  $2 \times 10^5$  splenocytes stimulated for 20 h/peptide (10  $\mu$ g/ml, final concentration). Splenocytes were treated with 9-mer peptides (a pool of four N-specific peptides, N-pool (n1, n2, n3, n4), a pool of three S-specific peptides, S-pool (s2, s5, s9), an E protein-derived peptide, e1 or a M protein-derived peptide, m1) or with 30-mer peptides pooled the same way as 9-mers (N-pool comprising peptides N1, N2, N3, and N4), S-pool comprising peptides S2, S5, and S9, and individual peptides E1 and M1. ELISpot assays were performed using Human IFN- $\gamma$  ELISpot PRO kit (ALP; ref 3321-4APT-2) from Mabtech for Hu-mice cohorts and Mouse IFN- $\gamma$  ELISpot PRO kit (ALP; ref 3321-4APT-10) from Mabtech for BALB/c mice cohorts, according to the manufacturer's instructions. Unstimulated (DMSO) assay control background spot forming unit (SFU) was subtracted from each data point and then the delta SFU (dSFU) was calculated. PMA/Ionomycin (Invitrogen) was used as a positive control.

*Ex vivo* FluoroSpot assays for convalescent donor testing were performed by Nexelis-IMXP (Belgium) as follows: IFN- $\gamma$ /IL-2 FluoroSpot plates were blocked with RPMI-10% FBS, then peptides (5  $\mu$ g/ml final concentration) or peptide pools (5  $\mu$ g/ml per peptide final concentration) were added to the relevant wells. PBMCs of *N* = 17 convalescent donors and *N* = 4 healthy controls were retrieved from cryogenic storage and thawed in culture medium. Then, 200,000 PBMC cells/well

were plated in triplicate (stimulation conditions) or 6-plicates (reference conditions) and incubated overnight at 37°C, 5% CO<sub>2</sub> before development. Development of the FluoroSpot plates was performed according to the manufacturer's recommendations. After removing cells, detection antibodies diluted in PBS containing 0.1% BSA were added to the wells and the FluoroSpot plates were incubated for 2 h at room temperature. Before read-out using the Mabtech IRIS<sup>TM</sup> automated FluoroSpot reader, the FluoroSpot plates were emptied and dried at room temperature for 24 h protected from light. All data were acquired with a Mabtech IRIS<sup>TM</sup> reader and analyzed using Mabtech Apex<sup>TM</sup> software. Unstimulated (DMSO) negative control, CEF positive control (T cell epitopes derived from CMV, EBV and influenza, Mabtech, Sweden), and a commercial SARS-CoV-2 peptide pool (SARS-CoV-2 S N M O defined peptide pool (3622-1)—Mabtech, Sweden) were included as assay controls. *Ex vivo* FluoroSpot results were considered positive when the test result was higher than DMSO negative control after subtracting non-stimulated control (dSFU).

Enriched FluoroSpot assays for convalescent donor testing were performed by Nexelis-IMXP (Belgium) as follows: PBMCs were retrieved from cryogenic storage and thawed in culture medium. The PBMCs of *N* = 17 convalescent donors and *N* = 5 healthy controls were seeded at 4,000,000 cells/24-well in presence of the peptide pools (5  $\mu$ g/ml per peptide) and incubated for 7 days at 37°C, 5% CO<sub>2</sub>. On days 1 and 4 of culture, the media were refreshed and supplemented



with 5 ng/mL IL-7 or 5 ng/mL IL-7 and 4 ng/mL IL-2 (R&D Systems), respectively. After 7 days of culture, the PBMCs were harvested and rested for 16 h. The rested PBMCs were then counted using Trypan Blue Solution, 0.4% (VWR) and the Cellometer K2 Fluorescent Viability Cell Counter (Nexcelom), and seeded on the IFN- $\gamma$ /Granzyme-B/TNF- $\alpha$  FluoroSpot plates (Mabtech) at 200,000 cells/well in RPMI 1640 with 10% Human Serum HI, 2 mM L-glutamin, 50  $\mu$ g/ml gentamycin, and 100  $\mu$ M  $\beta$ -ME into the relevant FluoroSpot wells containing peptide (5  $\mu$ g/mL), or peptide pool (5  $\mu$ g/mL per peptide), in triplicates. The FluoroSpot plates were incubated overnight at 37°C, 5% CO<sub>2</sub> before development. All data were acquired with a Mabtech IRIS<sup>TM</sup> reader and analyzed using Mabtech Apex<sup>TM</sup> software. DMSO, medium only, a commercial COVID peptide pool (SARS-CoV-2 S N M O defined peptide pool [3622-1]—Mabtech), and CEF were included as assay controls at a concentration of 1  $\mu$ g/ml. The positivity criterion was >1.5-fold the unstimulated control after subtracting the background (dSFU).

### Intracellular Cytokine Staining Assay

*Ex vivo* ICS assays for preclinical animal studies were performed as follows: splenocytes were removed from the ELISpot plates after 20 h of stimulation, transferred to a conventional 96-well flat bottom plate, and incubated for 4 h with BD GolgiStop<sup>TM</sup> according to the manufacturer's recommendations. Flow-cytometry was performed using a BD Cytofix/Cytoperm Plus Kit with BD GolgiStop<sup>TM</sup> protein transport inhibitor (containing monensin; Cat. No. 554715), following the manufacturer's instructions. Flow cytometry analysis and cytokine profile determination were performed on an Attune NxT Flow cytometer (Life Technologies). A total of  $2 \times 10^5$  cells were analyzed, gated for CD45<sup>+</sup>, CD3<sup>+</sup>, CD4<sup>+</sup>, or CD8<sup>+</sup> T cells. Counts below 25 were evaluated as 0. Spot counts  $\geq 25$  were background corrected by subtracting unstimulated (DMSO) control. PMA/Ionomycin (Invitrogen) was used as a positive control. As an assay control, Mann-Whitney test was used to compare negative control (unstimulated) and positive control (PMA/ionomycin) for each cytokine. When a statistical difference between controls was determined, the values corresponding to the other stimulation conditions were analyzed. The following stains were used for Hu-mice cohorts: MAb11 502932 (Biolegend), MP4-25D2 500836 (Biolegend), 4S.B3 502536 (Biolegend), HI30 304044 (Biolegend), SK7 344842 (Biolegend), JES6-5H4 503806 (Biolegend), VIT4 130-113-218 (Miltenyi), JES1-39D10 500904 (Biolegend), SK1 344744 (Biolegend), JES10-5A2 501914 (Biolegend), JES3-19F1 554707 (BD), and NA 564997 (BD). The following stains were used for BALB/c mice cohorts: 11B11 562915 (BD), MP6-XT22 506339 (Biolegend), XMG1.2 505840 (Biolegend), 30-F11 103151 (Biolegend), 145-2C11 100355 (Biolegend), JES6-5H4 503806 (Biolegend), GK1.5 100762 (Biolegend), JES1-39D10 500904 (Biolegend), 53-6.7 100762 (Biolegend), eBio13A 25-7133-82 (Thermo Scientific), JESS-16E3 505010 (Biolegend), and NA 564997 (BD).

*Ex vivo* ICS assays for convalescent donor testing were performed by Nexelis-IMXP (Belgium). Briefly, after thawing

200,000 PBMC cells/well, PBMCs were seeded in sterile round-bottom 96-well plates in RPMI total with 10% human serum HI, 2 mM L-glutamine, 50  $\mu$ g/mL gentamycin, and 100  $\mu$ M 2-ME in the presence of peptides (5  $\mu$ g/mL) or peptide pool (5  $\mu$ g/mL per peptide). After a 2-h incubation, BD GolgiPlug<sup>TM</sup> (BD Biosciences) was added to the 96-well plates at a concentration of 1  $\mu$ l/ml in culture medium. After a 10-h incubation, plates were centrifuged (800 g, 3 min, 8°C) and incubated for 10 min at 37°C and Zombie NIR Viability dye (Biolegend) was added to each well. Plates were incubated at room temperature for 15 min, shielded from the light. After incubation, PBS/0.1% BSA was added per well and the plates were centrifuged (800 g, 3 min, 8°C). Thereafter, cells were incubated in FcR blocking reagent at 4°C for 5 min, and then staining mixture (containing anti-CD3, Biolegend, anti-CD4, and anti-CD8 antibodies; BD Biosciences) was added to each well. After 30 min of incubation at 4°C, washing, and centrifugation (800 g, 3 min, 8°C), cells were permeabilized and fixed according to the manufacturer's recommendations (BD Biosciences). After fixation, cytokine staining mixture (containing anti-IFN- $\gamma$ , anti-IL-2, anti-IL-4, anti-IL-10 and anti-TNF- $\alpha$  antibodies, Biolegend) was added to each well. Plates were incubated at 4°C for 30 min and then washed twice before acquisition. All flow cytometry data were acquired with LSRFortessa<sup>TM</sup> X-20 and analyzed using FlowJo V10 software. DMSO negative control was subtracted from each data point obtained using test peptides or pools.

### Antibody ELISA

ELISAs for mouse studies were performed for the quantitative measurement of total mouse IgG production in plasma samples using IgG (Total) Mouse Uncoated ELISA Kit (Invitrogen, #88-50400-22) for BALB/c cohorts and IgG (Total) Human Uncoated ELISA Kit (Invitrogen, #88-50550-22) for Hu-mice cohorts according to the manufacturer's instructions. Analyses were performed using samples harvested at days 14, 21, and 28 ( $n = 6$  per group per time point). Absorbance were read on an Epoch Microplate Reader (Biotech) and analyzed using Gen5 software.

Euroimmune ELISA assays for convalescent donors were performed to determine S1-specific IgG levels via the independent medical research center, The Netherlands. The Anti-SARS-CoV-2 ELISA plates are coated with recombinant S-1 structural protein from SARS-CoV-2 to which antibodies against SARS-CoV-2 bind. This antigen was selected for its relatively low homology to other coronaviruses, notably SARS-CoV. The immunoassay was performed according to the manufacturer's instructions.

ELISAs were performed by Mikromikomed Kft (Budapest, Hungary) using a DiaPro COVID-19 IgM Enzyme Immunoassay for the determination of IgM antibodies to COVID-19 in human serum and plasma, DiaPro COVID-19 IgG Enzyme Immunoassay for the determination of IgG antibodies to COVID-19 in human serum and plasma, and DiaPro COVID-19 IgA Enzyme Immunoassay for the determination of IgA antibodies to COVID-19 in human serum and plasma, according to the manufacturer's instructions (Dia.Pro Diagnostic Bioprobes S.r.l., Italy). For the determination



of N-specific antibodies, Roche Elecsys® Anti-SARS-CoV-2 Immunoassay for the qualitative detection of antibodies (including IgG) against SARS-CoV-2 was used with a COBAS e411 analyzer (disk system; ROCHE, Switzerland) according to the manufacturer's instructions.

(Vero C1008 (ATCC No." should be replaced with "(Vero C1008, ATCC No.

## Pseudoparticle Neutralization Assay

Neutralizing antibodies in mice sera were assessed using a cell-based Pseudoparticle Neutralization Assay. Vero E6 cells expressing the ACE-2 receptor (Vero C1008 ATCC No. CRL-1586, US), were seeded at 20 000 cells/well to reach a cell confluence of 80%. Serum samples and controls (pool of human convalescent serum, internally produced) were diluted in duplicates in cell growth media at a starting dilution of 1/25 (for samples) or 1/100 (for controls), followed by a serial dilution (2-fold dilutions, five times). In parallel, SARS-CoV-2 pseudovirus (prepared by Nexelis, using Kerafast system), with Spike from Wuhan Strain (minus 19 C-terminal amino acids) was diluted as to reach the desired concentration (according to pre-determined TU/mL). Pseudovirus was then added to diluted sera samples and pre-incubated for 1 h at 37°C with CO<sub>2</sub>. The mixture was then added to the pre-seeded Vero E6 cell layers and plates were incubated for 18–24 h at 37°C with 5% CO<sub>2</sub>. Following incubation and removal of media, ONE-Glo EX Luciferase Assay Substrate, Promega, Cat. E8110) was added to cells and incubated for 3 min at room temperature with shaking. Luminescence was measured using a SpectraMax i3x microplate reader and SoftMax Pro v6.5.1 (Molecular Devices). Luminescence results for each dilution were used to generate a titration curve using a 4-parameter logistic regression (4PL) using Microsoft Excel (for Microsoft Office 365). The titer was defined as the reciprocal dilution of the sample for which the luminescence is equal to a pre-determined cut-point of 50, corresponding to 50% neutralization. This cut-point was established using linear regression using 50% flanking points.

## Statistical Analysis

Significance was compared between and among groups using *t*-tests, Mann-Whitney tests, or Permutation statistics using Montecarlo simulations, as appropriate. *p* < 0.05 was considered significant. Pearson's test and/or Spearman's test was performed to assess correlations. The correlation was considered strong if *R* > 0.7, moderate, if 0.5 < *R* ≤ 0.7 and weak, if 0.3 < *R* ≤ 0.5. Dependent variables were determined using Fisher Exact test for a 2 × 2 contingency table.

## RESULTS

### Tailoring PolyPEPI-SCoV-2 to Individuals' Genetic Profile

During the design of PolyPEPI-SCoV-2, we used the HLA genotype data of subjects in the *in silico* human cohort (Model Population) to determine the most immunogenic peptides (i.e., HLA class I PEPI hotspots, 9-mers) of the four selected

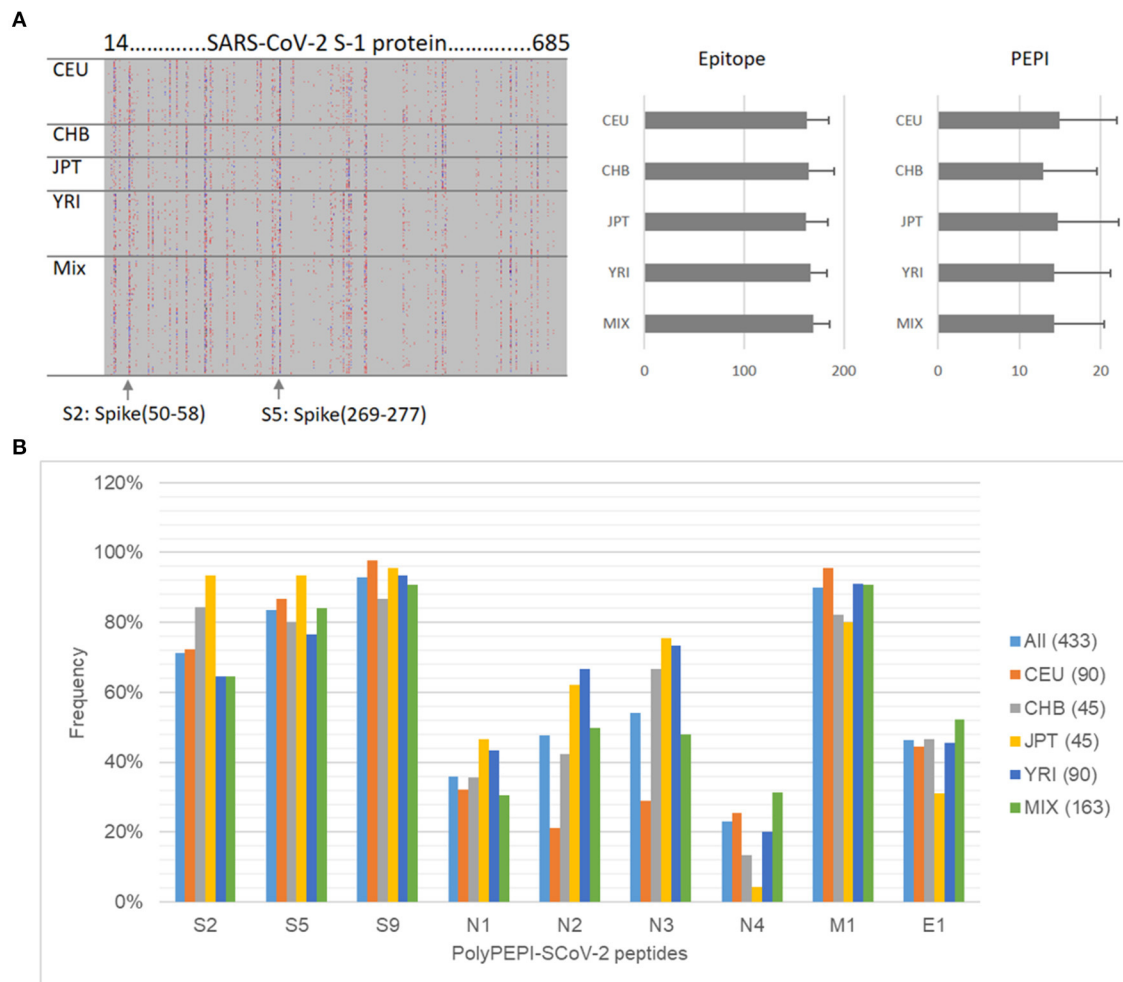
SARS-CoV-2 structural proteins aimed to induce CD8<sup>+</sup> T cell responses. The sequences of the identified 9-mer PEPI hotspots were then extended to 30-mers based on the viral protein sequences to maximize the number of HLA class II binding PEPIs (15-mers) aimed to induce CD4<sup>+</sup> T cell responses as detailed below.

First, we performed epitope predictions for each subject in the *in silico* human cohorts for each of their HLA class I and class II alleles (six HLA class I and class II alleles) for the AA sequence of the conserved regions of 19 known SARS-CoV-2 viral proteins using 9-mer (HLA class I) and 15-mer (HLA class II) frames, respectively (**Figure 1A**; section Materials and Methods). Then, we selected the epitopes restricted to multiple (≥3) autologous HLA alleles (PEPIs) to account for the most abundantly presented epitopes. We identified several HLA-restricted epitopes (average, 166 epitopes only for S1 protein) for each person. In contrast, PEPIs are represented at much lower level in all ethnicities (average, 14 multi-HLA binding epitopes, **Figure 1A**). Of note, we did not observe any difference in SARS-CoV-2 S1-protein specific epitope generation capacity of individuals with different ethnicities based on their complete HLA genotype, which does not seem to explain the higher infection and mortality rates observed in BAME. Instead, we observed heterogeneity in the frequency of the shared PEPIs in the different ethnic groups, especially for protein N, having high impact on the design of a potential global vaccine (**Figure 1B**). Combination of targets with different frequencies inside- and between ethnic groups into a vaccine candidate with high global coverage is feasible only by performing "*in silico* clinical trials" in large populations of real subjects.

Therefore, to maximize multi-antigenic immune responses at both the individual and population/ethnicity levels, and also considering the chemical and manufacturability properties of the peptides, we selected a total of nine 30-mer peptides from four structural proteins of SARS-CoV-2: three peptides from spike (S), four peptides from nucleoprotein (N), and one peptide from each matrix (M) and envelope (E). No peptides were included from the receptor-binding domain (RBD) of S protein. Overall, each member of the Model Population had HLA class I PEPIs for at least two of the nine peptides, and 97% had at least three (**Table 1**). Each subject had multiple class II PEPIs for the vaccine peptides (**Table 1**).

We identified experimentally confirmed linear B cell epitopes derived from SARS-CoV, with 100% sequence identity to the relevant SARS-CoV-2 antigen, to account for the potential B cell production capacity of the long peptides (Ahmed et al., 2020). Three overlapping epitopes located in N protein- and one epitope in S protein-derived peptides of PolyPEPI-SCoV-2 vaccine were reactive with the sera of convalescent patients with severe acute respiratory syndrome (SARS). This suggests that the above antigenic sites on the S and N protein are important in eliciting humoral immune response against SARS-CoV and likely against SARS-CoV-2, in humans.

None of the peptides involved in PolyPEPI-SCoV-2 composition are cross-reactive with the human genome at



**FIGURE 1 |** Design of PolyPEPI-SCoV-2. **(A)** Hotspot analysis of SARS-CoV-2 Spike-1 protein in the ethnically diverse *in silico* human cohort. Analysis was performed by predicting  $\geq 3$  HLA allele binding personal epitopes (PEPIs) for each subject. Left panel: Each row along the vertical axis represents one subject in the model population, while the horizontal axis represents the SARS-CoV-2 S-1 protein sequence. Vertical bands represent the most frequent PEPIs, i.e., the dominant immunogenic protein regions (hotspots). Colors represent the number of epitopes restricted to a subject's: red, 3; green/blue, 4; black,  $>5$  HLA class I allele. Right panel, average number of epitopes/PEPIs found for subjects of different ethnicities. **(B)** Heterogeneity of peptide frequencies in different ethnic groups. CEU, Central European; CHB, Chinese; JPT, Japanese; YRI, African; Mix, mixed ethnicity subjects.

minimal epitope level, as assessed by BLAST analysis (see section Methods). As expected, PolyPEPI-SCoV-2 contains several (eight out of nine) peptides that are cross-reactive with SARS-CoV due to high sequence homology between SARS-CoV-2 and SARS-CoV. Sequence similarity is low between the PolyPEPI-SCoV-2 peptides and common (seasonal) coronavirus strains belonging to alpha coronavirus (229E and NL63), beta coronavirus (OC43, HKU1) and MERS. Therefore, cross-reactivity between the vaccine and prior coronavirus-infected individuals remains low and might involve only the M1 peptide of the vaccine (See section Materials and Methods; **Supplementary Table 5**). However, none of the peptides involved in the PolyPEPI-SCoV-2 vaccine composition is affected by the emergent SARS-CoV-2 variants and mutations known to date (See Materials and Methods for the analysis).

## PolyPEPI-SCoV-2-Specific T Cell Responses Detected in COVID-19 Convalescent Donors

Next, we aimed to demonstrate that shared PEPIs identified for the *in silico* cohort are also present in the T cell repertoire of natural SARS-CoV-2 infection by investigating vaccine-specific T cells circulating in the blood of COVID-19 convalescent donors (donor information are reported in **Supplementary Tables 1, 2**).

First, the reactivity of vaccine peptides with convalescent immune components was investigated in 17 convalescent and four healthy donors using *ex vivo* FluoroSpot assay which can detect rapidly activating, effector phase T cell responses. Vaccine-reactive CD4<sup>+</sup> T cells were detected using the nine 30-mer vaccine peptides grouped in four pools according to their source

protein: S, N, M, and E peptides. CD8<sup>+</sup> T cell responses were measured using the 9-mer test peptides as well corresponding to the shared HLA class I PEPs defined for each of the nine vaccine peptides, grouped into four pools (s, n, m, and e peptides; **Table 1**, bold). Significant amount of vaccine-reactive, IFN- $\gamma$ -expressing T cells were detected with both 30-mer (average dSFU: 48.1) and 9-mer peptides (average dSFU: 16.5) compared with healthy subjects (**Figure 2A**). Detailed analysis of the four protein-specific peptide pools revealed that three out of the 17 donors reacted to all four structural proteins with the 30-mer vaccine peptides; 82% of donors reacted to two proteins and 59% to three proteins. Notably, highly specific 9-mer-detected CD8<sup>+</sup> T cell responses could be also identified against at least one of four proteins in all 17 donors and against at least two proteins in 53% (**Supplementary Table 6**).

As determined by ICS assay, stimulation with 9-mer test peptides resulted in an average T cell make up of 83% CD8<sup>+</sup> T cells, and 17% CD4<sup>+</sup> T cells (**Supplementary Figures 2A,B**). The 30-mer peptides reacted with both CD4<sup>+</sup> and CD8<sup>+</sup> T cells in average ratio of 50:50 (**Supplementary Figure 2B**). Functionality testing of the T cells revealed that CD8<sup>+</sup> T cells primarily produced IFN- $\gamma$ , TNF- $\alpha$ , and IL-2 (with small amounts of IL-4 and IL-10), while CD4<sup>+</sup> T cells were positive for mainly IL-2 and IFN- $\gamma$ . Recall responses demonstrated clear Th1 cytokine characteristics; Th2 responses were not present in the recall response detected with 30-mer vaccine peptides (**Supplementary Figure 2C**).

Next, we determined whether the *ex vivo* detected T cells could also expand *in vitro* in the presence of vaccine peptides. Using enriched FluoroSpot, significant numbers of vaccine-reactive, IFN- $\gamma$ -expressing T cells were detected with both 30-mer (average dSFU = 3,746) and 9-mer (average dSFU = 2,088) peptide pools compared with healthy subjects (**Figure 2B**). The intensity of the PolyPEPI-SCoV-2-derived T cell responses (30-mer pool) were also evaluated relative to the responses detected with a commercial, large SARS-CoV-2 peptide pool (SMNO) containing 47 long peptides derived from both structural (S, M, N) and non-structural (open reading frame ORF-3a and 7a) proteins. Interestingly, the magnitude of T cell responses were similar for the two peptide pools despite of the difference in their size, suggesting more prevalent responses for our peptide mix. In addition, the vaccine pool was favored by the COVID-19 donors, while healthy donors preferred the commercial peptide pool, confirming improved specificity of PolyPEPI-SCoV-2 peptides to SARS-CoV-2, in conformance with the result of cross-reactivity analysis with common coronavirus strains (**Figure 2B**).

To confirm and further delineate the multi-specificity of the PolyPEPI-SCoV-2-specific T cell responses of COVID-19 recovered individuals, we defined the distinctive peptides targeted by their T cells. We first deconvoluted the peptide pools and tested the CD8<sup>+</sup> T cell responses specific to each of the 9-mer HLA class I PEPs corresponding to each vaccine peptide using *in vitro* expansion (**Figure 2C**; **Supplementary Figure 3**). Analysis revealed that each 9-mer peptide was recognized by several subjects; the highest recognition rate in COVID-19 convalescent donors was observed for n4 and n3 (93%), s9 (87%), s2, n1, m1 (80%), e1 (60%), s5, n2 (40%) (**Figure 2C**). Detailed analysis of

the nine peptide-specific CD8<sup>+</sup> T cell responses revealed that 100% of COVID-19-recovered subjects had PolyPEPI-SCoV-2-specific T cells reactivated with at least one peptide, 93% with more than two, 87% with more than five, and 27% had T cell pools specific to all nine vaccine peptides. At the protein level, 87% of subjects had T cells against multiple (three) proteins and eight out of the 15 measured donors (53%) reacted to all four targeted viral proteins (**Figure 2C**). These data confirm that PolyPEPI-SCoV-2-peptides are dominant for an individual and shared between COVID-19 subjects. Convalescents' T cells recognizing PolyPEPI-SCoV-2-specific 9-mer peptides were fully functional, expressing IFN- $\gamma$  and/or TNF- $\alpha$  and/or Granzyme-B (**Supplementary Figure 4**). For our cohort of convalescent subjects, the breadth and magnitude of vaccine-specific T cell responses were independent of time from symptom onset, suggesting that these T cells are persistent (for at least 5 months) (**Supplementary Figure 5**).

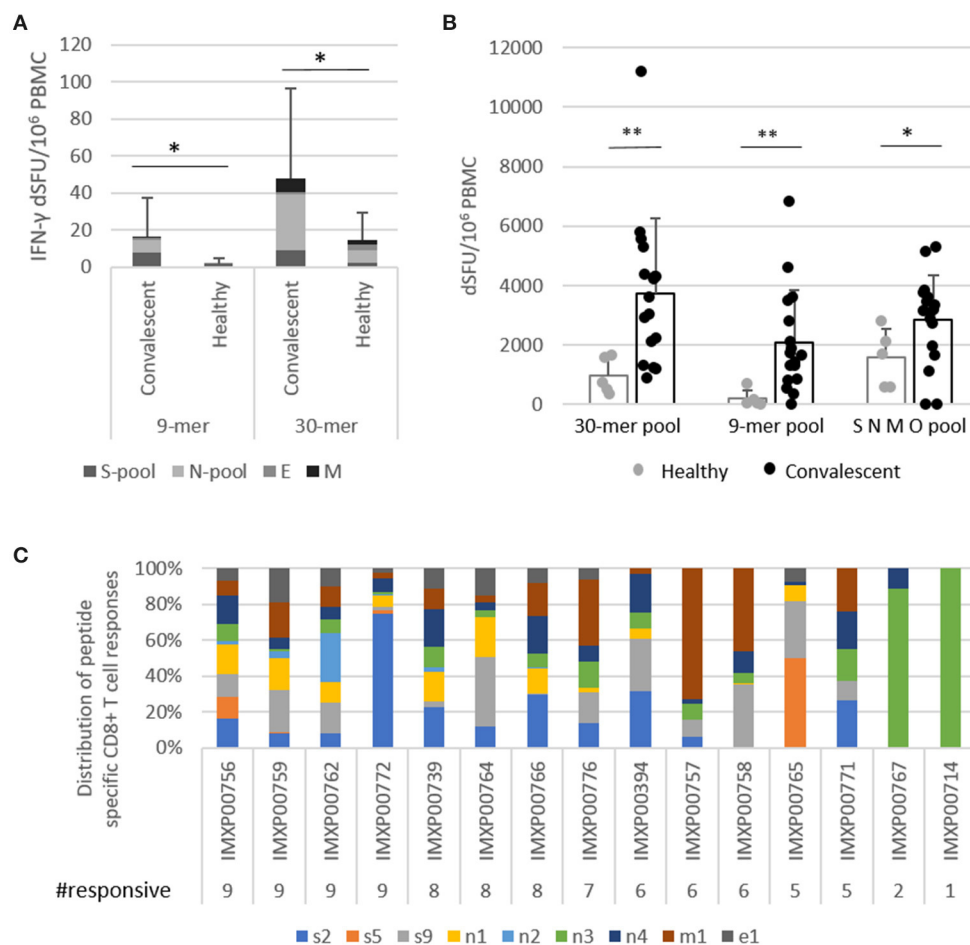
As an external validation, we determined the frequency of PolyPEPI-SCoV-2-specific T cell responses in a second convalescent cohort reported by Tarke et al. Uniquely, this study reports individual T cell response data to predicted MHC I and II-epitope pairs. Eight of our nine peptides (identical or overlapping sequences with at least eight amino acids) were tested for 42 convalescent subjects, with average 4.5 out of eight peptides being tested per subject.

We found that each of the eight peptides (100%) were shared between at least three subjects tested, peptide N3 was found for 13/27 (48%) of subjects (**Supplementary Figure 6**). Furthermore, 26/42 (62%) of subjects had T cell responses specific for at least one, 14/42 (33%) for at least two and 8/42 (19%) for at least three PolyPEPI-SCoV-2 peptides (**Supplementary Figure 6**). These data confirm the immunoprevalent nature of our peptides in an independent cohort of convalescents using an *ex vivo* T cell receptor dependent Activation Induced Marker (AIM) assay.

## Correlation Between PolyPEPI-SCoV-2-Reactive T Cells and SARS-CoV-2-Specific Antibody Responses

T cell-dependent B cell activation is required for antibody production. For each subject, different levels of antibody responses were detected against both S and N antigens of SARS-CoV-2 determined using different commercial kits (**Table 1**). All subjects tested positive with Euroimmune ELISA against viral S1 subunit (IgG-S1) and a Roche kit to measure N-related antibodies (IgG-N). All subjects tested positive for DiaPro IgG and IgM (except two donors), 7/17 for DiaPro IgA detecting mixed S1 and N protein-specific antibody responses (**Supplementary Table 1**).

We next evaluated the correlation between PolyPEPI-SCoV-2-specific CD4<sup>+</sup> T cell reactivities and antibody responses (**Figure 3**). The total amount of PolyPEPI-SCoV-2-reactive CD4<sup>+</sup> T cells correlated with IgG-S1 ( $R = 0.59$ ,  $p = 0.02$ , **Figure 3A**). Next, the subset of CD4<sup>+</sup> T cells reactive to specific S1 protein subunit-derived peptides of the PolyPEPI-SCoV-2 vaccine (S2 and S5) were analyzed and the correlation was similar ( $R = 0.585$ ,  $p = 0.02$ , **Figure 3B**). T cell responses detected with N protein derived PolyPEPI-SCoV-2 peptides



**FIGURE 2 |** PolyPEPI-SCoV-2-specific T cells detected for COVID-19 convalescent donors. **(A)** Highly specific vaccine-derived 9-mer peptide-reactive CD8<sup>+</sup> T cells and 30-mer peptide-reactive CD4<sup>+</sup> T cells detected by *ex vivo* FluoroSpot assay. Test conditions: S-pool contains the three peptides derived from S protein; N-pool contains the four peptides derived from N protein; M and E are the pepi/des derived from M and E proteins, respectively, in both the 9-mer and 30-mer pools. **(B)** IFN- $\gamma$  producing T cells activated with 30-mer peptides in one pool, 9-mer peptides in one pool, and a commercial SNMO peptide pool detected using enriched FluoroSpot assay. **(C)** IFN- $\gamma$  producing CD8<sup>+</sup> T cells activated by individual 9-mer peptides corresponding to each of the 30-mer peptides with the same name (**Table 1** bold), detected using enriched FluoroSpot assay. dSFU, delta spot forming units, calculated as background corrected spot counts per 10<sup>6</sup> PBMC. Significance was calculated using Permutation statistics with Montecarlo simulations; \* $p < 0.05$ , \*\* $p < 0.00005$ .

(N1, N2, N3, and N4) presented a weak but not significant correlation with IgG-N (**Figure 3C**). These data suggest a link between PolyPEPI-SCoV-2-specific CD4<sup>+</sup> T cell responses and subsequent IgG production for COVID-19 convalescent donors. Interestingly, IgA production correlated with PolyPEPI-SCoV-2-specific memory CD4<sup>+</sup> T cell responses ( $R = 0.63$ ,  $p = 0.006$ , **Figure 3D**, although Spearman test did not confirm the correlation). T cell responses reactive to the SMNO peptide pool exhibited no correlation with any of the antibody subsets. This suggests that not all CD4<sup>+</sup> T cells contributed to B cell responses, consequently to IgG production.

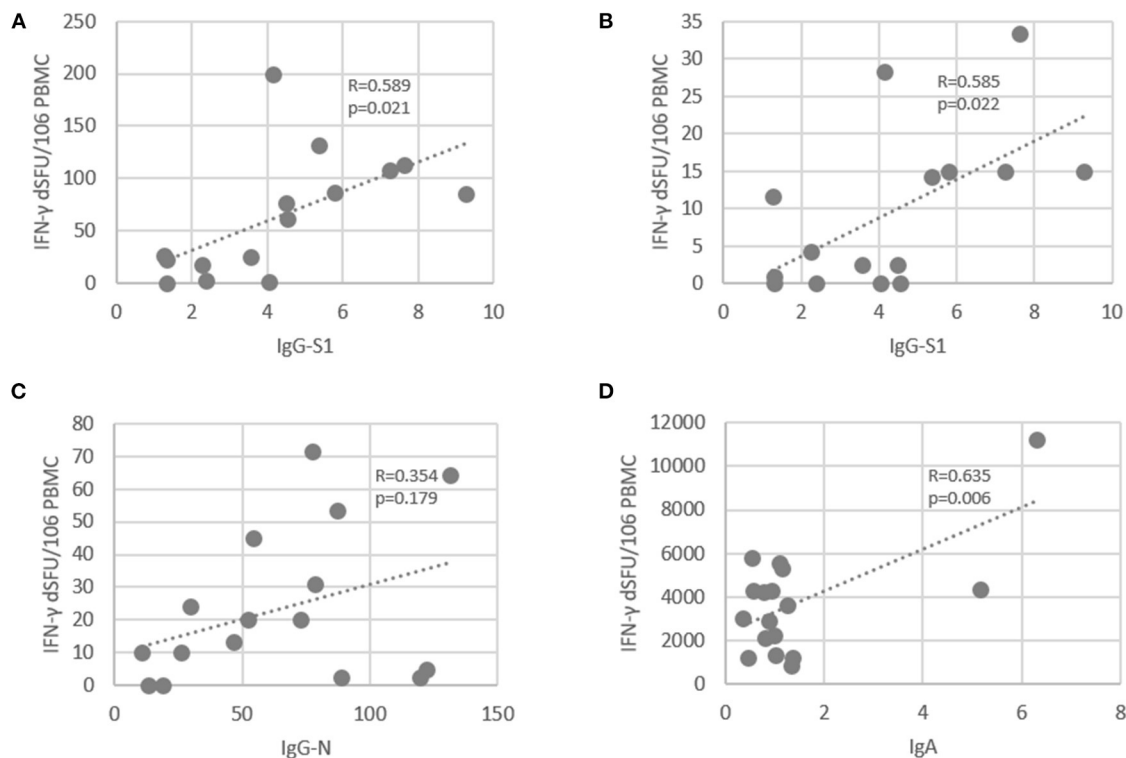
### Correlation Between Multiple Autologous Allele-Binding Epitopes (PEPIs) and CD8<sup>+</sup> T Cell Responses

We investigated the HLA-binding capacity of the immunogenic peptides detected for each subject.

First we determined the complete HLA class I genotype for each subject and then predicted the number of autologous HLA alleles that could bind to each of the nine shared 9-mer peptides used in the FluoroSpot assay. Then we matched the predicted HLA-binding epitopes to the CD8<sup>+</sup> T cell responses measured for each peptide in each patient (total  $15 \times 9 = 135$  data points, **Supplementary Figure 7**). The magnitude of CD8<sup>+</sup> T cell responses tended to correlate with epitopes restricted to multiple autologous HLA alleles ( $R_s = 0.188$ ,  $p = 0.028$ , **Figure 4A**). In addition, we observed that the magnitude of CD8<sup>+</sup> T cell responses generated by PEPIs (HLA  $\geq 3$ ) (median dSFU = 458) was significantly higher than those generated by non-PEPIs (HLA  $< 3$ ) (median dSFU = 110), ( $p = 0.008$ ) (**Figure 4B**).

Across the 135 data points there were 98 positive responses and 37 negative responses recorded. Among the 98 positive responses 37 were generated by PEPIs, while among the 37 negatives only seven were PEPIs, the others





**FIGURE 3 |** Correlation between SARS-CoV-2-specific antibody levels and PolyPEPI-SCoV-2-specific IFN- $\gamma$ -producing CD4<sup>+</sup> T cells in COVID-19 convalescent individuals. **(A)** T cell responses reactive to 30-mer pool of PolyPEPI-SCoV-2 peptides were plotted against the IgG-S1 (Euroimmune). **(B)** Average T cell responses reactive to S1 protein subunit-derived 30-mer peptides (S2 and S5) was plotted against IgG-S1 (Euroimmune). **(C)** T cell responses reactive to 30-mer N peptide pool comprising N1, N2, N3, and N4 was plotted against total IgG-N measured with Roche Elecsys<sup>®</sup> assay. **(D)** T cell responses reactive to 30-mer pool of PolyPEPI-SCoV-2 peptides were plotted against the IgA antibody amounts measured by DiaPro IgA ELISA assay. R: Pearson correlation coefficient.

were epitopes restricted to <3 autologous HLA alleles (Supplementary Figure 7). Overall, the  $2 \times 2$  contingency table revealed association of T cell responses with PEPIs ( $p = 0.041$ , Fisher Exact) but not with HLA-restricted epitopes ( $p = 1.000$ , Fisher Exact) (Figure 4C). For each subject between one and seven peptides out of nine proved to be PEPIs. Among the predicted PEPIs, 37/44 (84%) were confirmed by IFN- $\gamma$  FluoroSpot assay to generate specific T cell responses in the given subject (Figure 4D; Supplementary Figure 7).

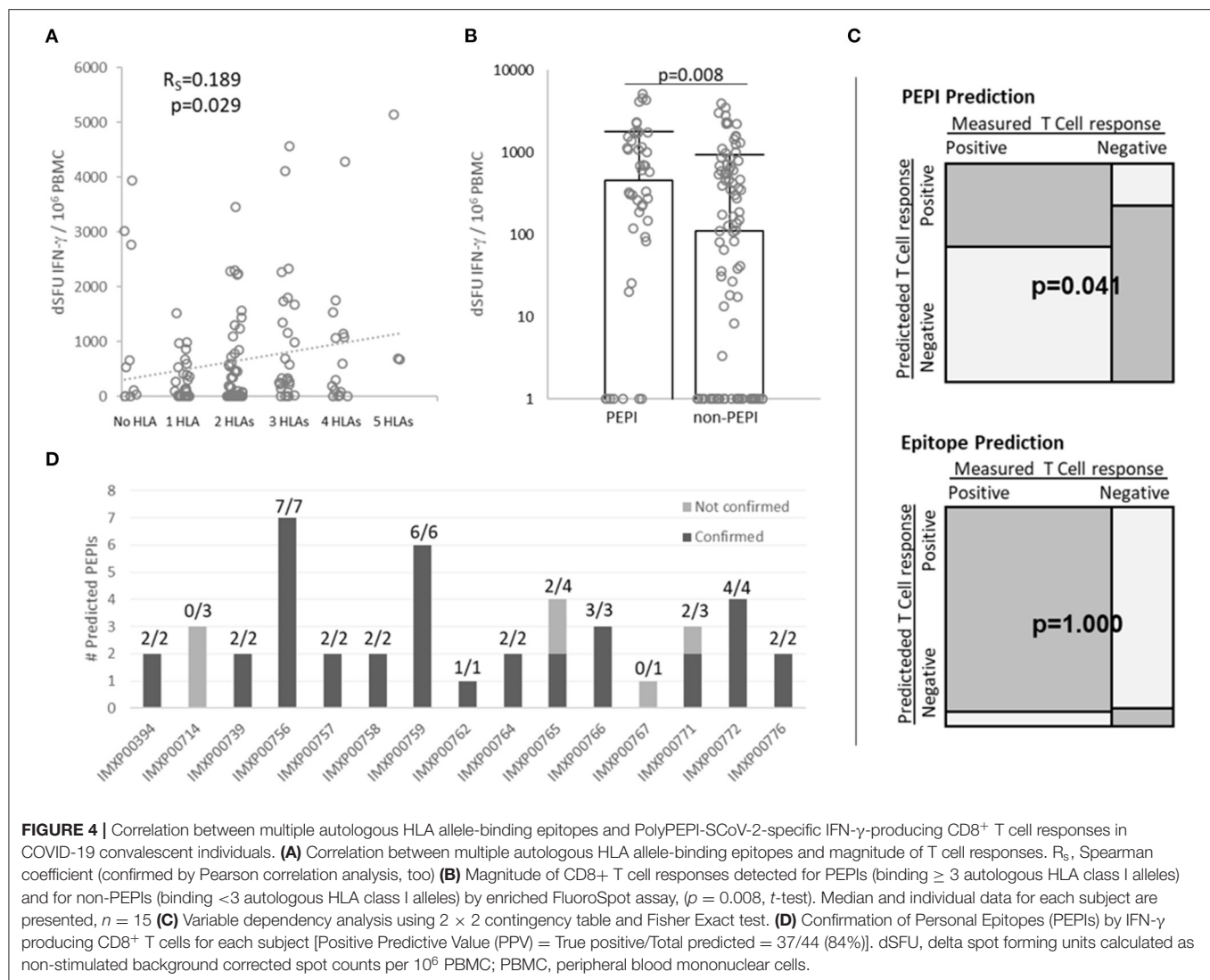
These data demonstrate that subjects' complete HLA-genotype influence their CD8<sup>+</sup> T cell responses and multiple autologous allele-binding capacity is a key feature of immunogenic epitopes. PEPIs in general underestimated the subject's overall T cell repertoire, however they precisely predicted subjects' PEPI-specific CD8<sup>+</sup> T cell responses.

## Predicted Immunogenicity in Different Ethnicities

Since the T cell responses detected in convalescents validated our hypothesis that PEPIs determined for an individual's HLA genotype generate CD8<sup>+</sup> T cell responses with high predictive value, we used this knowledge to determine the scalability of our approach and estimate the global coverage of our vaccine

candidate. As expected, the measured peptide-specific CD8<sup>+</sup> T cell frequencies obtained in the convalescent population were in good agreement with their predicted PEPI frequencies and also with the frequency of shared PEPIs of the Model Population ( $n = 433$ ) cohort used for the design (100% for at least one peptide for both predicted PEPIs and measured CD8<sup>+</sup> T cell frequencies; 93% measured T cell response vs. 100% predicted for at least two peptides) (Figure 5A; Table 1). The polypeptide-specific T cell responses were however underestimated by both the individual HLA-genotypes and the Model Population compared to measured T cell responses.

To estimate the scalability of our *in silico* model, we determined the PEPI frequencies for a large cohort of 16,000 HLA-genotyped subjects distributed among 16 different ethnic groups obtained from a US bone marrow donor database. The ethnic groups covered in this cohort are representative for the composition of the global population and involves 99.8% of the alleles cataloged in the CIWD database for > 8 million human subjects globally (compared to 97.4% in the Model Population) (see section Methods) (Hurley et al., 2020). The CIWD database contains frequent alleles (documented for  $\geq 5$  subjects) as well as rare alleles (documented for <5 subjects). The PEPI frequencies obtained for our Model Population ( $n = 433$ ) and this large US

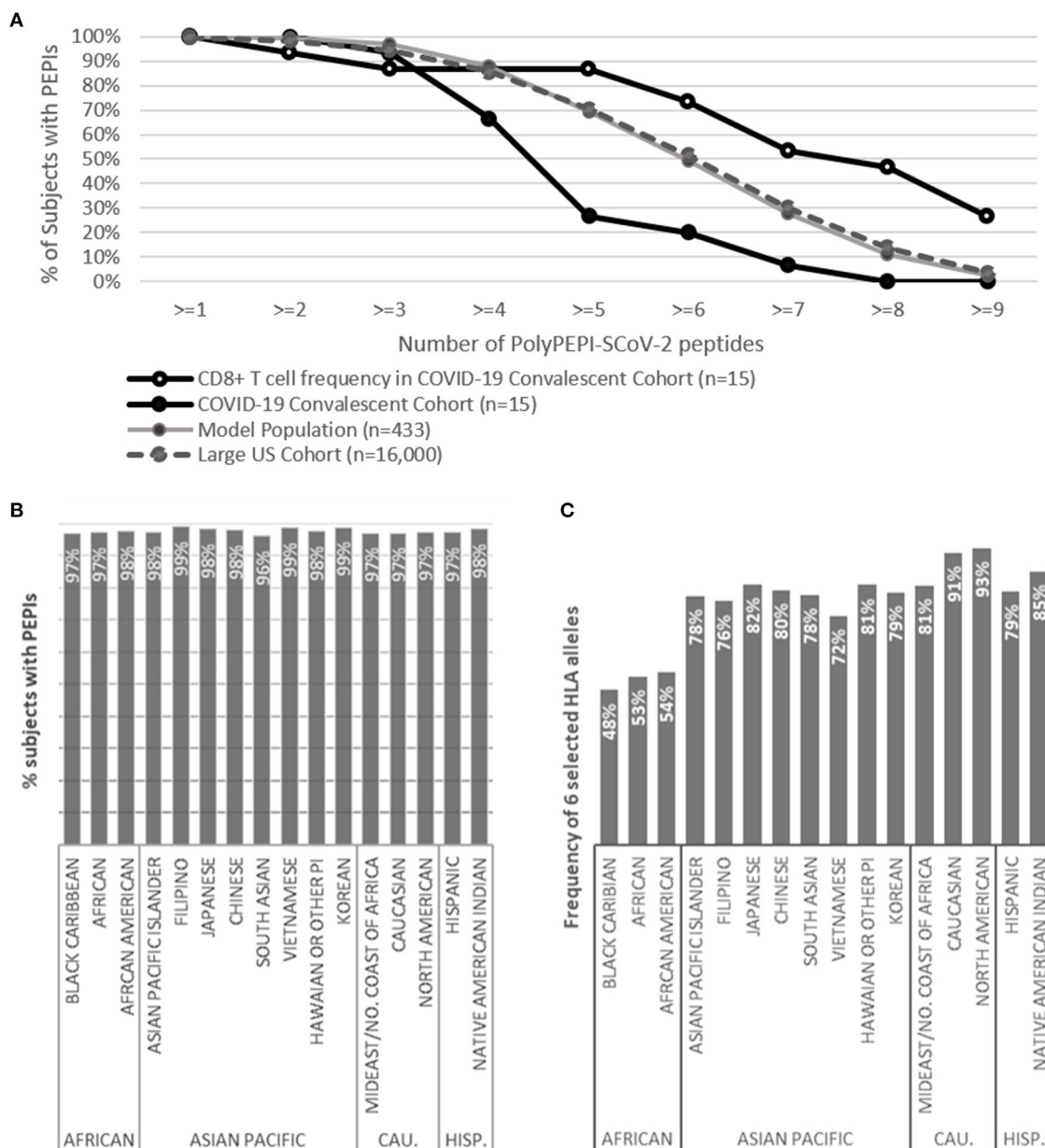


cohort ( $n = 16,000$ ) were in perfect alignment, suggesting high global coverage ensured by the high number of frequent alleles covered in the Model Population and an overall low impact of the rare alleles found in the individuals' HLA-genotype (Figure 5A; Supplementary Figures 8A–C). In the large US cohort, most subjects had a broad repertoire of predicted PEPIs that based on the above findings will likely be transformed to multiple virus-specific memory CD8<sup>+</sup> T cell clones: 98% of subjects were predicted to have PEPIs against at least two vaccine peptides, and 95, 86, and 70% against three, four, and five peptides, respectively (Figure 5A).

*In silico* testing revealed that 96–99% of subjects in each ethnic group will likely mount robust cellular responses, with both CD8<sup>+</sup> and CD4<sup>+</sup> T cell responses against at least two peptides in the vaccine (Figure 5B). This predicted high response rate was also true for the ethnicities reported to have worse clinical outcomes from COVID-19 (Black, Asian) (Pan et al., 2020). Based on these data, we expect that the vaccine

will provide global coverage, independent of ethnicity and geographic location.

We also used this cohort (and comprising ethnic groups) to assess theoretical global coverage as proposed by others, by filtering the sub-populations having at least one of the six prevalent HLA class I alleles considered to cover 95% of the global population (Maiers et al., 2007; Gonzalez-Galarza et al., 2019; Ferretti et al., 2020). Using this approach, we observed significant heterogeneity at the ethnicity level. While we confirmed that the selected six HLA alleles are prevalent in the Caucasian and North American cohorts (91–93%), the frequency of these alleles was lower in all other ethnic groups, especially in African populations (48–54%) (Figure 5C). We concluded that the proposed prevalent HLA allele set may cover the HLA frequency in an ethnically weighted global population, but epitope selection for vaccination purposes based only on these alleles would discriminate some ethnicities. Therefore, we propose using a representative model population that is sensitive



**FIGURE 5 |** Predicted global coverage in a large population with different ethnicities. **(A)** Proportion of subjects having predicted HLA class I PEPs against PolyPEPI-SCoV-2 peptides in different cohorts and the frequency of experimentally measured CD8<sup>+</sup> T cell responses in the COVID-19 convalescent cohort ( $n = 15$ ), obtained by FluoroSpot assay. **(B)** Proportion of subjects having both HLA class I and class II PEPs against at least two peptides in the PolyPEPI-SCoV-2 vaccine. **(C)** Theoretical global coverage estimated based on the frequency of six prevalent HLA alleles (A\*02:01, A\*01:01, A\*03:01, A\*11:01, A\*24:02, and B\*07:02), as proposed by Ferretti et al. (2020). CAU., Caucasian; HISP., Hispanic;  $n = 16,000$ .

to the heterogeneities in the human race and that allows selecting PEPs shared among individuals across ethnicities.

## PolyPEPI-SCoV-2 Vaccine Candidate Induced Broad T Cell Responses in Two Animal Models

Preclinical immunogenicity testing of PolyPEPI-SCoV-2 was performed to measure the induced immune responses after

one and two vaccine doses that were administered 2 weeks apart (days 0 and 14) in BALB/c and Hu-mouse models. After immunizations, no mice presented any clinical score at day 14, 21 or 28 (score 0, representing no deviation from normal), suggesting the absence of any side effects or immune aversion (**Supplementary Tables 7A,B**). In addition, the necropsies performed by macroscopic observation at each timepoint did not reveal any visible organ alteration in spleen, liver, kidneys, stomach and intestine (**Supplementary Table 7C**).

Repeated vaccine administration was also well-tolerated, and no signs of immune toxicity or other systemic adverse events were detected. Together, these data strongly suggest that PolyPEPI-SCoV-2 was safe in mice.

Vaccine-induced IFN- $\gamma$  producing T cells were measured after the first dose at day 14 and after the second dose at days 21 and 28. Vaccine-induced T cells were detected using the nine 30-mer vaccine peptides grouped in four pools according to their source protein: S, N, M, and E, to assess for the CD4<sup>+</sup> and CD8<sup>+</sup> T cell responses. CD8<sup>+</sup> T cell responses were also specifically measured using the short 9-mer test peptides corresponding to the shared HLA class I PEPIs defined above for each of the nine vaccine peptides, in four pools (s, n, m, and e peptides; **Table 1** bold).

In BALB/c mice at day 14, PolyPEPI-SCoV-2 vaccination did not induce more IFN- $\gamma$  production than the Vehicle (DMSO/Water emulsified with Montanide), this latter resulting in unusually high response probably due to Montanide mediated unspecific responses. Nevertheless, at days 21 and 28, the second dose of PolyPEPI-SCoV-2 increased IFN- $\gamma$  production compared to Vehicle control group by 6-fold and 3.5-fold for splenocytes detected with the 30-mer and 9-mer peptides, respectively (**Figure 6A**).

In immunodeficient Hu-mice at day 14, PolyPEPI-SCoV-2 vaccination increased IFN- $\gamma$  production by 2-fold with splenocytes specific for the 9-mer pool of peptides, but no increase was observed with 30-mer-stimulated splenocytes. At days 21 and 28, the second dose of PolyPEPI-SCoV-2 boosted IFN- $\gamma$  production by 4- and 2-fold with splenocytes detected with the 30-mer and 9-mer pools of peptides, respectively (**Figure 6B**). Importantly, both 9-mer-detected CD8<sup>+</sup> T cells and 30-mer-detected CD4<sup>+</sup> and CD8<sup>+</sup> T cell responses were directed against all four viral proteins targeted by the vaccine in both animal models (**Figures 6C,D; Supplementary Figures 9A–F**). Since the Hu-mouse model was developed by transplanting human CD34<sup>+</sup> hematopoietic stem cells to generate human antigen-presenting cells and T- and B-lymphocytes into NOD/Shi-scid/IL-2R $\gamma$  null immunodeficient mice, this model provides a real human immune system model (Brehm et al., 2013). Therefore, the robust multi-antigenic CD4<sup>+</sup> and CD8<sup>+</sup> T cell responses obtained in this model indicate that the vaccination resulted in properly processed and HLA-presented epitopes and subsequent antigen-specific T cell responses by the human immune cells of the Hu-mice.

ICS assay was performed to investigate the polarization of the T cell responses elicited by the vaccination. Due to the low frequency of T cells, individual peptide-specific T cells were more difficult to visualize by ICS than by ELISpot, but a clear population of CD4<sup>+</sup> and CD8<sup>+</sup> T cells producing Th1-type cytokines of TNF- $\alpha$  and IL-2 were detectable compared to animals receiving Vehicle in both BALB/c and Hu-mouse models (**Figures 6E,F; Supplementary Figures 10, 11**). For IL-4 and IL-13 Th2-type cytokines, analysis did not reveal any specific response at any time point. Low levels of IL-5 and/or IL-10 cytokine-producing CD4<sup>+</sup> T cells were detected for both models but it was significantly different from Vehicle control only for BALB/c mice at day 28. Even for this cohort the Th1/Th2 balance remained shifted toward Th1 for five out of six mice (one outlier)

confirming an overall Th1-skewed T cell response elicited by the vaccine (**Supplementary Figure 11**).

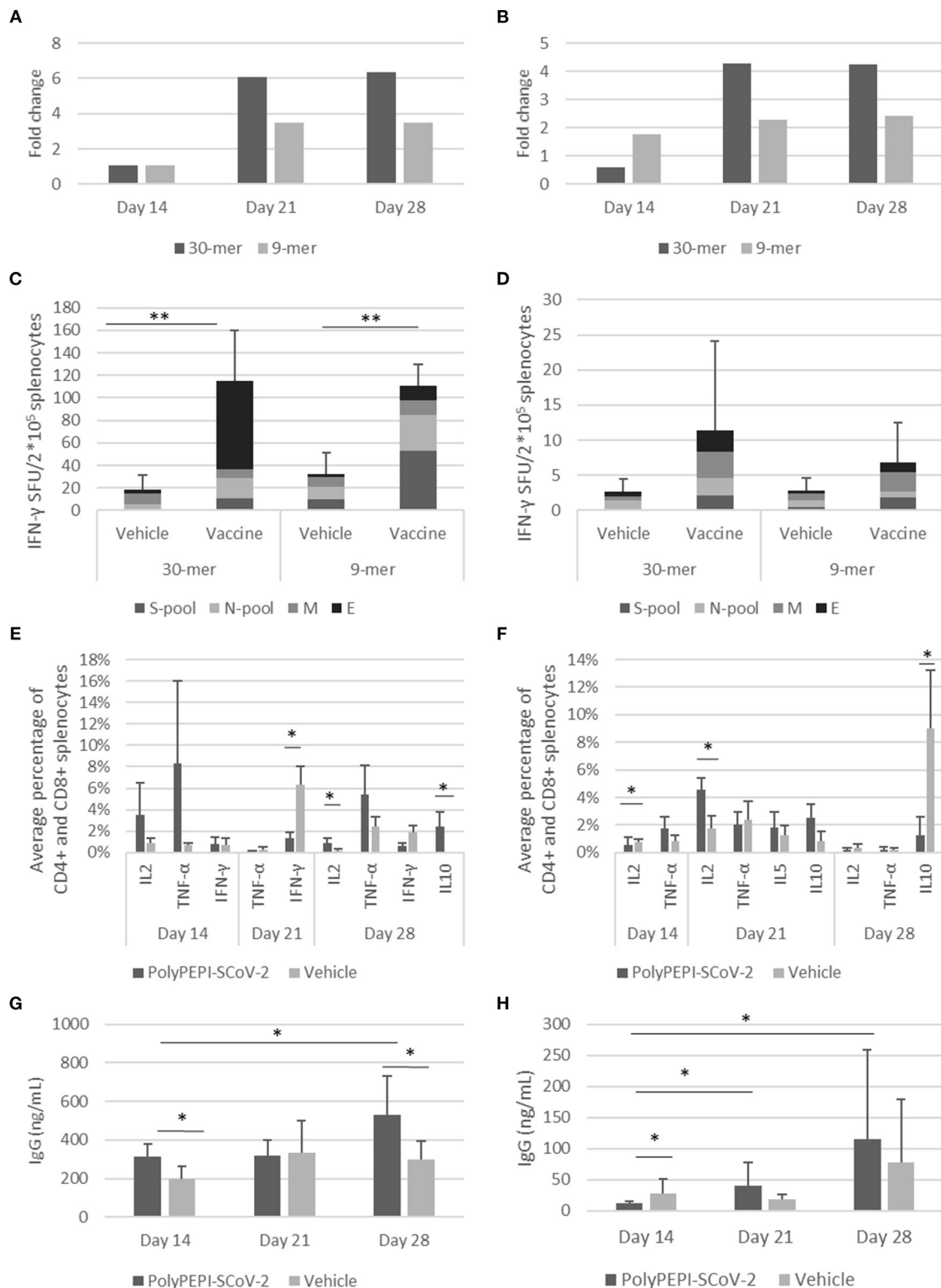
PolyPEPI-SCoV-2 vaccination also induced humoral responses, as measured by total mouse IgG for BALB/c and human IgG for Hu-mouse models. In BALB/c mice, vaccination resulted in vaccine-induced IgG production after the first dose (day 14) compared with Vehicle control group. IgG elevation were observed for both BALB/c and Hu-mouse models at later time points after the second dose (**Figures 6G,H**). IgG levels measured from the plasma of Hu-mice (average 115 ng/mL, **Figure 6H**) were lower than for BALB/c (average 529 ng/mL, **Figure 6G**) at D28. This is consistent with the known limitation of the NOD/Shi-scid/IL-2R $\gamma$  null immunodeficient mouse regarding its difficulty generating the human humoral responses that lead to class-switching and IgG production (Brehm et al., 2013). Humanization rate of ~50% in the Hu-mouse model further reduces the theoretically expected IgG levels. Despite these limitations, the dose-dependent human IgG production indicates vaccine-generated human humoral responses. As expected, given that PolyPEPI-SCoV-2 peptides do not contain conformational B cell epitopes, vaccination did not result in measurable neutralizing antibodies as assessed from the sera of Hu-mice using PNA assay. A 50% Neutralizing Antibody Titer (NT50) was undetectable at the assay detection limit of 1:25 dilution, for each tested samples (data not shown).

## DISCUSSION

We demonstrated that PolyPEPI-SCoV-2, a polypeptide vaccine candidate comprising nine synthetic long (30-mer) peptides derived from the four structural proteins of the SARS-CoV-2 (S, N, M, E) mimics the diversity of T cell immunity produced by natural SARS-CoV-2 infection, in each subject. The peptides were prospectively selected based on their frequency for an ethnically diverse, HLA-genotyped *in silico* cohort and their frequency was subsequently demonstrated in a group of convalescent subjects. Each (100%) selected peptide achieved an unprecedented recognition rate in 40–93% of convalescents, demonstrating their immunoprevalence in COVID-19. In comparison a comprehensive screening of 5,600 predicted epitopes restricted to 28 frequent HLA class I alleles in 99 COVID-19 convalescent subjects revealed 101/454 (22%) epitopes shared between at least two subjects (Tarke et al., 2021). As an external validation, T cells reactive to each of our peptides investigated (eight out of nine) were reported also for this larger cohort of convalescents, 62% of subjects having *ex vivo* recall responses specific to one or more PolyPEPI-SCoV-2 peptides (Tarke et al., 2021).

On the individual level, the PolyPEPI-SCoV-2-specific T cell repertoire used for recovery from asymptomatic/mild COVID-19 was extremely diverse: each donor had an average of seven different peptide-specific T cell pools, with multiple targets against SARS-CoV-2 proteins; 87% of donors had targets against at least three SARS-CoV-2 proteins and 53% against all four, 1–5 months after their disease onset. Despite 87% of subjects had CD8<sup>+</sup> T cells against S protein, we found that S-specific





**FIGURE 6 |** Induction of cellular and humoral immune responses by PolyPEPI-SCoV-2 vaccine in mouse models. Animals received PolyPEPI-SCoV-2 or Vehicle subcutaneously at days 0, 14. IFN- $\gamma$ -producing T cell responses elicited by PolyPEPI-SCoV-2 expressed as fold change in BALB/c (A) and Hu-mouse (B) models compared to the respective Vehicle cohorts; diversity of vaccine-induced T cell responses after two doses at day 28 in BALB/c (C) and Hu-mouse (D) models by ex vivo ELISpot. Test conditions: stimulation with 30-mer S-pool (three S-peptides), N-pool (four N-peptides), M-peptide, E-peptide, or 9-mer pools (s-pool, n-pool, e1, m1 peptides). For Fold change calculation the average dSFU values of the 30-mer and 9-mer stimulation conditions are pooled. (E,F) PolyPEPI-SCoV-2 induced Th1 (Continued)

**FIGURE 6 |** response and no significant Th2 cytokine induction shown as average of vaccine-specific CD4<sup>+</sup> and CD8<sup>+</sup> T cells producing IL-2, TNF- $\alpha$ , IFN- $\gamma$ , IL-5 or IL-10 in BALB/c (**E**) and Hu-mouse (**F**), using Intracellular cytokine staining. Mean  $\pm$  SEM are shown.  $2 \times 10^5$  cells were analyzed, gated for CD45<sup>+</sup>CD3<sup>+</sup>CD4<sup>+</sup>/CD8<sup>+</sup>. Average percentage was obtained by pooling the background-subtracted values of the 30-mer stimulation conditions for each cytokine for CD4<sup>+</sup> and CD8<sup>+</sup> splenocytes. IgG production measured from the plasma of BALB/c (**G**) and Hu-mice (**H**).  $N = 6$  animals at each time point, \* $p < 0.05$ , \*\* $p < 0.001$  (Mann-Whitney U).

(memory) T cells represented only 36% of the convalescents' total T cell repertoire detected with our peptides; the remaining 64% was distributed almost equally among N, M, and E proteins. These data support the increasing concern that S protein-based candidate vaccines are not harnessing the full potential of human anti-SARS-CoV-2 T cell immunity, especially since diversity of T cell responses was associated with mild/asymptomatic COVID-19 and they are vital for long-term immunity.

We demonstrated that individuals' anti-SARS-CoV-2 T cell responses reactive to the PolyPEPI-SCoV-2 peptide set are HLA genotype-dependent. Specifically, predicted, multiple autologous HLA binding epitopes (PEPIs) determine antigen-specific CD8<sup>+</sup> T cell responses with 84% accuracy. This suggests, that PEPIs overcome the unexplained high false positive rates generally observed using only the epitope-binding affinity as the T cell response predictor (Lorincz et al., 2019; Toke et al., 2019; Nelde et al., 2020; Wells et al., 2020; Tarke et al., 2021). Particularly, this predictive value compares favorably to the 10–25% positive epitope-specific T cell tests obtained in HLA-matched COVID-19 subjects reported by two recent publications (Nelde et al., 2020; Tarke et al., 2021).

Our vaccine design concept, targeting multi-antigenic immune responses at both the individual and population level, represents a novel target identification strategy that has already been used successfully in cancer vaccine development to achieve unprecedented immune response rates correlating with initial efficacy in the clinical setting (Hubbard et al., 2019). For COVID-19, we focused on selecting fragments of the SARS-CoV-2 proteins that contain overlapping HLA class I and II T cell epitopes that can generate diverse and broad immune responses against the whole virus. Therefore, we selected long 30-mer fragments to favor generation of multi-antigenic effector responses (B cells and cytotoxic T cells) and helper T cell responses.

PolyPEPI-SCoV-2 vaccine elicits the desired humoral responses as well as the CD8<sup>+</sup> and CD4<sup>+</sup> T cells responses against all four SARS-CoV-2 proteins in vaccinated BALB/c and humanized mice. Particularly, the robust, truly vaccine-induced immune responses obtained in the humanized mice suggest that immune responses obtained in mice are relevant also in humans.

The interaction between T and B cells is a well-known mechanism toward both antibody-producing plasma cell production and generation of memory B cells (Parker, 1993). During the analysis of convalescents' antibody subsets, we found correlations between antigen-specific IgG levels and corresponding peptide-specific CD4<sup>+</sup> T cell responses. This correlation might represent the link between CD4<sup>+</sup> T cells and antibody production, a concept also supported by total IgG production in the animal models. Binding IgG antibodies can act in cooperation with the vaccine induced CD8<sup>+</sup> killer T cells upon later SARS-CoV-2 exposure of the vaccinees. This interplay

might result in effective CD8<sup>+</sup> T cell mediated direct killing of infected cells and IgG-mediated killing of virus-infected cells and viral particles, inhibiting Th2-dependent immunopathologic processes, too.

In this way, it is expected that both intracellular and extracellular virus reservoirs are attacked to help viral clearance in the early stage of infection blocking progression to severe COVID-19, even in the absence of neutralizing antibodies (Parker, 1993; Kar et al., 2020).

This hypothesis may be supported by previous animal challenge studies demonstrating that reactivated T cells provided protection from lethal dose infection with SARS (Zhao et al., 2010; Channappanavar et al., 2014a). Moreover, a study reported that CD8<sup>+</sup> T cells contribute to the protection of convalescent macaques against re-challenge with SARS-CoV-2 in the setting of waning and subprotective antibody titers (McMahan et al., 2020). For mRNA-based COVID-19 vaccines it was suggested that binding antibodies and T cell responses are responsible for early protection against COVID-19 and lack of neutralizing antibodies indicate they are not absolutely required for protection (Kalimuddin et al., 2021). As of yet, the role of T cell responses in the protection against SARS-CoV-2 infection or COVID-19 has not been directly demonstrated. We acknowledge that, historically, T cell-focused vaccines represent an uncharted territory in the development of highly effective vaccines where antibody-based vaccines already demonstrated major role. However, the pandemic is still evolving and it would be important to understand the body's response to infection and to vaccines in order to develop the most effective vaccine or vaccination strategy.

Although PEPIs generally underestimated the subject's overall T cell repertoire, they are precise target identification "tools" and predictors of PEPI-specific immune responses. In addition both predicted PEPI frequencies and related T cell response frequencies obtained for the convalescent cohort were in good alignment with the predicted PEPI frequencies obtained for the *in silico* model population used for vaccine design. Therefore, our findings could be extrapolated to large cohorts of 16,000 HLA-genotyped individuals and 16 human ethnicities, representative for global population. Based on this, PolyPEPI-SCoV-2 will likely generate meaningful, multi-antigenic CD8<sup>+</sup> and CD4<sup>+</sup> T cell responses in  $\sim 98\%$  of the global population, independent of ethnicity. In comparison, a T cell epitope-based vaccine design approach based on globally frequent HLA alleles, as proposed by others, would miss generation of immune responses for  $\sim 50\%$  of Black Caribbean, African, African-American, and Vietnamese ethnicities. We propose, HLA-genotypes should be taken into consideration during the development of widely desired, second-generation, "universal" vaccines focusing not only on humoral but cellular responses, too (Dai and Gao, 2021). We believe, focusing on several targets in each subject would

better recapitulate the natural T cell immunity induced by the virus, potentially leading to long-term memory responses and protection against mutations.

The present study has limitations. The limited number of donors studied did not allow validation of the performance of our approach. Further statistically powered studies would be required to demonstrate immunoprevalence of the selected peptides in large cohorts of convalescents with different ethnic background, immune status, age, etc. Nevertheless, the study presents a novel *in silico* approach for the selection of immunogenic epitopes for an individual or a population, and promising initial confirmation at both individual and population level in two independent convalescent cohorts. Translation of predicted anti-SARS-CoV-2-specific T cell responses based on HLA- genotypes to T cell responses obtained upon vaccination should be carefully interpreted even in the light of immunogenicity data obtained in vaccinated animals modeling human immune system. HLA-genotype-dependent vaccine-specific T cell responses can be validated in a clinical study involving HLA-genotyped individuals.

Due to the well-known limitations of the NOD/Shi-scid/IL-2R $\gamma$  null immunodeficient mouse model for producing robust IgG antibodies, we opted to measure total IgG as demonstration of the vaccine's capacity to induce humoral responses in human immune systems. Therefore, split to individual SARS-CoV-2 antigen-specific antibody responses need to be confirmed in further (preferably human-like or human) models. Similarly, this study does not provide evidence for the pre-clinical efficacy of the vaccine. A challenge study will be performed in rodent model to investigate the impact of the vaccine-induced T and B cell responses on functional immunity and on the disease pathology upon SARS-CoV-2 exposure.

Synthetic polypeptide-based platform technology is considered a safe and immunogenic subunit vaccination strategy with several advantages over platforms using whole antigens: limits unwanted antigenicity, induces robust cellular responses and can be less reactogenic (Wu et al., 2008; Atsmon et al., 2012; Crooke et al., 2020; Kanduc and Shoenfeld, 2020; Poland, 2020; Vojdani and Kharrazian, 2020). Synthetic peptide manufacturing at multi-kilogram scale is relatively inexpensive and peptides are generally stable for years (>6 months stability demonstrated for PolyPEPI-SCoV-2). Therefore, the manufacturing and distribution of peptide vaccines could benefit from the well-established processes of the existing multinational or nation-sized facilities. Peptide-based vaccines have had only limited success to date, but this can be attributed to lack of knowledge regarding which peptides to use. Such uncertainty is reduced by an understanding of how an individual's genetic background is able to respond to specific peptides. As we demonstrated here, this knowledge drives to the desired and predicted immune responses, both on individual and population level.

In conclusion, our multi-antigen targeting peptide set has the potential to lead to a versatile second-generation tool against COVID-19. Potential clinical opportunities include: PolyPEPI-SCoV-2 used either alone or in combination with other vaccines focused on neutralizing-antibody responses in

COVID-naïve subjects or used as a booster agent to broaden or strengthen immune responses in vaccinated or COVID-convalescent subjects, or used in early infection or in "long COVID" (therapeutic setting) or as a diagnostic tool in monitoring SARS-CoV-2-specific T cell responses. In addition, "*in silico* clinical trial" in large, ethnically diverse cohorts allows for continuous and rapid monitoring of the global coverage and cross-protection with the appearance of new viral variants, potentially de-risking the success of clinical trials and likely an indispensable tool for global post-vaccination surveillance.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

## ETHICS STATEMENT

Blood samples were collected from convalescent individuals (n = 15) at an independent medical research center in The Netherlands under an approved protocol (NL57912.075.16.) or collected by PepTC Vaccines Ltd (n = 2). All donors including the non-exposed individuals (n = 10) provided written informed consent to participate in this study. The study was conducted in accordance with the Declaration of Helsinki. The animal study was reviewed and approved by the French Ethical Committee (CEEAG) and validated by the French Ministry of Research.

## AUTHOR CONTRIBUTIONS

ES designed and coordinated the preclinical experiments and participated in data evaluations. ZC designed the PolyPEPI-SCoV-2 vaccine and participated in preclinical data evaluations. LM and JT performed the *in silico* analyses and prepared the Figures and Tables of the manuscript. SP, JS, and AM performed the *in vitro* experiments using COVID-19 donors' specimen and participated in the analysis of these data. OL had leading role in the manufacturing and quality control of the PolyPEPI-SCoV-2 peptides and development of vaccine formulation. LM and IM performed the statistical analysis. KP and PP participated in data mining and literature search. ET participated in the design of the experiments and interpretation of data as well in the preparation of the manuscript. All authors reviewed the manuscript.

## ACKNOWLEDGMENTS

We are grateful to all COVID-19 convalescent donors for the biospecimens; to Gabor Illes (Treos Bio Group) and to Florent Arbogast (TransCure Bioservices) who supported and facilitated the timely execution of the experiments.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.684152/full#supplementary-material>

## REFERENCES

- Ahmed, S. F., Quadeer, A. A., and McKay, M. R. (2020). Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses* 12:254. doi: 10.3390/v12030254
- Aldridge, R. W., Lewer, D., Katikireddi, S. V., Mathur, R., Pathak, N., Burns, R., et al. (2020). Black, Asian, and Minority Ethnic groups in England are at increased risk of death from COVID-19: indirect standardisation of NHS mortality data. *Wellcome Open Res.* 5, 88–107. doi: 10.12688/wellcomeopenres.15922.1
- Altmann, D. M., and Boyton, R. J. (2020). SARS-CoV-2 T cell immunity: specificity, function, durability, and role in protection. *Sci. Immunol.* 5:eabd6160. doi: 10.1126/sciimmunol.abd6160
- Anderson, E. J., Roupael, N. G., Widge, A. T., Jackson, L. A., Roberts, P. C., Makhene, M., et al. (2020). Safety and Immunogenicity of SARS-CoV-2 mRNA-1273 vaccine in older adults. *N. Engl. J. Med.* 383, 2427–2438. doi: 10.1056/NEJMoa2028436
- Atsmon, J., Kate-Ilovitz, E., Shaikevich, D., Singer, Y., Volokhov, I., Haim, K. Y., et al. (2012). Safety and immunogenicity of multimeric-001—a novel universal influenza vaccine. *J. Clin. Immunol.* 32, 595–603. doi: 10.1007/s10875-011-9632-5
- Brehm, M. A., Shultz, L. D., Luban, J., and Greiner, D. L. (2013). Overcoming current limitations in humanized mouse research. *J. Infect. Dis.* 208(Suppl. 2), S125–S130. doi: 10.1093/infdis/jit319
- Callaway, E. (2021). Fast-spreading COVID variant can elude immune responses. *Nature* 589, 500–501. doi: 10.1038/d41586-021-00121-z
- Channappanavar, R., Fett, C., Zhao, J., Meyerholz, D. K., and Perlman, S. (2014a). Virus-specific memory CD8 T cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. *J. Virol.* 88, 11034–11044. doi: 10.1128/JVI.01505-14
- Channappanavar, R., Zhao, J., and Perlman, S. (2014b). T cell-mediated immune response to respiratory coronaviruses. *Immunol. Res.* 59, 118–128. doi: 10.1007/s12026-014-8534-z
- Consortium, The UniProt (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049
- Crooke, S. N., Ovsyannikova, I. G., Kennedy, R. B., and Poland, G. A. (2020). Immunoinformatic identification of B cell and T cell epitopes in the SARS-CoV-2 proteome. *Sci. Rep.* 10:14179. doi: 10.1038/s41598-020-70864-8
- Dai, L., and Gao, G. F. (2021). Viral targets for vaccines against COVID-19. *Nat. Rev. Immunol.* 21, 73–82. doi: 10.1038/s41577-020-00480-0
- Dan, J. M., Mateus, J., Kato, Y., Hastie, K. M., Yu, E. D., Faliti, C. E., et al. (2021). Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science* 371:eabf4063. doi: 10.1126/science.abf4063
- Diao, B., Wang, C., Tan, Y., Chen, X., Liu, Y., Ning, L., et al. (2020). Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19). *Front. Immunol.* 11:827. doi: 10.3389/fimmu.2020.00827
- Ewer, K. J., Barrett, J. R., Belij-Rammerstorfer, S., Sharpe, H., Makinson, R., Morter, R., et al. (2020). T cell and antibody responses induced by a single dose of ChAdOx1 nCoV-19 (AZD1222) vaccine in a phase 1/2 clinical trial. *Nat. Med.* 27, 270–278. doi: 10.1038/s41591-020-01194-5
- Ferretti, A. P., Kula, T., Wang, Y., Nguyen, D. M. V., Weinheimer, A., Dunlap, G. S., et al. (2020). Unbiased screens show CD8+ T cells of Covid-19 patients recognize shared epitopes in SARS-CoV-2 that largely reside outside the spike protein. *Immunity* 53, 1095–1107. doi: 10.1016/j.immuni.2020.10.006
- Forni, G., Mantovani, A., and Rome Covid-19 Commission of Accademia Nazionale dei Lincei (2021). COVID-19 vaccines: where we stand and challenges ahead. *Cell Death Differ.* 28, 626–639. doi: 10.1038/s41418-020-00720-9
- Gonzalez-Galarza, F. F., McCabe, A., Santos, E. J. M. D., Jones, J., Takeshita, L., Ortega-Rivera, N. D., et al. (2019). Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* 48, D783–D88. doi: 10.1093/nar/gkz1029
- Gragert, L., Madbouly, A., Freeman, J., and Maier, M. (2013). Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum. Immunol.* 74, 1313–1320. doi: 10.1016/j.humimm.2013.06.025
- Green, D. R. (2020). SARS-CoV2 vaccines: slow is fast. *Sci. Adv.* 6:eabc7428. doi: 10.1126/Sciadv.abc7428
- Grifoni, A., Weiskopf, D., Ramirez, S. I., Mateus, J., Dan, J. M., Moderbacher, C. R., et al. (2020). Targets of T Cell Responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell* 181, 1–13. doi: 10.1016/j.cell.2020.05.015
- Guo, J. P., Petric, M., Campbell, W., and McGeer, P. L. (2004). SARS corona virus peptides recognized by antibodies in the sera of convalescent cases. *Virology* 324, 251–256. doi: 10.1016/j.virol.2004.04.017
- He, Y., Zhou, Y., Wu, H., Kou, Z., Liu, S., and Jiang, S. (2004). Mapping of antigenic sites on the nucleocapsid protein of the severe acute respiratory syndrome coronavirus. *J. Clin. Microbiol.* 42, 5309–5314. doi: 10.1128/JCM.42.11.5309-5314.2004
- Hellerstein, M. (2020). What are the roles of antibodies versus a durable, high quality T-cell response in protective immunity against SARS-CoV-2? *Vaccine* 38:100076. doi: 10.1016/j.vaccine.2020.100076
- Helmberg, W., Duniwin, R., and Feolo, M. (2004). The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. *Nucleic Acids Res.* 32, W173–W175. doi: 10.1093/nar/gkh424
- Huang, A., Bange, E., Han, N., Wileyto, E. P., Kim, J., Gouma, S., et al. (2021). CD8 T cells compensate for impaired humoral immunity in COVID-19 patients with hematologic cancer. *Res Sq.* doi: 10.21203/rs.3.rs-162289/v1
- Hubbard, J. M., C., Cremolini, R. P., Graham, R., Moretto, J., Mitchell, J., et al. (2019). P329 PolyPEP1018 off-the shelf vaccine as add-on to maintenance therapy achieved durable treatment responses in patients with microsatellite-stable metastatic colorectal cancer patients (MSS mCRC). *J. Immunother. Cancer* 7:282. doi: 10.1186/s40425-019-0763-1
- Hurley, C. K., Kempenich, J., Wadsworth, K., Sauter, J., Hofmann, J. A., Schefzyk, D., et al. (2020). Common, intermediate, and well-documented HLA alleles in world populations: CIWD version 3.0.0. *HLA* 95, 516–531. doi: 10.1111/tan.13811
- Jackson, L. A., Anderson, E. J., Roupael, N. G., Roberts, P. C., Makhene, M., Coler, R. N., et al. (2020). An mRNA Vaccine against SARS-CoV-2—preliminary report. *N. Engl. J. Med.* 383, 1920–1931. doi: 10.1056/NEJMoa2022483
- Kalimuddin, S., Tham, C. Y., Qui, M., de Alwis, R., Sim, J. X., Lim, J. M., et al. (2021). Early T cell and binding antibody responses are associated with Covid-19 RNA vaccine efficacy onset. *Med (N. Y.)* 2, 1–7. doi: 10.1016/j.medj.2021.04.003
- Kanduc, D., and Shoenfeld, Y. (2020). On the molecular determinants of the SARS-CoV-2 attack. *Clin. Immunol.* 215:108426. doi: 10.1016/j.clim.2020.108426
- Kar, T., Narsaria, U., Basak, S., Deb, D., Castiglione, F., Mueller, D. M., et al. (2020). A candidate multi-epitope vaccine against SARS-CoV-2. *Sci. Rep.* 10:10895. doi: 10.1038/s41598-020-67749-1
- Le Bert, N., Tan, A. T., Kunasegaran, K., Tham, C. Y. L., Hafezi, M., Chia, A., et al. (2020). SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, uninfected controls. *Nature* 584, 457–462. doi: 10.1038/s41586-020-2550-z
- Liu, S. J., Leng, C. H., Lien, S. P., Chi, H. Y., Huang, C. Y., Lin, C. L., et al. (2006). Immunological characterizations of the nucleocapsid protein based SARS vaccine candidates. *Vaccine* 24, 3100–3108. doi: 10.1016/j.vaccine.2006.01.058
- Lorincz, O., Toth, J., Megyesi, M., Pántya, K., Miklos, I., Somogyi, E., et al. (2019). 1935PComputational model to predict response rate of clinical trials. *Ann. Oncol.* 30:V780. doi: 10.1093/annonc/mdz268.062
- Maier, M., Gragert, L., and Klitz, W. (2007). High-resolution HLA alleles and haplotypes in the United States population. *Hum. Immunol.* 68, 779–788. doi: 10.1016/j.humimm.2007.04.005
- McMahan, K., Yu, J., Mercado, N. B., Loos, C., Tostanoski, L. H., Chandrashekar, A., et al. (2020). Correlates of protection against SARS-CoV-2 in rhesus macaques. *Nature* 590, 630–634. doi: 10.1038/s41586-020-03041-6
- Mohammadpour, S., Torshizi Esfahani, A., Halaji, M., Lak, M., and Ranjbar, R. (2021). An updated review of the association of host genetic factors with susceptibility and resistance to COVID-19. *J. Cell. Physiol.* 236, 49–54. doi: 10.1002/jcp.29868
- Nelde, A., Bilich, T., Heitmann, J. S., Maringer, Y., Salih, H. R., Roerden, M., et al. (2020). SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nat. Immunol.* 22, 74–85. doi: 10.1038/s41590-020-00808-x
- Nguyen, A., David, J. K., Maden, S. K., Wood, M. A., Weeder, B. R., Nellore, A., et al. (2020). Human leukocyte antigen susceptibility map for severe



- acute respiratory syndrome coronavirus 2. *J. Virol.* 94, e00510–e0051020. doi: 10.1128/JVI.00510-20
- O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., et al. (2021a). "B.1.351 report 2021-02-15." Available online at: [https://cov-lineages.org/global\\_report\\_B.1.351.html](https://cov-lineages.org/global_report_B.1.351.html) (accessed February 15, 2021).
- O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., et al. (2021b). "P.1 report 2021-02-15." Available online at: [https://cov-lineages.org/global\\_report\\_P.1.html](https://cov-lineages.org/global_report_P.1.html) (accessed February 15, 2021).
- Pan, D., Sze, S., Minhas, J. S., Bangash, M. N., Pareek, N., Divall, P., et al. (2020). The impact of ethnicity on clinical outcomes in COVID-19: a systematic review. *EClinicalMedicine* 23:100404. doi: 10.1016/j.eclinm.2020.100404
- Parker, D. C. (1993). T cell-dependent B cell activation. *Annu. Rev. Immunol.* 11, 331–360. doi: 10.1146/annurev.iv.11.040193.001555
- Peiris, M., and Leung, G. M. (2020). What can we expect from first-generation COVID-19 vaccines? *Lancet* 396, 1467–1469. doi: 10.1016/S0140-6736(20)31976-0
- Peng, Y., Mentzer, A. J., Liu, G., Yao, X., Yin, Z., Dong, D., et al. (2020). Broad and strong memory CD4(+) and CD8(+) T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. *Nat. Immunol.* 21, 1336–1345. doi: 10.1038/s41590-020-0782-6
- Poland, G. A. (2020). Tortoises, hares, and vaccines: a cautionary note for SARS-CoV-2 vaccine development. *Vaccine* 38, 4219–4220. doi: 10.1016/j.vaccine.2020.04.073
- Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., et al. (2020). Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations.
- Sahin, U., Muik, A., Vogler, I., Derhovanessian, E., Kranz, L. M., Vormehr, M., et al. (2020). BNT162b2 induces SARS-CoV-2-neutralising antibodies and T cells in humans. *medRxiv [Preprint]* doi: 10.1101/2020.12.09.20245175
- Schwarzkopf, S., Krawczyk, A., Knop, D., Klump, H., Heinold, A., Heinemann, F. M., et al. (2021). Cellular immunity in COVID-19 convalescents with PCR-confirmed infection but with undetectable SARS-CoV-2-specific IgG. *Emerg. Infect. Dis.* 27, 122–129. doi: 10.3201/2701.203772
- Sekine, T., Perez-Potti, A., Rivera-Ballesteros, O., Stralin, K., Gorin, J. B., Olsson, A., et al. (2020). Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. *Cell* 183, 158 e14–168 e14. doi: 10.1101/2020.06.29.174888
- Tada, T., Zhou, H., Dcosta, B. M., Samanovic, M. I., Mulligan, M. J., and Landau, N. R. (2021). The spike proteins of SARS-CoV-2 B.1.617 and B.1.618 variants identified in India provide partial resistance to vaccine-elicited and therapeutic monoclonal antibodies. *bioRxiv [Preprint]*. doi: 10.1101/2021.05.14.444076
- Tarke, A., Sidney, J., Kidd, C. K., Dan, J. M., Ramirez, S. I., Yu, E. D., et al. (2021). Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *Cell Rep. Med.* 2:100204. doi: 10.1016/j.xcrm.2021.100204
- Thomson, E. C., Rosen, L. E., Shepherd, J. G., Spreafico, R., da Silva Filipe, A., Wojcechowskyj, J. A., et al. (2021). Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* 184, 1171–1187. doi: 10.1016/j.cell.2021.01.037
- Toke, E. R., Megyesi, M., Molnar, L., Tóth, J., Lorincz, O., van der Burg, S. H., et al. (2019). Prediction the clinical outcomes of cancer patients after peptide vaccination. *J. Clin. Oncol.* 37: e14295. doi: 10.1200/JCO.2019.37.15\_suppl.e14295
- Vojdani, A., and Kharrazian, D. (2020). Potential antigenic cross-reactivity between SARS-CoV-2 and human tissue with a possible link to an increase in autoimmune diseases. *Clin. Immunol.* 217:108480. doi: 10.1016/j.clim.2020.108480
- Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., and Wei, G. W. (2021). Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *Commun. Biol.* 4:228. doi: 10.1038/s42003-021-01867-y
- Wells, D. K., van Buuren, M. M., Dang, K. K., Hubbard-Lucey, V. M., Sheehan, K. C. F., Campbell, K. M., et al. (2020). Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* 183, 818 e13–834 e13. doi: 10.1016/j.cell.2020.09.015
- WHO (2020). *Draft Landscape of COVID-19 Candidate Vaccines*. Available online at: <https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines> (Retrieved June 03, 2021).
- Williams, T. C., and Burgers, W. A. (2021). SARS-CoV-2 evolution and vaccines: cause for concern? *Lancet Respir. Med.* 9, 333–335. doi: 10.1016/S2213-2600(21)00075-8
- Wu, Y., Ellis, R. D., Shaffer, D., Fontes, E., Malkin, E. M., Mahanty, S., et al. (2008). Phase 1 trial of malaria transmission blocking vaccine candidates Pfs25 and Pvs25 formulated with montanide ISA 51. *PLoS ONE* 3:e2636. doi: 10.1371/journal.pone.0002636
- Zhang, Y., Zeng, G., Pan, H., Li, C., Hu, Y., Chu, K., et al. (2021). Safety, tolerability, and immunogenicity of an inactivated SARS-CoV-2 vaccine in healthy adults aged 18–59 years: a randomised, double-blind, placebo-controlled, phase 1/2 clinical trial. *Lancet Infect. Dis.* 21, 181–192. doi: 10.1016/S1473-3099(20)30843-4
- Zhao, J., Zhao, J., and Perlman, S. (2010). T cell responses are required for protection from clinical disease and for virus clearance in severe acute respiratory syndrome coronavirus-infected mice. *J. Virol.* 84, 9318–9325. doi: 10.1128/JVI.01049-10
- Zou, L., Ruan, F., Huang, M., Liang, L., Huang, H., Hong, Z., et al. (2020). SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N. Engl. J. Med.* 382, 970–971. doi: 10.1056/NEJMc2001737

**Conflict of Interest:** ES, ZC, LM, OL, JT, IM, KP, PP, and ET hold shares in Treos Bio Ltd. and are employed by Treos Bio Zrt. SP, JS and AM are employed by ImmunXperts SA, a Nexelis company. Authors at Treos Bio Ltd. are listed as inventors of the following patents: US10973909B1 and PCT/GB2021/050829.

Copyright © 2021 Somogyi, Csiszovszki, Molnár, Lőrincz, Tóth, Pattijn, Schockaert, Mazy, Miklós, Pántya, Páles and Töke. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Structural Analysis of SARS-CoV-2 ORF8 Protein: Pathogenic and Therapeutic Implications

Antonio Valcarcel<sup>1</sup>, Antonio Bensussen<sup>1</sup>, Elena R. Álvarez-Buylla<sup>2,3\*</sup> and José Díaz<sup>1\*</sup>

<sup>1</sup> Laboratorio de Dinámica de Redes Genéticas, Centro de Investigación en Dinámica Celular, Universidad Autónoma del Estado de Morelos, Cuernavaca, Mexico, <sup>2</sup> Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Ciudad de México, Mexico, <sup>3</sup> Laboratorio de Genética Molecular, Epigenética, Desarrollo y Evolución de Plantas, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

## OPEN ACCESS

### Edited by:

Dariusz Plewczynski,  
Warsaw University of Technology,  
Poland

### Reviewed by:

Sandra Paulina Smieszek,  
Vanda Pharmaceuticals Inc.,  
United States  
Pragati Agnihotri,  
Advanced BioScience Laboratories,  
Inc. (ABL), United States

### \*Correspondence:

José Díaz  
biofisica@yahoo.com  
Elena R. Álvarez-Buylla  
elenabuylla@protonmail.com

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 10 April 2021

**Accepted:** 29 July 2021

**Published:** 06 September 2021

### Citation:

Valcarcel A, Bensussen A,  
Álvarez-Buylla ER and Díaz J (2021)  
Structural Analysis of SARS-CoV-2  
ORF8 Protein: Pathogenic  
and Therapeutic Implications.  
Front. Genet. 12:693227.  
doi: 10.3389/fgene.2021.693227

Current therapeutic strategies and vaccines against SARS-CoV-2 are mainly focused on the Spike protein despite there are other viral proteins with important roles in COVID-19 pathogenicity. For example, ORF8 restructures vesicular trafficking in the host cell, impacts intracellular immunity through the IFN-I signaling, and growth pathways through the mitogen-activated protein kinases (MAPKs). In this mini-review, we analyze the main structural similarities of ORF8 with immunological molecules such as IL-1, contributing to the immunological deregulation observed in COVID-19. We also propose that the blockage of some effector functions of ORF8 with Rapamycin, such as the mTORC1 activation through MAPKs 40 pathway, with Rapamycin, can be a promising approach to reduce COVID-19 mortality.

**Keywords:** SARS-CoV-2, COVID-19, ORF8, structural biology, COVID-19 therapeutics

## INTRODUCTION

The SARS-CoV-2 appeared in Wuhan at the end of December 2019 with the consequent crisis in the health systems due to the lack of an effective treatment to face a then unknown disease with a mortality of 10%. The implementation of physical distancing leads to an overall reduction in incidence by 13% (Islam et al., 2020). At the time of writing, 190 million infections and 4.14 million deaths have been reported. Although the mortality rate is reducing the number of infected patients is increasing, and there is still no effective pharmacological protocol against the disease. More than a year after the start of the pandemic, available vaccines are still uncertain since the virus genome has shown high genetic variability (Islam et al., 2020). Therefore, new drug strategies are needed for the prevention and treatment of the infection caused by this virus and the aftereffects of the disease. One of the most relevant characteristics of the virus is the strong immune response in some patients, in addition to some long-lasting pathologic consequences observed in convalescence patients.

The proteins encoded in the nine open reading frames (ORFs) of SARS-CoV-2 do not appear to be necessary for viral replication. However, they participate in the modulation of the metabolism of the infected host cells, the vesicular trafficking and packing of new viral particles, and the modification of the innate immunity (Gordon et al., 2020). From this group of proteins, ORF8 is the most connected hub with 47 links, and one of these links is the Tor1a (Torsin-1a) protein, that is involved in the quality control of protein folding in the ER (Hill et al., 2018; Gordon et al., 2020). ORF8 acts on ER to modulate the unfolded protein response (UPR) by up regulation of the ER-resident chaperones GRP78 and GRP94 leading to stimulate ATF6 and IRE1 pathways. Although,

it does not seem to have any influence on the PERK pathway (Rashid et al., 2021; **Figure 1**). Thus, during SARS-CoV-2 infection, ORF8 takes the role of a central organizer of the activity of the virus-host hybrid network (the interactome model of viral components with the host proteins) toward the production of new virions (Díaz, 2020).

Coronaviruses show high genetic variability, and the structure of the SARS-CoV-2 genome consist of a set of conserved genes with an exceptionally low or null rate of mutation, together with a set of genes with high rate of variation. For example, of the 11,113 ORF8 sequences analyzed by Pereira (2020), the L84S substitution is the mutation that has been positively selected during the course of the pandemic. In 58 sites with this mutation the change in position 84 from leucine (observed in 85% of the sequences) to serine (observed in 15% of the sequences) stands up (Vilar and Isom, 2021; Zinzula, 2021). In the last group, the gene ORF8 (*ORF8*) has a notable tendency to recombine and undergo deletions that exceed the evolutionary capacity of its analogs in other coronaviruses, facilitating SARS-CoV-2 adaptability to new reservoirs and hosts (Abdelrahman et al., 2020; Zinzula, 2021). Despite the fact that truncations in ORF8 become more common as the pandemic progresses, and that these changes have apparently no influence on the replication of the virus, they are associated with non-synonymous mutations that increases the affinity of protein S for its receptor producing genetic variants with greater contagion capacity and an increased epidemiological persistence (Pereira, 2020).

During the first 6 months of the 2020 pandemic, 240 different non-synonymous mutations and 2 deletions in *ORF8* have been found in 45,400 sequences. Approximately, 50% of these mutations are detrimental to the ORF8 protein, and 25% of them are among the conserved amino acids of other variants of coronavirus in animals. These mutations, regardless of their effects on ORF8 itself, can influence the biology of SARS-CoV-2 and slow down the discovery of new drugs, vaccines, and diagnostics against this coronavirus (Chan et al., 2020; Velazquez-Salinas et al., 2020; Alkhansa et al., 2021). An observational cohort study made in Singapore in the first 3 months of 2020, highlighted that an infection process with the D382 ORF8 variant induced late onset of pneumonia with milder symptoms, compared to the patients infected with the wild type (WT) ORF8. This result was associated with a lower probability of developing hypoxia and a better recovery from the disease (Young et al., 2020), possibly due to an elicited immune response in the absence of a fully functional ORF8. The most distinctive characteristic of severe COVID-19 is the accumulation of high levels of pro-inflammatory cytokines, chemokines, and growth factors that are systemically released and are associated with lung injury. However, in patients infected with the D382 ORF8 variant all these molecules were found in lower concentrations together with high levels of gamma interferon (IFN- $\gamma$ ), and other cytokines responsible for the activation of T cells, in contrast with patients infected with WT (Su et al., 2020). In 2018, a 29-nucleotide deletion in *ORF8* was reported in the SARS-CoV genome, which was acquired during the first stage of person-to-person transmission. These observations point to the fact that these genomic changes relate to the deletion mutations of *ORF8*,

given SARS-CoV-2 some advantage in its process of adaptation to humans (Muth et al., 2018).

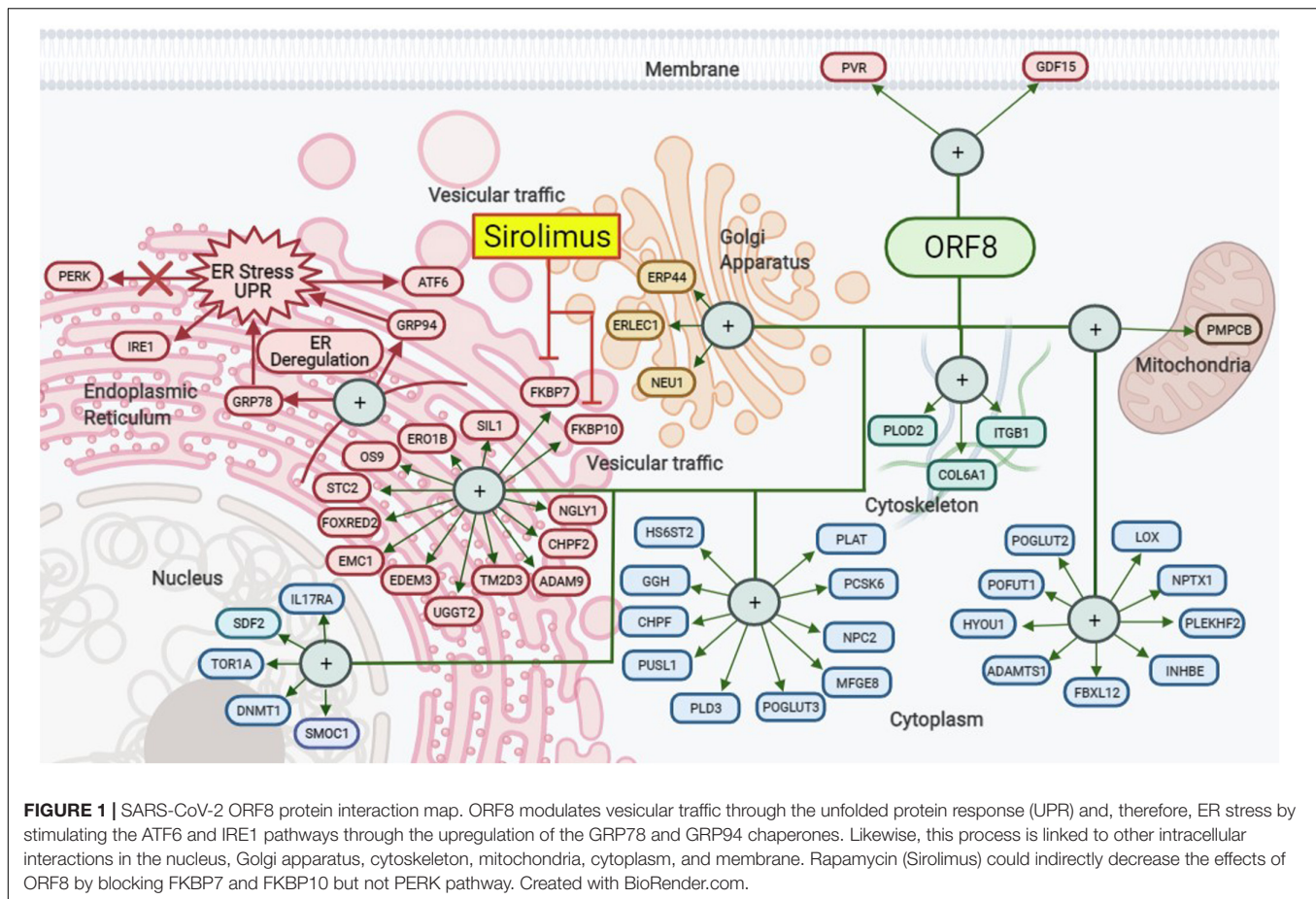
The new *ORF8* encodes a 121 amino acid secretory protein with 55.4% nucleotide similarity, and 30% protein identity with SARS-CoV counterpart. However, despite this genomic divergence, they share structural similarities as they both present a cavity with adequate electrostatic charges for protein-protein interaction (Neches et al., 2021). Structurally, SARS-CoV-2 ORF8 is a dimer in which each chain is made up of an alpha helix, followed by six-stranded chain  $\beta$  sheet, and an N-terminal hydrophobic signal peptide (1–15 aa of length) that promotes its import into the ER lumen where it can interact with a wide range of host proteins (Gordon et al., 2020; Rashid et al., 2021). This, together with two dimerization interfaces, means that ORF8 has a high possibility of forming unique complexes that can take part in immunological activity. This dimerization is probably an adaptative characteristic absent in homologs from other coronaviruses (Flower et al., 2020).

Among the functions that ORF8 plays in the evasion of the immune system are the activation of IL-17 signaling pathway, and the promotion of the expression of pro-inflammatory factors, supporting the lower intensity and late response to pneumonia caused by the D382 ORF8 variant. Additionally, association of ORF8 deletion variant (D382 variant) with milder disease outcome strongly supports the importance of ORF8 protein as a therapeutic target against SARS-CoV-2 (Sharma et al., 2021). However, the search for direct inhibition drugs of ORF8 is difficult due to the globular structure and high variability of this viral component. Another distinct function of ORF8 protein, different from SARS-CoV 29 nucleotide deleted versions- ORF8a and ORF8b (Pereira, 2021), is the regulation of the amount of MHC-I on the surface of the infected cell through a mechanism of lysosomal degradation dependent on autophagy. This results in dysregulated and deficient antigen presentation, hindering the recognition and elimination of infected cells (Zhang et al., 2020; de Sousa et al., 2020). Recently, computational experiments of homology modeling and molecular coupling suggested that a high expression of ORF8 and the surface glycoprotein may interact with heme porphyrin in the 1-beta chain of hemoglobin, resulting in a significant decrease in gas exchange processes and aggravating hypoxia in patients with severe disease (Liu and Li, 2020). However, these observations are still under investigation because of their clinical implications.

The joint action of ORF8, Nsp1, and Nsp6 results in a significant decrease in the production of IFN-I through different mechanisms to suppress signaling and produce failures and incorrect immune response, which favors the replication and transmission of the virus, to other host cells (Xia et al., 2020). Another example of this synergy is Nsp5, Nsp7, ORF3b, and M that can act together with ORF8 in more than one cell organelle, as in the case of stress-induced to ER (**Figure 1**; Gordon et al., 2020).

The diverse immune response evasion strategies generating an adaptative advantage for SARS-CoV-2 survival and propagation could be a result of functional mimicry that intensifies the host-pathogen interaction. An example of this functional mimicry comes from *in silico* simulations of ORF8-substrate complexes





with F1 and C3b. The results of the coupling suggest that ORF8 can have interactions based on its mimicry with host targets inside and outside of the ER. Even a high extracellular concentration of ORF8 could have unknown interactions with other cell types different from lung alveolar type 2 cells based on possible putative functions conferred by its Ig-like structure. According to the structural alignment of the monomer of SARS-CoV-2 ORF8 (PDB: 7JTL) with SARS-CoV ORF7a (PDB: 1XAK) (Dali Server, Z-score = 4.6, RMSD = 2.4) they share two sets of structural disulfide bonds generating a fold like Ig conformation (Flower et al., 2020). Using the Dali server (Holm, 2020), result in more than one hundred immunoglobulins reporting a Z score higher than 3.9, and RMSD values between 2.7 and 3.6. The most outstanding results are shown in **Table 1**, in which it is possible to identify their role in mimicking possible host factors.

ORF8 mimics ALCAM (CD166), which is a structural protein that can activate ERK (Ibáñez et al., 2006). Once activated, ERK stimulates cell growth through the indirect activation of mTORC1 (Saxton and Sabatini, 2017). Some observations *in vitro* from MERS-CoV replication determine that mTORC1 activity is crucial for viral replication, and that the drug Rapamycin can abrogate 60% of the production of new virions (Kindrachuk et al., 2015). Additionally, ORF8 also mimics DNAM-1 (CD266), which is an important molecule that activates Natural Killer (NK) cells (Zhang et al., 2015; Wang et al., 2019), and has been implicated

in the regulation of T CD8+ activation (Gilfillan et al., 2008), which can be used by the virus as a potent mechanism to evade the immune response. Moreover, ORF8 also has similarities with OX-2 (CD200) (Hatherley et al., 2013), which is an inhibitory molecule of macrophages (Gordon et al., 2020). Other effect of the structural mimicry of ORF8 is its ability to activate the immune response by itself due to its similarity with the soluble IL-1 $\beta$  receptor and IL-1RA agonists, stimulating the inflammation process. ORF8 also mimics CD79B (3KG5-A) and CD80 (1DR9-A), which are antigens required to activate B and T cell effector functions, respectively (Vasile et al., 1994; Trzupek et al., 2019). However, ORF8 is not precisely equal to such antigens, and can produce an incomplete stimulation of the receptors. In a biological context, incomplete stimulation produces anergy (Rollins and Gibbons, 2017), which may be used by SARS-CoV-2 to enhance its replication.

## DISCUSSION

Current pharmacological strategies to control SARS-CoV-2 infection are mainly focused on inhibiting spike-ACE2 interaction, and to block viral RNA synthesis. Some examples of these drugs are Remdesivir, Lopinavir and Ritonavir, which have been tested on many clinical trials around the world



**TABLE 1** | Highlights of PDB 7JTL comparative studies using dali server.

PDB_Chain	Description	Z	Rmsd	%id	Reference	Overlap
-----------	-------------	---	------	-----	-----------	---------



<b>5A2F_A</b>	CD166 ANTIGEN	5.9	3.1	5	Chappell et al., 2015	
5A2F_A 1	MESK <b>GASS</b> CRL <b>L</b> FCL <b>L</b> ISAT <b>V</b> FR <b>P</b> GL <b>G</b> W <b>Y</b> T <b>N</b> SAY <b>G</b> DT <b>I</b> I <b>P</b> CR <b>L</b> D <b>V</b> P <b>Q</b> N <b>L</b> M <b>F</b> G <b>K</b> W-- <b>K</b> Y <b>E</b> K <b>P</b> D <b>G</b> S <b>P</b> V <b>F</b> -- <b>I</b> A <b>F</b> R <b>S</b> S <b>T</b> K <b>K</b>					76
7JTL_A 1	--- <b>S</b> N <b>A</b> Q <b>E</b> C <b>S</b> L <b>Q</b> S <b>C</b> ----- <b>T</b> Q <b>H</b> Q <b>P</b> --- <b>Y</b> V <b>V</b> D <b>D</b> ----- <b>P</b> C <b>P</b> I <b>H</b> ----- <b>F</b> Y <b>S</b> K <b>W</b> <b>Y</b> I <b>R</b> V <b>G</b> A <b>R</b> K <b>S</b> A <b>P</b> L <b>I</b> E <b>L</b> C <b>V</b> D <b>E</b> A <b>G</b> S <b>K</b> S					55
5A2F_A 77	<b>S</b> V <b>Q</b> Y <b>D</b> D <b>V</b> P <b>E</b> Y <b>K</b> D <b>R</b> L <b>N</b> L <b>S</b> E <b>N</b> Y <b>T</b> L <b>S</b> I <b>S</b> N <b>A</b> R <b>I</b> S <b>D</b> E <b>K</b> R <b>F</b> V <b>C</b> M <b>L</b> V <b>T</b> E <b>D</b> N <b>V</b> F <b>E</b> A <b>P</b> T <b>I</b> V <b>K</b> V <b>F</b> K <b>Q</b> P <b>S</b> K <b>P</b> E <b>I</b> V <b>S</b> K <b>A</b> L <b>F</b> L <b>E</b> T <b>E</b> Q <b>L</b> K <b>K</b> L <b>G</b> D					156
7JTL_A 56	<b>P</b> I <b>Q</b> Y <b>I</b> D <b>I</b> G----- <b>N</b> Y <b>T</b> V <b>S</b> C <b>L</b> P <b>F</b> T <b>I</b> N <b>C</b> Q <b>E</b> P <b>K</b> L <b>G</b> S <b>L</b> V <b>R</b> C <b>S</b> F <b>Y</b> E <b>D</b> ----- <b>F</b> L <b>E</b> Y <b>H</b> D <b>V</b> R <b>V</b> L <b>D</b>					105
5A2F_A 157	<b>C</b> I <b>S</b> E <b>D</b> S <b>Y</b> P <b>D</b> G <b>N</b> I <b>T</b> W <b>Y</b> R <b>N</b> G <b>K</b> V <b>L</b> H <b>P</b> L <b>E</b> G <b>A</b> W <b>I</b> I <b>F</b> K <b>K</b> E <b>M</b> D <b>P</b> V <b>T</b> Q <b>L</b> Y <b>T</b> M <b>T</b> S <b>T</b> L <b>E</b> Y <b>K</b> T <b>T</b> K <b>A</b> D <b>I</b> Q <b>M</b> P <b>F</b> T <b>C</b> S <b>V</b> T <b>Y</b> Y <b>G</b> P <b>S</b> G <b>Q</b> K <b>T</b> I <b>H</b> S <b>E</b>					236
7JTL_A 106	<b>F</b> I-----					107

PDB_Chain	Description	Z	Rmsd	%id	Reference	Overlap
-----------	-------------	---	------	-----	-----------	---------

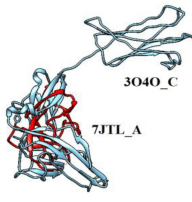


<b>1IRA_Y</b>	INTERLEUKIN-1 RECEPTOR ANTAGONIST	5.3	3.0	7	Schreuder et al., 1997	
1IRA_A 1	R <b>P</b> S <b>G</b> R <b>K</b> S <b>S</b> K <b>M</b> Q <b>A</b> F <b>R</b> I <b>W</b> D <b>V</b> N <b>Q</b> K <b>T</b> F <b>Y</b> L <b>R</b> N <b>N</b> Q <b>L</b> V <b>A</b> G <b>Y</b> L <b>Q</b> G <b>P</b> N <b>V</b> N <b>L</b> E <b>E</b> K <b>I</b> D <b>V</b> P <b>I</b> E <b>P</b> A <b>L</b> F <b>L</b> G <b>I</b> H <b>G</b> G <b>K</b> M <b>C</b> L <b>S</b> C <b>V</b> K <b>S</b> G <b>D</b> E <b>T</b> R <b>L</b> Q <b>L</b>					80
7JTL A 1	----- <b>S</b> N <b>A</b> Q <b>E</b> C <b>S</b> L <b>Q</b> S <b>C</b> T <b>Q</b> H <b>Q</b> P <b>Y</b> V <b>V</b> D <b>D</b> <b>P</b> C <b>P</b> I <b>H</b> F <b>Y</b> S <b>K</b> W <b>Y</b> I <b>R</b> V <b>G</b> A <b>R</b> K <b>S</b> A <b>P</b> L <b>I</b> E----- <b>L</b> C <b>V</b> D <b>E</b> A <b>G</b> S <b>K</b> S <b>P</b> I <b>Q</b> Y					59
1IRA_A 81	<b>E</b> A <b>V</b> N <b>I</b> T <b>D</b> L <b>S</b> E <b>N</b> R <b>K</b> Q <b>D</b> K <b>R</b> F <b>A</b> F <b>I</b> R <b>S</b> D <b>S</b> G <b>P</b> T <b>T</b> S <b>F</b> E <b>S</b> A <b>A</b> C <b>P</b> G <b>W</b> F <b>L</b> C <b>T</b> A <b>M</b> E <b>A</b> D <b>Q</b> P <b>V</b> S <b>L</b> T <b>N</b> M <b>P</b> D <b>E</b> G <b>V</b> M <b>T</b> K <b>F</b> Y <b>F</b> Q <b>E</b> D <b>E</b> -----					152
7JTL A 60	<b>I</b> D <b>I</b> G----- <b>N</b> Y <b>T</b> V <b>S</b> C <b>L</b> P <b>F</b> T <b>I</b> N <b>C</b> Q <b>E</b> ----- <b>P</b> K <b>L</b> G <b>S</b> L <b>V</b> R <b>C</b> S <b>F</b> Y <b>E</b> D <b>F</b> L <b>E</b> Y <b>H</b> D <b>V</b> R <b>V</b>					102

PDB_Chain	Description	Z	Rmsd	%id	Reference	Overlap
-----------	-------------	---	------	-----	-----------	---------


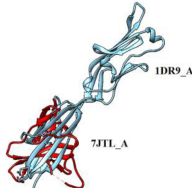
  



<b>3O4O_C</b>	INTERLEUKIN-1 $\beta$ RECEPTOR	5.0	2.7	14	Wang et al., 2010	
3O4O_C 161	Y <b>M</b> G <b>C</b> Y <b>K</b> I <b>Q</b> N <b>F</b> N <b>N</b> V <b>I</b> E <b>G</b> M <b>N</b> L <b>S</b> F <b>L</b> I <b>A</b> L <b>I</b> S <b>N</b> G <b>N</b> Y <b>T</b> C <b>V</b> V <b>T</b> P <b>E</b> N <b>G</b> R <b>T</b> F <b>H</b> L <b>T</b> R <b>T</b> L <b>T</b> V <b>K</b> V <b>G</b> S <b>P</b> K <b>N</b> A <b>V</b> P <b>P</b> V <b>I</b> H <b>S</b> P <b>N</b> D <b>H</b> V <b>V</b> Y <b>E</b> K <b>E</b>					240
7JTL_A 1	----- <b>S</b> N <b>A</b> Q <b>E</b> C <b>S</b> L <b>Q</b> S <b>C</b> T <b>Q</b> H <b>Q</b> P <b>Y</b> V <b>V</b> D					20
3O4O_C 241	<b>P</b> G <b>E</b> L <b>L</b> I <b>P</b> C <b>T</b> V <b>Y</b> F <b>S</b> F <b>L</b> M <b>D</b> S <b>R</b> N <b>E</b> V <b>W</b> T <b>I</b> D <b>G</b> K <b>K</b> P <b>D</b> D <b>I</b> T <b>I</b> D <b>V</b> T <b>I</b> N <b>E</b> S <b>I</b> S <b>H</b> S <b>R</b> T <b>E</b> D <b>T</b> R <b>T</b> Q <b>I</b> L <b>S</b> I <b>K</b> K <b>V</b> T <b>S</b> E <b>D</b> L <b>K</b> R <b>S</b> Y <b>V</b> C <b>H</b> A <b>R</b> S <b>A</b>					320
7JTL_A 21	<b>D</b> ----- <b>P</b> C <b>P</b> I <b>H</b> F <b>Y</b> ----- <b>S</b> K <b>W</b> Y <b>I</b> R <b>V</b> G <b>A</b> R <b>K</b> S <b>A</b> P <b>L</b> I <b>E</b> L <b>C</b> V <b>D</b> E <b>A</b> G <b>S</b> K <b>S</b> P <b>I</b> ----- <b>Q</b> Y <b>I</b> D <b>I</b> G <b>N</b> Y <b>T</b> V <b>S</b> C <b>L</b> P <b>F</b> T <b>I</b> N <b>C</b> Q <b>E</b> P <b>K</b> L					81
3O4O_C 321	<b>K</b> G <b>E</b> V <b>A</b> K <b>A</b> A <b>K</b> V <b>K</b> Q <b>K</b> H <b>H</b> H <b>H</b> H <b>H</b> -----			339		
7JTL_A 82	<b>G</b> S <b>L</b> V <b>R</b> C <b>S</b> F <b>Y</b> E <b>D</b> F <b>L</b> E <b>Y</b> H <b>D</b> V <b>R</b> V <b>L</b> D <b>F</b> I			107		

(Continued)

TABLE 1 | Continued

PDB_Chain	Description	Z	Rmsd	%id	Reference	Overlap
						
<b>3KG5_A</b>	B-CELL ANTIGEN RECEPTOR COMPLEX-ASSOCIATED	4.2	3.9	9	Radaev et al., 2010	
3KG5_A 1	VPAARSEDYRN <b>PKGSACSRI</b> WQSP <b>RFIARKRGFTVKMH</b> CMNSASGNV <b>SWLWKQEMDENPQQLKL</b> ---EKGRMEESQNE					77
7JTL_A 1	-----S <b>NAQECSL</b> -----Q <b>SCTQHQP</b> YVDDPC <b>PIHFYS</b> ---KWYIRVGARK <b>SAPLIEL</b> CVDEAGSK <b>SPIQYI</b>					60
3KG5_A 78	SLAT <b>LTIQ</b> GIRFEDNGI <b>YFCQ</b> QKC <b>NNTSEVYQ</b> CGTEL <b>RMGFSTLAQLKQRNTLKD</b> --					134
7JTL_A 61	DIGN <b>YTVS</b> -----CL <b>PFTINCQ</b> EPKLG-SL <b>VVRC</b> SFYED <b>FL</b> EYHD <b>VVRVLD</b> FI					107
PDB_Chain	Description	Z	Rmsd	%id	Reference	Overlap
						
<b>1DR9_A</b>	T LYMPHOCYTE ACTIVATION ANTIGEN	4.1	3.8	10	Ikemizu et al., 2000	
1DR9_A 1	VIHVTKEVKEVATL <b>SCGHNV</b> VEELAQ <b>TRIYWQ</b> KEKK <b>MLTMM</b> SGDMNIWPEY <b>KNRTIFDITN</b> NLSIVILAL <b>RPSDEGTY</b>					80
7JTL_A 1	-----S <b>NAQECSL</b> Q <b>SCTQHQP</b> YVDDPC <b>PIHFYS</b> KWYIRVGAR-K <b>SAPLIEL</b> CVDEAGSK <b>SPIQYIDIGNY</b>					65
1DR9_A 81	ECV <b>LKYE</b> -----K <b>DAFKRE</b> HLAEV <b>TL</b> SV <b>KAD</b> FP <b>TP</b> SISDFE <b>IP</b> TSNIRRI <b>IC</b> STSGGF <b>PE</b> PHLSW <b>LENGEE</b>					147
7JTL_A 66	TVSCL <b>PFTINCQ</b> EPKLGSL <b>VVRC</b> SFYED <b>FL</b> EYHD <b>VVRVLD</b> FI-----					107

Match Protein overlap in cyan and ORF8 in red. Alignment conservation setting: 2 Bits.

(McKee et al., 2020). Unfortunately, these drugs have little or no effect on patients of COVID-19 (WHO, 2021). Therefore, it is urgent to find novel viral therapeutic targets to control COVID-19. In this regard, evidence shows that cells in which ORF8 is expressed, MHC-I molecules selectively target lysosomal degradation by autophagy and hinder antigen presentation by reducing the recognition and clearance of infected cells. Other pathways of recognition interrupted by the presence of ORF8 are IFN-I signaling, and NF- $\kappa$ B functions. ORF8 also activates the ERK pathway through CD166 signaling (Bouhaddou et al., 2020) and it stimulates growth pathways directly as reported by Gordon et al. (2020). Likewise, ORF8 mimics immune molecules such as IL-1 $\beta$ , activating immunological effector signals from B cells and inhibitory molecules from immune cells such as macrophages, CD8+ T lymphocytes and NK cells (Table 1). These facts pointed out the multi-organizational role of ORF8 inside the host cells. The central question that remains is how all these functions are contributing to ensure viral replication. It seems that all ORF8 interactions are focused on provide an intracellular favorable environment to viral seems that all ORF8 interactions favor a suitable environment for viral replication through activated growth pathways and downregulated immune system, all through the inactivation of

macrophages, NK cells, B cells and CD8+ T lymphocytes, making ORF8 a feasible therapeutic target (Supplementary Figure 1). This complex network of interactions contributes to worsen the immune deregulation observed in severe cases of COVID-19 (Pasirja and Naime, 2021).

Consequently, ORF8 is a feasible therapeutic target to simultaneously shut-down viral replication and host immune downregulation. However, ORF8 is a highly mutating region of the SARS-CoV-2 genome, which decreases the feasibility of ORF8 as a good therapeutic target (Zinzula, 2021), hindering the search for inhibitory drugs. A valid approach to overcome this obstacle is either targeting ORF8 immunological functions or its growth-promoting functions. In general, several RNA viruses such as hepatitis C virus, influenza A virus, Zika virus, and MERS-CoV require a specific metabolic environment and have their own activator mechanisms that ensure the intracellular proliferation of the virus (Karam et al., 2021). In the particular case of SARS-CoV-2, ORF8 can activate the mTOR-PI3K-AKT signaling pathway with a MAPK-dependent process to ensure a proliferative environment. This favorable environment can be blocked with inhibitors of mTORC1 like Rapamycin. It has been reported for the case of MERS-CoV replication that Rapamycin was able to reduce 60% of this virus (Kindrachuk et al., 2015).

Thus, the blockage of growth pathways to prevent ORF8 binding interactions can be a better option than the targeting of several immune cells to stop viral infection.

It is possible that Rapamycin, as a co-adjuvant treatment, can improve clinical outcome because it is able to block viral interactions that promotes cell growth, and viral replication. Moreover, Rapamycin can reduce pro-inflammatory cytokines decreasing cell damage in patients with severe COVID-19 (Bischof et al., 2021). The metabolic changes conferred by SARS-CoV-2 infection in renal epithelial cells and lung air-fluid interface (ALI) cultures, showed that SARS-CoV-2 infection reduces the oxidative metabolism of glutamine while maintaining reductive carboxylation, increasing the activity of mTORC1. The work of Mullen et al. (2021) provide evidence of mTORC1 activation in lung tissue from COVID-19 patients, and that mTORC1 inhibitors reduce viral replication in renal epithelial cells and lung ALI cultures. These results suggest that targeting mTORC1 can be a feasible treatment strategy for COVID-19 patients, although more studies are required to determine the mechanism of inhibition and potential efficacy in patients. Rapamycin (Sirolimus) was chosen as it can interact through its methoxy group with the immunophilin binding protein FK506 (FKBP12) forming the rapamycin-FKBP12 complex that is highly specific to the mTOR protein, inhibiting effector processes such as antigen-induced T cell proliferation and cytokine-induced proliferative responses. From the family of polyketide macrolide drugs, Rapamycin (Sirolimus) it is the most studied and unlike Tacrolimus, it does not inhibit calcineurin (PP2B). Despite the fact that the effectiveness of Rapamycin has already been proven as a promising anti-covid drug, the interaction effects with another anti-inflammatory compounds are still to be discovered and open the possibility to have better therapeutic results with lower doses, avoiding toxic effects during the treatment.

The actual evidence shows that the variations observed in the most unstable region of the SARS-CoV-2 genome result in changes in the structure and functions of a set of proteins that counteract the immune response of the host (**Supplementary Figure 1**). However, SARS-CoV-2 seems not to have a mechanism that allows viral replication under non-permissive conditions (**Figure 1**). In consequence, the blockage of the activation of cell growth pathway through the inhibition of mTORC1 activity can be a therapeutic strategy that the virus possibly cannot counteract. Nonetheless, more research is necessary to explore the therapeutic use of Rapamycin against the SARS-CoV-2 infection.

## REFERENCES

- Abdelrahman, Z., Mengyuan, L., and Xiaosheng, W. (2020). Comparative review of SARS-CoV-2, SARS-CoV, MERS-CoV, and influenza a respiratory viruses. *Front. Immunol.* 11:552909. doi: 10.3389/fimmu.2020.552909
- Alkhansa, A., Ghayas, L., and Loubna, E. Z. (2021). Mutational analysis of SARS-CoV-2 ORF8 during six months of COVID-19 pandemic. *Gene Rep.* 23:101024. doi: 10.1016/j.genrep.2021.101024
- Bischof, E., Siow, R. C., Zhavoronkov, A., and Kaeberlein, M. (2021). The potential of rapalogs to enhance resilience against SARS-CoV-2 infection and reduce

## CONCLUSION

ORF8 is the most linked protein in the virus-host hybrid molecular network formed during the SARS-CoV-2 infection. The structural properties of ORF8 suggest functional mimicry with several immunological molecules such as the IL-1 $\beta$  receptor, resulting in immune system evasion that helps the virus to adapt to new hosts. Additionally, ORF8 restructures the vesicular trafficking in the host cell, and enhances the activity of the growth pathway through the mitogen-activated protein kinases (MAPKs). However, the high mutation rate of ORF8 decreases its feasibility as a good therapeutic target. In consequence, the blockage of the activation of cell growth pathway through the inhibition of mTORC1 activity with Rapamycin can be a therapeutic strategy that the virus possibly cannot counteract.

## AUTHOR CONTRIBUTIONS

AV and AB wrote the manuscript and contributed equally to this work. AV, AB, EÁ-B, and JD conceived the study and discussed the content of the review. JD coordinated the study. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by UNAM-PAPIIT IN211721. AV and AB were supported by CONACYT postdoctoral grants Modality 3 of ProNacEs.

## ACKNOWLEDGMENTS

We thank PRODEP-UAEM and CONACYT for the financial support for this work. JD thanks to Erika Juárez Luna for her logistic support.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.693227/full#supplementary-material>

the severity of COVID-19. *Lancet Healthy Longev.* 2, e105–e111. doi: 10.1016/s2666-7568(20)30068-4

- Bouhaddou, M., Memon, D., Meyer, B., White, K. M., Rezeli, V. V., Correa Marrero, M., et al. (2020). The global phosphorylation landscape of SARS-CoV-2 infection. *Cell* 182, 685–712.e19. doi: 10.1016/j.cell.2020.06.034
- Chan, A. P., Choi, Y., and Schork, N. J. (2020). Conserved genomic terminals of SARS-COV-2 as co-evolving functional elements and potential therapeutic targets. *BioRxiv* [Preprint]. doi: 10.1101/2020.07.06.190207
- Chappell, P. E., Garner, L. I., Yan, J., Metcalfe, C., Hatherley, D., Johnson, S., et al. (2015). Structures of CD6 and its ligand CD166 give insight into their interaction. *Structure* 23, 1426–1436. doi: 10.1016/j.str.2015.05.019

- de Sousa, E., Ligeiro, D., Lérias, J. R., Zhang, C., Agrati, C., Osman, M., et al. (2020). Mortality in COVID-19 disease patients: correlating the association of major histocompatibility complex (MHC) with severe acute respiratory syndrome 2 (SARS-CoV-2) variants. *Int. J. Infect. Dis.* 98, 454–459. doi: 10.1016/j.ijid.2020.07.016
- Díaz, J. (2020). SARS-CoV-2 molecular network structure. *Front. Physiol.* 11:870. doi: 10.3389/fphys.2020.00870
- Flower, T. G., Buffalo, C. Z., Hooy, R. M., Allaire, M., Ren, X., and Hurley, J. H. (2020). Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. *BioRxiv* [Preprint]. doi: 10.1101/2020.08.27.270637
- Gilfillan, S., Chan, C. J., Cella, M., Haynes, N. M., Rapaport, A. S., Boles, K. S., et al. (2008). DNAM-1 promotes activation of cytotoxic lymphocytes by nonprofessional antigen-presenting cells and tumors. *J. Exp. Med.* 205, 2965–2973. doi: 10.1084/jem.20081752
- Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468. doi: 10.1038/s41586-020-2286-9
- Hatherley, D., Lea, S. M., Johnson, S., and Barclay, A. N. (2013). Structures of CD200/CD200 receptor family and implications for topology, regulation, and evolution. *Structure* 21, 820–832. doi: 10.1016/j.str.2013.03.008
- Hill, A., Niles, B., Cuyekeng, A., and Powers, T. (2018). Redesigning TOR kinase to explore the structural basis for TORC 1 and TORC 2 assembly. *Biomolecules* 8:36. doi: 10.3390/biom8020036
- Holm, L. (2020). DALI and the persistence of protein shape. *Prot. Sci.* 29, 128–140. doi: 10.1002/pro.3749
- Ibáñez, A., Sarrias, M., Farnós, M., Gimferrer, I., Serra-Pagès, C., Vives, J., et al. (2006). Mitogen-activated protein kinase pathway activation by the CD6 lymphocyte surface receptor. *J. Immunol.* 177, 1152–1159. doi: 10.4049/jimmunol.177.2.1152
- Ikemizu, S., Robert, J. C., Gilbert, R. J. C., Fennelly, J. A., Collins, A. V., Harlos, K., et al. (2000). Structure and dimerization of a soluble form of B7-1. *Immunity* 12, 51–60. doi: 10.1016/S1074-7613(00)80158-2
- Islam, M. R., Hoque, M. N., Rahman, M. S., Rubayet Ul Alam, A. S. M., Akther, M., Akter Puspo, J., et al. (2020). Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. *Sci. Rep.* 10, 1–9. doi: 10.1038/s41598-020-70812-6
- Karam, B. S., Morris, R. S., Bramante, C. T., Puskarich, M., Zolfaghari, E. J., Lotfi-Emran, S., et al. (2021). MTOR inhibition in COVID-19: a commentary and review of efficacy in RNA viruses. *J. Med. Virol.* 93, 1843–1846. doi: 10.1002/jmv.26728
- Kindrachuk, J., Ork, B., Hart, B. J., Mazur, S., Holbrook, M. R., Frieman, M. B., et al. (2015). Antiviral potential of ERK/MAPK and PI3K/AKT/MTOR signaling modulation for middle east respiratory syndrome coronavirus infection as identified by temporal kinome analysis. *Antimicrob. Agents Chemother.* 59, 1088–1099. doi: 10.1128/AAC.03659-14
- Liu, W., and Li, H. (2020). COVID-19: attacks the 1-beta chain of hemoglobin and captures the porphyrin to inhibit human heme metabolism. *ChemRxiv* [Preprint]. doi: 10.26434/chemrxiv.11938173.v9
- McKee, D. L., Sternberg, A., Stange, U., Laufer, S., and Naujokat, C. (2020). Candidate drugs against SARS-CoV-2 and COVID-19. *Pharmacol. Res.* 157:104859. doi: 10.1016/j.phrs.2020.104859
- Mullen, P. J., Garcia, G., Purkayastha, A., Matulionis, N., Schmid, E. W., Momcilovic, M., et al. (2021). SARS-CoV-2 infection rewires host cell metabolism and is potentially susceptible to MTORC1 inhibition. *Nat. Commun.* 12:1876. doi: 10.1038/s41467-021-22166-4
- Muth, D., Corman, V. M., Roth, H., Binger, T., Dijkman, R., Theresa Gottula, I., et al. (2018). Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-018-33487-8
- Neches, R. Y., Kyripides, N. C., and Ouzounis, C. A. (2021). Atypical divergence of SARS-CoV-2 Orf 8 from Orf 7a within the coronavirus lineage suggests potential stealthy viral strategies in immune evasion. *mBio* 12:e03014-20. doi: 10.1128/mBio.03014-20
- Pasrija, R., and Naime, M. (2021). The deregulated immune reaction and cytokines release storm (CRS) in COVID-19 disease. *Int. Immunopharmacol.* 90:107225. doi: 10.1016/j.intimp.2020.107225
- Pereira, F. (2020). Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. *Infection, genetics and evolution. J. Mol. Epidemiol. Evol. Genet. Inf. Dis.* 85:104525. doi: 10.1016/j.meegid.2020.104525
- Pereira, F. (2021). SARS-CoV-2 variants combining spike mutations and the absence of ORF8 may be more transmissible and require close monitoring. *Biochem. Biophys. Res. Commun.* 550, 8–14. doi: 10.1016/j.bbrc.2021.02.080
- Radaev, S., Zou, Z., Tolar, P., Nguyen, K., Nguyen, A., Krueger, P. D., et al. (2010). Structural and functional studies of Igαβ; and its assembly with the B cell antigen receptor. *Structure* 18, 934–943. doi: 10.1016/j.str.2010.04.019
- Rashid, F., Dzakah, E. E., Wang, H., and Tang, S. (2021). The ORF8 protein of SARS-CoV-2 induced endoplasmic reticulum stress and mediated immune evasion by antagonizing production of interferon beta. *Virus Res.* 296:198350. doi: 10.1016/j.virusres.2021.198350
- Rollins, M. R., and Gibbons, J. R. M. (2017). CD80 expressed by CD8 + T cells contributes to PD-L1-induced apoptosis of activated CD8 + T Cells. *J. Immunol. Res.* 2017:7659462. doi: 10.1155/2017/7659462
- Saxton, R. A., and Sabatini, D. M. (2017). MTOR signaling in growth, metabolism, and disease. *Cell* 169, 361–371. doi: 10.1016/j.cell.2017.03.035
- Schreuder, H., Tardif, C., Trump-Kallmeyer, S., Soffientini, A., Sarubbi, E., Akeson, A., et al. (1997). A new cytokine-receptor binding mode revealed by the crystal structure of the IL-1 receptor with an antagonist. *Nature* 386, 194–200. doi: 10.1038/386194a0
- Sharma, H. B., Panigrahi, S., Sarmah, A. K., and Dubey, B. K. (2021). ORF8 contributes to cytokine storm during SARS-CoV-2 infection by activating IL-17 pathway. *iScience* 24:102293. doi: 10.1016/j.isci.2021.102293
- Su, Y. C. F., Anderson, D. E., Young, B. E., Linster, M., Zhu, F., Jayakumar, J., et al. (2020). Discovery and genomic characterization of a 382-nucleotide deletion in ORF7B and Orf8 during the early evolution of SARS-CoV-2. *MBio* 11, 1–9. doi: 10.1128/mBio.01610-20
- Trzupek, D., Dunstan, M., Cutler, A. J., Lee, M., Godfrey, L., Jarvis, L., et al. (2019). Discovery of CD80 and CD86 as recent activation markers on regulatory T cells by protein-RNA single-cell analysis. *BioRxiv* [Preprint]. doi: 10.1101/706275
- Vasile, S., Coligan, J. E., Yoshida, M., and Seon, B. K. (1994). Isolation and chemical characterization of the human B29 and Mb-1 proteins of the B cell antigen receptor complex. *Mol. Immunol.* 31, 419–427. doi: 10.1016/0161-5890(94)90061-2
- Velazquez-Salinas, L., Zarate, S., Eberl, S., Gladue, D. P., Novella, I., and Borca, M. V. (2020). Positive selection of ORF3a and ORF8 genes drives the evolution of SARS-CoV-2 during the 2020 COVID-19 pandemic. *BioRxiv* [Preprint]. doi: 10.1101/2020.04.10.035964
- Vilar, S., and Isom, D. G. (2021). One Year of SARS-CoV-2: how much has the virus changed? *Biology* 10:91. doi: 10.3390/biology10020091
- Wang, D., Zhang, S., Li, L., Liu, X., Mei, K., and Wang, X. (2010). Structural insights into the assembly and activation of IL-1β with its receptors. *Nat. Immunol.* 11, 905–911. doi: 10.1038/ni.1925
- Wang, H., Qi, J., Zhang, S., Li, Y., Tan, S., and Gao, G. F. (2019). Binding mode of the side-by-side two-IgV molecule CD226/DNAM-1 to its ligand CD155/Necl-5. *Proc. Natl. Acad. Sci.* 116, 988–996. doi: 10.1073/pnas.1815716116
- WHO (2021). Repurposed antiviral drugs for Covid-19 — interim WHO solidarity trial results. *N. Engl. J. Med.* 384, 497–511. doi: 10.1056/nejmoa2023184
- Xia, H., Cao, Z., Xie, X., Zhang, X., Chen, J. Y., Wang, H., et al. (2020). Evasion of type I interferon by SARS-CoV-2. *Cell Rep.* 33:108234. doi: 10.1016/j.celrep.2020.108234
- Young, B. E., Fong, S. W., Chan, Y. H., Mak, T. M., Ang, L. W., Anderson, D. E., et al. (2020). Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *Lancet* 396, 603–611. doi: 10.1016/S0140-6736(20)31757-8



- Zhang, Y., Zhang, J., Chen, Y., Luo, B., Yuan, Y., Huang, F., et al. (2020). The ORF8 protein of SARS-CoV-2 mediates immune evasion through potently downregulating MHC-I. *BioRxiv* [Preprint]. doi: 10.1101/2020.05.24.111823
- Zhang, Z., Wu, N., Lu, Y., Davidson, D., Colonna, M., and Veillette, A. (2015). DNAM-1 controls NK cell activation via an ITT-like Motif. *J. Exp. Med.* 212, 2165–2182. doi: 10.1084/jem.20150792
- Zinzula, L. (2021). Lost in deletion: the enigmatic ORF8 protein of SARS-CoV-2. *Biochem. Biophys. Res. Commun.* 530, 116–124. doi: 10.1016/j.bbrc.2020.10.045

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Valcarcel, Bensussen, Álvarez-Buylla and Díaz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# Correlation Between SARS-Cov-2 Vaccination, COVID-19 Incidence and Mortality: Tracking the Effect of Vaccination on Population Protection in Real Time

Kiyoshi F. Fukutani<sup>1,2</sup>, Mauricio L. Barreto<sup>3</sup>, Bruno B. Andrade<sup>1,2†</sup> and Artur T. L. Queiroz<sup>1,2,3\*†</sup>

<sup>1</sup> KAB Group, Goncalo Moniz Institute, Oswaldo Cruz Foundation, Salvador, Brazil, <sup>2</sup> Multinational Organization Network Sponsoring Translational and Epidemiological Research Initiative, Salvador, Brazil, <sup>3</sup> Center of Data and Knowledge Integration for Health, Oswaldo Cruz Foundation, Salvador, Brazil

## OPEN ACCESS

### Edited by:

Nimisha Ghosh,  
Siksha O Anusandhan University, India

### Reviewed by:

Gustavo Fioravanti Vieira,  
Universidade La Salle Canoas, Brazil  
Ruchi Tiwari,  
Pranveer Singh Institute  
of Technology PSIT, India

### \*Correspondence:

Artur T. L. Queiroz  
arturlopo@gmail.com

<sup>†</sup> These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 14 March 2021

**Accepted:** 06 May 2021

**Published:** 02 June 2021

### Citation:

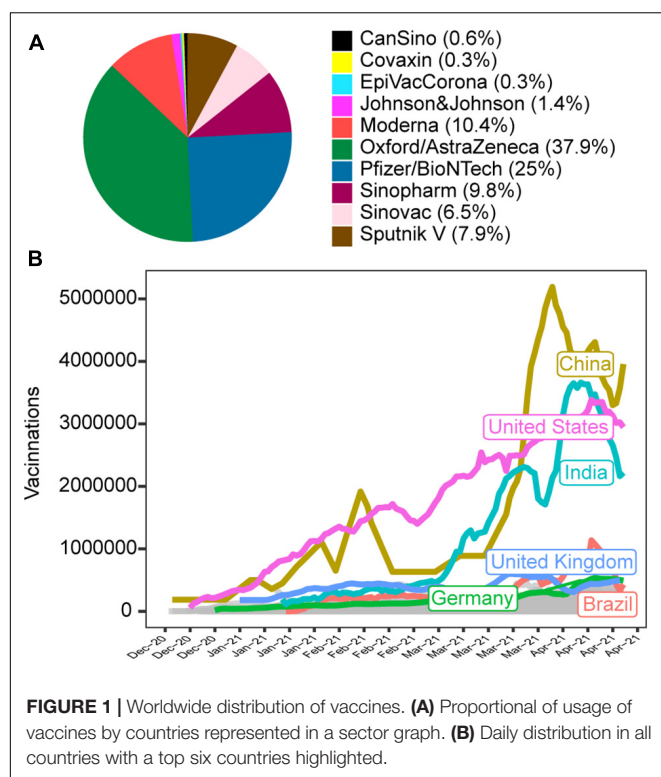
Fukutani KF, Barreto ML,  
Andrade BB and Queiroz ATL (2021)  
Correlation Between SARS-Cov-2  
Vaccination, COVID-19 Incidence  
and Mortality: Tracking the Effect  
of Vaccination on Population  
Protection in Real Time.  
Front. Genet. 12:679485.  
doi: 10.3389/fgene.2021.679485

Coronavirus disease 19 (COVID-19) has struck the world since the ending of 2019. Tools for pandemic control were scarce, limited only to social distance and face mask usage. Today, upto 12 vaccines were approved and the rapid development raises questions about the vaccine efficiency. We accessed the public database provided by each country and the number of death, active cases, and tests in order to evaluate how the vaccine is influencing the COVID-19 pandemic. We observed distinct profiles across the countries and it was related to the vaccination start date and we are proposing a new way to manage the vaccination.

**Keywords:** COVID19, vaccine, worldwide, epidemiology, virosis

## INTRODUCTION

A new SARS-Cov-2 associated disease is commonly known as coronavirus disease 19 (COVID-19) and present as a spectrum of clinical manifestations ranging from asymptomatic, minor flu-like symptoms to acute respiratory distress syndrome, pneumonia, and death (Sharma et al., 2020). Rapidly, the COVID-19 became a worldwide public health emergency and several attempts to control its dissemination were proposed by non- pharmacological interventions. The most used interventions were social distancing and the use of face masks, since there was no antiviral treatment or any effective vaccine (Randolph and Barreiro, 2020). In the last year, several vaccine candidates were in development, as a result of the great effort to contain the pandemic. However, due to the rapid vaccine development, uncertain questions have been raised in common media, such as the vaccine production capacity to attempt the global demand and its efficacy (Chen, 2020). The emergencial development of COVID-19 vaccines occurred extremely fast, integrating various tools and vaccine platforms. In the future, this technology will be useful to quickly develop vaccines against other new emerging diseases (Hodgson, 2020). Each government must have its own platform for vaccination tracking, in order to perform the monitoring of vaccine coverage and to early identification of possible adverse effects (Hanney et al., 2020). In 2020, we developed a



recursive sub-typing screening surveillance system able to perform automated genomic surveillance accessing all the sequences deposited in different repositories for mining, subtyping and performing a genomic surveillance. This system was also able to evaluate the vaccination profile in Brazil by accessing the global vaccination program dataset. As a result the system was able to identify new zika lineage occurrences (Kasprzykowski et al., 2020) and revealed a decrease in children vaccination in the last years in Brazil (Césaire et al., 2020). Given the relevance of the SARS-Cov-2 pandemic, we adapted our system to track the association between implementation of vaccines, occurrence of new cases and mortality over time.

## MATERIALS AND METHODS

To evaluate the COVID-19 vaccination, we developed an application of this tool to real-time access a public access COVID-19 database provided in a cross-country database of COVID-19 (Hasell et al., 2020). CaVaCo (Cases, Vaccinations, and COVID-19) tool allows us to retrieve the COVID-19 cases, deaths and vaccination data to compare and correlate countries vaccination coverage with other parameters. The tool was developed in R (Wickham and Golemund, 2016), powered to download and standardize the data automatically. As a result the correlation between number of daily vaccines by number of new cases, number of new deaths and number of tests is performed, using the spearman correlation. To access the real-time tool, access: <http://kaiju.bahia.fiocruz.br/sample-apps/CaVaCo/>.

**TABLE 1 |** Correlation between the numbers of vaccines against the number of new cases and new deaths in the country have started the vaccination.

	Cases		Deaths		N_of_days
	Rho coefficient	P-value	Rho coefficient	P-value	
Afghanistan	0.737	9.67876E-11	0.297	0.026230439	56
Albania	-0.702	1.5955E-15	-0.132	0.200294507	96
Algeria	-0.355	0.124947226	-0.023	0.923070238	20
Andorra	0.171	0.127808994	-0.123	0.2738076	81
Angola	0.762	2.04949E-10	0.074	0.610931829	49
Antigua and Barbuda	0.596	3.99133E-05	0.247	0.119180346	41
Argentina	0.274	0.003181829	0.035	0.71201508	114
Australia	0.472	0.000138566	0.139	0.289253571	60
Austria	0.581	8.1842E-12	-0.413	4.09028E-06	116
Azerbaijan	0.665	2.0045E-13	0.428	1.52276E-05	95
Bahamas	-0.084	0.817442415	0.432	0.213058411	10
Bahrain	0.821	1.53122E-30	0.562	2.29498E-11	120
Bangladesh	-0.129	0.235845636	-0.098	0.369853974	86
Barbados	0.769	1.76922E-13	0.249	0.049489179	63
Belarus	-0.581	2.97413E-09	-0.495	9.68988E-07	88
Belgium	0.553	1.765E-10	-0.339	0.000228344	114
Belize	-0.152	0.422511874	-0.247	0.187607437	30
Bolivia	-0.131	0.239039865	-0.47	7.18512E-06	83
Botswana	-0.738	0.262135213	-0.632	0.367544468	4
Brazil	0.306	0.002733153	0.641	3.29746E-12	94
Bulgaria	0.602	2.1644E-12	0.376	4.43204E-05	112
Cambodia	0.645	2.88954E-05	0.407	0.015027248	35
Canada	0.111	0.211429123	-0.676	1.46437E-18	129
Chile	0.521	1.6939E-09	0.267	0.003674764	117
China	-0.548	2.11955E-11	-0.363	2.52497E-05	128
Colombia	0.763	3.73228E-13	0.53	8.10472E-06	63
Costa Rica	-0.005	0.968928062	-0.133	0.302256683	62
Cote d'Ivoire	-0.667	1.21673E-07	-0.092	0.52432257	50
Croatia	0.422	3.31679E-06	-0.2	0.033482549	113
Cyprus	0.592	4.55219E-11	-0.309	0.001502789	103
Czechia	-0.276	0.002743769	-0.03	0.745564613	116
Denmark	-0.156	0.096400077	-0.711	5.64755E-19	115
Dominican Republic	-0.197	0.139137365	-0.214	0.106540076	58
Ecuador	0.17	0.111648671	0.125	0.24157091	89
Egypt	0.334	0.007902683	-0.343	0.006353965	62
El Salvador	0.065	0.658542705	-0.45	0.001318082	48
Equatorial Guinea	0.411	0.209233119	0.181	0.594070448	11
Estonia	0.403	7.97801E-06	0.528	1.37048E-09	115
Eswatini	0.05	0.839790752	0.012	0.961032401	19
Finland	0.496	3.53742E-08	-0.002	0.984810853	110
France	0.419	3.22325E-06	-0.097	0.302673079	115
Gabon	0.017	0.964546145	0.152	0.696613433	9
Gambia	-0.363	0.183775848	-0.038	0.891645336	15
Georgia	0.714	3.3297E-07	0.298	0.06509552	39
Germany	0.077	0.409459266	-0.672	1.45921E-16	116
Ghana	0.536	0.0027202	0.586	0.00083082	29
Greece	0.807	1.20134E-27	0.564	5.35719E-11	115
Guatemala	0.162	0.24210615	0.089	0.521713212	54
Guinea	0.354	0.14947453	0.323	0.191635298	18
Guyana	0.593	3.10376E-07	0.463	0.000131385	63
Honduras	0.22	0.184078772	0.133	0.427631584	38
Hungary	0.719	1.5585E-19	0.742	2.16927E-21	115
India	0.881	1.2858E-32	0.681	1.60918E-14	97
Indonesia	-0.874	1.8398E-32	-0.721	2.69614E-17	100
Iran	0.501	6.28703E-06	0.578	8.77196E-08	73
Iraq	0.756	1.25194E-09	0.52	0.00021189	46
Ireland	-0.801	8.60646E-26	-0.379	4.51725E-05	110
Israel	0.711	4.70919E-20	0.696	5.44762E-19	122
Italy	0.349	0.000123357	-0.146	0.117036098	116
Jamaica	0.132	0.449153159	0.101	0.563670349	35
Japan	0.729	8.62366E-12	-0.583	4.23837E-07	64
Jordan	0.578	5.89729E-10	0.819	1.29465E-24	97
Kazakhstan	0.788	4.2308E-18	0.258	0.020992955	80
Kenya	0.152	0.302319625	0.645	7.30516E-07	48

(Continued)

TABLE 1 | Continued

	Cases		Deaths		N_of_days
	Rho coefficient	P-value	Rho coefficient	P-value	
Kuwait	0.92	5.90966E-38	0.825	9.00643E-24	91
Kyrgyzstan	0.364	0.200594765	0.444	0.111919261	14
Latvia	-0.334	0.00093721	-0.383	0.000139038	94
Lebanon	-0.176	0.151851246	-0.394	0.000887308	68
Liechtenstein	-0.027	0.799263967	-0.209	0.044703081	93
Lithuania	-0.24	0.009513178	-0.78	5.82682E-25	116
Luxembourg	0.384	0.000573533	-0.127	0.270387032	77
Malawi	0.057	0.744125998	-0.091	0.604507315	35
Malaysia	-0.321	0.015926702	-0.189	0.163141237	56
Maldives	-0.25	0.027255853	-0.146	0.20234993	78
Mali	0.382	0.198295213	0.377	0.203554459	13
Malta	-0.423	2.18047E-05	-0.164	0.113838248	94
Mauritania	0.328	0.274642718	0	1	13
Mexico	-0.6	7.14745E-13	-0.388	1.41634E-05	118
Moldova	-0.482	0.000520045	-0.075	0.614125617	48
Monaco	0.057	0.562869113	0.177	0.072673617	104
Mongolia	0.596	4.87635E-06	0.312	0.027613831	50
Montenegro	-0.83	1.39318E-16	-0.002	0.984983613	61
Morocco	-0.405	0.000130867	-0.151	0.171129017	84
Mozambique	0.414	0.125247712	0.468	0.078739952	15
Myanmar	0.534	2.23273E-05	0.617	4.17262E-07	56
Namibia	-0.233	0.199540931	-0.07	0.703377868	32
Nepal	-0.536	1.24587E-07	-0.208	0.056620063	85
Netherlands	0.563	3.62332E-09	-0.578	1.09621E-09	94
Nigeria	-0.47	0.000658627	-0.323	0.023420163	49
North Macedonia	-0.184	0.141337842	0.308	0.012558504	65
Norway	0.194	0.037338479	-0.193	0.0391789	115
Oman	0.744	1.70361E-15	0.727	1.54321E-14	81
Pakistan	0.85	1.10786E-16	0.275	0.039973983	56
Palestine	0.375	0.078301362	0.018	0.933625998	23
Panama	-0.607	1.15805E-10	-0.655	1.03082E-12	93
Papua New Guinea	-0.134	0.694830743	-0.304	0.364208929	11
Paraguay	0.532	2.39756E-05	0.735	1.10315E-10	56
Peru	0.152	0.26207624	0.087	0.522670769	56
Philippines	0.743	2.96125E-10	0.508	0.000120473	52
Poland	0.584	7.38759E-12	0.219	0.018833951	115
Portugal	-0.785	1.84683E-25	-0.801	3.53171E-27	116
Qatar	0.953	1.57075E-63	0.76	5.05664E-24	121
Romania	0.295	0.001391266	0.488	3.06984E-08	115
Russia	-0.942	1.35517E-61	-0.731	1.17779E-22	128
Rwanda	-0.183	0.16240758	0.152	0.247662124	60
Saint Lucia	0.09	0.540947937	-0.121	0.41287011	48
Saint Vincent and the Grenadines	0.201	0.336375569	0.261	0.207778613	25
San Marino	-0.599	3.45821E-05	-0.096	0.54909611	41
Sao Tome and Principe	-0.154	0.600169081	-0.379	0.182026033	14
Saudi Arabia	0.718	4.51089E-18	0.767	8.35156E-22	106
Senegal	-0.218	0.096446802	-0.153	0.245958698	59
Serbia	0.19	0.057375795	-0.016	0.87312687	101
Seychelles	0.011	0.921462236	0.113	0.324517823	78
Singapore	-0.107	0.298611556	0.044	0.670464492	97
Slovakia	-0.489	7.71715E-08	-0.127	0.189701311	108
Slovenia	-0.323	0.000406873	-0.803	2.20051E-27	116
South Africa	0.049	0.717938189	0.019	0.891011593	56
South Korea	0.506	7.03271E-05	0.033	0.808115573	56
Spain	-0.663	1.22683E-10	-0.644	5.86039E-10	74
Sri Lanka	0.151	0.203361052	0.089	0.452643727	73
Suriname	-0.254	0.056169252	-0.25	0.060694914	57
Sweden	0.222	0.080308246	-0.565	1.38924E-06	63
Switzerland	0.338	0.008830615	-0.604	4.1967E-07	59
Taiwan	0.018	0.927252954	-0.203	0.290198758	29
Thailand	0.531	0.000104799	-0.034	0.819580758	48
Togo	-0.299	0.09095591	-0.03	0.869058406	33
Trinidad and Tobago	0.393	0.002273168	0.239	0.071135107	58
Tunisia	0.829	3.90157E-11	0.796	8.3898E-10	40

(Continued)

TABLE 1 | Continued

	Cases		Deaths		N_of_days
	Rho coefficient	P-value	Rho coefficient	P-value	
Turkey	0.383	9.18204E-05	-0.005	0.96031409	99
Uganda	0.369	0.022487201	0.11	0.509511091	38
Ukraine	0.679	6.5518E-09	0.669	1.2839E-08	57
United Arab Emirates	0.311	0.001169005	0.035	0.722547459	106
United Kingdom	-0.557	3.83837E-10	-0.521	7.28931E-09	108
United States	-0.78	1.94261E-26	-0.774	8.82158E-26	123
Uruguay	0.799	7.33099E-13	0.886	1.18631E-18	53
Uzbekistan	0.815	1.19672E-05	0.124	0.60357372	20
Venezuela	0.484	0.000278054	0.558	1.74225E-05	52
Zambia	-0.6	0.208	-0.676	0.140357387	6
Zimbabwe	0.056	0.664808883	-0.269	0.034558192	62

\*Countries without enough number of observations to perform a correlation analysis were excluded

The up to date table is available in <http://kaiju.bahia.fiocruz.br/sample-apps/CaVaCo/>

List of countries: Afghanistan, Albania, Algeria, Andorra, Angola, Anguilla, Antigua and Barbuda, Argentina, Armenia, Aruba, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Bermuda, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Cambodia, Cameroon, Canada, Cape Verde, Cayman Islands, Chile, China, Colombia, Congo, Costa Rica, Cote d'Ivoire, Croatia, Curacao, Cyprus, Czechia, Denmark, Djibouti, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, England, Equatorial Guinea, Estonia, Eswatini, Ethiopia, Faeroe Islands, Falkland Islands, Fiji, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Gibraltar, Greece, Greenland, Grenada, Guatemala, Guernsey, Guinea, Guyana, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Isle of Man, Israel, Italy, Jamaica, Japan, Jersey, Jordan, Kazakhstan, Kenya, Kosovo, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Lesotho, Libya, Liechtenstein, Lithuania, Luxembourg, Macao, Malawi, Malaysia, Maldives, Mali, Malta, Mauritania, Mauritius, Mexico, Moldova, Monaco, Mongolia, Montenegro, Montserrat, Morocco, Mozambique, Myanmar, Namibia, Nauru, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, North Macedonia, Northern Cyprus, Northern Ireland, Norway, Oman, Pakistan, Palestine, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Rwanda, Saint Helena, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, San Marino, Sao Tome and Principe, Saudi Arabia, Scotland, Senegal, Serbia, Seychelles, Sierra Leone, Singapore, Slovakia, Slovenia, Solomon Islands, Somalia, South Africa, South Korea, South Sudan, Spain, Sri Lanka, Sudan, Suriname, Sweden, Switzerland, Syria, Taiwan, Thailand, Timor, Togo, Tonga, Trinidad and Tobago, Tunisia, Turkey, Turks and Caicos Islands, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Venezuela, Vietnam, Wales, Zambia, and Zimbabwe.

## PERSPECTIVE

So far (April 23, 2021), there are 10 vaccines approved and being used worldwide (until: CanSino, Covaxin, EpiVacCorona, Johnson & Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm, Sinovac, and Sputnik V). From the 193 countries that started vaccination (List of countries below) the majority have started the vaccination program using Oxford/AstraZeneca vaccine ( $n = 135$ , 37.9%) while 25% had chosen the Pfizer/BioNTech and 10.4% Moderna and the remaining 26.7% used CanSino, Covaxin, EpiVacCorona, Johnson & Johnson, Sinopharm, Sinovac, and Sputnik V (Figure 1). Using the date available up to April 23, 2021, we performed a correlation analysis between the numbers of new cases with the daily vaccinations. As a result, 60 countries presented positive correlations (Table 1) and 27 countries with negative correlation (Table 1). Despite the vaccination,



the number of new cases has been still increasing in these countries. This finding reinforces the need to keep social distance and the use of face masks recommendations to reduce the virus transmission. In other hand the decreasing number of vaccinations and cases can depict a positive correlation and the number of days and the percent of vaccine population could inform how successfully the vaccination program is going. These recommendations should be employed until at least the immunization starts to show a significant reduction in the number of cases (Ahmed et al., 2021). The countries with negative correlation started to have a reduction in the number of new cases and the vaccination should maintain the decreasing number of cases, since the isolation alone is not able to control the COVID-19 (Hellewell et al., 2020). The same approach has employed with the number of new deaths and we observed 37 countries with positive correlations and 33 countries have negative correlations (Table 1). These results show that implementation of vaccines is not the final solution and the maintenance of the non-pharmacological interventions should not be abandoned once the increase of new cases and deaths are indicating the population remains vulnerable to SARS- COV2 infection (Billon-Denis and Tournier, 2020). On the other hand, the negative correlation in certain countries point to a success en route to the vaccination program in reducing both the COVID-19 cases and related deaths. Only 5 countries have positive correlation between the number of vaccination and the number of tests positive for COVID-19 in February 2, 2021 (This data was discontinued). These countries remained testing the population even though the vaccination started. Only Sweden presented a negative correlation (Supplementary Table 1). This approach is useful for pandemic surveillance and the stop of population testing is dangerous and does not prevent the identification of new waves (Holt, 2021). The correlation between the cases/deaths and the vaccination numbers could be a powerful indicator of disease control, since a certain coverage is required for population protection. The continuous follow up of the correlation patterns from the beginning of the vaccination can be used to track the immunization program in each country. Additionally with the genomic surveillance can reveal how the vaccine responds

against the introduction of new COVID-19 variants, as previously described (Korber et al., 2020). The present study has some limitations, such as the heterogeneity of strategies applied by the different countries indicated that an individual analysis of specific countries should be performed to evaluate in more granularities the distinct epidemiologic situations, to minimize this effect the number of days used in the correlation analysis are depicted in the table. Some countries displayed substantial missing data or discontinue measuring few variables, like the number of test to COVID-19 in their database. This analysis uses numerical measurements and it cannot reflect the entire national behavior or public politics. Also the present analysis cannot handle or correct numeric bias or outlier interferences. However, taking together these data and applying statistics methods allowed us to monitor the vaccination process in countries or in sub national units. Recursive evaluation of immunization and COVID-19 morbimortality has potential to provide a unique tool to aid decision-making strategies to overcome the current pandemic.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

KF and AQ performed data acquisition and analysis. KF, BA, and AQ performed the results interpretation. All authors wrote the manuscript, contributed to the article, and approved the submitted version.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.679485/full#supplementary-material>

## REFERENCES

- Ahmed, I., Ahmad, M., Rodrigues, J. J. P. C., Jeon, G., and Din, S. (2021). A deep learning-based social distance monitoring framework for COVID-19. *Sustain. Cities Soc.* 65:102571. doi: 10.1016/j.scs.2020.102571
- Billon-Denis, E., and Tournier, J.-N. (2020). [COVID-19 and vaccination: a global disruption]. *Med. Sci.* 36, 1034–1037.
- Césaire, N., Mota, T. F., Lopes, F. F. L., Lima, A. C. M., Luzardo, R., Quintanilha, L. F., et al. (2020). Longitudinal profiling of the vaccination coverage in Brazil reveals a recent change in the patterns hallmarked by differential reduction across regions. *Int. J. Infect. Dis.* 98, 275–280. doi: 10.1016/j.ijid.2020.06.092
- Chen, W. (2020). Promise and challenges in the development of COVID-19 vaccines. *Hum. Vaccin. Immunother.* 16, 2604–2608. doi: 10.1080/21645515.2020.1787067
- Hanney, S. R., Wooding, S., Sussex, J., and Grant, J. (2020). From COVID-19 research to vaccine application: why might it take 17 months not 17 years and what are the wider lessons?. *Health Res. Policy Syst.* 18:61.
- Hasell, J., Mathieu, E., Beltekian, D., Macdonald, B., Giattino, C., Ortiz-Ospina, E., et al. (2020). A cross-country database of COVID-19 testing. *Sci. Data* 7:345.
- Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., et al. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob. Health* 8, e488–e496.
- Hodgson, J. (2020). The pandemic pipeline. *Nat. Biotechnol.* 38, 523–532. doi: 10.1038/d41587-020-00005-z
- Holt, E. (2021). COVID-19 testing in Slovakia. *Lancet Infect. Dis.* 21:32. doi: 10.1016/s1473-3099(20)30948-8
- Kasprzykowski, J. I., Fukutani, K. F., Fabio, H., Fukutani, E. R., Costa, L. C., Andrade, B. B., et al. (2020). A recursive sub-typing screening surveillance system detects the appearance of the ZIKV African lineage in Brazil: is there a risk of a new epidemic?. *Int. J. Infect. Dis.* 96, 579–581. doi: 10.1016/j.ijid.2020.05.090
- Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Tracking Changes in SARS-CoV-2 Spike: evidence

- that D614G Increases Infectivity of the COVID-19 Virus. *Cell* 182, 812–827.e19.
- Randolph, H. E., and Barreiro, L. B. (2020). Herd Immunity: understanding COVID-19. *Immunity* 52, 737–741. doi: 10.1016/j.immuni.2020.04.012
- Sharma, O., Sultan, A. A., Ding, H., and Triggle, C. R. (2020). A Review of the Progress and Challenges of Developing a Vaccine for COVID-19. *Front. Immunol.* 11:585354. doi: 10.3389/fimmu.2020.585354
- Wickham, H., and Grolemund, G. (2016). *R for Data Science: import, Tidy, Transform, Visualize, and Model Data*. United States: O'Reilly Media, Inc.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Fukutani, Barreto, Andrade and Queiroz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



# ORF8 and Health Complications of COVID-19 in Down Syndrome Patients

Antonio Bensussen<sup>1</sup>, Antonio Valcarcel<sup>1</sup>, Elena R. Álvarez-Buylla<sup>2,3\*</sup> and José Díaz<sup>1\*</sup>

<sup>1</sup>Laboratorio de Dinámica de Redes Genéticas, Centro de Investigación en Dinámica Celular, Universidad Autónoma del Estado de Morelos, Cuernavaca, Mexico, <sup>2</sup>Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Ciudad de México, Mexico, <sup>3</sup>Laboratorio de Genética Molecular, Epigenética, Desarrollo y Evolución de Plantas, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

**Keywords: SARS-CoV-2, COVID-19, ORF8, Down syndrome, Cardiac damage**

This manuscript intends to be a “letter to the editor” with some complementary commentaries to the manuscript by Valcarcel et al. (2021). In that work, ORF8 was characterized as a notable SARS-CoV-2 protein able to downregulate the function of MHC-I (Zhang et al., 2021) and which shares structural similarities with human immunoglobulins (including interleukins) that can eventually produce immune dysregulation (Valcarcel et al., 2021). It is not still clear if all COVID-19 patients are equally susceptible to this ORF8-mediated immune dysregulation, but Down syndrome (Ds) patients with COVID-19 have more health complications, such as cardiac diseases, and higher rate of mortality than the general population, especially in those over 40 years old (Hüls et al., 2021). Ds is an important comorbidity since these patients have an extra copy of the *TMPRSS2* gene, which probably produces enhanced levels of the transmembrane *TMPRSS2* protease for S protein priming, facilitating the SARS-CoV-2 infection of the target cells (Hoffmann et al., 2020; De Toma and Dierssen, 2021).

Therefore, we proposed a minimal mathematical model of the effect of the extra copy of *TMPRSS2* on ORF8 production and persistence in the infected cells (**Figure 1A**), which reasonably fits with the experimental data reported in literature. According to the model results, we found that systemic levels of ORF8 are considerably higher and persists up to 40 days in patients with Ds (**Figures 1B, C**) in contrast with patients without Ds. These results support our hypothesis that the high susceptibility of people with Ds to be infected by SARS-CoV-2 is a consequence of the overproduction of *TMPRSS2*, which produces high systemic levels of ORF8 with the subsequent immune dysregulation, lung inflammatory effects, and cardiac damage that worsen the disease (Espinosa, 2020).

Additional consequences of the overproduction of ORF8 in Ds patients with COVID-19 are as follows: 1) the several structural similarities of this viral protein with the nitric oxide synthase can alter the serum concentrations of NO, reducing the protective function of this gas against arrhythmias (Burger and Feng, 2011); 2) ORF8 can also be an important factor to aggravate the cytokine storm due its high degree of structural mimicry with immunoglobulins and their receptors (Valcarcel et al., 2021), with the subsequent small protoembolic events that cause a cardiovascular damage similar to that of older Ds patients never infected with Covid-19 (Colvin and Yeager, 2017; De Toma and Dierssen, 2020). 3) Taking into consideration that chromosome 21 also harbors multiple genes involved in the immune response, and their overexpression induces the dysregulation of interleukins IL-10, IL-22, and IL-26 prior to infection (De Weerd and Nguyen, 2020), the presence of high levels of ORF8 could also be an important factor to aggravate the cytokine storm in Ds patients with COVID-19.

However, it is necessary to do more theoretical, experimental, and clinical research to elucidate the precise role of ORF8 in the immune dysregulation, lung inflammatory effects, and cardiac damage in this group of patients.

## OPEN ACCESS

### Edited by:

Dariusz Plewczynski,  
Warsaw University of Technology,  
Poland

### Reviewed by:

Roopa Biswas,  
Uniformed Services University of the  
Health Sciences, United States

### \*Correspondence:

Elena R. Álvarez-Buylla  
elenabuylla@protonmail.com  
José Díaz  
biofisica@yahoo.com

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

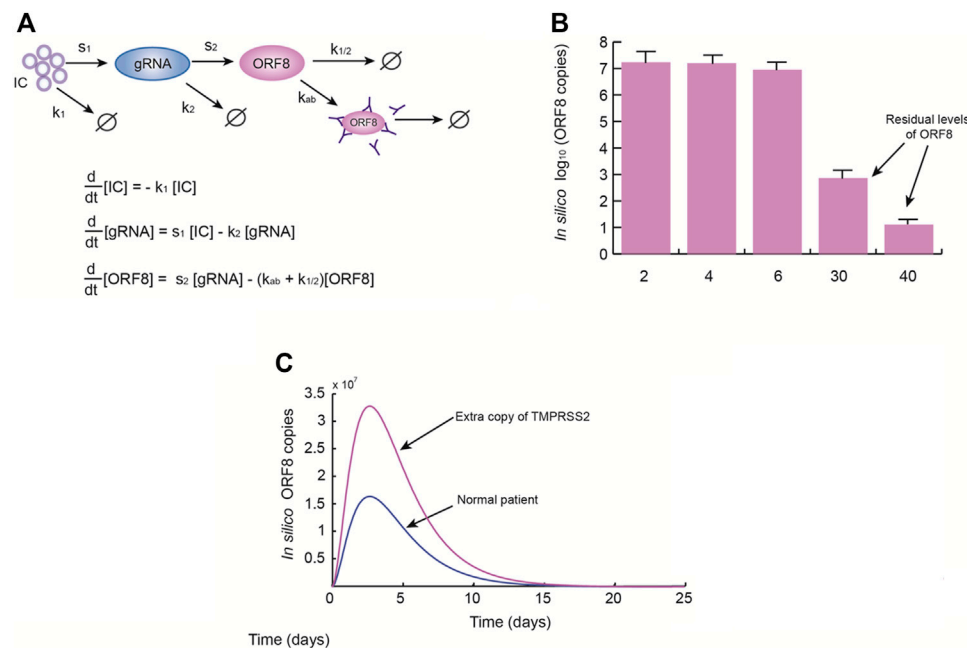
**Received:** 07 December 2021

**Accepted:** 10 January 2022

**Published:** 09 February 2022

### Citation:

Bensussen A, Valcarcel A,  
Álvarez-Buylla ER and Díaz J (2022)  
ORF8 and Health Complications of  
COVID-19 in Down  
Syndrome Patients.  
Front. Genet. 13:830426.  
doi: 10.3389/fgene.2022.830426



**FIGURE 1 |** Down syndrome patients have higher levels of ORF8 during SARS-CoV-2 infection. **(A)** Once the SARS-CoV-2 ACE2-Spike complexes enter the cells (IC), the production of genomic RNA (gRNA) originates ORF8 antigens by several editing steps. These copies of ORF8 may be naturally degraded or eliminated by antibodies. **(A)** The simplified model of ORF8 production used in this work. In the model  $k_1 = 0.4 \text{ days}^{-1}$ ,  $s_1 = 100 \text{ days}^{-1}$ ,  $k_2 = 1.3 \text{ days}^{-1}$ ,  $s_2 = 20 \text{ days}^{-1}$ ,  $k_{ab} = 0.3 \text{ days}^{-1}$ , and  $k_{1/2} = 0.5 \text{ days}^{-1}$ . **(B)** Mean levels of ORF8 obtained from the model simulations at 2, 4, 6, 30, and 40 days. These results suggest that ORF8 persists even if the patient is discharged 15 days after viral onset. **(C)** Effect of the presence of an extra copy of TMPRSS2. These simulations show that an extra copy of TMPRSS2 gene is able to dramatically increase systemic levels of ORF8, which implies that Down syndrome patients are more susceptible to medical complications produced by ORF8.

## AUTHOR CONTRIBUTIONS

AB and AV have contributed equally to this work and share first authorship. AB made the *in silico* analysis of the model. All authors participated equally in the writing of the manuscript.

## REFERENCES

- Bar-On, Y. M., Flamholz, A., Phillips, R., and Milo, R. (2020). Sars-CoV-2 (Covid-19) by the Numbers. *eLife* 9, e57309. doi:10.7554/eLife.57309
- Can, H., Köseoğlu, A. E., Erkunt Alak, S., Güvendi, M., Döşkaya, M., Karakavuk, M., et al. (2020). In Silico discovery of Antigenic Proteins and Epitopes of SARS-CoV-2 for the Development of a Vaccine or a Diagnostic Approach for COVID-19. *Sci. Rep.* 10, 22387. doi:10.1038/s41598-020-79645-9
- Colvin, K. L., and Yeager, M. E. (2017). What People with Down Syndrome Can Teach Us about Cardiopulmonary Disease. *Eur. Respir. Rev.* 26, 160098. doi:10.1183/16000617.0098-2016
- De Toma, I., and Dierssen, M. (2021). Network Analysis of Down Syndrome and SARS-CoV-2 Identifies Risk and Protective Factors for COVID-19. *Sci. Rep.* 11, 1930. doi:10.1038/s41598-021-81451-w
- Espinosa, J. M., (2020). Down Syndrome and COVID-19: A Perfect Storm?. *Cell Reports Medicine* 1, 1–8. doi:10.1016/j.crm.2020.100019
- Burger, D. E., and Feng, Q. (2011). Protective Role of Nitric Oxide against Cardiac Arrhythmia - an Update. *Tonaj* 3 (Suppl. 1-M6), 38–47. doi:10.2174/1875042701103010038
- Hachim, A., Kavian, N., Cohen, C. A., Chin, A. W. H., Chu, D. K. W., Mok, C. K. P., et al. (2020). ORF8 and ORF3b Antibodies Are Accurate Serological Markers of

## ACKNOWLEDGMENTS

AB and AV thank CONACYT for the postdoctoral fellowship. We thank Erika Juarez Luna for her logistical support.

- Early and Late SARS-CoV-2 Infection. *Nat. Immunol.* 21, 1293–1301. doi:10.1038/s41590-020-0773-7
- Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181, 271–280. doi:10.1016/j.cell.2020.02.052
- Hüls, A., Costa, A. C. S., Dierssen, M., Baksh, R. A., Bargagna, S., Baumer, N. T., et al. (2021). Medical Vulnerability of Individuals with Down Syndrome to Severe COVID-19-Data from the Trisomy 21 Research Society and the UK ISARIC4C Survey. *EclinicalMedicine* 33, 100769. doi:10.1016/j.eclinm.2021.100769
- Peng, Y., Mentzer, A. J., Liu, G., Yao, X., Yin, Z., Dong, D., et al. (2020). Broad and strong Memory CD4+ and CD8+ T Cells Induced by SARS-CoV-2 in UK Convalescent Individuals Following COVID-19. *Nat. Immunol.* 21, 1336–1345. doi:10.1038/s41590-020-0782-6
- Valcarcel, A., Bensussen, A., Álvarez-Buylla, E. R., and Díaz, J. (2021). Structural Analysis of SARS-CoV-2 ORF8 Protein: Pathogenic and Therapeutic Implications. *Front. Genet.* 12, 1–8. doi:10.3389/fgene.2021.693227
- Weerd, N. A., and Nguyen, T. (2012). The Interferons and Their Receptors-Distribution and Regulation. *Immunol. Cel Biol* 90, 483–491. doi:10.1038/icb.2012.9



Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., et al. (2020). Virological Assessment of Hospitalized Patients with COVID-2019. *Nature* 581, 465–469. doi:10.1038/s41586-020-2196-x

Zhang, Y., Chen, Y., Li, Y., Huang, F., Luo, B., Yuan, Y., et al. (2021). The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through Down-Regulating MHC-I. *Proc. Natl. Acad. Sci. USA* 118, e2024202118. doi:10.1073/pnas.2024202118

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Bensussen, Valcarcel, Álvarez-Buylla and Díaz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

## APPENDIX

The model we used in this work takes into account the levels of the Spike-ACE2-TMPRSS2 complex (*IC*) that allows the internalization of the SARS-CoV-2 genomic RNA (*gRNA*) and the systemic levels of ORF8. To model ORF8 systemic levels, we considered the half-life of the protein (Can et al., 2020) as well as the effect of neutralizing antibodies against

ORF8 (Hachim et al., 2020) (**Figure 1A**). Next, we used *ex vivo* data obtained from COVID-19 patients (Bar-On et al., 2020; Hachim et al., 2020; Peng et al., 2020; Wölfel et al., 2020) to estimate parameters and to calibrate the model congruently with previous observations (Peng et al., 2020). The model was numerically solved using the Runge–Kutta 4-5 method and MATLAB software.

# Advantages of publishing in Frontiers



## OPEN ACCESS

Articles are free to read  
for greatest visibility  
and readership



## FAST PUBLICATION

Around 90 days  
from submission  
to decision



## HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,  
and constructive  
peer-review



## TRANSPARENT PEER-REVIEW

Editors and reviewers  
acknowledged by name  
on published articles

## Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne | Switzerland

**Visit us:** [www.frontiersin.org](http://www.frontiersin.org)

**Contact us:** [frontiersin.org/about/contact](http://frontiersin.org/about/contact)



## REPRODUCIBILITY OF RESEARCH

Support open data  
and methods to enhance  
research reproducibility



## DIGITAL PUBLISHING

Articles designed  
for optimal readership  
across devices



## FOLLOW US

@frontiersin



## IMPACT METRICS

Advanced article metrics  
track visibility across  
digital media



## EXTENSIVE PROMOTION

Marketing  
and promotion  
of impactful research



## LOOP RESEARCH NETWORK

Our network  
increases your  
article's readership