Computers in Biology and Medicine 129 (2021) 104131

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

DBCOVP: A database of coronavirus virulent glycoproteins

Susrita Sahoo [a], Soumya Ranjan Mahapatra [a], Bikram Kumar Parida [c], Satyajit Rath [c], Budheswar Dehury [a], Vishakha Raina [a], Nirmal Kumar Mohakud [d], Namrata Misra [a,b], Mrutyunjay Suar [a,b,*]

[a] School of Biotechnology, Kalinga Institute of Industrial Technology (KIIT), Deemed to be University, Bhubaneswar, Odisha, India
[b] KIIT-Technology Business Incubator (KIIT-TBI), Kalinga Institute of Industrial Technology (KIIT), Deemed to be University, Bhubaneswar, Odisha, India
[c] Informatics Lab, CSIR-Institute of Minerals and Materials Technology (CSIR-IMMT), Bhubaneswar, Odisha, India
[d] Department of Pediatrics, Kalinga Institute of Medical Sciences, KIIT Deemed to Be University, Bhubaneswar, Odisha, India

ARTICLE INFO

Keywords:
Coronavirus
Glycoproteins
Database
Bioinformatics
Immunoinformatics
COVID-19

ABSTRACT

Since the emergence of SARS-CoV-1 (2002), novel coronaviruses have emerged periodically like the MERS- CoV (2012) and now, the SARS-CoV-2 outbreak which has posed a global threat to public health. Although, this is the third zoonotic coronavirus breakout within the last two decades, there are only a few platforms that provide information about coronavirus genomes. None of them is specific for the virulence glycoproteins and complete sequence-structural features of these virulence factors across the betacoronavirus family including SARS-CoV-2 strains are lacking. Against this backdrop, we present DBCOVP (http://covp.immt.res.in/), the first manually-curated, web-based resource to provide extensive information on the complete repertoire of structural virulent glycoproteins from coronavirus genomes belonging to *betacoronavirus* genera. The database provides various sequence-structural properties in which users can browse and analyze information in different ways. Furthermore, many conserved T-cell and B-cell epitopes predicted for each protein are present that may perform a significant role in eliciting the humoral and cellular immune response. The tertiary structure of the epitopes together with the docked epitope-HLA binding-complex is made available to facilitate further analysis. DBCOVP presents an easy-to-use interface with in-built tools for similarity search, cross-genome comparison, phylogenetic, and multiple sequence alignment. DBCOVP will certainly be an important resource for experimental biologists engaged in coronavirus research studies and will aid in vaccine development.

1. Introduction

Coronaviruses belonging to the *Coronaviridae* family is the causative agent of neurologic, enteric, hepatic, and upper respiratory tract diseases in a wide range of hosts including human, cattle, camels, swine, bats, cats, dogs, rabbits, snake, and several other wild animals and avian host species [1]. The genome comprises a single positive-stranded RNA genome, with size ranging from 26 to 32 Kilo bases in length, with G + C contents varying from 32 to 43% [1,2]. Among the various coronaviruses that are infecting humans, the majority are associated with mild clinical symptoms unlike the Severe Acute Respiratory Syndrome (SARS) coronavirus (SARS-CoV-1) and Middle East Respiratory Syndrome (MERS) coronavirus (MERS-CoV) [3], which cause high morbidity and mortality in human populations. SARS-CoV-1 incidence was initially reported in November 2002 in Guangdong, Southern China, and resulted in around 8000 cases of human infections with 744 deaths, around 9.5% mortality rate [4,5]. Later on, a similar epidemic outbreak (MERS-CoV) was first detected in Saudi Arabia in September 2012 which resulted in a higher incidence of mortality rate [6–8]. Recently, in late December 2019, patients with viral pneumonia symptoms due to an unidentified etiology were reported first in Wuhan City, China [9]. A novel coronavirus was later identified as the causative pathogen, provisionally named as 2019-nCoV, and later renamed as SARS-CoV-2, has been declared as the Public Health Emergency of International Concern by the World Health Organization (WHO) on 30 January 2020 [9,10]. As of 1st August 2020, the virus has spread worldwide affecting 213 countries with more than 6 million cases of infected patients. According to comparative genomic analysis, SARS-CoV-2 shares 79.5% nucleotide identity with SARS-CoV-1; and 96% identity with bat-CoV-RaTG13. Therefore, SARS-CoV-2 is considered as SARS related coronavirus, and

* Corresponding author. School of Biotechnology, Kalinga Institute of Industrial Technology (KIIT), Deemed to be University Bhubaneswar, 751024, Odisha, India.
E-mail addresses: msuar@kiitbiotech.ac.in, mrutyunjay@kiitincubator.in (M. Suar).

https://doi.org/10.1016/j.compbiomed.2020.104131
Received 10 August 2020; Received in revised form 31 October 2020; Accepted 13 November 2020
Available online 21 November 2020

bats as the most probable source of infection [11]. The SARS-CoV-2, SARS-CoV-1, and MERS-CoV show several similarities regarding the clinical presentations with pneumonia-like symptoms, evidence of zoonotic transmission as the route of disease origin, and human to human transmission [12]. Furthermore, all three coronaviruses belong to the genus *betacoronavirus* which is further classified into five sub-genus, namely *Sarbecovirus*, *Embecovirus*, *Hibecovirus*, *Merbecovirus*, and *Nobecovirus*. The SARS-CoV-2, SARS-CoV-1belongs to the *Sarbecovirus* subgenus [9]. Despite the great threats to public health around the world and global concern to combat the spread of the ongoing outbreak, to date, there are no clinically approved vaccines available for either SARS-CoV-2 or SARS, MERS, and therefore further research is imperative for identifying appropriate therapeutic targets for the development of safe, stable vaccines for combating human coronavirus infections [12,13].

Advances in molecular biology and the use of bioinformatics resources, particularly the immunoinformatics approach have resulted in a deluge of genomic data that can provide prior information on the efficacy of potential vaccine targets worthy of subsequent validation through wet-lab experiments, thus saving a lot of time and effort in the vaccine discovery process [13,14]. The prediction and characterization of immunogenic epitopes that can induce antibody production from B-cells and cellular response and cytokine secretion from T-cells is a critical step in silico identification and assessment of potential vaccine targets. The epitope-driven vaccine concept has already been successfully employed against many infectious diseases in recent years [15–17]. As the first step in this direction, it is essential to find proteins that play a definite role in the pathogenesis of any virus. The primary goal of any viral infection is to pinpoint a receptor on the host cell surface for effective binding which would pave the entry of the virus into the host cell. In most cases, glycoproteins are involved in host binding and subsequent virus-host membrane fusion to establish the pathogenesis of the virus [18]. The four important glycoproteins that majorly contribute to the structure of all coronaviruses are the spike protein (S), small envelope protein (E), membrane protein (M), and nucleocapsid (N) protein [13]. The S protein mediates receptor binding and membrane fusion and is vital for identifying host tropism and transmission capacity [19–21]. Mutations in the gene encoding spike protein have resulted in altered pathogenesis and virulence in other coronaviruses [22]. It is believed that three molecules of spike proteins form the characteristic ‘spikes’ or the crown-like appearance specific of this virus family [13].The majority of the candidate vaccine that is being developed against coronaviruses, targets the spike protein as they are the major inducer of neutralizing antibodies [23,24]. It is seen that the association of the spike with the membrane protein is crucial in the formation of the viral envelope and the accumulation of both the glycoproteins at the site of virus assembly [22]. The gene encoding the nucleocapsid protein in the SARS-CoV-1 virus is believed to possess a novel nuclear function, which could play a role in pathogenesis. Additionally, the basic nature of this protein implies that it may assist in RNA binding [22,23]. Lastly, the envelope protein has been shown to play an important role in the assembly of the virion and its replication [25,26]. These structural proteins have a diverse functional role in the viral pathogenesis; therefore, a dedicated database on all the four discussed major structural glycoproteins will provide a timely and valuable source of detailed sequence-structural properties about these virulence factors to the scientific community that will aid in the development of vaccines against coronavirus.

Despite the constant emerging and re-emerging of the deadly coronavirus since the last two decades, to date, there are only a few dedicated web resources exclusively available to study coronaviruses genes and proteins. For instance, the Comprehensive Database for Comparative Analysis of Coronavirus Genes and Genomes (CoVDB) that performs fast, and precise batch sequence retrieval, the basis for comparative gene or genome analysis [27]. CoVDB has not been updated since 2007 and provides limited annotation features including cleavage sites, genome information, tandem repeat sequences, transcription regulatory sequences, and RNA structures. Virus Pathogen Database and Analysis Resource (VipR) covers a huge plethora of human pathogenic viruses but includes knowledge on sequence records, a few genome and protein annotations, tertiary protein structures, immune epitope, surveillance, and clinical metadata derived from comparative genomics analysis [28]. Although very useful, VipR doesn’t hold any information specific to the virulence glycoprotein and further lacks details on secondary structure properties, subcellular location, molecular function, biological process, domain, cluster, Super family, Physicochemical properties, Epitope conservancy, Allergenicity, Antigenicity, Toxicity, 3D epitope structure, Population coverage analysis. Similarly, ViralZone (https://viralzone.expasy.org/), a web-resource for viral genus and families, hosted by the Swiss Institute of Bioinformatics provides general molecular and epidemiological information of viruses [29]. Since March 2020, ViralZone holds Covid-19 genome expression details, protein sequence records, host-virus interaction, and general information on coronaviruses belonging to *betacoronavirus* genera. A GenBank submissions tool, Viral Annotation DefineR (VADR, https://github.com/nawrockie/vadr), was specifically designed to validate and annotate viral sequences. VADR has been used to check sequence submissions norovirus (May 2018), dengue virus (January 2019), and SARS-CoV-2 (March 2020) sequence submissions [30]. Likewise, the Viral Bioinformatics Resource Center (VBRC, https://4virology.net/) funded by the National Institute of Allergy and Infectious Diseases, holds information on curated viral genomes (belonging to the family Coronaviridae, Asfarviridae, Poxviridae) and a plethora of bioinformatics tools to perform genome analysis [31]. Presently, VBRC redirects to various exclusive SARS-CoV-2 resources viz., genome, scientific literature, Worldometers, case trackers, and COVID-19 specific news. Earlier developed, Rfam, an online resource providing access to families of structural RNAs, where each family is characterized by a covariance model and multiple sequence alignment [32]. Its current special release RFAM 14.2 includes details on Untranslated regions of all the five families of coronavirus. Recently made available, CORona Drug InTERactions database (CORDITE, https://cordite.mathematik.unimarburg.de/#/) collects and aggregates details on in vitro, computational, or case analyses on promising drugs for COVID-19 from PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), chemRxiv (https://www.chemrxiv.org/), bioRxiv (https://www.biorxiv.org/), and medRxiv (https://www.medrxiv.org/) to further perform meta-analyses and new clinical trials [33]. To find putative drug targets and further explore the molecular mechanisms of pathogenicity, sadegh et al., developed a CoronaVirus Explorer (CoVeX, https://exbio.wzw.tum.de/covex/) that includes information on drug candidates and experimentally validated virus–human interaction data for both SARS-CoV-2 and SARS-CoV-1 with human interactome [34]. Apart from the above mentioned web resources, few other platforms are recently made available exclusively focused on coronavirus research like Coronavirus Database V3 (http://covdb.popgenetics.net/v3/ [35]), that contains only genomic data; COVID-Profiler (http://genomics.lshtm.ac.uk/), analyses Sars-Cov-2 sequencing and a few immunological data; COVIEdb (http://biopharm.zju.edu.cn/coviedb/help/ [36]), holds only some potential B/T cell epitopes for SARS CoV-2, RaTG13-CoV, SARS-CoV and MERS-CoV; CoVIDep (https://covidep.ust.hk/ [37]), consists of genetic data for SARS-CoV-2 and immunological data for the 2003 SARS virus, to identify B-cell and T-cell epitopes; CoV3D (https://cov3d.ibbr.umd.edu/cov3d), contains structures of SARS-CoV-2, SARS-CoV, and MERS-CoV proteins, without any other structural details; CoronaVIR (https://webs.iiitd.edu.in/raghava/coronavir/index.html [38]) contains a few genomic, proteomic, diagnostic and therapeutic knowledge about novel SARS-CoV-2 coronaviruses. Moreover, none of them is specific for the virulence proteins encompassing the spike protein, small envelope protein, membrane protein, and nucleocapsid protein and an in-depth investigation of complete sequence-structural features of these virulence factors across the betacoronavirus family including the newly identified SARS-CoV-2 strains is lacking. Although sequence efforts have resulted in a marked increase in emerging SARS-CoV-2 sequenced data; however, functional annotation of the encoded proteins in primary databases such as GenBank and UniProt knowledgebase remains limited. To address this issue, we developed a specifically designed web-accessible resource DBCOVP (http://covp.immt.res.in/) to integrate in-depth functional annotation of coronavirus virulence glycoproteins (Fig. 1). DBCOVP is the first manually curated data repository that provides comprehensive details on the entire repertoire of structural glycoproteins from coronavirus genomes of betacoronavirus genera including the SARS-CoV-1, MERS-CoV, and SARS-CoV-2 strains. The database provides complete functional annotation of the proteins highlighting fourteen sequence-structural properties. A comparative overview between DBCOVP and other platforms are presented in Table 1

Furthermore, since computational identification of antigenic epitopes require a complex analysis with a combination of several different tools and is a time-consuming and complex process. Therefore, to enable researchers to have a better understanding of the immunological properties and identify suitable vaccine candidates in the coronaviruses, we have mapped the potential conserved T-cell and B-cell epitopes on all the antigenic protein sequences along with information on the conservancy of the epitopes, potential immunogenicity, allergenicity, toxicity, and allergenicity analysis. Since HLA allele distribution differs among diverse geographic regions and ethnic groups around the world, population coverage analysis is an important factor in vaccine development. Thus, the cumulative percentage of population coverage across the world was estimated for the predicted epitopes and these results are freely available in the database. Besides, we determined the 3D structure of the epitopes and its binding interaction with the HLA molecules using in silico docking techniques. To our knowledge, DBCOVP is the first database with a special focus on SARS and MERS betacoronavirus virulence proteins containing detailed physicochemical, and structural information on the spike, envelope, membrane, and nucleocapsid

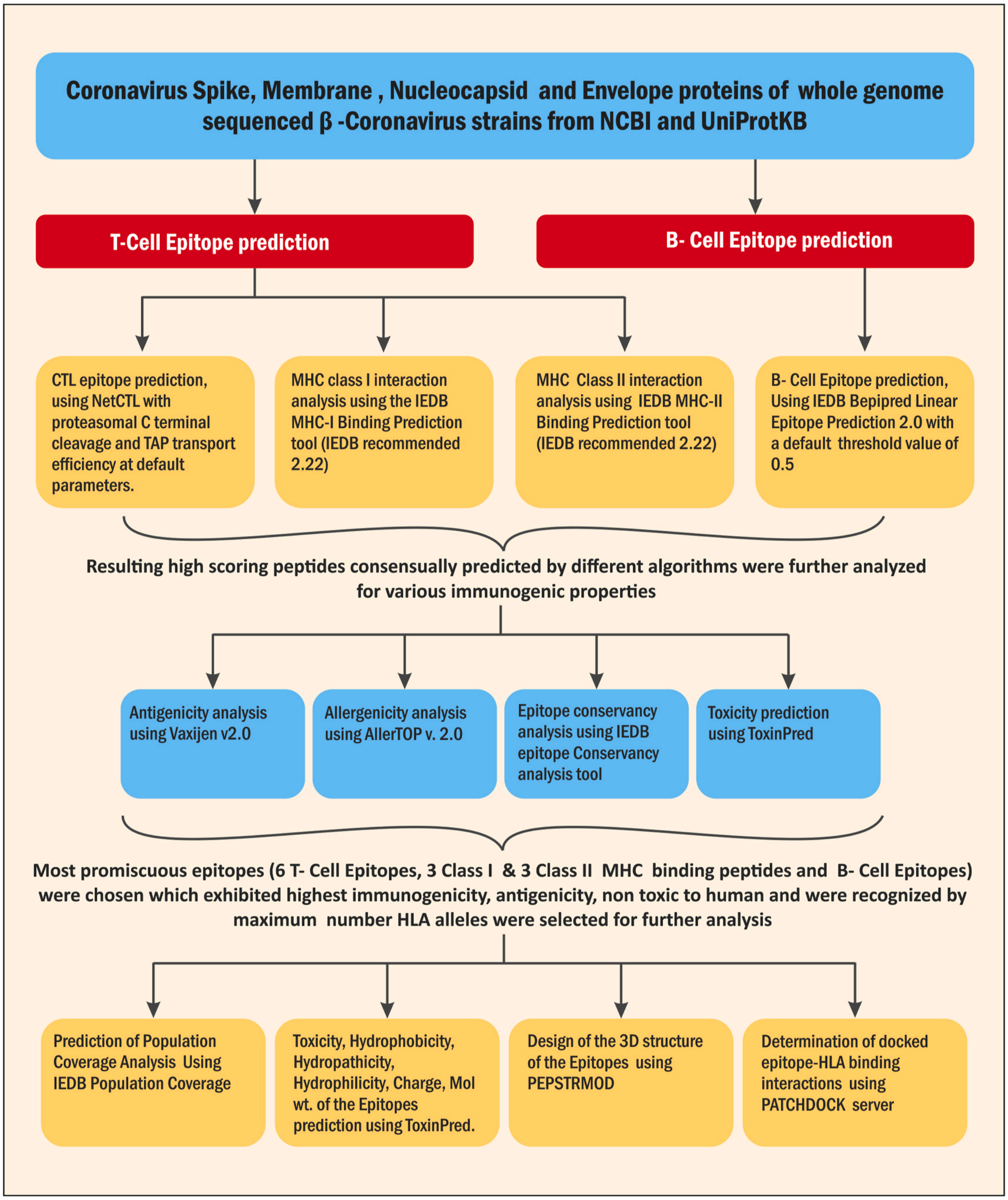

Coronavirus Spike, Membrane , Nucleocapsid and Envelope proteins of whole genome sequenced β -Coronavirus strains from NCBI and UniProtKB
T-Cell Epitope prediction
B- Cell Epitope prediction
CTL epitope prediction, using NetCTL with proteasomal C terminal cleavage and TAP transport efficiency at default parameters.
MHC class I interaction analysis using the IEDB MHC-I Binding Prediction tool (IEDB recommended 2.22)
MHC Class II interaction analysis using IEDB MHC-II Binding Prediction tool (IEDB recommended 2.22)
B- Cell Epitope prediction, Using IEDB Bepipred Linear Epitope Prediction 2.0 with a default threshold value of 0.5
Resulting high scoring peptides consensually predicted by different algorithms were further analyzed for various immunogenic properties
Antigenicity analysis using Vaxijen v2.0
Allergenicity analysis using AllerTOP v. 2.0
Epitope conservancy analysis using IEDB epitope Conservancy analysis tool
Toxicity prediction using ToxinPred
Most promiscuous epitopes (6 T- Cell Epitopes, 3 Class I & 3 Class II MHC binding peptides and B- Cell Epitopes) were chosen which exhibited highest immunogenicity, antigenicity, non toxic to human and were recognized by maximum number HLA alleles were selected for further analysis
Prediction of Population Coverage Analysis Using IEDB Population Coverage
Toxicity, Hydrophobicity, Hydropathicity, Hydrophilicity, Charge, Mol wt. of the Epitopes prediction using ToxinPred.
Design of the 3D structure of the Epitopes using PEPSTRMOD
Determination of docked epitope-HLA binding interactions using PATCHDOCK server

Fig. 1. Schematic representation of complete protocol employed for theidentification of promiscuous epitope-based vaccine candidates present in DBCOVP.

Table 1
Comparison of DBCOVP with the existing coronavirus web repositories.

	Covdb	*VipR*	*CoronaVIR*	*Covdb(CoronavirusDatabaseV3)*	*COVID – Profiler*	*Coviedb*	*Covidep*
URLrowhead	http://covdb.microbiology.hku.hk	https://www.viprbrc.org/brc/home.spg?Decorator=vipr	https://webs.iiitd.edu.in/raghava/coronavir/	http://covdb.popgenetics.net/v3/index	http://genomics.lshtm.ac.uk/	http://biopharm.zju.edu.cn/coviedb/	https://covidep.ust.hk/
Specificityrowhead	Includes annotated coronavirus genes and genomes belonging to six coronavirus species	ViPR contains information for human pathogenic viruses	Contains genomic, proteomic, diagnostic and therapeutic knowledge about novel SARS-CoV-2 coronaviruses	Contains coronavirus genomic data belonging to 32 organisms	Allows to analyze Sars-Cov-2 sequencing and immunological data	potential B/T cell epitopes for SARS CoV-2, RaTG13-CoV, SARS-CoV, and MERS-CoV	Consists genetic data for SARS-CoV-2 and immunological data for the 2003 SARS virus, to identify B-cell and T-cell epitopes
Strain Informationrowhead							
Description, Isolation Source, Collection Date, Host, Countryrowhead	×	Available	×	Available	×	×	×
Transmission, Epidemiology, Clinical symptomsrowhead	×	Available	×	×	×	×	×
Associated Glycoproteinrowhead	×	×	Available	×	×	×	×
Toolsrowhead							
Search and Advanced Searchrowhead	×	Available	×	×	×	×	×
BLASTrowhead	Available	Available	×	×	×	×	×
Phylogenyrowhead	×	Available	×	Available	×	×	×
Comparerowhead	×	×	×	×	×	×	×
MSArowhead	×	Available	×	Available	×	×	×
Covid-19 Trackerrowhead	×	×	×	×	×	×	×
Proteins Detailsrowhead							
Taxonomic lineagerowhead	×	Available	×	×	×	×	×
Subcellular locationrowhead	×	×	×	Available	×	×	×
Molecular Functionrowhead	×	×	×	×	×	×	×
Biological Processrowhead	×	×	×	×	×	×	×
Domainrowhead	×	Available	×	×	×	×	×
Clusterrowhead	×	×	×	×	×	×	×
Super familyrowhead	×	×	×	×	×	×	×
Protein Fasta Sequencerowhead	×	Available		×	×	×	×
Secondary Structure detailsrowhead	×	×	×	×	×	×	×
Disordered Regionrowhead	×	×	×	×	×	×	×
Disulfide Bondrowhead	×	×	×	×	×	×	×
Transmembrane Helixrowhead	×	Available	×	Available	×	×	×
Ubiquitination Siterowhead	×	×	×		×	×	×
Proteinase Clevage Sitesrowhead	Available	×	×	×	×	×	×
Internal repeatsrowhead	Available	×	×	×	×	×	×
3D protein structurerowhead	×	Available	12 protein structures are present	×	Available	×	×
Physicochemical propertiesrowhead	×	×	×	×	×	×	×
Epitope Details (MHC-I; MHC-II & B-Cell Epitope)rowhead	×	Available	Available	×	×	×	Available
Associated allelesrowhead	×	Available		×	×	×	×
Epitope Conservancyrowhead	×	×	×	×	×	×	×
Allerginicityrowhead	×	×	×	×	×	×	×
Antigenicityrowhead	×	×	×	×	×	×	×
Toxicityrowhead	×	×	×	×	×	×	×
Hydropathicityrowhead	×	×	×	×	×	×	×
Hydrophilicityrowhead	×	×	×	×	×	×	×
Chargerowhead	×	×	×	×	×	×	×
Molecular Weightrowhead	×	×	×	×	×	×	×
3D epitope structurerowhead	×	×	×	×	×	×	×
Population Coveragerowhead	×	×	×	×	×	×	×
Links to external databaserowhead	Available	Available	Available	×	×	×	×

Cov3d	*ViralZone*	*VADR*	*VBRC*	*Rfam*	*CORDITE*	*CoVex*	*DBCOVP*
https://cov3d.ibbr.umd.edu/cov3d	https://viralzone.expasy.org/	https://github.com/nawrocki/vadr	https://4virology.net/	https://rfam.org/covid-19	https://cordite.mathematik.unimarburg.de/#/	https://exbio.wzw.tum.de/covex/	http://covp.immt.res.in/
Includes Structures of SARS-CoV-2, SARS-CoV, and MERS-CoV proteins	Contains Covid-19 genome expression details, protein sequence records, host-virus interaction, and general information on coronaviruses belonging to betacoronavirus genera.	Designed to validate and annotate viral sequences	Validates and annotates viral sequences in GenBank submissions	providing access to families of structural RNAs	collects and aggregates details on in vitro, computational, or case analyses on promising drugs for COVID-19 from PubMed, bioRxiv, medRxiv.	Contains information on drug candidates and experimentally validated virus–human interaction data for both SARS-CoV-2 and SARS-CoV-1 with human interactome	The only database of structural glycoproteins from coronavirus genomes belonging to 137 strains from betacoronavirus genera.
Strain Informationrowhead							
×	x	x	x	X	x	x	Available
×	x	x	x	X	x	x	Available
×	x	x	x	X	x	x	Available
Toolsrowhead							
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
Proteins Detailsrowhead							
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
AVAILABLE	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	x	x	x	x	x	x	Available
×	Available	Available	Available	Available	Available	x	Available

protein sequences derived from 137 strains belonging to diverse host organisms. Most importantly, it is the only database to provide computed high-confidence complete immunological data of the coronavirus antigenic proteins in one platform. All the annotation data were manually curated from public databases and published literature but also computationally predicted using various bioinformatics tools and databases for complete functional annotation of each protein. Additionally, to facilitate further comparative data analysis, DBCOVP supports multiple search and browsing options, with integrated tools for multiple sequence alignment, phylogenetic tree construction, local BLAST alignment search, and in house developed compare tool for comparative genomic analysis. To promote its usability, '*Exclusive Entries for COVID-19*' has been included, which consists of proteomic, genomic, and immunoinformatics details of virulent glycoproteins specific to SARS-CoV-2. Moreover, DBCOVP maintains a 'Data Submission Form' that enables users to submit a protein sequence in FASTA format to proceed with the sequence-structure analysis. With the rapidly increasing global demand for the development of a vaccine against SARS-CoV-2, this database will certainly act as a one-stop resource for virologist and vaccinologists for understanding the pathogenesis of this epidemic disease and also for accelerating rational vaccine design by subsequent in vitro and in vivo experimental validation of the identified promiscuous vaccine targets.

2. Database contents and web interface

Currently, DBCOVP contains 185 proteins sequences including spike proteins (47), envelope proteins (43), membrane (46), and nucleocapsid proteins (49) in 137 strains originating from eight species (human, bat, murine, bovine, rat, rabbit, equine, hedgehog) across all the five subgenera of the betacoronavirus viz., *Sarbecovirus*, *Embecovirus*, *Hibecovirus*, *Merbecovirus*, and *Nobecovirus*. Sequences were collected from the National Centre for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) and UniProt Knowledgebase (UniProtKB; https://www.uniprot.org/).

All backend data are organized into a set of relational tables in a SQL server database. Stored procedures were implemented to improve the scalability and efficiency of the database. The graphical interface was developed using HTML5.0, ASP.Net(C#), CSS 3.0, JavaScript, and AJAX to obtain a rich user experience. DBCOVP provides user-friendly browsing, searching and data download functionalities which are made highly interactive to facilitate data extraction on each coronavirus virulence proteins. The "Search" option on the homepage enables users to search for information easily by a variety of keywords, including host species, proteins, and subgenus. Also, an "Advanced" search is provided on the search page for more specific requirements where users could obtain desired information by entering multiple combinations of keywords (Fig. 2a) with AJAX-driven auto-suggestions for users. The database can be browsed by visiting the 'Browse' tab either from the navigation menu or home page, where multiple options such as Browse by Host species, Proteins, and Epitopes are available to retrieve results (Fig. 2b). Selecting any of the strains in the list will bring up the corresponding strains details page containing information on taxonomic lineage, genome size, etc as shown in Fig. 2c. The virus-strains and the encoded protein sequences present in the database are also classified based on sequence similarity and phylogeny into the five subgenera of *Betacoronavirus* including *Sarbecovirus*, *Embecovirus*, *Hibecovirus*, *Merbecovirus*, and *Nobecovirus* which can be viewed by clicking the Browse by "Subgenus" option (Fig. 2d). If users want to view the detailed information of any particular protein occurring in the search results, they can click the corresponding UniProt Id or the hyperlink named 'READ MORE', which links to the detailed annotation page of that protein (Fig. 2e). Also, users can search for both T-cell and B-cell Epitopes present in the database by first selecting the epitope type, and then either one or multiple proteins from the four categories namely spike, envelope, membrane, and nucleocapsid, and finally selecting the host strain. The resulting page will contain detailed immunogenic information of the desired protein sequence as shown in Fig. 3. Furthermore, as a publicly released scientific database, the full dataset of DBCOVP is available for batch download in several forms including the FASTA Sequences and tabular (Excel) files.

3. Database content and annotation features

Each protein entry in the database has five important annotation components as discussed below. The detailed annotation has been manually predicted using various tools and databases as described in Supplementary Table 1.

a. Summary: This section presents general information about the protein sequence as retrieved from UniProt and NCBI GenBank like accession ID, strain name, host species, taxonomic lineage, subcellular localization, genomic location, Gene ontology, Pfam domain, family description, cross-referenced links to external databases like NCBI, KEGG, UniProt and protein and nucleotide sequence in fast format (Fig. 4a).

b. Structural Details: This includes the secondary and tertiary structure details of each protein stating the no of helices, beta-sheet, and turns, predicted disorder region, disulfide bond position, transmembrane helices, presence of signal peptides and cleavage sites, ubiquitination Site details, the position of repeat sequences, and crystal structure of protein sequences if available in the protein data bank. Users can view the 3D structure with the help of the Jmol program integrated and can also download the structure in PDB format (Fig. 4b).

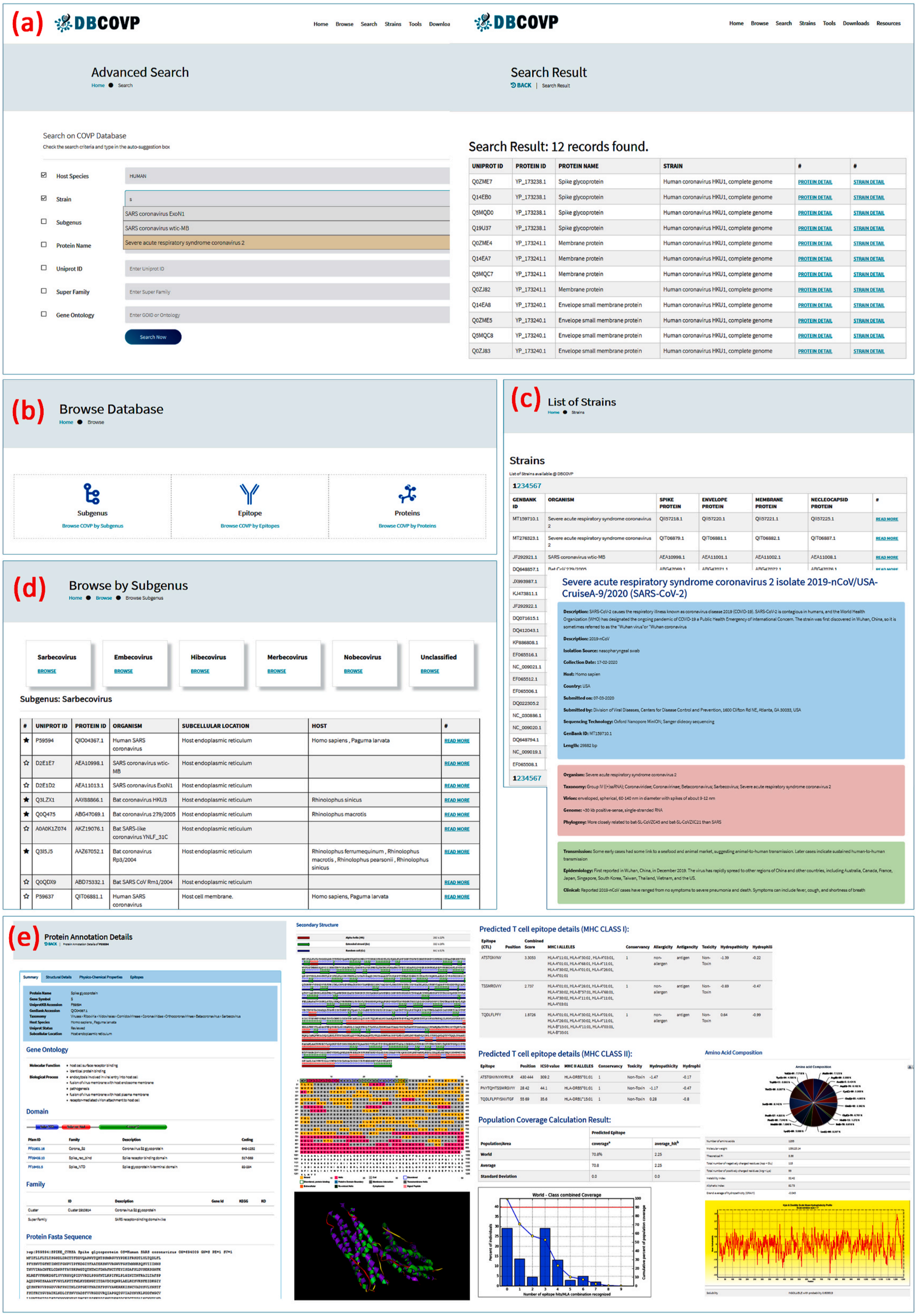
c. Physicochemical properties: physicochemical properties of proteins comprising of pI, no of positive/negatively charged amino acids, instability index, aliphatic index, GRAVY, hydropathy plot, and solubility (Fig. 4c).

d. Epitopes: Each spike, membrane, envelope, and nucleocapsid protein sequences were analyzed to identify the highest immunogenic, and antigenic T-cell epitopes along with B-cell epitopes. We have also predicted the binding Class I and Class II HLA alleles, conservancy score, allergenicity, antigenicity, toxicity, hydropathicity, hydrophilicity, charge, molecular weight of the predicted peptides. In addition, the population coverage analysis of the promiscuous epitopes is also available in the database. Furthermore, the 3D structure of the epitopes along with the docked complex of the epitope and binding HLA have been developed and users can also download the structures for further analysis. The detailed immunogenic results obtained for epitope analysis is described in the next section (Fig. 4d).

4. Immunoinformatics data

The immunoinformatics data are organized in the database for easy analysis and retrieval. Users can retrieve these resources either from the annotation detail page of each protein entry or by clicking the epitope option from the browse section of the homepage of the database. Each of the 185 sequences encompassing coronavirus spike, membrane, envelope, and nucleocapsid proteins were analyzed with several Immunoinformatics algorithms and tools displayed in Supplementary Table 2. The complete protocol employed for the identification of promiscuous epitope-based vaccine candidates is shown in Fig. 1.

For each protein sequence, most promiscuous T-cell epitopes and B-cell epitopes were selected which were recognized by a considerable number of HLA alleles and contained the highest immunogenicity, antigenicity value, and were nontoxic to human and hence, considered as the most potential epitopes to induce a strong immune response. Furthermore, the epitopes were selected based on the consensus matching results of all the employed tools. HLA allele distribution differs among diverse geographic regions and ethnic groups around the world. Therefore, population coverage analysis of the epitopes is a very important factor that must be taken into consideration during the development of an effective vaccine. Therefore, for all the predicted


(a)
DBCOVP
Home Browse Search Strains Tools Downloads Resources
(b)
(c)
(d)
(e)
Secondary Structure
Alpha helix (Hh)
Extended strand (Ee)
Random coil (Cc)
Amino Acid Composition
World - Class combined Coverage
Percent of individuals
Cumulative percent of population coverage
Number of epitope hits/HLA combination recognized

Advanced Search
Home ● Search

Search on COVP Database
Check the search criteria and type in the auto-suggestion box

☑	Host Species	HUMAN
☑	Strain	s SARS coronavirus ExoN1 SARS coronavirus wtic-MB Severe acute respiratory syndrome coronavirus 2
☐	Subgenus	
☐	Protein Name	
☐	Uniprot ID	Enter Uniprot ID
☐	Super Family	Enter Super Family
☐	Gene Ontology	Enter GOID or Ontology

Search Now

Search Result
BACK | Search Result

Search Result: 12 records found.

UNIPROT ID	PROTEIN ID	PROTEIN NAME	STRAIN	#	#
Q0ZME7	YP_173238.1	Spike glycoprotein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q14EB0	YP_173238.1	Spike glycoprotein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q5MQD0	YP_173238.1	Spike glycoprotein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q19U37	YP_173238.1	Spike glycoprotein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q0ZME4	YP_173241.1	Membrane protein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q14EA7	YP_173241.1	Membrane protein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q5MQC7	YP_173241.1	Membrane protein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q0ZJ82	YP_173241.1	Membrane protein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q14EA8	YP_173240.1	Envelope small membrane protein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q0ZME5	YP_173240.1	Envelope small membrane protein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q5MQC8	YP_173240.1	Envelope small membrane protein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL
Q0ZJ83	YP_173240.1	Envelope small membrane protein	Human coronavirus HKU1, complete genome	PROTEIN DETAIL	STRAIN DETAIL

Browse Database
Home ● Browse

Subgenus
Browse COVP by Subgenus

Epitope
Browse COVP by Epitopes

Proteins
Browse COVP by Proteins

List of Strains
Home ● Strains

Strains
List of Strains available @ DBCOVP

1234567

GENBANK ID	ORGANISM	SPIKE PROTEIN	ENVELOPE PROTEIN	MEMBRANE PROTEIN	NECLEOCAPSID PROTEIN	#
MT159710.1	Severe acute respiratory syndrome coronavirus 2	QII57218.1	QII57220.1	QII57221.1	QII57225.1	READ MORE
MT276323.1	Severe acute respiratory syndrome coronavirus 2	QIT06879.1	QIT06881.1	QIT06882.1	QIT06887.1	READ MORE
JF292921.1	SARS coronavirus wtic-MB	AEA10998.1	AEA11001.1	AEA11002.1	AEA11008.1	READ MORE
DQ648857.1	Bat CoV 279/2005	ABG47069.1	ABG47071.1	ABG47072.1	ABG47076.1	READ MORE
JX993987.1						
KJ473811.1						
JF292922.1						
DQ071615.1						
DQ412043.1						
KP886808.1						
EF065516.1						
NC_009021.1						
EF065512.1						
EF065506.1						
DQ022305.2						
NC_030886.1						
NC_009020.1						
DQ648794.1						
NC_009019.1						
EF065508.1						

1234567

Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-CruiseA-9/2020 (SARS-CoV-2)

Description: SARS-CoV-2 causes the respiratory illness known as coronavirus disease 2019 (COVID-19). SARS-CoV-2 is contagious in humans, and the World Health Organization (WHO) has designated the ongoing pandemic of COVID-19 a Public Health Emergency of International Concern. The strain was first discovered in Wuhan, China, so it is sometimes referred to as the "Wuhan virus"or "Wuhan coronavirus

Description: 2019-nCoV

Isolation Source: nasopharyngeal swab

Collection Date: 17-02-2020

Host: Homo sapien

Country: USA

Submitted on: 07-03-2020

Submitted by: Division of Viral Diseases, Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta, GA 30033, USA

Sequencing Technology: Oxford Nanopore MinION; Sanger dideoxy sequencing

GenBank ID: MT159710.1

Length: 29882 bp

Organism: Severe acute respiratory syndrome coronavirus 2

Taxonomy: Group IV ((+)ssRNA); Coronaviridae; Coronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome coronavirus 2

Virion: enveloped, spherical, 60-140 nm in diameter with spikes of about 9-12 nm

Genome: ~30 kb positive-sense, single-stranded RNA

Phylogeny: More closely related to bat-SL-CoVZC45 and bat-SL-CoVZXC21 than SARS

Transmission: Some early cases had some link to a seafood and animal market, suggesting animal-to-human transmission. Later cases indicate sustained human-to-human transmission

Epidemiology: First reported in Wuhan, China, in December 2019. The virus has rapidly spread to other regions of China and other countries, including Australia, Canada, France, Japan, Singapore, South Korea, Taiwan, Thailand, Vietnam, and the US.

Clinical: Reported 2019-nCoV cases have ranged from no symptoms to severe pneumonia and death. Symptoms can include fever, cough, and shortness of breath

Browse by Subgenus
Home ● Browse ● Browse Subgenus

Sarbecovirus BROWSE
Embecovirus BROWSE
Hibecovirus BROWSE
Merbecovirus BROWSE
Nobecovirus BROWSE
Unclassified BROWSE

Subgenus: Sarbecovirus

#	UNIPROT ID	PROTEIN ID	ORGANISM	SUBCELLULAR LOCATION	HOST	#
★	P59594	QIO04367.1	Human SARS coronavirus	Host endoplasmic reticulum	Homo sapiens , Paguma larvata	READ MORE
☆	D2E1E7	AEA10998.1	SARS coronavirus wtic-MB	Host endoplasmic reticulum		READ MORE
☆	D2E1D2	AEA11013.1	SARS coronavirus ExoN1	Host endoplasmic reticulum		READ MORE
★	Q3LZX1	AAY88866.1	Bat coronavirus HKU3	Host endoplasmic reticulum	Rhinolophus sinicus	READ MORE
★	Q0Q475	ABG47069.1	Bat coronavirus 279/2005	Host endoplasmic reticulum	Rhinolophus macrotis	READ MORE
☆	A0A0K1Z074	AKZ19076.1	Bat SARS-like coronavirus YNLF_31C	Host endoplasmic reticulum		READ MORE
★	Q3I5J5	AAZ67052.1	Bat coronavirus Rp3/2004	Host endoplasmic reticulum	Rhinolophus ferrumequinum , Rhinolophus macrotis , Rhinolophus pearsonii , Rhinolophus sinicus	READ MORE
☆	Q0QDX9	ABD75332.1	Bat SARS CoV Rm1/2004	Host endoplasmic reticulum		READ MORE
☆	P59637	QIT06881.1	Human SARS coronavirus	Host cell membrane.	Homo sapiens, Paguma larvata	READ MORE

Protein Annotation Details
BACK | Protein Annotation Details of P59594

Summary | Structural Details | Physico-Chemical Properties | Epitopes

Protein Name	Spike glycoprotein
Gene Symbol	S
UniprotKB Accession	P59594
GenBank Accession	QIO04367.1
Taxonomy	Viruses › Riboviria › Nidovirales › Cornidovirineae › Coronaviridae › Orthocoronavirinae › Betacoronavirus › Sarbecovirus
Host Species	Homo sapiens , Paguma larvata
Uniprot Status	Reviewed
Subcellular Location	Host endoplasmic reticulum

Gene Ontology

Molecular Function
- host cell surface receptor binding
- identical protein binding

Biological Process
- endocytosis involved in viral entry into host cell
- fusion of virus membrane with host endosome membrane
- pathogenesis
- fusion of virus membrane with host plasma membrane
- receptor-mediated virion attachment to host cell

Domain

Pfam ID	Family	Description	Coding
PF01601.16	Corona_S2	Coronavirus S2 glycoprotein	648-1252
PF09408.10	Spike_rec_bind	Spike receptor binding domain	317-569
PF16451.5	Spike_NTD	Spike glycoprotein N-terminal domain	82-284

Family

	ID	Description	Gene Id	KEGG	KO
Cluster	Cluster 2910614	Coronavirus S2 glycoprotein			
Super Family		SARS receptor-binding domain-like			

Protein Fasta Sequence

>sp|P59594|SPIKE_CVHSA Spike glycoprotein OS=Human SARS coronavirus OX=694009 GN=S PE=1 SV=1

Predicted T cell epitope details (MHC CLASS I):

Epitope (CTL)	Position	Combined Score	MHC I ALLELES	Conservancy	Allergicity	Antigencity	Toxicity	Hydropathicity	Hydrophili
ATSTGNYNY		3.3053	HLA-A*11:01, HLA-A*30:02 , HLA-A*03:01, HLA-A*01:01, HLA-A*68:01, HLA-A*11:01, HLA-A*30:02, HLA-A*01:01, HLA-A*26:01, HLA-A*01:01	1	non-allergen	antigen	Non-Toxin	-1.39	-0.22
TSSMRGVYY		2.737	HLA-A*01:01, HLA-A*26:01, HLA-A*01:01, HLA-A*30:02, HLA-B*57:01, HLA-A*68:01, HLA-A*30:02, HLA-A*11:01, HLA-A*11:01, HLA-A*03:01	1	non-allergen	antigen	Non-Toxin	-0.69	-0.47
TQDLFLPFY		1.8726	HLA-A*01:01, HLA-A*30:02, HLA-A*01:01, HLA-A*26:01, HLA-A*30:02, HLA-A*11:01, HLA-B*15:01, HLA-A*11:01, HLA-A*03:01, HLA-B*35:01	1	non-allergen	antigen	Non-Toxin	0.64	-0.99

Predicted T cell epitope details (MHC CLASS II):

Epitope	Position	IC50 value	MHC II ALLELES	Conservancy	Toxicity	Hydropathicity	Hydrophi
ATSTGNYNYKYRYLR	430 444	309.2	HLA-DRB5*01:01	1	Non-Toxin	-1.47	-0.17
PNYTQHTSSMRGVYY	28 42	44.1	HLA-DRB5*01:01	1	Non-Toxin	-1.17	-0.47
TQDLFLPFYSNVTGF	55 69	35.6	HLA-DRB1*15:01	1	Non-Toxin	0.28	-0.8

Population Coverage Calculation Result:

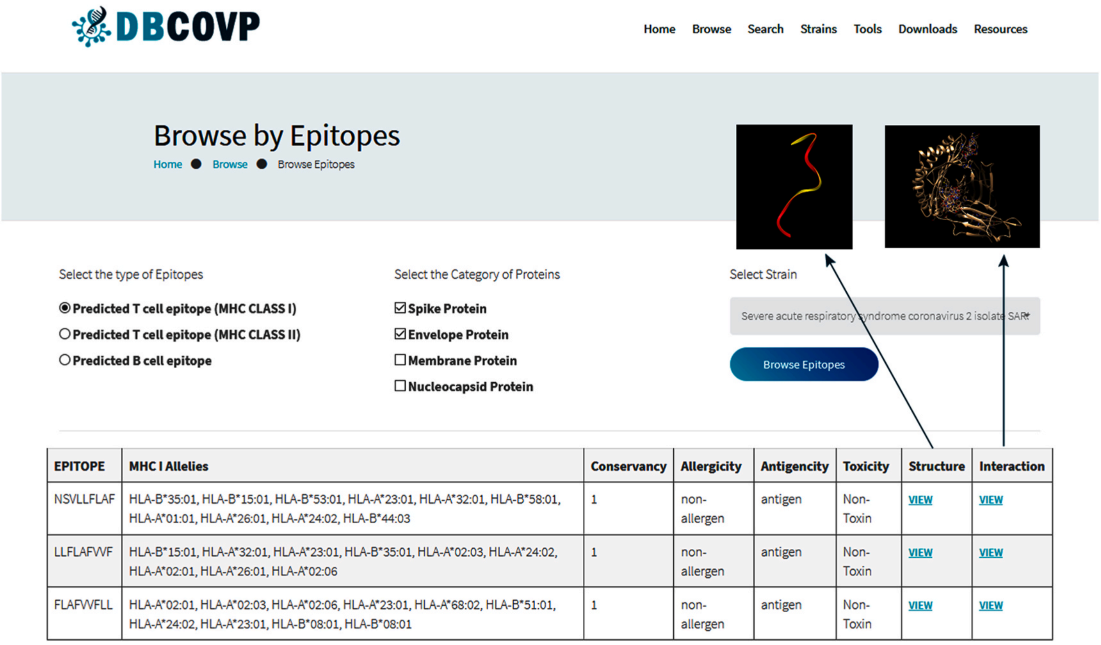
	Predicted Epitope	
Population/Area	coverage^{a}	average_hit^{b}
World	70.8%	2.25
Average	70.8	2.25
Standard Deviation	0.0	0.0

Amino Acid Composition

Number of amino acids	1255
Molecular weight	139125.14
Theoretical PI	5.56
Total number of negatively charged residues (Asp + Glu)	115
Total number of positively charged residues (Arg + Lys)	99
Instability index	32.42
Aliphatic index	82.73
Grand average of hydropathicity (GRAVY)	-0.045

Fig. 2. Screenshot of DBCOVP Web-interface: a) 'Advanced Search' provides users to input multiple search queries simultaneously to retrieve specific proteins of interest. b) Browse by 'Subgenus', 'Epitope' and 'Proteins'. c) Details page for a coronavirus strain d) Result page from Browse by 'Subgenus'. e) The detailed annotation page of a protein.

epitopes, the cumulative percentage of population coverage across the world was measured and the results are displayed in a graphical format as shown in Fig. 2d. The results indicate that all the predicted epitopes and their binding HLA alleles covered more than 80% of the world's population, which is a very important factor for a vaccine candidate since the emerging SARS-CoV-2 strain has affected the human population across the world. Besides, the three-dimensional structure of each of the predicted epitopes was determined and the binding interaction with



DBCOVP

Home Browse Search Strains Tools Downloads Resources

Browse by Epitopes

Home ● Browse ● Browse Epitopes

Select the type of Epitopes

● Predicted T cell epitope (MHC CLASS I)
○ Predicted T cell epitope (MHC CLASS II)
○ Predicted B cell epitope

Select the Category of Proteins

☑ Spike Protein
☑ Envelope Protein
☐ Membrane Protein
☐ Nucleocapsid Protein

Select Strain

Severe acute respiratory syndrome coronavirus 2 isolate SARS

Browse Epitopes

EPITOPE	MHC I Allelies	Conservancy	Allergicity	Antigencity	Toxicity	Structure	Interaction
NSVLLFLAF	HLA-B*35:01, HLA-B*15:01, HLA-B*53:01, HLA-A*23:01, HLA-A*32:01, HLA-B*58:01, HLA-A*01:01, HLA-A*26:01, HLA-A*24:02, HLA-B*44:03	1	non-allergen	antigen	Non-Toxin	VIEW	VIEW
LLFLAFVVF	HLA-B*15:01, HLA-A*32:01, HLA-A*23:01, HLA-B*35:01, HLA-A*02:03, HLA-A*24:02, HLA-A*02:01, HLA-A*26:01, HLA-A*02:06	1	non-allergen	antigen	Non-Toxin	VIEW	VIEW
FLAFVVFLL	HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*23:01, HLA-A*68:02, HLA-B*51:01, HLA-A*24:02, HLA-A*23:01, HLA-B*08:01, HLA-B*08:01	1	non-allergen	antigen	Non-Toxin	VIEW	VIEW

Fig. 3. Detailed immunogenic information obtained for proteins using Browse by Epitope.

the most conserved HLA allele was studied using the docking technique. The PDB structures are available for download. The ribbon representation of the structures was prepared and visualized by the PyMOL molecular graphics system.

5. Integrated tools

To facilitate further in-depth analysis of virulence proteins from coronavirus, four analysis tools have been integrated. Sequence similarity search of both nucleotide and amino acid sequences can be performed using the basic local alignment search tool (BLAST) algorithm through an integrated Blast module within the database. The BLAST interface allows alignment of a user-provided sequence against a customized BLAST library containing all sequences present in the DBCOVP database. This helps to identify the sequence similarity of any unknown sequence to known annotated proteins. The user may specify BLAST parameters and upload or paste the query sequences. The output is given in the standard format with the blast score and ordered by ascending e-value. Each hit is hyperlinked to that entry's browser page. As the analysis of variability of virulence proteins is important for understanding the emergence of novel strains and to decipher sequence level variations leading to changes in pathogenicity, therefore to facilitate cross-genome comparative analysis a COMPARE Tool has been integrated by which users can analyze the variations in targeted sequences across multiple strains belonging to same or different host species. Additionally, multiple sequence alignment and phylogenetic tree can be constructed using embedded MUSCLE tool and PhyML tool, respectively in the database.

6. Discussion and future directions

The COVID-19 pandemic has resulted in an exponential increase in

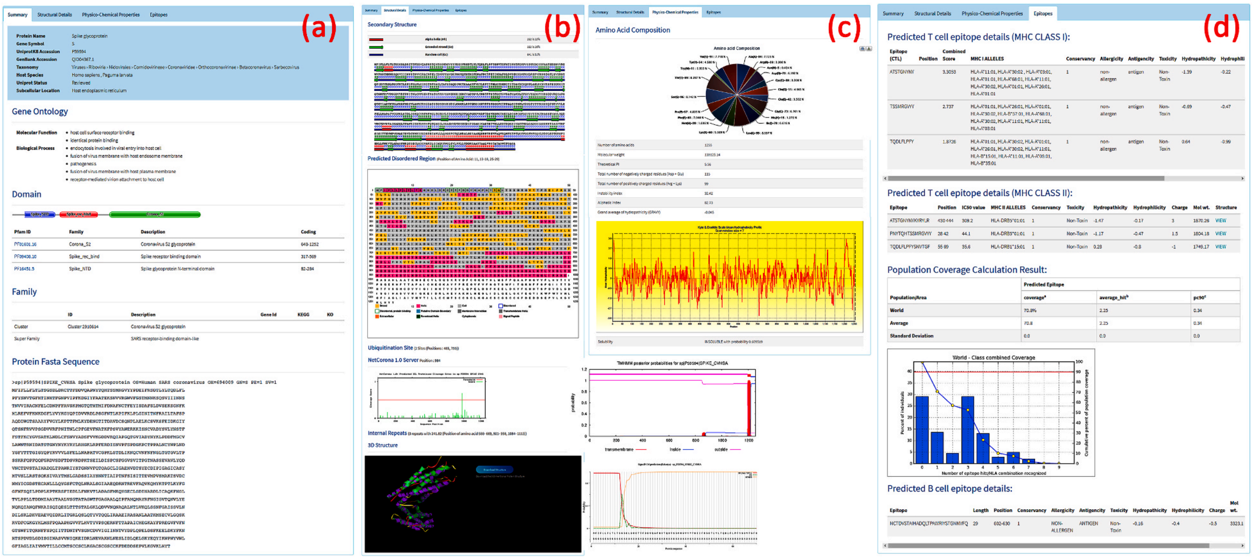

(a)
Summary
Structural Details
Physico-Chemical Properties
Epitopes
Gene Ontology
Domain
Family
Protein Fasta Sequence
(b)
Secondary Structure
Predicted Disordered Region
Ubiquitination Site
NetCorona 1.0 Server
Internal Repeats
3D Structure
(c)
Amino Acid Composition
(d)
Predicted T cell epitope details (MHC CLASS I):
Predicted T cell epitope details (MHC CLASS II):
Population Coverage Calculation Result:
World - Class combined Coverage
Predicted B cell epitope details:

Fig. 4. Schematic representation of database content and annotation features: a) *Summary* tab of protein annotation page. b) *Structural Details* tab of protein annotation page. c) *Physicochemical properties* tab of protein annotation page. d) *Epitopes* tab of protein annotation page.

the number of novel SARS-CoV-2 coronaviruses genomes being sequenced. Therefore, computational methods and databases are needed to organize, explore and analyze large volumes of the biological data to aid in understanding the mechanisms of disease pathogenesis and, most importantly, to speed up the vaccines development process by providing adequate information on the efficacy and immunogenicity of potential molecular targets critical for subsequent clinical validation. Increasing studies have shown that the four major structural glycoproteins namely spike protein, envelope protein, membrane protein and nucleocapsid protein have important functions and play vital roles in viral infection and particularly spike protein has been shown to elicit T-cell responses suggesting as potential vaccine candidates against SARS infection [39].

In this study, we developed the DBCOVP, the first manually curated database to provide comprehensive information on the entire repertoire of structural glycoproteins from coronavirus genomes of betacoronavirus genera including the newly sequenced SARS-CoV-2 strains which are majorly responsible for the atypical severe acute respiratory syndrome. As compared to few existing databases on coronaviruses research, DBCOVP is a specialized database focussed on coronavirus spike, envelope, membrane, and nucleocapsid proteins and excels in the following aspects: (i) Substantially extended data volume consisting of a total of 185 structural proteins from 137 strains including sequences from the recently deposited SARS-CoV-2 strains in NCBI. (ii) Complete functional annotation of the proteins highlighting 14 sequence-structural properties which are partially addressed in some of the existing coronavirus sequence data resources. Basic information about each protein includes manually curated information from known databases while more specific and source-dependent annotation features have been computationally predicted using various bioinformatics tools and methods. (iii) The major purpose of the database is to enable users to perform knowledge discovery from coronavirus antigen data with particular emphasis on applications in immunology and vaccinology. Each spike, membrane, envelope, and nucleocapsid protein sequences have been mapped to highlight the most promiscuous epitopic regions (T-cell and B-cell) along with conservancy score, allergenicity, antigenicity, toxicity, hydropathicity, hydrophilicity, charge, molecular weight, and population coverage analysis of the predicted peptides. In addition, the 3D structure of the epitopes along with the docked epitope-HLA binding complex is available for further analysis. This is the first database containing the aforementioned immunogenic data specific for coronavirus virulent glycoproteins on one single platform. (iv) Multiple searches and browse options to facilitate data extraction. (v) Links to resources pertinent to coronavirus research. (vi) DBCOVP provides a user-friendly interface, incorporating an application for BLAST similarity search and integrating many useful tools for cross genome comparison, phylogenetic, and multiple sequence alignment to facilitate further studies on structural glycoproteins and their functional role in virulence.

Research on viable therapeutics and vaccine targets against human coronavirus infection is probably only beginning to unfold. In the future, we will continue to update the database and include sequences from other coronavirus strains as well as with more valuable resources constantly integrated into the database. Furthermore, we will also try to combine all the complex steps and tools employed in this study for epitope analysis into one automated tool which would be particularly useful for researchers with little knowledge in bioinformatics to rapidly analyze the immunogenic properties of uncharacterized sequences in one platform without moving data between different analysis tools. DBCOVP will certainly be an important resource when prioritizing vaccine candidates against coronavirus infection.

Declaration of competing interest

Authors declare there is no conflict of interest.

Acknowledgments

The authors acknowledge the School of Biotechnology, Kalinga Institute of Industrial Technology (KIIT), Deemed to be University, Bhubaneswar 751024, India for providing the necessary infrastructure to carry out this work. The authors also acknowledge Informatics Lab, CSIR-Institute of Minerals and Materials Technology (CSIR-IMMT), Bhubaneswar-751013, India for providing the computational facilities and hosting the database.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2020.104131.

Author contributions

Conception and design: MS, NM; Computational work: SS, SM; Data Analysis and Curation: NM, BD, VR; Original Draft Preparation: NM; Writing- Reviewing and Editing; MS, VR, SS. The manuscript has been read and approved by all authors.

Funding statement

The author received no funding from an external source.

References

[1] Y.S. Malik, S. Sircara, S. Bhata, K. Sharunb, K. Dhamac, M. Dadard, et al., Emerging novel coronavirus (SARS-CoV-2)—current scenario, evolutionary perspective based on genome analysis and recent developments, Vet. Q. 40 (2020) 68–76.

[2] S. Su, G. Wong, W. Shi, J. Liu, A.C.K. Lai, J. Zhou, et al., Epidemiology, genetic recombination, and pathogenesis of coronaviruses, Trends Microbiol. 24 (2016) 490–502.

[3] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Lancet 395 (2020) 565–574.

[4] J.S. Peiris, Y. Guan, K.Y. Yuen, Severe acute respiratory syndrome, Nat. Med. 10 (suppl 12) (2004) S88–S97.

[5] M. Chan-Yeung, R.H. Xu, SARS: epidemiology, Respirology 8 (suppl) (2003) S9–S14.

[6] A.M. Zaki, S. van Boheemen, T.M. Bestebroer, A.D. Osterhaus, R.A. Fouchier, Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia, N. Engl. J. Med. 367 (2012) 1814–1820.

[7] J. Lee, G. Chowell, E. Jung, A dynamic compartmental model for the Middle East respiratory syndrome outbreak in the Republic of Korea: a retrospective analysis on control interventions and superspreading events, J. Theor. Biol. 408 (2016) 118–126.

[8] J.Y. Lee, Y.J. Kim, E.H. Chung, et al., The clinical and virological features of the first imported case causing MERS-CoV outbreak in South Korea, 2015, BMC Infect. Dis. 17 (2017) 498.

[9] F. Wu, S. Zhao, B. Yu, Y.M. Chen, W. Wang, Z.G. Song, et al., A new coronavirus associated with human respiratory disease in China, Nature 580 (2020) E7.

[10] J.Y. Li, Z. You, Q. Wang, Z.J. Zhou, Y. Qiu, R. Luo, et al., The epidemic of 2019-novel-coronavirus (SARS-CoV-2) pneumonia and insights for emerging infectious diseases in the future, Microb. Infect. 22 (2020) 80–85.

[11] P. Zhou, X.L. Yang, X.G. Wang, B. Hu, L. Zhang, W. Zhang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273.

[12] J. Liu, X. Zheng, Q. Tong, W. Li, B. Wang, K. Sutter, et al., Overlapping and discrete aspects of the pathology and pathogenesis of the emerging human pathogenic coronaviruses SARS-CoV, MERS-CoV, and SARS-CoV-2, J. Med. Virol. 92 (2020) 491–494.

[13] G. Li, Y. Fan, Y. Lai, T. Han, Z. Li, P. Zhou, et al., Coronavirus infections and immune responses, J. Med. Virol. 92 (2020) 424–432.

[14] P.C.Y. Woo, Y. Huang, S.K.P. Lau, K.Y. Yuen, Coronavirus genomics and bioinformatics analysis, Viruses 2 (2010) 1804–1820.

[15] D.N. Bourdette, E. Edmonds, C. Smith, et al., A highly immunogenic trivalent T cell receptor peptide vaccine for multiple sclerosis, Mult. Scler. 11 (2005) 552–561.

[16] J.A. Lopez, C. Weilenman, R. Audran, et al., A synthetic malaria vaccine elicits a potent CD8(+) and CD4(+) T lymphocyte immune response in humans. Implications for vaccination strategies, Eur. J. Immunol. 31 (2001) 1989–1998.

[17] K.L. Knutson, K. Schiffman, M.L. Disis, Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T-cell immunity in cancer patients, J. Clin. Invest. 107 (2001) 477–484.

[18] N. Banerjee, S. Mukhopadhyay, Viral glycoproteins: biological role and application in diagnosis, Virus Dis 27 (2016) 1–11.

[19] F. Li, Structure, function, and evolution of coronavirus spike proteins, Ann. Rev. Virol. 3 (2016) 237–261.

[20] G. Lu, Q. Wang, G.F. Gao, Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond, Trends Microbiol. 23 (2015) 468–478.

[21] Q. Wang, G. Wong, G. Lu, J. Yan, G.F. Gao, MERS-CoV spike protein: targets for vaccines and therapeutics, Antivir. Res. 133 (2016) 165–177.

[22] M.A. Marra, S.J. Jones, C.R. Astell, R.A. Holt, A. Brooks-Wilson, Y.S. Butterfield, et al., The Genome sequence of the SARS-associated coronavirus, Science 300 (2003) 1399–1404.

[23] S. Jiang, Y. He, S. Liu, SARS vaccine development, Emerg. Infect. Dis. 11 (2005) 1016–1020.

[24] G. Salvatori, L. Luberto, M. Maffei, L. Aurisicchio, G. Roscilli, F. Palombo, E. Marra, SARS-CoV-2 SPIKE PROTEIN: an optimal immunological target for vaccines, J. Transl. Med. 18 (2020 Dec) 1–3.

[25] D. Schoeman, B.C. Fielding, Coronavirus envelope protein: current knowledge, Virol. J. 16 (2019) 69.

[26] T.R. Ruch, C.E. Machamer, The coronavirus E protein: assembly and beyond, Viruses 4 (2012) 363–382.

[27] Y. Huang, S.K. Lau, P.C. Woo, K.Y. Yuen, CoVDB: a comprehensive database for comparative analysis of coronavirus genes and genomes, Nucleic Acids Res. 36 (2008) D504–D511.

[28] B.E. Pickett, D.S. Greer, Y. Zhang, L. Stewart, L. Zhou, G. Sun, et al., Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community, Viruses 4 (2012) 3209–3226.

[29] C. Hulo, E. De Castro, P. Masson, L. Bougueleret, A. Bairoch, I. Xenarios, P. Le Mercier, ViralZone: a knowledge resource to understand virus diversity, Nucleic Acids Res. 39 (suppl_1) (2011) D576–D582.

[30] A.A. Schäffer, E.L. Hatcher, L. Yankie, L. Shonkwiler, J.R. Brister, I. Karsch-Mizrachi, E.P. Nawrocki, VADR: validation and annotation of virus sequence submissions to GenBank, BMC Bioinf. 21 (2020) 1–23.

[31] D. Amgarten, C. Upton, Bioinformatic approaches for comparative analysis of viruses, InComparative Genomics (2018) 401–417. Humana Press, New York, NY.

[32] I. Kalvari, J. Argasinska, N. Quinones-Olvera, E.P. Nawrocki, E. Rivas, S.R. Eddy, A. Bateman, R.D. Finn, A.I. Petrov, Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families, Nucleic Acids Res. 46 (D1) (2018) D335–D342.

[33] R. Martin, H.F. Löchel, M. Welzel, G. Hattab, A.C. Hauschild, D. Heider, CORDITE: the curated CORona drug InTERactions database for SARS-CoV-2, Iscience 23 (7) (2020) 101297.

[34] S. Sadegh, J. Matschinske, D.B. Blumenthal, G. Galindez, T. Kacprowski, M. List, R. Nasirigerdeh, M. Oubounyt, A. Pichlmair, T.D. Rose, M. Salgado-Albarrán, Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing, Nat. Commun. 11 (2020). Article number: 3518.

[35] Zhenglin Zhu, Kaiwen Meng, Meng Geng, A database resource for Genome-wide dynamics analysis of Coronaviruses on a historical and global scale. https://doi.org/10.1101/2020.02.05.920009, 2020.

[36] J. Wu, W. Chen, J. Zhou, W. Zhao, S. Chen, Z.* Zhou, COVIEdb : a database for potential immune epitopes of coronaviruses, bioRxiv vol. 5 (2020), 096164, https://doi.org/10.1101/2020.05.24.096164.

[37] S.F. Ahmed, A.A. Quadeer, M.R. McKay, COVIDep: a web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2, Nat. Protoc. (2020).

[38] S. Patiyal, D. Kaur, H. Kaur, N. Sharma, A. Dhall, S. Sahai, et al., A web-based platform on COVID-19 to maintain Predicted Diagnostic, Drug and Vaccine candidates, OSF Preprints (2020), https://doi.org/10.31219/osf.io/xegzu.

[39] J. Huang, Y. Cao, J. Du, X. Bu, R. Ma, C. Wu, Priming with SARS CoV S DNA and boosting with SARS CoV S epitopes specific for CD4+ and CD8+ T cells promote cellular immune responses, Vaccine 25 (2007) 6981–6991.