Contents

Original Papers

Structural Basis for Designing Multiepitope Vaccines Against COVID-19 Infection: In Silico Vaccine Design and Validation (e19371)
Sukrit Srivastava, Sonia Verma, Mohit Kamthania, Rupinder Kaur, Ruchi Badyal, Ajay Saxena, Ho-Joon Shin, Michael Kolbe, Kailash Pandey.

2

The Novel Coronavirus Enigma: Phylogeny and Analyses of Coevolving Mutations Among the SARS-CoV-2 Viruses Circulating in India (e20735)

24

Tutorial

Molecular Docking Using Chimera and Autodock Vina Software for Nonbioinformaticians (e14232)
Sania Butt, Yasmin Badshah, Maria Shabbir, Mehak Rafiq.

39
Original Paper

Structural Basis for Designing Multiepitope Vaccines Against COVID-19 Infection: In Silico Vaccine Design and Validation

Sukrit Srivastava1,2, PhD; Sonia Verma3, MSc; Mohit Kamthania4, PhD; Rupinder Kaur5, PhD; Ruchi Kiran Badyal6, MPhil; Ajay Kumar Saxena2, PhD; Ho-Joon Shin7, PhD; Michael Kolbe8,9, PhD; Kailash C Pandey3, PhD

1Infection Biology Group, Department of Biotechnology, Mangalayan University, Aligarh, India
2Molecular Medicine Laboratory, School of Life Science, Jawaharlal Nehru University, New Delhi, India
3Parasite-Host Biology Group, Protein Biochemistry and Engineering Lab, ICMR-National Institute of Malaria Research, New Delhi, India
4Department of Biotechnology, Institute of Applied Medicines and Research, Ghaziabad, India
5Department of Chemistry, Guru Nanak Dev University, Amritsar, India
6Department of Economics, Mangalayatan University, Aligarh, India
7Department of Microbiology, School of Medicine, Ajou University, Suwon, Gyeonggi-do, Republic of Korea
8Centre for Structural Systems Biology, Department for Structural Infection Biology, Helmholtz-Centre for Infection Research, Hamburg, Germany
9Faculty of Mathematics, Informatics and Natural Sciences, University of Hamburg, Hamburg, Germany

Corresponding Author:
Sukrit Srivastava, PhD
Infection Biology Group
Department of Biotechnology
Mangalayan University
Aligarh-Mathura Highway
Aligarh, 202145
India
Phone: 91 8178718421
Email: srivastav.sukrit@gmail.com

Abstract

Background: The novel coronavirus disease (COVID-19), which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to the ongoing 2019-2020 pandemic. SARS-CoV-2 is a positive-sense single-stranded RNA coronavirus. Effective countermeasures against SARS-CoV-2 infection require the design and development of specific and effective vaccine candidates.

Objective: To address the urgent need for a SARS-CoV-2 vaccine, in the present study, we designed and validated one cytotoxic T lymphocyte (CTL) and one helper T lymphocyte (HTL) multi-epitope vaccine (MEV) against SARS-CoV-2 using various in silico methods.

Methods: Both designed MEVs are composed of CTL and HTL epitopes screened from 11 Open Reading Frame (ORF), structural and nonstructural proteins of the SARS-CoV-2 proteome. Both MEVs also carry potential B-cell linear and discontinuous epitopes as well as interferon gamma–inducing epitopes. To enhance the immune response of our vaccine design, truncated (residues 10-153) Onchocerca volvulus activation-associated secreted protein-1 was used as an adjuvant at the N termini of both MEVs. The tertiary models for both the designed MEVs were generated, refined, and further analyzed for stable molecular interaction with toll-like receptor 3. Codon-biased complementary DNA (cDNA) was generated for both MEVs and analyzed in silico for high level expression in a mammalian (human) host cell line.

Results: In the present study, we screened and shortlisted 38 CTL, 33 HTL, and 12 B cell epitopes from the 11 ORF protein sequences of the SARS-CoV-2 proteome. Moreover, the molecular interactions of the screened epitopes with their respective human leukocyte antigen allele binders and the transporter associated with antigen processing (TAP) complex were positively validated. The shortlisted screened epitopes were utilized to design two novel MEVs against SARS-CoV-2. Further molecular models of both MEVs were prepared, and their stable molecular interactions with toll-like receptor 3 were positively validated. The codon-optimized cDNAs of both MEVs were also positively analyzed for high levels of overexpression in a human cell line.

Conclusions: The present study is highly significant in terms of the molecular design of prospective CTL and HTL vaccines against SARS-CoV-2 infection with potential to elicit cellular and humoral immune responses. The epitopes of the designed
MEVs are predicted to cover the large human population worldwide (96.10%). Hence, both designed MEVs could be tried in vivo as potential vaccine candidates against SARS-CoV-2.


KEYWORDS
COVID-19; severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); coronavirus; human transporter associated with antigen processing (TAP); toll-like receptor (TLR); epitope; simulation; multiepitope vaccine

Introduction

The novel coronavirus disease (COVID-19), which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has resulted in the ongoing outbreak of a severe form of respiratory disease leading to death with a mortality rate of 3.4% [1]. SARS-CoV-2 is a novel coronavirus associated with a respiratory disease that initiated in the city of Wuhan in Hubei province, China. The disease is highly contagious; as of March 21, 2020, it had spread to 182 countries and territories since its outbreak in China in December 2019. Worldwide, as of March 21, 2020, the total number of confirmed cases was reported to be 266,073, and the total death count was reported to be 11,184 [2]. Overall, SARS-CoV-2 infection has created a global emergency. The economic impact of COVID-19 is even harsher and has placed the world at economic risk. As of March 9, 2020, the worst case scenario was a US $2 trillion shortfall in global income, with a $220 billion impact on developing countries. The COVID-19 shock will cause a recession in several countries and depress global annual growth this year to below 2.5%, which is the recessionary threshold for the world economy [3].

The infection mechanism and pathogenesis of SARS-CoV-2 are currently largely unknown. According to the National Center for Biology Information (NCBI) protein sequence database [4], the proteome of SARS-CoV-2 is composed of 11 Open Reading Frame (ORF), structural and non-structural proteins. These include a polyprotein (ORF1ab), surface protein (S protein), ORF3, envelope protein (E protein), membrane protein (M protein), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid protein (N protein), and ORF10. The actual functions and pathogenic or proliferative roles of these SARS-CoV-2 coronavirus proteins are currently largely unknown.

The SARS-CoV-2 polyprotein (ORF1ab), with a length of 7096 amino acids (AA), is composed of 16 different expressed proteins, namely leader protein (nsp1, location: 1-180 AA); nsp2 (location: 181-818 AA); nsp3 (former nsp1, carries proteins, namely leader protein (nsp1, location: 1-180 AA); nsp4 (contains transmembrane domain 1 and adenosine diphosphate-ribose 1’-phosphatase, location: 819-2763 AA); nsp4 (contains transmembrane domain 2, location: 2764-3263 AA); 3C-like proteinase (nsp5, main proteinase, mediating cleavage downstream of nsp4, location: 3264-3569 AA); nsp6 (putative transmembrane domain, location: 3570-3859 AA); nsp7 (location: 3860-3942 AA); nsp8 (location: 3943-4140 AA); nsp9 (ssRNA-binding protein, location: 4141-4253 AA); nsp10 (formerly known as growth-factor-like protein, location: 4254-4392 AA); nsp11 (location: 4393-4405 AA); RNA-dependent RNA polymerase (nsp12, location: 4393-5324 AA); helicase (nsp13; zinc-binding domain, NTPase/helicase domain, RNA 5’-triphosphatase, location: 5325-5925 AA); 3’-to-5’ exonuclease (nsp14, location: 5926-6452 AA); endo RNAse (nsp15, location: 6453-6798 AA); and 2’-O-ribose methyltransferase (nsp16; location: 6799-7096 AA).

The SARS-CoV-2 coronavirus S protein is a structural protein that acts as a spike protein; its location is 21563-25384 AA, and its length is 1273 AA. The ORF3a protein is located at 25393-26220 AA, and its length is 275 AA. The E protein (ORF4) is a structural protein; its location is 26245-26472 AA, and its length is 75 AA. The M protein (ORF5) is a structural protein; its location is 26523-27191 AA, and its length is 222 AA. The ORF6 protein is located at 27202-27387 AA, and its length is 61 AA. The ORF7a protein is located at 27394-27759 AA, and its length is 121 AA. The ORF7b protein is located at 27756-27887 AA, and its length is 43 AA. The SARS-CoV-2 coronavirus ORF8 protein is located at 27894-28259 AA, and its length is 121 AA. The N protein (ORF9) is a structural protein; its location is 28274-29533 AA, and its length is 419 AA. The ORF10 protein is located at 29558-29674 AA and has a length of 38 AA [4].

Although the exact mechanisms and roles of the abovementioned proteins of the SARS-CoV-2 coronavirus proteome are not well known, these proteins are potential candidates for use in vaccines against SARS-CoV-2 coronavirus infection. In this study, we screened high-potential epitopes from all the abovementioned proteins; further, we designed and proposed cytotoxic T lymphocyte (CTL) and helper T lymphocyte (HTL) multiepitope-based vaccine candidates against SARS-CoV-2 coronavirus infection.

Methods

Background

In this study on SARS-CoV-2 coronavirus, we screened potential epitopes and designed and proposed two multiepitope vaccines (MEVs) composed of screened CTL and HTL epitopes with overlapping regions of B cell epitopes. Hence, the proposed MEVs have the potential to elicit both humoral and cellular immune response. To enhance immune response, truncated (residues 10-153) Onchocerca volvulus activation-associated secreted protein-1 (Ov-ASP-1) was utilized as an adjuvant at the N-termini of both MEVs. The truncated Ov-ASP-1 was chosen due to its potential to activate antigen-processing cells (APCs) [5-7]. All the SARS-CoV-2 proteins mentioned in the introduction were utilized to screen potential CTL, HTL, and...
B cell epitopes. The screened epitopes were further studied to identify overlapping consensus regions among them. The epitopes showing regions of partial or complete overlap were chosen for further detailed studies.

The chosen CTL and HTL epitopes were analyzed for their molecular interactions with their respective human leukocyte antigen (HLA) allele binders. Moreover, the molecular interactions of the chosen CTL epitopes were analyzed for with the transporter associated with antigen processing (TAP) cavity to observe their smooth passage from the cytoplasm to the endoplasmic reticulum (ER) lumen [8,9]. Tertiary models of both MEVs were generated and refined. Both MEV models were further utilized to screen B cell linear and discontinuous epitopes as well as interferon gamma (IFNγ)-inducing epitopes.

Molecular signaling by multiple toll-like receptors is an essential component of the innate immune response against SARS-CoV-2. Because Ov-ASP-1 primarily binds APCs among human peripheral blood mononuclear cells and triggers proinflammatory cytokine production via toll-like receptor 3 (TLR3), the molecular interactions of both the CTL and HTL MEV models with TLR3 were further analyzed by molecular docking studies [10-13]. Furthermore, the codon-optimized cDNAs of both MEVs were analyzed and were found to have high levels of expression in a mammalian (human) cell line, which would facilitate in vivo expression, experimentation, and trials (see Supplementary Figure S1 in Multimedia Appendix 1).

### Screening of Potential Epitopes

#### T cell Epitope Prediction

#### Screening of CTL Epitopes

The CTL epitopes were screened using the Immune Epitope Database (IEDB) tools MHC (major histocompatibility complex)-I Binding Predictions and MHC-I Processing Predictions [14-16]. These two tools use six different methods (consensus, NN-align, SMM-align, combinatorial library, Sturniolo, and NetMHCIIPan), and they generate a percentile rank and a total score, respectively.

The screening is based on the total number of cleavage sites in the protein. The TAP score estimates an effective –log value of the half maximal inhibitory concentration (IC50) for binding to the TAP of a peptide or its N-terminal prolonged precursors. The MHC binding prediction score is the –log(IC50) value for binding to the MHC of a peptide [17]. The IC50 values (nanomolar) for each epitope and MHC allele binding pair were also obtained using the MHC-I Binding Predictions IEDB tool. Epitopes with high, intermediate, and low affinities of binding to their HLA allele binders have IC50 values of <50 nM, <500 nM, and <5000 nM, respectively.

The immunogenicities of all the screened CTL epitopes were also obtained using the MHC I Immunogenicity IEDB tool [17] with all parameters set to the default to analyze the first, second, and C-terminus amino acids of each screened epitope. The tool predicts the immunogenicity of a given peptide-MHC complex based on the physiochemical properties of its constituent amino acids and their positions within the peptide sequence.

#### Screening of HTL Epitopes

To screen out the HTL epitopes from the SARS-CoV-2 proteins, the IEDB tool MHC-II Binding Predictions was used. This tool generates a percentile rank for each potential peptide. The lower the percentile rank, the higher the affinity. This percentile rank is generated by the combination of three different methods, namely combinatorial library, SMM_align, and Sturniolo, and by comparing the score of the peptide against the scores of five million other random 15-mer peptides in the SWISS-PROT database [18-21]. The rank from the consensus of all three methods was generated by the median percentile rank of the three methods.

#### Population Coverage by CTL and HTL Epitopes

The IEDB Population Coverage tool was used to elucidate the world human population coverage by the shortlisted 38 CTL and 33 HTL epitopes derived from 9 SARS-CoV-2 proteins [22]. T cells recognize the complex between a specific major MHC molecule and a particular pathogen-derived epitope. The given epitope will only elicit a response in an individual who expresses an MHC molecule that is capable of binding that particular epitope. This denominated MHC restriction of T cell responses and the MHC polymorphism provides the basis for population coverage study. The MHC types are expressed at dramatically different frequencies in different ethnicities. Hence, a vaccine with larger population coverage could be of greater importance [21]. Clinical administration of multiple epitopes, including both CTL and HTL epitopes, is predicted here to have a higher probability of larger human population coverage worldwide.

#### B Cell Epitope Prediction

#### Sequence-Based B Cell Epitope Prediction

The protein sequence–based Bepipred Linear Epitope Prediction method [23] was utilized to screen linear B cell epitopes from 11 different SARS-CoV-2 protein ORFs. The B Cell Epitope Prediction Tools of the IEDB server were utilized. In this screening, parameters such as the hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity, and antigenic propensity of the polypeptides are correlated with their location in the protein. This enables a search for continuous epitopes predicted from a protein sequence. The prediction is based on the propensity scales for each of the 20 amino acids. For a window size n, i – (n – 1)/2 neighboring residues on each side of residue i are used to compute the score for residue i. The Bepipred Linear Epitope Prediction method used here is based on the propensity scale method as well as the physiochemical properties of the given antigenic sequence to screen potential epitopes [23].

#### Characterization of Potential Epitopes

#### Epitope Conservation Analysis

The shortlisted CTL, HTL, and B cell epitopes screened from eleven SARS-CoV-2 proteins were analyzed for the conservancy of their amino acid sequences using the IEDB Epitope Conservancy Analysis tool. The epitope conservancy is the number of protein sequences retrieved from the NCBI protein database that contains that particular epitope. The analysis was...
performed against the entire respective source protein sequences of SARS-CoV-2 proteins retrieved from the NCBI protein database [24].

**Epitope Toxicity Prediction**

The ToxinPred tool was used to analyze the toxicity of the shortlisted CTL, HTL, and B cell epitopes. The tool enables the identification of highly toxic or nontoxic short peptides. The toxicity check analysis was performed using the support vector machine-based ToxinPred method using a dataset of 1805 positive sequences and 3593 negative sequences from SWISS-PROT as well as an alternative dataset comprising the same 1805 positive sequences and 12,541 negative sequences from the Translated European Molecular Biology Laboratory (TriEMBL) database [25].

**Overlapping Residue Analysis**

The overlapping residue analysis for the shortlisted 38 CTL, 33 HTL, and 12 B cell linear epitopes was performed using multiple sequence alignment analysis with the European Bioinformatics Institute’s Clustal Omega tool [26]. The Clustal Omega multiple sequence alignment tool virtually aligns any number of protein sequences and delivers an accurate alignment.

**Selection of Epitopes for Molecular Interaction Studies With HLA Alleles and the TAP Transporter**

Based on the overlapping residue analysis of the shortlisted CTL, HTL, and linear B cell epitopes, a few CTL and HTL epitopes were chosen for further analysis. The chosen epitopes are circled in Supplementary Figure S10 (Multimedia Appendix 1). These epitopes were chosen based on their partial or full overlapping sequence regions among all three types of epitopes (CTL, HTL, and B cell). The chosen epitopes were further analyzed for their interactions with their respective HLA allele binders and TAP cavity interactions.

**Molecular Interaction Analysis of the Selected Epitopes with HLA Alleles and the TAP Transporter**

**Tertiary Structure Modeling of HLA Alleles and Selected T Cell Epitopes**

SWISS-MODEL [27] was used for homology modeling of the HLA class I and II allele binders of the chosen epitopes. The amino acid sequences of the HLA allele binders were retrieved from the Immuno Polymorphism Database (IPD-IMGT/HLA). Templates for homology modeling were chosen based on the highest amino acid sequence similarity. All the generated HLA allele models had acceptable QMEAN values (cutoff -4.0) (Supplementary Table S1, Multimedia Appendix 1). The QMEAN value gives a composite quality estimate involving both global and local analysis of the model [28].

PEP-FOLD 2.0 [29], a de novo structure prediction tool available at RPBS Web Portal, was utilized to generate tertiary structures for the chosen CTL and HTL epitopes.

**Molecular Interaction Analysis of Chosen CTL and HTL Epitopes With HLA Alleles**

The PatchDock tool was utilized for in silico molecular docking studies of the selected CTL and HTL epitopes with their respective HLA class I and II allele binders. PatchDock utilizes an algorithm for unbound (real-life) docking of molecules for protein-protein complex formation. The algorithm carries out rigid docking, and the surface variability/flexibility is implicitly addressed through liberal intermolecular penetration. The algorithm focuses on the initial molecular surface fitting on localized, curvature-based surface patches, the use of geometric hashing and pose clustering for initial transformation detection, computation of shape complementarity utilizing Distance Transform, efficient steric clash detection and geometric fit scoring based on multiresolution shape representation, and utilization of biological information by focusing on hotspot-rich surface patches [30-32].

**Molecular Interaction Analysis of Selected CTL Epitopes With the TAP Transporter**

The TAP transporter plays an important role in the presentation of a CTL epitope. From the cytosol after proteasome processing, the fragmented peptide of the foreign protein is transported to the ER through the TAP transporter. From the ER, these short peptides reach the Golgi bodies and are then presented on the cell surface [9]. Molecular interaction studies of the chosen CTL epitopes within the TAP cavity were performed by molecular docking using the PatchDock tool. For accurate prediction, the cryo-EM structure of the TAP transporter (PDB ID: 5u1d) was used by removing the antigen from the TAP cavity of the original structure [8].
from CD4+ T cells. This module generates overlapping peptides from the query sequence and predicts IFN-γ-inducing peptides. For the screening, the IEDB database was used with 3705 IFN-γ-inducing and 6728 non–IFN-γ-inducing MHC class II binders [39,40].

**MEV Allergenicity and Antigenicity Prediction**

Both the designed MEVs were further analyzed for allergenicity and antigenicity prediction using the AlgPred [41] and VaxiJen [42] tools, respectively. The AlgPred prediction is based on the similarity of an already known epitope with any region of the submitted protein. To screen allergenicity, the SWISS-PROT data set consisting of 101,725 non-allergens and 323 allergens was used. VaxiJen utilizes an alignment-free approach that is based solely on the physicochemical properties of the query amino acid sequence. To predict the antigenicity, VaxiJen uses bacterial, viral, and tumor protein datasets to derive models for the prediction of the antigenicity of a whole protein. Every set consists of known 100 antigens and 100 nonantigens.

**Tertiary Structure Modeling, Refinement, and Validation of the MEVs**

The tertiary structures of the designed CTL and HTL MEVs were generated by homology modeling using the I-TASSER modeling tool [43]. I-TASSER is a protein structure prediction tool that is based on the sequence-to-structure-to-function paradigm. The tool generates 3D atomic models from multiple threading alignments and iterative structural assembly simulations for a submitted AA sequence. I-TASSER is based on the structure templates identified by LOMETS, a metaserver from the Protein Data Bank (PDB) library. I-TASSER only uses the template with the highest Z-score, which is the difference between the raw and average scores in the unit of standard deviation. For each target model, the I-TASSER simulations generate a large ensemble of structural conformations, called decoys. To select the final models, I-TASSER uses the SPICKER program to cluster all the decoys based on their pairwise structure similarity and reports up to 5 models. A normalized Z-score >1 indicates a good alignment and vice versa. The Cov represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of the query protein. Ranking of template proteins is based on the TM-score of the structural alignment between the query structure model and known structures. The root mean square deviation (RMSD) is the RMSD between template residues and query residues that are structurally aligned by the TM-align algorithm.

Both the generated MEV models were refined using the ModRefiner [44] and GalaxyRefine [45] tools. The TM-score generated by ModRefiner indicates the structural similarity of the refined model to the original input model. The closer the TM-score to 1, the greater the similarity of the original and refined models. The RMSD of the refined model shows the conformational deviation from the original input models.

The GalaxyRefine tool refines the query tertiary structure by repeated structure perturbation and by using the subsequent structural relaxation by the molecular dynamics simulation. The GalaxyRefine tool generates reliable core structures from multiple templates and then rebuilds unreliable loops or termini using an optimization-based refinement method [46,47]. To avoid any breaks in the 3D model, GalaxyRefine uses the triaxial loop closure method. The MolProbity score generated for a given refined model indicates the log-weighted combination of the clash score, the percentage of Ramachandran unfavored residues and the percentage of bad side chain rotamers.

**Validation of the Refined Models of the CTL and HTL MEVs**

The refined CTL and HTL MEV 3D models both were further validated by the RAMPAGE analysis tool [48,49]. The generated Ramachandran plots for the MEV models show the sterically allowed and disallowed residues along with their dihedral psi (ψ) and phi (φ) angles.

**Linear and Discontinuous B-cell Epitope Prediction of the MEVs**

The ElliPro antibody epitope prediction tool available at the IEDB was used to screen the linear and discontinuous B cell epitopes from the MEV models. The ElliPro method analyses are based on the location of a residue in the 3D structure of a protein. For example, the residues lying outside an ellipsoid covering 90% of the inner core protein residues score the highest protrusion index (PI) of 0.9. The discontinuous epitopes predicted by the ElliPro tool are clustered based on the distance R in angstroms between the centers of mass of two residues lying outside the largest possible ellipsoid. A larger value of R indicates that more distant residues (residue discontinuity) are screened in the epitopes [50,51].

**Molecular Interaction Analysis of MEVs With an Immunological Receptor**

**Molecular Docking Studies of the MEVs and TLR3**

Molecular interaction analysis of both designed MEVs with TLR3 was performed by molecular docking and molecular dynamics simulations. Molecular docking was performed using the PatchDock server [32]. PatchDock utilizes an algorithm for unbound docking of molecules (mimicking the real world environment) for protein-protein complex formation, as explained earlier [30,31]. For molecular docking, the 3D structure of the human TLR3 ectodomain was retrieved from the PDB (PDB ID: 2A0Z). The study provides the dynamical properties of the designed system with the MEV-TLR3 complexes and guesses at the interactions between the molecules; also, it gives exact predictions of bulk properties, including hydrogen bond formation and the conformation of the molecules forming the complex.

**Molecular Dynamics Simulation Studies of the MEVs and the TLR3 Complex**

The MEV-TLR3 molecular interactions were further evaluated using molecular dynamics simulations. The molecular dynamics simulations were performed for 10 nanoseconds using YASARA (Yet Another Scientific Artificial Reality Application) [52]. The simulations were performed in an explicit water environment in a dodecahedron simulation box at a constant temperature (298 kelvin) and pressure (1 atmosphere) at pH 7.4 with a periodic cell boundary condition. The solvated systems were...
neutralized with counterions (sodium chloride, concentration 0.9 molar). The AMBER14 force field was applied to the systems during the simulations [53,54]. Long-range electrostatic energies and forces were calculated using the particle mesh–based Ewald method [55]. The solvated structures were minimized by the steepest descent method at a temperature of 298 K and a constant pressure. Then, the complexes were equilibrated for a period of 1 nanosecond. After equilibration, a production molecular dynamics simulation was run for 10 ns at a constant temperature and pressure, and time frames were saved every 10 picoseconds for each simulation. The RMSD and root mean square fluctuation (RMSF) values for the alpha carbon (Cα) atoms, backbone atoms, and all the atoms of both MEV complexes were analyzed for each simulation conducted.

**Generation and Analysis of cDNA of the MEVs**

cDNAs of both MEVs, codon-optimized for expression in a mammalian (human) cell line, were generated using the Java Codon Adaptation Tool. The generated cDNAs of both the MEVs were further analyzed by the GenScript Rare Codon Analysis Tool. This tool analyzes the GC content, codon adaptation index (CAI) and tandem rare codon frequency for a given cDNA [56,57]. The CAI indicates the possibility of cDNA expression in a chosen expression system. The tandem rare codon frequency indicates the presence of low-frequency codons in a given cDNA.

**Results**

**Screening of Potential Epitopes**

**T Cell Epitope Prediction**

**Screening of CTL Epitopes**

CTL epitopes were screened using the MHC-I Binding Predictions and MHC-I Processing Predictions IEDB tools. These epitopes were shortlisted based on the total number of cleavage sites in the protein, low IC₅₀ (nM) values for epitope-HLA class I allele pairs, and binding to the TAP cavity. The 38 epitopes predicted by the MHC-I Binding Predictions tool with the highest percentile ranks were shortlisted for MEV design and are listed in Table 1. The remaining 101 epitope-HLA I allele pairs are listed in Supplementary Table S8 (Multimedia Appendix 1). The 67 epitope-HLA I allele pairs predicted by the MHC-I Processing Predictions tool with the highest total scores are listed in Supplementary Table S9 (Multimedia Appendix 1).

The immunogenicities of the shortlisted CTL epitopes were also determined and are noted in Table 1 and in Supplementary Tables S8 and S9 (Multimedia Appendix 1). A higher immunogenicity score indicates greater immunogenic potential of the given epitope.
Table 1. Characteristics of the shortlisted high-percentile-ranking SARS-CoV-2 CTL epitopes and their respective HLA allele binders.

<table>
<thead>
<tr>
<th>SARS-CoV-2 protein</th>
<th>Peptide sequence</th>
<th>Peptide length, amino acids</th>
<th>Conservancy (%)</th>
<th>Immunogenicity</th>
<th>Toxicity</th>
<th>Allele</th>
<th>Methods used</th>
<th>Percentile rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>E protein</td>
<td>LLFLAFVVF</td>
<td>480/482 (99.59)</td>
<td>0.2341</td>
<td>Nontoxic</td>
<td>B15:01</td>
<td>Consensus (ann/smm/comblib_siddney 2008)</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>M protein</td>
<td>LTALRLCAY</td>
<td>478/482 (99.17)</td>
<td>0.01886</td>
<td>Nontoxic</td>
<td>A01:01</td>
<td>Consensus (ann/smm)</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>M protein</td>
<td>YTFAVLRFA</td>
<td>474/477 (99.37)</td>
<td>0.19709</td>
<td>Nontoxic</td>
<td>A33:01</td>
<td>ann</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>N protein</td>
<td>ATKRTLSYYKE</td>
<td>472/477 (98.95)</td>
<td>-0.13563</td>
<td>Nontoxic</td>
<td>A11:01</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>N protein</td>
<td>MEVTSGTGW</td>
<td>485/498 (97.39)</td>
<td>-0.06279</td>
<td>Nontoxic</td>
<td>B44:02</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf10</td>
<td>MGYINNFAL</td>
<td>477/480 (99.38)</td>
<td>-0.09452</td>
<td>Nontoxic</td>
<td>B35:01</td>
<td>Consensus (ann/smm/comblib_siddney 2008)</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>orf10</td>
<td>GINVFAFPFRF</td>
<td>232/236 (98.31)</td>
<td>0.20158</td>
<td>Nontoxic</td>
<td>A23:01</td>
<td>Consensus (ann/smm)</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>SEMVCGGSLY</td>
<td>452/456 (99.12)</td>
<td>0.32633</td>
<td>Nontoxic</td>
<td>B44:02</td>
<td>ann</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>FYWFFS-NYLKR</td>
<td>455/456 (99.78)</td>
<td>0.37766</td>
<td>Nontoxic</td>
<td>A33:01</td>
<td>ann</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>ISNSWLMW</td>
<td>454/456 (99.56)</td>
<td>-0.24791</td>
<td>Nontoxic</td>
<td>B58:01</td>
<td>ann</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>ETILAGSYK</td>
<td>455/456 (99.78)</td>
<td>0.08174</td>
<td>Nontoxic</td>
<td>A68:01</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>QIEILGTVSW</td>
<td>455/456 (99.78)</td>
<td>0.27341</td>
<td>Nontoxic</td>
<td>B44:02</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>STFNVPEMK</td>
<td>456/456 (100.00)</td>
<td>-0.32016</td>
<td>Nontoxic</td>
<td>A11:01</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>RMYFFAFSY</td>
<td>456/456 (100.00)</td>
<td>0.21107</td>
<td>Nontoxic</td>
<td>A30:02</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>FLVAAAFYLY</td>
<td>454/456 (99.56)</td>
<td>-0.19814</td>
<td>Nontoxic</td>
<td>A20:01</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>RYFRLTLGTVY</td>
<td>456/456 (100.00)</td>
<td>0.03976</td>
<td>Nontoxic</td>
<td>A30:02</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>FLNCGSCGV</td>
<td>456/456 (100.00)</td>
<td>-0.20585</td>
<td>Nontoxic</td>
<td>A20:03</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>CTDGNALAY</td>
<td>476/479 (99.37)</td>
<td>0.32004</td>
<td>Nontoxic</td>
<td>A01:01</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>CT- DDNALAYY</td>
<td>476/479 (99.37)</td>
<td>0.28694</td>
<td>Nontoxic</td>
<td>A01:01</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>MYKGLPL-WNVVR</td>
<td>456/456 (100.00)</td>
<td>-0.11151</td>
<td>Nontoxic</td>
<td>A33:01</td>
<td>ann</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>SINNTVYTYK</td>
<td>456/456 (100.00)</td>
<td>0.15936</td>
<td>Nontoxic</td>
<td>A11:01</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>LPVNVAFELW</td>
<td>450/456 (98.68)</td>
<td>-0.00254</td>
<td>Nontoxic</td>
<td>B53:01</td>
<td>Consensus (ann/smm)</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>DEWSDMATYY</td>
<td>455/456 (99.78)</td>
<td>0.07355</td>
<td>Nontoxic</td>
<td>B44:03</td>
<td>Consensus (ann/smm)</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>orf-lab</td>
<td>YILFTRFFYV</td>
<td>454/456 (99.56)</td>
<td>-0.02845</td>
<td>Nontoxic</td>
<td>A02:06</td>
<td>Consensus (ann/smm)</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>ORF3a</td>
<td>YLLYAVYFYFL</td>
<td>456/456 (100.00)</td>
<td>0.40924</td>
<td>Nontoxic</td>
<td>A02:01</td>
<td>Consensus (ann/smm/comblib_siddney 2008)</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>ORF3a</td>
<td>IPYNSVTSSI</td>
<td>454/456 (99.56)</td>
<td>0.13772</td>
<td>Nontoxic</td>
<td>B51:01</td>
<td>Consensus (ann/smm)</td>
<td>0.11</td>
<td></td>
</tr>
</tbody>
</table>
The screening of HTL epitopes from 11 different SARS-CoV-2 ORF proteins was performed based on percentile rank. The smaller the percentile rank, the higher the affinity of the peptide to its respective HLA allele binders. The 33 epitopes with high percentile ranking were shortlisted (Table 2). An additional 180 potential HTL cell epitope-HLA allele II pairs with high percentile ranks screened in our study are listed in Supplementary Table S10 (Multimedia Appendix 1).
<table>
<thead>
<tr>
<th>SARS-CoV-2(^a) protein</th>
<th>Peptide</th>
<th>Conservancy (%)</th>
<th>Toxicity</th>
<th>Alleles</th>
<th>Methods used(^b)</th>
<th>Percentile rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>E protein(^c)</td>
<td>LLFLAVVFLVTLA</td>
<td>480/482 (99.59)</td>
<td>Nontoxic</td>
<td>DPA1-03:01/DPB1-04:02</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.02</td>
</tr>
<tr>
<td>E protein</td>
<td>VLLFLAVVFLVTLT</td>
<td>480/482 (99.59)</td>
<td>Nontoxic</td>
<td>DPA1-03:01/DPB1-04:02</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.02</td>
</tr>
<tr>
<td>M protein(^d)</td>
<td>GLMWLSFYIASFRLF</td>
<td>465/477 (97.48)</td>
<td>Nontoxic</td>
<td>DPA1-01:03/DPB1-02:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.05</td>
</tr>
<tr>
<td>M Protein</td>
<td>LMWLSFYIASFRLFA</td>
<td>466/477 (97.69)</td>
<td>Nontoxic</td>
<td>DPA1-01:03/DPB1-02:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.05</td>
</tr>
<tr>
<td>M protein</td>
<td>LSYYKLGSQRVAGD(^e)</td>
<td>472/477 (99.95)</td>
<td>Nontoxic</td>
<td>DRB1-09:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.06</td>
</tr>
<tr>
<td>N protein(^f)</td>
<td>AQFAPSASAFFGMSR</td>
<td>486/498 (97.59)</td>
<td>Nontoxic</td>
<td>DRB1-09:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.01</td>
</tr>
<tr>
<td>N protein</td>
<td>IAQFAPSASAFFGMS</td>
<td>485/498 (97.39)</td>
<td>Nontoxic</td>
<td>DRB1-09:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.01</td>
</tr>
<tr>
<td>N protein</td>
<td>PQIAQFAPSASAFFG</td>
<td>485/498 (97.39)</td>
<td>Nontoxic</td>
<td>DRB1-09:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.01</td>
</tr>
<tr>
<td>ORF1ab</td>
<td>AIILASFSTSAFV</td>
<td>456/456 (100.00)</td>
<td>Nontoxic</td>
<td>DRB1-09:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.01</td>
</tr>
<tr>
<td>ORF1ab</td>
<td>ESPFVMMSAPPAQYE(^e)</td>
<td>456/456 (100.00)</td>
<td>Nontoxic</td>
<td>DRB1-01:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.01</td>
</tr>
<tr>
<td>ORF1ab</td>
<td>IILASFSTSAFVE</td>
<td>456/456 (100.00)</td>
<td>Nontoxic</td>
<td>DRB1-09:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.01</td>
</tr>
<tr>
<td>ORF1ab</td>
<td>QESPFVMMSAPPAQY</td>
<td>456/456 (100.00)</td>
<td>Nontoxic</td>
<td>DRB1-01:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.01</td>
</tr>
<tr>
<td>ORF1ab</td>
<td>SPFVMMMSAPPAQYEL</td>
<td>456/456 (100.00)</td>
<td>Nontoxic</td>
<td>DRB1-01:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.01</td>
</tr>
<tr>
<td>ORF3a</td>
<td>FVRATATIPIQASLP</td>
<td>478/481 (99.37)</td>
<td>Nontoxic</td>
<td>DPA1-02:01/DPB1-14:01</td>
<td>NetMHCIIPan</td>
<td>0.12</td>
</tr>
<tr>
<td>ORF3a</td>
<td>LLFVTLYSHLLLVA</td>
<td>467/481 (97.08)</td>
<td>Nontoxic</td>
<td>DRB1-01:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.1</td>
</tr>
<tr>
<td>ORF6</td>
<td>FKVISWNLDYINLI</td>
<td>478/481 (99.38)</td>
<td>Nontoxic</td>
<td>DQA1-01:01/DQB1-05:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.02</td>
</tr>
<tr>
<td>ORF6</td>
<td>KVISWNLDYINLI</td>
<td>478/481 (99.38)</td>
<td>Nontoxic</td>
<td>DQA1-01:01/DQB1-05:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.02</td>
</tr>
<tr>
<td>ORF6</td>
<td>TFKVISWNLDYINLI(^e)</td>
<td>478/481 (99.38)</td>
<td>Nontoxic</td>
<td>DQA1-01:01/DQB1-05:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.02</td>
</tr>
<tr>
<td>ORF7a</td>
<td>IILFLALILTACEL</td>
<td>479/480 (99.79)</td>
<td>Nontoxic</td>
<td>DRB1-01:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.16</td>
</tr>
<tr>
<td>ORF7a</td>
<td>ILFLALILATCLEY</td>
<td>479/480 (99.79)</td>
<td>Nontoxic</td>
<td>DRB1-01:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.16</td>
</tr>
<tr>
<td>ORF7b</td>
<td>CFLAFLLFLLMLI</td>
<td>231/236 (97.88)</td>
<td>Nontoxic</td>
<td>DPA1-03:01/DPB1-04:02</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.03</td>
</tr>
<tr>
<td>ORF7b</td>
<td>LCFLAFLLFLLML</td>
<td>231/236 (97.88)</td>
<td>Nontoxic</td>
<td>DPA1-03:01/DPB1-04:02</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.02</td>
</tr>
<tr>
<td>ORF7b</td>
<td>YLCFLAFLLFLLML</td>
<td>231/236 (97.88)</td>
<td>Nontoxic</td>
<td>DPA1-03:01/DPB1-04:02</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.02</td>
</tr>
<tr>
<td>ORF8</td>
<td>CTQHPYVVDPCPI</td>
<td>476/480 (99.17)</td>
<td>Nontoxic</td>
<td>DRB3-01:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.08</td>
</tr>
<tr>
<td>ORF8</td>
<td>HQPYVVDDPCPIHFY</td>
<td>476/480 (99.17)</td>
<td>Nontoxic</td>
<td>DRB3-01:01</td>
<td>Consensus (comb.lib/smm/nn)</td>
<td>0.08</td>
</tr>
</tbody>
</table>
**Population Coverage by CTL and HTL Epitopes**

The population coverage by the shortlisted epitopes was also studied, particularly in China, France, Italy, the United States, South Asia, East Asia, Northeast Asia, and the Middle East. From this study, we can conclude that the combined use of all the shortlisted CTL and HTL epitopes would have an average worldwide population coverage as high as 96.10% (SD 23.74) (Supplementary Table S12, Supplementary Table S12, Multimedia Appendix 1).

**B Cell Epitope Prediction**

**Sequence-Based B Cell Epitope Prediction**

To screen B cell epitopes, we used the Bepipred Linear Epitope Prediction method. In our study, we screened 12 B cell epitopes from 11 SARS-CoV-2 ORF proteins which show partial or complete overlap with the shortlisted CTL and HTL epitopes (Table 3). An additional 206 B cell epitopes with epitope lengths of at least four AAs and a maximum of 20 AAs were screened and are listed in Supplementary Table S11, Supplementary Table S11, Multimedia Appendix 1.

---

### Table 3: Shortlisted Epitopes

<table>
<thead>
<tr>
<th>SARS-CoV-2 protein</th>
<th>Peptide</th>
<th>Conservancy (%)</th>
<th>Toxicity</th>
<th>Alleles</th>
<th>Methods used&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Percentile rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF8</td>
<td>QPYVDDPCPIHFYS</td>
<td>476/480 (99.17)</td>
<td>Nontoxic</td>
<td>DRB3-01:01</td>
<td>Consensus (comb.lib./smm/nn)</td>
<td>0.07</td>
</tr>
<tr>
<td>ORF10</td>
<td>INVFAFPFTYSSL</td>
<td>476/480 (99.17)</td>
<td>Nontoxic</td>
<td>HLA-DPA1-01:03/DPB1-02:01</td>
<td>Consensus (comb.lib./smm/nn)</td>
<td>0.29</td>
</tr>
<tr>
<td>ORF10</td>
<td>YINVFAFPFTYSSL</td>
<td>476/479 (99.37)</td>
<td>Nontoxic</td>
<td>DPA1-01:03/DPB1-02:01</td>
<td>Consensus (comb.lib./smm/nn)</td>
<td>0.29</td>
</tr>
<tr>
<td>S protein&lt;sup&gt;f&lt;/sup&gt;</td>
<td>KTQSLIVNATNVV</td>
<td>472/472 (100.00)</td>
<td>Nontoxic</td>
<td>DRB1-13:02</td>
<td>Consensus (smm/nn/sturniolo)</td>
<td>0.01</td>
</tr>
<tr>
<td>S protein</td>
<td>LLIVNATNVVKVC</td>
<td>469/472 (99.36)</td>
<td>Nontoxic</td>
<td>DRB1-13:02</td>
<td>Consensus (smm/nn/sturniolo)</td>
<td>0.01</td>
</tr>
<tr>
<td>S protein</td>
<td>QSLLIVNATNVVK</td>
<td>471/472 (99.79)</td>
<td>Nontoxic</td>
<td>DRB1-13:02</td>
<td>Consensus (smm/nn/sturniolo)</td>
<td>0.01</td>
</tr>
<tr>
<td>S protein</td>
<td>SLLIVNATNVVK&lt;sup&gt;e&lt;/sup&gt;</td>
<td>471/472 (99.79)</td>
<td>Nontoxic</td>
<td>DRB1-13:02</td>
<td>Consensus (smm/nn/sturniolo)</td>
<td>0.01</td>
</tr>
<tr>
<td>S protein</td>
<td>TQSLIVNATNVV</td>
<td>471/472 (99.79)</td>
<td>Nontoxic</td>
<td>DRB1-13:02</td>
<td>Consensus (smm/nn/sturniolo)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

---

<sup>a</sup>SARS-CoV-2: severe acute respiratory syndrome coronavirus 2.


<sup>c</sup>E protein: envelope protein.

<sup>d</sup>M protein: membrane protein.

<sup>e</sup>Matches a recently published epitope, indicating consensus with results [58].

<sup>f</sup>N protein: nucleocapsid protein

<sup>g</sup>S protein: surface protein.
Table 3. Characteristics of the shortlisted SARS-CoV-2 linear B cell epitopes obtained by the BepiPred method.

<table>
<thead>
<tr>
<th>SARS-CoV-2a protein</th>
<th>Peptide length, amino acids</th>
<th>Conservancy (%)</th>
<th>Overlapping B cell epitope</th>
<th>Toxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>M proteinb</td>
<td>12</td>
<td>471/477 (98.74)</td>
<td>KLGASQRVAGDS</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>N proteinc</td>
<td>42</td>
<td>483/498 (96.99)</td>
<td>RNLQLESKMSGKQQQGQTVTKKSAEASK KPRKRTATKA</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>ORF1ab</td>
<td>20</td>
<td>455/456 (99.78)</td>
<td>GTTQTACDTDNALAYYNTTK</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>ORF3a</td>
<td>12</td>
<td>478/481 (99.37)</td>
<td>QGEIKDATPDSF</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>ORF3a</td>
<td>6</td>
<td>471/481 (97.92)</td>
<td>PYNSVT</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>ORF7a</td>
<td>9</td>
<td>479/480 (99.79)</td>
<td>LHYQECVR</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>ORF7a</td>
<td>26</td>
<td>470/480 (97.92)</td>
<td>VHKVQLRARSVSPKLFIRQEEVQEL</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>ORF8</td>
<td>23</td>
<td>460/480 (95.83)</td>
<td>QSTCQHQPYYVDDCPHIHYSKW</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>ORF8</td>
<td>9</td>
<td>476/480 (99.17)</td>
<td>RVGARKSAP</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>S proteind</td>
<td>11</td>
<td>470/472 (99.58)</td>
<td>TPGDSSSGWTA</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>S protein</td>
<td>35</td>
<td>470/472 (99.58)</td>
<td>FPNITNLCPFGEVNFATRFASYWNYRNKRISNCVA</td>
<td>Nontoxic</td>
</tr>
<tr>
<td>S protein</td>
<td>62</td>
<td>454/472 (96.19)</td>
<td>NLDSKVGGNYYNYRLFRKSNLKPFERDISTSIEY QAGSTPCNGVEGFNCYYPPLQSYGFQPTNN</td>
<td>Nontoxic</td>
</tr>
</tbody>
</table>

aSARS-CoV-2: severe acute respiratory syndrome coronavirus 2.
bM protein: membrane protein.
cN protein: nucleocapsid protein.
dS protein: surface protein.

Characterization of Potential Epitopes

Epitope Conservation Analysis

Sequence conservation analysis of the screened CTL, HTL, and B cell epitopes showed the highly conserved nature of the shortlisted epitopes. The amino acid sequences of both the CTL epitopes and the HTL epitopes were found to be significantly conserved among the NCBI-retrieved protein sequences of SARS-CoV-2 (the CTL epitopes were 96.88%-100% conserved and the HTL epitopes were 97.08%-100% conserved; see Tables 1, 2, and 4 and Supplementary Tables S8, S9, S10, and S11, Multimedia Appendix 1).

Epitope Toxicity Prediction

Toxicity analyses of all the screened CTL, HTL, and B cell epitopes were also performed. The ToxinPred study of all the shortlisted epitopes showed that they all are nontoxic (Tables 1, 2, and 4; Supplementary Tables S8, S9, S10, and S11, Multimedia Appendix 1).

Overlapping Residue Analysis

The AA sequence overlap among the shortlisted CTL, HTL, and B cell epitopes from 11 SARS-CoV-2 ORF proteins was analyzed using the Clustal Omega multiple sequence alignment analysis tool. The analysis showed that several CTL, HTL, and B cell epitopes had overlapping AA sequences. The CTL, HTL, and B cell epitopes with two or more overlapping AA residues are shown in Supplementary Figure S3 (Multimedia Appendix 1).

Selection of Epitopes for Molecular Interaction Studies with HLA Alleles and the TAP Transporter

The epitopes showing overlap among the CTL, HTL, and B cell epitopes were circled in Supplementary Figure S10 (Multimedia Appendix 1) and were chosen for further study of their interactions with HLA alleles and the TAP transporter.

Molecular Interaction Analysis of Selected Epitopes With HLA Alleles and the TAP Transporter

Molecular Interaction Analysis of the Chosen CTL and HTL Epitopes With HLA Alleles

Molecular docking studies of the chosen CTL and HTL epitopes with their respective HLA class I and II allele binders were performed using the PatchDock tool. Images were generated by PyMOL [59]. The study revealed significant molecular interactions between all the chosen epitopes and their HLA allele binders, showing the formation of multiple hydrogen bonds (Figure 1). Furthermore, B-factor analysis of all the epitope–HLA allele complexes showed that the epitope ligand had a stable (blue) binding conformation in complex with the HLA allele molecule (Supplementary Figure S4, Multimedia Appendix 1). The violet-indigo-blue-green-yellow-orange-red (VIBGYOR) color presentation was used, where blue is very stable.
**Figure 1.** Molecular docking analysis of SARS-CoV-2 CTL epitopes and HLA alleles. Molecular docking of the chosen CTL and HTL epitopes (cyan sticks) binding the amino acid residues of their respective HLA class I and class II allele binders (magenta sticks). The study shows that the docked complexes are stable, with the formation of multiple hydrogen bonds (green dots, lengths in angstroms). CTL: cytotoxic T lymphocyte. HLA: human leukocyte antigen.

**Molecular Interaction Analysis of Selected CTL Epitopes With the TAP Cavity**

The molecular docking interaction analysis of the chosen CTL epitopes with the TAP cavity showed significantly strong molecular interactions with the formation of several hydrogen bonds at different sites of the TAP cavity. Two sites of interaction were of particular interest: one closer to the cytoplasmic end and another closer to the ER lumen (Figure 2). This study confirms the feasibility of transportation of the chosen CTL epitopes from the cytoplasm to the ER lumen, which is an essential event for the representation of an epitope by HLA allele molecules on the surface of antigen-presenting cells.
Figure 2. Molecular docking analysis of two CTL epitopes within the TAP transporter cavity. The molecular interactions of the CTL epitopes (cyan sticks) within the TAP cavity (gray ribbons/sticks) are shown. Detailed interactions between the residues of the epitopes and the TAP transporter residues are shown, with hydrogen bond formation indicated with green dots. H bonds are shown in green dots with lengths in angstroms. TAP: transporter associated with antigen processing.

Epitope ATSRTLSTKY through TAP cavity

Epitope HFYSKFWYIR through TAP cavity

Characterization and Molecular Interaction Analysis of the Designed MEVs with Immune Receptors

Characterization of the Designed MEVs

Physicochemical Property Analysis of the Designed MEVs

ProtParam analysis of both the CTL and HTL MEVs was performed to analyze their physiochemical properties. The empirical physiochemical properties of the CTL and HTL MEVs are given in Table 4. The aliphatic indices and grand averages of hydrophaticity of both MEVs indicate their globular and hydrophilic natures. The instability index scores of both MEVs indicates the stable nature of the protein molecules.

Table 4. Physicochemical property analysis based on the amino acid sequences of the designed CTL and HTL MEVs.

<table>
<thead>
<tr>
<th>Property</th>
<th>Cytotoxic T lymphocyte multiepitope vaccine</th>
<th>Helper T lymphocyte multiepitope vaccine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (amino acids)</td>
<td>704</td>
<td>810</td>
</tr>
<tr>
<td>Molecular weight (kilodaltons)</td>
<td>72.62</td>
<td>82.80</td>
</tr>
<tr>
<td>Theoretical protrusion index</td>
<td>9.70</td>
<td>8.64</td>
</tr>
<tr>
<td>Expected half-life (hours)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Yeast</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Mammalian cell</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Aliphatic index</td>
<td>61.09</td>
<td>96.43</td>
</tr>
<tr>
<td>Grand average of hydrophaticity</td>
<td>−0.090</td>
<td>0.501</td>
</tr>
<tr>
<td>Instability index</td>
<td>44.31</td>
<td>40.28</td>
</tr>
</tbody>
</table>

IFNγ-Inducing Epitope Prediction

IFNγ-inducing epitopes are involved in both the adaptive and the innate immune response. IFNγ-inducing 15mer peptide epitopes were screened from the amino acid sequences of the CTL and HTL MEVs using the IFNepitope server. A total of 20 CTL MEV and 20 HTL MEV IFNγ-inducing positive epitopes with a score ≥1 were shortlisted (Supplementary Table S2, Multimedia Appendix 1).
Allergenicity and Antigenicity Prediction of the MEVs

Both the CTL and HTL MEVs were found to be nonallergenic by the AlgPred analysis (scores of –0.95185601 and –1.1293352, respectively; the threshold was –0.4). The CTL and HTL MEVs were also indicated by VaxJen analysis to be probable antigens (prediction scores of 0.4485 and 0.4215, respectively; the default threshold is 0.4). Hence, with the mentioned analysis tools, both the CTL and HTL MEVs are predicted to be nonallergic and antigenic in nature.

Tertiary Structure Modeling, Refinement, and Validation of the MEVs

3D homology models were generated for both the CTL and HTL MEVs using the I-TASSER modeling tool (Figure 3). The models were generated for the CTL MEV (PDB ID: 5n8pA, normal Z-score of 1.49, Cov of 0.92, TM-score of 0.916, and RMSD of 1.04 Å) and the HTL MEV (PDB ID: 5n8pA, normal Z-score of 1.52, Cov of 0.97, TM-score of 0.916, and RMSD of 1.04 Å).

Figure 3. Tertiary structure modelling of the CTL and HTL multi-epitope vaccines. The epitopes are shown in cyan. The adjuvant (Ov-ASP-1) is shown in orange. The linkers are shown in gray, and the 6xHis tag is shown in magenta. Cartoon and surface presentations of both the MEVs are shown. CTL: cytotoxic T lymphocyte. HTL: helper T lymphocyte.

The generated CTL and HTL 3D models were both further refined by ModRefiner to repair any gaps, followed by GalaxyRefine refinement. The refinement by ModRefiner showed TM-scores of 0.9189 and 0.9498 for the CTL and HTL models, respectively; because these values are close to 1, the initial and refined models were structurally similar. After refinement, the RMSDs for the CTL and HTL models with respect to the initial model were 3.367 Å and 2.318 Å, respectively. Further, both the CTL and HTL MEV models were refined with GalaxyRefine, and model 1 was chosen based on the best scoring parameters. The CTL MEV model refinement output model (Ramachandran favored 83.6%, GDT-HA 0.9371, RMSD 0.459, MolProbity 2.539, clash score 23.2, and poor rotamers 1.8) and the HTL MEV model refinement output model (Ramachandran favored 87.7%, GDT-HA 0.9552, RMSD 0.402, MolProbity 2.537, clash score 27.9, and poor rotamers 1.6)
show that well-refined and acceptable models were generated for both the MEVs. After refinement, all the mentioned parameters were found to be significantly improved in comparison to the initial CTL and HTL MEV models (Supplementary Table S3, Multimedia Appendix 1).

Validation of the Refined Models of the CTL and HTL MEVs

Both the CTL and HTL models were analyzed with the RAMPAGE analysis tool after refinement. The refined CTL MEV model was found to have 85.8% residues in the favored region, 11.3% residues in the allowed region, and only 3.0% residues in the outlier region; meanwhile, the refined HTL MEV model was found to have 88.9% residues in the favored region, 8.9% residues in the allowed region, and only 2.2% residues in the outlier region (Supplementary Figure S5, Multimedia Appendix 1).

Linear and Discontinuous B-cell Epitope Prediction From the MEVs

Linear and discontinuous B-cell epitope prediction was performed to identify potential linear and discontinuous epitopes in the refined 3D models of the CTL and HTL MEVs utilizing the ElliPro tool available on the IEDB server. The screening revealed that the CTL MEV carries 17 linear and 2 potential discontinuous B cell epitopes and the HTL MEV carries 17 linear and 4 potential discontinuous epitopes. The wide range of the PI scores of the linear and discontinuous epitopes in the CTL and HTL MEVs show the high potential of the epitopes to cause humoral immune response (PI scores: CTL MEV linear and discontinuous B cell epitopes: 0.511-0.828 and 0.664-0.767, respectively; HTL MEV linear and discontinuous B cell epitopes: 0.518-0.831 and 0.53-0.776, respectively) (Supplementary Tables S4, S5, S6, and S7, Multimedia Appendix 1).

Molecular Interaction Analysis of the MEVs With Immunological Receptors

Molecular Docking Studies of the MEVs With TLR3

The refined models of both the CTL and HTL MEVs were further studied for their molecular interactions with the ectodomain of human TLR3. Therefore, molecular docking of the CTL and HTL MEV models with the TLR3 crystal structure model (PDB ID: 2A0Z) was performed utilizing the PatchDock tool. The generated docking conformations with the highest scores of 20776 and 20350 for the CTL and HTL MEVs, respectively, were chosen for further study. The highest docking score indicates the best geometric shape complementarity fitting conformation of the MEV and the TLR3 receptor as predicted by the PatchDock tool. Both the CTL and HTL MEVs fit into the ectodomain region of TLR3 after docking, involving numerous molecular interactions with active site residues of the TLR3 cavity region (Figure 4A, C, D, and F). As shown in Figure 4A and 4D, an entire patch of the TLR3 cavity surface is involved in the molecular interactions with the MEVs, favoring the formation of molecular complexes between the MEVs and the TLR3 ectodomain cavity. Particular residues involved in this interaction are shown in Figure 4C and 4F (CTL:TLR3: Y496:D437, K467:H359, A521:K416, P547:K416, S545:N361, G544:K330, V565:Y307, Y538:H129, V537:N105, Y634:H108, HTL:TLR3: S629:K416, S649:Y307, G668:K330, H810:E533, H809:R484, H805:H359, N801:R325, H613:N230, N252:N718, Y701:Q107). The CTL and HTL MEVs showed the formation of multiple hydrogen bonds within the ectodomain cavity region of TLR3.

B-factor analysis of the MEV-TLR3 complexes was also performed. The B-factor indicates the displacement of the atomic positions from an average (mean) value, as in, the more flexible the atom, the larger its displacement from the mean position (mean-squares displacement) (Figure 4B, 4D). PDBsum [60] was used to calculate patches on the TLR3 receptor indicating the region of binding sites. The B-factor analysis of the CTL and HTL MEVs bound to the TLR3 receptor shows that most of the regions of the MEVs bound to TLR3 are stable. The B-factor analysis is represented by a VIBGYOR color presentation, where blue represents a low B-factor and red represents a high B-factor (Figure 4B, 4D). These results suggest tendencies toward stable complex formation for both the CTL and HTL MEVs with the ectodomain of the human TLR3 receptor.
Figure 4. Molecular docking studies of the CTL and HTL MEVs with TLR3. (A), (D): The docking complexes of CTL-TLR3 and HTL-TLR3 with patches on the TLR3 receptor indicating the region of binding sites calculated by PDBsum [60]. (C), (F): Detailed molecular interactions between the binding site residues of the CTL and HTL MEVs and TLR3 (CTL, HTL: cyan; TLR3: magenta). Hydrogen bond formation is shown by orange dotted lines. (B), (E): B-factors of the docked MEVs to the TLR3 receptor. The presentation is in VIBGYOR color, with blue showing a low B-factor and red showing a high B-factor. Most of the MEV regions are blue, showing low B-factors; this indicates the formation of stable complexes with the TLR3 receptor. CTL, cytotoxic T lymphocyte. HTL, helper T lymphocyte. TLR3, toll-like receptor 3.

Molecular Dynamics Simulation Study of the Complexes of the MEVs with TLR3

Both the complexes CTL-TLR3 and HTL-TLR3 were further subjected to molecular dynamics simulation analysis to investigate the stability of the molecular interactions involved. Both the MEV-TLR3 complexes showed very convincing and reasonably stable RMSD values for the Cα backbone, and all atoms (CTL-TLR3 complex: approximately 4-7.5 Å; HTL-TLR3 complex: approximately 3.0-9.8 Å) which stabilized toward the end (Figure 5A and 5C). The RMSDs of both complexes remained in the abovementioned RMSD range for a given time window of 10 ns at reasonably invariable temperature.
In an acceptable range (approximately 2-6 Å) (Figure 5B and 5D). These results indicate that both the CTL-TLR3 and HTL-TLR3 complexes are stable, with acceptable molecular interaction tendencies.

Figure 5. Molecular dynamics simulations of the CTL and HTL MEVs with TLR3. (A), (C): Root mean square deviations for the Cα, backbone, and all atoms for the CTL MEV-TLR3 complex and the HTL MEV-TLR3 complex. (B), (D): Root mean square fluctuations of all the amino acid residues of the CTL MEV and the HTL MEV in complex with the TLR3 immune receptor. Å: angstroms. COVID-19: coronavirus disease. CTL: cytotoxic T lymphocyte. HTL: helper T lymphocyte. MEV: multiepitope vaccine. TLR3: toll-like receptor 3. RMSD: root mean square deviation. RMSD Ca: root mean square deviation for the alpha carbon atoms. RMSD Bb: root mean square deviation for the backbone atoms. RMSD All: root mean square deviation for all atoms. RMSF: root mean square fluctuation.

In Silico Analysis of cDNA of the MEVs for Cloning and Expression Potency in a Mammalian Host Cell Line

cDNA optimized for CTL and HTL expression in a mammalian (human) host cell line was generated using the Java Codon Adaptation Tool. Further, the generated optimized cDNAs for both the MEVs were analyzed using the GenScript Rare Codon Analysis Tool. The analysis revealed that the codon-optimized cDNAs of both the CTL and HTL MEVs have crucial and favorable compositions for high-level expression in a mammalian cell line (CTL MEV: GC content 70.40%, CAI score 1.00, and 0% tandem rare codons; HTL MEV: GC content 69.26%, CAI score 1.00, and 0% tandem rare codons). Ideally, the GC content of cDNA should be 30%-70%; a CAI score that indicates the possibility of cDNA expression in a chosen expression system should be between 0.8 and 1.0; and the tandem rare codon frequency that indicates the presence of low-frequency codons in cDNA should be <30%. Tandem rare codons may hinder proper expression of the cDNA or even interrupt the translational machinery of the chosen expression system. Therefore, as per the GenScript Rare Codon analysis, the cDNAs of both the MEVs satisfy all the mentioned parameters and are predicted to have high expression in the mammalian (human) host cell line.

Discussion

Principal Findings

In the present study, we have reported the design of CTL and HTL multiepitope-based vaccine candidates against SARS-CoV-2 infection. These MEVs are composed of multiple CTL and HTL epitopes with truncated Ov-ASP-1 as an adjuvant at the N termini of both the MEVs. To design the
We have designed and proposed two MEVs derived from multiple CTL and HTL epitopes against SARS-CoV-2 (COVID-19). The chosen CTL and HTL epitopes show significant sequence overlap with screened linear B cell epitopes. The shortlisted CTL and HTL epitopes were used to design CTL and HTL MEVs. Tertiary models of both the generated CTL and HTL MEVs were shown to contain potential linear and discontinuous B cell epitopes as well as potential INFγ epitopes. Therefore, the designed MEVs are predicted to be capable of eliciting humoral and cellular immune responses. Because Ov-ASP-1 binds to APCs and triggers pro-inflammatory cytokine production via TLR3, truncated Ov-ASP-1 was used as an adjuvant at the N termini of both the CTL and HTL MEV models. The molecular interactions of the chosen overlapping clustering epitopes with their respective HLA allele binders were validated by molecular docking studies. The molecular interactions of the chosen CTL epitopes with the TAP transporter cavity were also analyzed. Analysis of the average world population coverage by both the shortlisted CTL and HTL epitopes combined revealed coverage of 96.10% of the world population. The molecular interaction analysis of both the CTL and HTL MEVs with the immunoreceptor TLR3 showed very convincing structural fitting of the MEVs into the ectodomain of the TLR3 cavity. This result was further confirmed by molecular dynamics simulation studies of both the CTL-MEV-TLR3 and HTL-MEV-TLR3 complexes, indicating tendencies toward stable molecular complex formation of both MEVs with TLR3. cDNAs for both MEVs were generated considering codon-biasing for expression in a mammalian (human) host cell line. Both cDNAs were optimized with respect to their GC content and zero tandem rare codons to increase their possibility of high expression in the mammalian host cell line (human). Therefore, for further studies, both the designed CTL and HTL MEVs could be cloned, expressed, and tested for in vivo validation and animal trials as potential vaccine candidates against SARS-CoV-2 infection.

**Conclusion**

We are thankful to the Indian Council for Medical Research (ICMR) for providing a SRF fellowship to SV (ISRM/11(98)/2019). We are also thankful to the ICMR National Institute of Malaria Research for providing infrastructural and computational facilities for the experiments. We also thank Mangalayatan University for providing infrastructural facilities for the experiments.
Authors' Contributions
The protocol was designed by SS and MK. The methodology was performed by SS, SV, MK, and RK. The global economic risk analysis was performed by RKB. Data analysis, scientific writing, and revision of the article were performed by SS, SV, MK, RK, RKB, AKS, HJS, MK, and KCP.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Supplementary material.
[PDF File (Adobe PDF File), 1840 KB - bioinform_v1i1e19371_appl1.pdf ]

References


Abbreviations

AA: amino acid
Cα: alpha carbon
cDNA: complementary DNA
COVID-19: coronavirus disease
CTL: cytotoxic T lymphocyte
E protein: envelope protein
ER: endoplasmic reticulum
HLA: human leukocyte antigen
HTL: helper T lymphocyte
IC50: half maximal inhibitory concentration
IEDB: Immune Epitope Database
IFNγ: interferon gamma
MERCI: Motif-EmeRging and with Classes-Identification
MEV: multiepitope vaccine
MHC: major histocompatibility complex
M protein: membrane protein
NCBI: National Center for Biology Information
N protein: nucleocapsid protein
ORF: Open Reading Frame
Ov-ASP-1: Onchocerca volvulus activation-associated secreted protein-1
PDB: Protein Data Bank
PI: protrusion index
RMSD: root mean square deviation
RMSF: root mean square fluctuation
SARS-CoV-2: severe acute respiratory syndrome coronavirus 2
S protein: surface protein
TAP: transporter associated with antigen processing
TLR: toll-like receptor
TLR3: toll-like receptor 3
TrEMBL: Translated European Molecular Biology Laboratory
VIBGYOR: violet-indigo-blue-green-yellow-orange-red
YASARA: Yet Another Scientific Artificial Reality Application
Original Paper

The Novel Coronavirus Enigma: Phylogeny and Analyses of Coevolving Mutations Among the SARS-CoV-2 Viruses Circulating in India

Anindita Banerjee1, PhD; Rakesh Sarkar1*, MSc; Suvrotoa Mitra1, MSc; Mahadeb Lo1, MSc; Shanta Dutta1, PhD; Mamta Chawla-Sarkar1, PhD

Indian Council of Medical Research-National Institute of Cholera and Enteric Diseases, Kolkata, India

* these authors contributed equally

Corresponding Author:
Mamta Chawla-Sarkar, PhD
Indian Council of Medical Research-National Institute of Cholera and Enteric Diseases
P-33, CIT Road, Scheme-XM, Beliaghata,
Kolkata, 700010
India
Phone: 91 9830660999
Email: chawlam70@gmail.com

Abstract

Background: The RNA genome of the emerging novel coronavirus is rapidly mutating, and its human-to-human transmission rate is increasing. Hence, temporal dissection of their evolutionary dynamics, the nature of variations among different strains, and understanding the single nucleotide polymorphisms in the endemic settings are crucial. Delineating the heterogeneous genomic constellations of this novel virus will help us understand its complex behavior in a particular geographical region.

Objective: This is a comprehensive analysis of 95 Indian SARS-CoV-2 genome sequences available from the Global Initiative on Sharing All Influenza Data (GISAID) repository during the first 6 months of 2020 (January through June). Evolutionary dynamics, gene-specific phylogeny, and the emergence of the novel coevolving mutations in 9 structural and nonstructural genes among circulating SARS-CoV-2 strains across 12 different Indian states were analyzed.

Methods: A total of 95 SARS-CoV-2 nucleotide sequences submitted from India were downloaded from the GISAID database. Molecular Evolutionary Genetics Analysis, version X software was used to construct the 9 phylogenetic dendrograms based on nucleotide sequences of the SARS-CoV-2 genes. Analyses of the coevolving mutations were done in comparison to the prototype SARS-CoV-2 from Wuhan, China. The secondary structure of the RNA-dependent RNA polymerase/nonstructural protein NSP12 was predicted with respect to the novel A97V mutation.

Results: Phylogenetic analyses revealed the evolution of “genome-type clusters” and adaptive selection of “L”-type SARS-CoV-2 strains with genetic closeness to the bat severe acute respiratory syndrome–like coronaviruses. These strains were distant to pangolin or Middle East respiratory syndrome–related coronavirus strains. With regard to the novel coevolving mutations, 2 groups have been seen circulating in India at present, the “major group” (66/95, 69.4%) and the “minor group” (21/95, 22.1%) harboring 4 and 5 coexisting mutations, respectively. The “major group” mutations fall in the A2a clade. All the minor group mutations, except 11083G>T (L37F, NSP6 gene), were unique to the Indian isolates.

Conclusions: This study highlights the rapidly evolving SARS-CoV-2 virus and the cocirculation of multiple clades and subclades. This comprehensive study is a potential resource for monitoring the novel mutations in the viral genome, interpreting changes in viral pathogenesis, and designing vaccines or other therapeutics.


KEYWORDS
SARS-CoV-2; Indian isolates; phylogeny; nucleotide homology; novel coevolving mutations; NSP12/RdRP secondary structure; genetics; virus; evolution; mutation; COVID-19; genome; epidemiology; infectious disease
Introduction

The COVID-19 pandemic caused by the novel SARS-CoV-2 was initially reported from Wuhan, China in December 2019, but it spread across the world within 3 months [1]. As of July 21, 2020, more than 14.9 million people have been found to be infected by SARS-CoV-2, with a death toll of approximately 615,939 in more than 210 countries. Phylogenetic analyses reveal that SARS-CoV-2 clusters within the subgenus Sarbecovirus under the genus Betacoronavirus and has probably undergone zoonotic transmission from the bats through the possible intermediate host Malayan pangolins, culminating among humans [2]. The positive sense, single-stranded RNA genome of SARS-CoV-2 is continuously mutating and generating multiple clades within a short time span (December 2019 to June 2020). Hence, there is a need to dissect the complex evolutionary characteristics of this novel coronavirus, identifying the single nucleotide polymorphisms (SNPs) and other mutations among strains circulating across different parts of the world. Previous reports on the genetic and evolutionary dynamics of the SARS-CoV-2 virus have tried to deduce the mode of transmission that this virus made its way into humans from bats during the early phase of the pandemic, but many questions remain unanswered even though more sequence data has been made available. Therefore, studying the heterogeneous genomic constellations within specific geographical settings will help to understand its complex epidemiology and formulate region specific strategies to curb its spread and severity.

The first 3 cases from India with travel history to Wuhan were reported in Kerala during January 2020 [3]; subsequently, 4,02,529 active cases, 724,577 recovered cases, and 28,084 deaths have been officially recorded in India as of July 21, 2020, 8 AM India Standard Time GMT +5:30 [4]. India ranks third worldwide according to the number of COVID-19 infections and is geographically vulnerable to this novel virus, as it accounts for almost 6% of global and 3.5% of COVID-19–attributable mortality. In spite of high population density, poor hygiene conditions, and an overburdened health care system, the proportion of the total infected population is much lower when compared to other western countries, that is, 0.05% in India versus 0.87% in the United States, 0.73% in Brazil, 0.46% in Russia, and 0.4% in Italy. Though the average death rate due to SARS-CoV-2 infection in India (2.46%) is comparable to that of world (eg, the United States 3.88%, Europe 6.6%), 50% of deaths in India are attributable to the age group 40-64 years [5]. Thus, to understand the phylodynamics of circulating strains in India, this study was initiated to analyze the complete viral genome sequences submitted in the Global Initiative on Sharing All Influenza Data (GISAID) [6] from 95 SARS-CoV-2 representative strains circulating across 12 differentially affected states within India. To elucidate the possible ancestry, gene-wise phylogeny of these Indian strains has been deciphered with respect to other isolates reported from Europe, the United States, and China along with coronavirus strains belonging to other genera infecting humans and other animal hosts. The novel coexisting mutations among the Indian SARS-CoV-2 strains have also been analyzed.

Through this genome analyses and phylogenetic approach, we have attempted to focus on the natural evolution of SARS-CoV-2 from its existing ancestors within the zoonotic reservoir. Furthermore, analyzing the novel mutations accumulated within the viral genome over the period with reference to the Wuhan strains (clade O) will underscore their impact on the structure and function of viral proteins.

Methods

Sequence Mining

A total of 95 SARS-CoV-2 nucleotide sequences submitted from India from January to June 2020 were downloaded from the GISAID database for phylogenetic analyses and screening of novel mutations. Several other reference gene sequences of SARS-CoV-2 as well as other types of coronaviruses were downloaded from the GenBank database submitted from several other countries for dendrogram construction and further lineage analyses.

Phylogenetic Analyses and Screening of Mutations

Nine phylogenetic dendrograms were constructed with respect to 2 structural genes (spike and nucleocapsid) and 7 nonstructural genes (nonstructural protein [NSP]2, NSP3, NSP4, NSP6, NSP7, NSP8, and NSP12). Multiple sequence alignment for all the respective set of gene sequences was done using MUSCLE v3.8.31 (drive5). Amino acid sequences were deduced through TRANSEQ (EMBL-EBI). Phylogenetic dendrograms were constructed by Molecular Evolutionary Genetics Analysis, version X (MEGAX), using the maximum-likelihood statistical method (at 1000 bootstrap replicates) and using the best fit nucleotide substitution models for each dendrogram. The best fit models were determined through model testing parameter of MEGAX. Different novel coexisting mutations in the Indian strains were identified and analyzed in comparison to the prototype SARS-CoV-2 strain from Wuhan (MN908947.3/SARS-CoV-2 Wuhan-Hu-1).

Secondary Structure Prediction of RNA-Dependent RNA Polymerase Having A97V Mutation

We used the Chou and Fasman Secondary Structure Prediction (CFSSP) online server to predict the secondary structure of RNA-dependent RNA polymerase (RdRP)/NSP12 with novel A97V mutation [7].

Results

Phylogenetic Analysis of the Structural and Nonstructural Genes

Spike Gene

Among the 95 Indian study isolates, 93 strains clustered among themselves within the same lineage of Betacoronavirus SARS-CoV-2 in 3 different subclusters (A=53, B=12, and C=28 strains), while 2 strains, 1 from Telangana (EPI_ISL_431101) and the other from Maharashtra (EPI_ISL_479550), extruded out separately, close to subcluster C containing the clade-specific strains A1, A3, B1, B2, B4-1, and B4-2 in the phylogenetic dendrogram for the spike (S) gene. The prototype SARS-CoV-2
strain belonging to the O clade (MN908947.3/SARS-CoV-2/HUMAN/CHN/Wuhan-Hu-1/2019) was present in the same lineage with the Indian strains within subcluster C (>99% identity). Subcluster A comprised of the clade-specific A2 and A2a strains along with a tiger strain from the New York zoo and another carnivorous mammal, mink SARS-CoV-2. No specific pattern of temporal distribution of strains was observed among the 3 subclusters. All the representative Indian strains had 99%-100% nucleotide sequence homology among themselves. The Indian strains had 92.8%-93% and 83.5% homology with Bat (EPI_ISL_402131 /COV /BAT /YUNNAN / RATG13 /2013) and Pangolin coronavirus (EPI_ISL_410540 /COV /PANGOLIN /GUANGXI /P5L/2017), respectively. Homology was much less (75.8%-76.7%) with other bat severe acute respiratory syndrome (SARS)-like coronavirus strains (eg, MG772933.1/ SARS-LIKE-COV/BAT /BAT-SL-COVZC45 /2017 and MG772934.1/ SARS-LIKE-COV/ BAT /BAT-SL-COVZXC21/2015), while Middle East respiratory syndrome–related coronavirus (MERS-CoV; KJ713299.1 /MERS-COV /CAMEL /SAU /KSA-CAMEL-376 /2013 and KU308549.1/MERS-COV /HUMAN /KOR /SEOUL-SNU1-035/2015) were distantly related to the Indian SARS-CoV-2 strains (52.5%-52.9% identity; Multimedia Appendix 1).

Nucleocapsid Gene

The phylogenetic dendrogram for the nucleocapsid (N) gene revealed that out of 95 Indian study isolates, 92 strains clustered within the same lineage of Betacoronavirus SARS-CoV-2 in 3 different subclusters (A=33 strains, B=47 strains, and C=12 strains), while 3 strains from Tamil Nadu (EPI_ISL_ 458040), Gujarat (EPI_ISL_ 458107), and Delhi (EPI_ISL_ 435111) extruded out separately, close to subcluster C strains. Subcluster A comprised of strains from the earlier 3 months (January, February, and March), while subcluster B contained strains from the later 3 months. Subcluster C had mixed strains. The clade-specific strains (A1, A1a, A2, A2a, A5, B1, and B4-1) as well as the prototype SARS-CoV-2 strain O clade (MN908947.3/SARS-CoV-2/HUMAN/CHN/Wuhan-Hu-1/2019) clustered near subcluster B, while B4-2, A3, and B2 clade strains were close to subcluster C. All the representative Indian strains had >99.8% nucleotide identity among themselves as well as with the different clade-specific strains. The Indian strains had 91%-97% sequence identity with bat coronaviruses (EPI_ISL_402131 /COV /BAT /YUNNAN / RATG13 /2013, MG772933.1/ SARS-LIKE-COV /BAT /BAT-SL-COVZC45 /2017, and MG772934.1/ SARS-LIKE-COV /BAT /BAT-SL-COVZXC21/2015) and 91% similarity with pangolin strains (EPI_ISL_410540 /COV /PANGOLIN /GUANGXI /P5L/2017). In contrast to bat strains, the MERS-CoV strains (KJ713299.1/MERS-COV/CAMEL /SAU /KSA-CAMEL-376 /2013 and KU308549.1/MERS-COV/HUMAN /KOR /SEOUL-SNU1-035/2015) were genetically distant to the Indian SARS-CoV-2 strains (56.9%-57.3% identity; Multimedia Appendix 2).

RNA-Dependent RNA Polymerase Gene (RdRP/NSP12)

The phylogenetic dendrogram for the RdRP/NSP12 gene depicted that, among the 95 Indian study isolates, 93 strains clustered within the same lineage of Betacoronavirus SARS-CoV-2 into 3 subclusters (A=64 strains, B=22 strains, and C=7 strains). Two strains, one from Kerala (EPI_ISL_ 413523) and the other from Delhi (EPI_ISL_ 435111), were placed distant to these 3 subclusters in the dendrogram and were close to A1a, A2, A3, A5, B1, B2, B4-1, B4-2, and the prototype O clade strains. Subcluster A strains clustered with the A2a clade-specific strain while subcluster C clustered with A1. No temporal specificity was observed among the 3 subcluster strain distributions. All the Indian strains had >99.8% nucleotide identity among themselves as well as the different clade-specific strains. The prototype SARS-CoV-2 strain O clade (MN908947.3 /SARS-CoV-2 /HUMAN /CHN /Wuhan-Hu-1 /2019) was distant to all the 3 subclusters. The Indian strains had 97.8% sequence homology with bat coronavirus (EPI_ISL_402131 /COV /BAT /YUNNAN / RATG13/2013) and 86.7%-88.6% similarity with both pangolin (EPI_ISL_410540 /COV /PANGOLIN /GUANGXI /P5L/2017) and other bat SARS-like coronavirus strains (MG772933.1/ SARS-LIKE-COV /BAT /BAT-SL-COVZC45 /2017 and MG772934.1 /SARS-LIKE-COV /BAT /BAT-SL-COVZXC21/2015). MERS-CoV strains (KJ713299.1 /MERS-COV /CAMEL /SAU /KSA-CAMEL-376/2013 and KU308549.1/MERS-COV /HUMAN /KOR /SEOUL-SNU1-035/2015) were distantly related to the Indian SARS-CoV-2 strains (68.1% identity; Multimedia Appendix 2).

NSP2, NSP3, NSP4, NSP6, NSP7, and NSP8 Genes

The dendrograms of all these 6 genes showed a similar pattern. All the 95 Indian strains clustered in 2 subclusters (A=39 and B=56 strains) within the Betacoronavirus lineage of SARS-CoV-2. Principally, subcluster A strains were from the first 3 months, whereas B contained strains from the next 3 months of 2020. Strains of subcluster A and B had 99.9%-100% DNA homology among themselves. All the clade-specific strains (A1, A1a, A2, A2a, A3, A5, B1, B2, B4-1, and B4-2) along with the prototype SARS-CoV-2 strain clade O (MN908947.3/SARS-CoV-2 /HUMAN /CHN /Wuhan-Hu-1/2019) clustered close to subcluster A (99.9% identity), except NSP7 and NSP8 where the prototype clade O strain was present within subcluster B strains. SARS-CoV-2 strains isolated from carnivorous mammals like mink and tiger also grouped close to the subcluster A strains in all the dendrograms (99.9% identity). Subcluster A and B strains revealed 95.4%-98.1% nucleotide sequence similarity with bat coronavirus (EPI_ISL_402131 /COV /BAT /YUNNAN / RATG13 /2013, while the pangolin-derived strain EPI_ISL_410540 /COV /PANGOLIN /GUANGXI /P5L/2017 showed less identity (83%-87.5%). MERS-CoV strains NC_019843.3/MERS-COV/HUMAN/NLD/HCOV-EMC/2012 and KU740200.1 /MERS-COV /CAMEL /EGYPT /NRCE-NC163/2014 exhibited a significant phylogenetic distance (only 49.6%-60.8% homology) from the Indian isolates (Multimedia Appendices 4-9).

L- and S-Type of SARS-CoV-2

SNPs at positions 8782 (NSP4 gene) and 28,144 (open reading frame [ORF]8) showed complete linkage among the Indian
isolates under study. At these two sites, 93 strains showed a “CT” haplotype (designated as “L” type as T28,144 falls in the codon position which encodes amino acid leucine in the 84th position of ORF8 protein), while only 2 strains (1 from Kerala, EPI_ISL_413523, and 1 from Delhi, EPI_ISL_435111) revealed a “TC” haplotype (called as “S” type as C28,144 falls in the codon encoding serine at the 84th position of the ORF8 protein).

Analyses of Synonymous and Nonsynonymous Mutations

The Common Mutations in SARS-CoV-2 Indian Isolates

To explore the mutations among the 95 SARS-CoV-2 strains, we performed in-depth sequence analyses both at the genome level and at the corresponding amino acid level in different proteins, especially S glycoprotein, N protein, NSP2, NSP3, NSP4, NSP6, NSP7, NSP8, and RdRP/NSP12 with reference to the prototype SARS-CoV-2 strain (MN908947.3/SARS-CoV-2/HUMAN/CHN/Wuhan-Hu-1/2019). Out of 95 samples, 2 (2.1%) were found to have no significant “L”-type mutations (EPI_ISL_435111 and EPI_ISL_413523). Out of 93 “L”-type samples, 6 (6.3%; EPI_ISL_481156, EPI_ISL_476840, EPI_ISL_476023, EPI_ISL_458080, EPI_ISL_431101, and EPI_ISL_413522) harbored none of the mutations and were wild-type–like. Mutational analysis of the remaining 87 strains revealed circulation of two predominant “groups,” namely, the “major group” and the “minor group,” across India. The “major group,” which, of 95 isolates, was comprised of 66 (69.4%), revealed 4 coexisting SNPs: 241C>T in the five prime untranslated region (5′ UTR), 3037C>T (F106F) in the NSP3 gene, 14408C>T (P323L) in the NSP12 gene, and 23403A>G (D614G) in the S gene (Table 1). This “major group” of SARS-CoV-2 was predominantly found to circulate in regions like Delhi, Maharashtra, West Bengal, Odisha, Telangana, and Gujarat. The other 21 (22.1%) samples, which represent the “minor group,” harbored 5 coexisting mutations: 23929C>T (Y789Y) in the S gene, 28311C>T (P13L) in the N gene, 6312C>A (T1198K) in the NSP3 gene, 11083G>T (L37F) in the NSP6 gene, and 13730C>T (A97V) in the NSP12/RdRP gene (Table 2). Needless to say, the 5 coexisting mutations of the “minor group” and the 4 coexisting mutations of the “major group” did not overlap among the same SARS-CoV-2 strains. The “minor group” of SARS-CoV-2 predominated across Tamil Nadu (South) and Uttar Pradesh (Central/North).
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Delhi (n=13)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_435061, EPI_ISL_435062, EPI_ISL_482665 (n=3)</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>EPI_ISL_482498 (n=1)</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>EPI_ISL_435065- EPI_ISL_435069 (n=5)</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>EPI_ISL_435070, EPI_ISL_435071 (n=2)</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>EPI_ISL_435063, EPI_ISL_435064 (n=2)</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td><strong>Tamil Nadu (n=4)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_458032, EPI_ISL_458033, EPI_ISL_458044, EPI_ISL_458040 (n=4)</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td><strong>Maharashtra (n=13)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Single nucleotide polymorphisms associated with the major group SARS-CoV-2 strains (n=66) across India from January to June 2020.⁵
<table>
<thead>
<tr>
<th>State and accession number</th>
<th>Spike glycoprotein (21,563-25,384 nts/1273 amino acids)</th>
<th>RdRP(^b) protein (13,442-16,236 nts/932 amino acids)</th>
<th>NSP(^3) protein (2720-8554 nts/1945 amino acids)</th>
<th>5(^-)UTR(^d) (1-265 nt)</th>
<th>NSP(^2) protein (806-2719 nts/638 amino acids)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q271R(^f)</td>
<td>D64C(^f)</td>
<td>G12A(^h)</td>
<td>P323L(^i)</td>
<td>241C(^T)</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EPI-ISL_479493, EPI-ISL_479510, EPI-ISL_479553, EPI-ISL_479533, EPI-ISL_479538, EPI-ISL_479497, EPI-ISL_479550, EPI-ISL_479554, EPI-ISL_479557, EPI-ISL_479560, EPI-ISL_479562, EPI-ISL_479571, EPI-ISL_479564 (N=13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>West Bengal (n=5)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI-ISL_430466, EPI-ISL_430467 (n=2)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EPI-ISL_430465 (n=1)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EPI-ISL_430468, EPI-ISL_430464 (n=2)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Gujarat (n=11)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI-ISL_458107, EPI-ISL_483878, EPI-ISL_476869, EPI-ISL_469036, MT576031 (n=5)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

(N=13)
<table>
<thead>
<tr>
<th>State and accession number</th>
<th>Spike glycoprotein (21,563-25,384 nts/1273 amino acids)</th>
<th>RdRP protein (13,442-16,236 nts/932 amino acids)</th>
<th>NSP3 protein (2720-8554 nts/1945 amino acids)</th>
<th>5'UTR蛋白 (1-265 nts/500 coding)</th>
<th>N° protein (28,274-29,533 nts/419 amino acids)</th>
<th>NSP2 protein (806-2719 nts/638 amino acids)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPI_ISL_461484,</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_476864 (n=2)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_471637 (n=1)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_475058 (n=1)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_426414,</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_426415 (n=2)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Odisha (n=7)**

| EPI_ISL_481154,           | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_481157 (n=2)      | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_481115,           | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_463078,           | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_481177,           | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_481180,           | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_481186 (n=5)      | ✓                                                       |                                               |                                             |                                  |                                               |                                  |

**Madhya Pradesh (n=2)**

| EPI_ISL_476884,           | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_476842 (n=2)      | ✓                                                       |                                               |                                             |                                  |                                               |                                  |

**Telangana (n=5)**

<p>| EPI_ISL_458080,           | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_431101 (n=2)      | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_431117 (n=1)      | ✓                                                       |                                               |                                             |                                  |                                               |                                  |
| EPI_ISL_471588 (n=1)      | ✓                                                       |                                               |                                             |                                  |                                               |                                  |</p>
<table>
<thead>
<tr>
<th>State and accession number</th>
<th>Spike glycoprotein (21,563-25,384 nts/1273 amino acids)</th>
<th>RdRP&lt;sup&gt;b&lt;/sup&gt; protein (13,442-16,236 nts/932 amino acids)</th>
<th>NSP&lt;sup&gt;c&lt;/sup&gt;3 protein (2720-8554 nts/1945 amino acids)</th>
<th>5'-UTR&lt;sup&gt;d&lt;/sup&gt; (1-265 nt) non-coding</th>
<th>N&lt;sup&gt;e&lt;/sup&gt; protein (28,274-29,533 nts/419 amino acids)</th>
<th>NSP2 protein (806-2719 nts/638 amino acids)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Karnataka (n=5)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_471629&lt;sup&gt;a&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_477207&lt;sup&lt;b&gt;&lt;/b&gt;&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_477250&lt;sup&gt;n&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_477255&lt;sup&gt;n&lt;/sup&gt;</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_477237&lt;sup&gt;n&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_477239&lt;sup&gt;n&lt;/sup&gt;</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Uttar Pradesh (n=1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPI_ISL_435060&lt;sup&gt;n&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Mutations were analyzed with compared to Wuhan-Hu-1 (MN908947.3).

<sup>b</sup>RdRP: RNA-dependent RNA polymerase.

<sup>c</sup>NSP: nonstructural protein.

<sup>d</sup>5' UTR: five prime untranslated region.

<sup>e</sup>N: nucleocapsid.

<sup>f</sup>22374A>G.

<sup>g</sup>23403A>G.

<sup>h</sup>24933G>T.

<sup>i</sup>24444C>T.

<sup>j</sup>14408C>T.

<sup>k</sup>3037C>T.

<sup>l</sup>5700C>A.

<sup>m</sup>6466A>G.

<sup>n</sup>28854C>T.

<sup&o=</sup>28881-28883 GGG>AAC.

<sup>p</sup>28396G>A.

<sup>q</sup>29451C>T.

<sup>r</sup>1059C>T.
Table 2. Single nucleotide polymorphisms associated with the minor group SARS-CoV-2 strains (n=21) across India during January to June 2020.a

<table>
<thead>
<tr>
<th>State and accession number</th>
<th>Spike glycoprotein (21,563-25,384 nts/1273 amino acids)</th>
<th>RdRPb protein (13,442-16,236 nts/932 amino acids)</th>
<th>Nc protein (28,274-29,533 nts/419 amino acids)</th>
<th>NSP3 protein (2720-8554 nts/1945 amino acids)</th>
<th>NSP6 protein (10,973-11,842 nts/290 amino acids)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamil Nadu (n=8)</td>
<td>Y789Y (23929C&gt;T)</td>
<td>A97V (13730C&gt;T)</td>
<td>P13L (28311C&gt;T)</td>
<td>S1197R (6310C&gt;A)</td>
<td>T1198K (6312C&gt;A)</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>EPI_ISL_435093</td>
<td>EPI_ISL_435096</td>
<td>EPI_ISL_435084</td>
<td>EPI_ISL_435087 (n=6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPI_ISL_435091</td>
<td>EPI_ISL_435092 (n=2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maharashtra (n=1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odisha (n=2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPI_ISL_463017</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPI_ISL_463010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madhya Pradesh (n=1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telangana (n=1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karnataka (n=4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uttar Pradesh (n=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bihar (n=1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPI_ISL_435110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPI_ISL_435099</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPI_ISL_435082 (n=3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[a] Mutations were analyzed with compared to Wuhan-Hu-1 (MN908947.3).
[c] N: nucleocapsid.

The Unique Mutations in SARS-CoV-2 Indian Isolates

In addition to 23403A>G (D614G), 3 uncommon mutations, 23374A>G (Q271R), 24933G>T (G1124V), and 22444C>T (D294D), were also observed in the S gene of the “major group” (Table 1). Out of the 67 isolates of the major group, 28 revealed 4 novel mutations: 28854C>T (S194L; n=13), 28881-28883GGG>AAC (R203K and G204R; n=13), and coevolving mutation 29451C>T (T393I) and 28395G>A (R41R; n=2) in the N gene (Table 1). Intriguingly, 28854C>T (S194L) in the N gene was found to coevolve with the 22444C>T (D294D) mutation in the S gene of 11 samples in the major group (Table 1). We also observed 1059T>A (T85I) change within the NSP2 gene (n=2) and 6466A>G (K1249K) change in the NSP3 gene (n=2). With the 3 samples of the minor group, 6310C>A (S1197R) was found to be associated. No mutations were found within the NSP7 and NSP8 genes.

Effect of Missense Mutation A97V on the Secondary Structure of NSP12/RdRP

RdRP is the crucial enzyme for both viral RNA replication and maintenance of genomic fidelity. Thus, any significant change in RdRP structure could affect its functions, leading to an increase in the rate of mutagenesis in the genome. We have identified 2 missense mutations in the RdRP protein: P323L associated with the “major group” isolates and A97V associated with the “minor group” isolates. The effect of P323L on the secondary structure of RdRP has already been described [8].
Therefore, we analyzed the effect of novel mutation A97V on the secondary structure of RdRP by using the CFSSP server. The A97V mutation resulted in substitution of α-helices at positions 94, 95, and 96 within the β-sheets in the RdRP secondary structure, which may alter its tertiary conformation and affect functionality (Multimedia Appendix 10).

Discussion

Principal Findings

The molecular and genetic characterization of SARS-CoV-2 pandemic strains worldwide has been studied by several scientific groups based on whole-genome sequencing [9,10]. Through this comprehensive analysis, we aimed to closely investigate the ancestry, evolutionary dynamics, accumulation of rapid mutations, and cross-genetic translation among the emerging SARS-CoV-2 strains across India. Rapid accumulation of several point mutations across the genome of SARS-CoV-2 since its origin is a prime driving force behind the evolution of different monophyletic clades. As depicted through the different phylogenetic dendrograms in our study, a monophyletic clade of all SARS-CoV-2 strains was seen with the prototype strain (Wuhan IME-WH01/2019). Clustering of all the Indian isolates with other SARS-CoV-2 strains reported worldwide (99.8%–100% nucleotide sequence identity) suggests the introduction of this virus in India was from several countries. The clustering pattern of the prototype strain from Wuhan in the phylogenetic dendrogram underscores the fact that China might have served as the origin of this zoonotic virus, which was eventually transmitted worldwide [11-13].

The origin of SARS-CoV-2 is still undetermined, but identification of its intermediate host is much needed to prevent further dissemination and interspecies transmission in the near future. Hence, we initiated this study as one of the first in India to decipher the gene-wise phylogenetics of SARS-CoV-2 strains circulating in this endemic setting. The results depicted genome-type clusters of the 95 Indian isolates, for the structural genes S and N, and the nonstructural gene RdRP/NSP12. Clustering of the study isolates with different clade-specific genome-type clusters of the 95 Indian isolates, for the structural genes S and N, and the nonstructural gene RdRP/NSP12.

Convoluted mutational analysis also revealed cocirculation of 2 groups of mutated SARS-CoV-2 strains in India. The “major group” of SARS-CoV-2 strains (66/95, 69.4%) represents the A2a clade reported previously from Africa, South America, Oceania, and South and West Asia, comprising of strains with coevolving mutations like 241 C>T (5’ UTR), 3037 C>T (F106F, NSP3), 14403 C>T (P323L, RdRP/NSP12), and 23403 A>G (D614G, S glycoprotein) [18-20]. Certain strains in the “major group” displayed 22374A>G (Q271R), 24933G>T (G1124V), and 22444C>T (D294D) changes in the S gene, which were unique to India. Missense mutations, Q271R and G1124V in the S protein, were found to reside around the N-linked glycosylation sites 282 and 1134, respectively, and these might affect the protein function [21]. It was not surprising to observe the triple site mutation 28881-28883 GGG>AAC (R203K and G204R) in the N gene of 13 SARS-CoV-2 strains of the “major group.” This has previously been reported from Mexico, South America, Australia, New Zealand, and a few Asian countries [22]. The 203/204 region is part of the SR dipeptide domain of the N protein (SRNSSRNSTPGSSRGTSPARMA) and changes in arginine at position 203 to lysine; and glycine at position 204 to arginine resulted in the insertion of a lysine residue between serine and arginine (SRNSSRNSTPGSSRKTSPARMA), which might interfere with the phosphorylation at serine residue required for normal functioning of the N protein [23]. This mutation demands particular attention as reduced pathogenicity has been observed previously in SARS-related coronavirus on deletion of the SR domain [24]. Mutations observed in the NSP3 gene at positions 6310 C>A (S197R), 7392 C>T (P1558L), and 6466 A>G (K1249K) were completely unique to Indian strains. Few infrequent mutations at position 1059 T>A (T85I) in NSP2 and 8782 C>T (S76S) in NSP4 observed here have also been reported to be prevalent in other countries [19,22,25].

The “minor group” of Indian SARS-CoV-2 (21/95, 22.1%) was comprised of strains with 5 coevolving mutations: 13730C>T (L144L), 14403 C>T (P323L, RdRP/NSP12), 23403 A>G (D614G, S glycoprotein) [18-20], 241 C>T (5’ UTR), and 3037 C>T (F106F, NSP3). All the “minor group” mutations were novel among the Indian SARS-CoV-2 strains. The “minor group” of Indian SARS-CoV-2 (21/95, 22.1%) was comprised of strains with 5 coevolving mutations: 13730C>T (L144L), 14403 C>T (P323L, RdRP/NSP12), and 23403 A>G (D614G, S glycoprotein) [18-20], 241 C>T (5’ UTR), and 3037 C>T (F106F, NSP3). The “minor group” of Indian SARS-CoV-2 (21/95, 22.1%) was comprised of strains with 5 coevolving mutations: 13730C>T (L144L), 14403 C>T (P323L, RdRP/NSP12), and 23403 A>G (D614G, S glycoprotein) [18-20], 241 C>T (5’ UTR), and 3037 C>T (F106F, NSP3). All the “minor group” mutations were novel among the Indian SARS-CoV-2 strains. The “minor group” of Indian SARS-CoV-2 (21/95, 22.1%) was comprised of strains with 5 coevolving mutations: 13730C>T (L144L), 14403 C>T (P323L, RdRP/NSP12), and 23403 A>G (D614G, S glycoprotein) [18-20], 241 C>T (5’ UTR), and 3037 C>T (F106F, NSP3). All the “minor group” mutations were novel among the Indian SARS-CoV-2 strains.
The L37F mutation strongly implies positive selection toward evolution of Betacoronaviruses, indicating a possible origin of the "minor group" out of this positive selection, with subsequent acquisition of mutations among the strains already harboring the 11083G>T change [25,26]. The interaction of NSP6 with NSP3 and NSP4 has been described to be essential for the formation of double membrane vesicles [25,26]. Hence, it is interesting to note the presence of a coexisting mutation 6312 C>A (T1198K) in NSP3 of the "minor group" strains, though the significance of this coexistence (L37F and T1198K) in context to the NSP6-NSP3 interaction can only be confirmed through association studies. The functional accuracy of RdRP is challenged due to the presence of the 13730 C>T (A97V) change, which was predicted to have significant effect on the secondary structure of RdRP [8]. In addition, A97V was found to be located in the nidovirus RdRP-associated nucleotidyl transferase domain whose function remains unknown [27]. The P13L mutation is located in the intrinsically disordered region of the N protein and might affect RND-binding activity of the N-terminal domain and C-terminal domain of the N protein [28,29].

Any significant mutation in the RdRP/NSP12 protein might alter replication machinery, thereby compromising the fidelity of viral RNA replication and subsequent accumulation of plausible novel mutations. The missense mutation 14408C>T (P323L) in RdRP was first observed in Italy (Lombardy) in February 2020. Few strains from Europe and North America since February 2020 have shown the emergence of mutations like 3037C>T (F106F, NSP3), 23403A>G (D614G), and 28881-28883GGG>AAC (R203K and G204R, N) in the SARS-CoV-2 genome harboring the 14408C>T (P323L) mutation within the RdRP gene, suggesting a probable association or coexistence of 14408C>T (P323L) and the emerging higher number of novel point mutations compared to viral genomes from Asia [30]. Therefore, we can assume that two mutations, 14408C>T (P323L) and 13730C>T (A97V), which were found to have significant influence on the secondary structure of RdRP, could play key roles in the simultaneous establishment of “two groups” of SARS-CoV-2 with several characteristic “co-evolving mutations” in India (Asia). However, this needs to be validated experimentally. A recent study reported that the frequency of mutations within the SARS-CoV-2 genome varies in different geographical areas, as SARS-CoV-2 gene sequences from Europe and North America present an overwhelming mutation frequency compared to that of Asia. Their study identified few recurrent mutations among isolates from Europe that were not detected among the viruses circulating within Asian countries, such as 3037C>T (F106F/NSP3 gene), 14408C>T (P323L/RdRP gene), 28881-28883GGG>AAC (R203K and G204R/N gene), and 23403A>G (D614G/S gene) [30]. In contrast, our analyses revealed that all these mutations accumulated over time beyond Europe and were profoundly seen among the “major group” of SARS-CoV-2 strains circulating across India (Asia).

The free availability of genome sequences in the publicly available servers like National Center for Biotechnology Information and GISAID has revolutionized the genome studies, resulting in continuous monitoring of mutations, recombination events, development of molecular diagnostics, identification of vaccine strains, etc. The ongoing deadly pandemic requires recording the complete patient metadata along with full genome sequences of the SARS-CoV-2 strains for better understanding of the epidemiology and virulence of this virus. Exploiting newer technologies that could help in recording additional information such as specific disease traits (comorbidity, respiratory scores, essential blood parameters), treatment, requirement of hospitalization or outpatient treatment, treatment outcome, life-threatening complication, or mortality in addition to the full viral genome sequences. This would also help in geographical region-based decisions regarding treatment modalities as well as inclusion of highly virulent subtypes of strains in vaccine formulations.

**Conclusion**

India harbors a greater risk of community transmission of COVID-19 due to high population density, a large population below the poverty line, and overburdened health care facilities. Hence, stringent surveillance and monitoring of the viral epidemiology and genetic diversity of a novel virus can pave way for better health care strategies and vaccine designing. This study provides comprehensive analysis of the ancestry, evolutionary dynamics, clade-specific genetic variations, as well as development of unique coevolving mutations among SARS-CoV-2 strains circulating across different regions in India. Owing to the lack of patient metadata, the impact of novel mutations on the clinical outcome or the difference in virulence of the two distinct groups of circulating strains in India could not be determined.

**Acknowledgments**

This study was supported by an intramural grant from the Indian Council of Medical Research, New Delhi, India. The authors acknowledge the hard work and dedication of scientists and laboratory staff in COVID-19 testing and next generation genome sequencing labs.

RS and ML were supported by fellowships from the University Grants Commission and Council of Scientific and Industrial Research, India, respectively.

**Conflicts of Interest**

None declared.
Multimedia Appendix 1
Molecular phylogenetic analysis by the maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of S gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. The scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model with gamma distribution having invariant sites (GTR+G+I). S: spike.

[PNG File, 4745 KB - bioinform_v1i1e20735_app1.png]

Multimedia Appendix 2
Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of N gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model with gamma distribution having invariant sites (GTR+G+I). N: nucleocapsid.

[PNG File, 4999 KB - bioinform_v1i1e20735_app2.png]

Multimedia Appendix 3
Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of RdRP/NSP12 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general time reversible model with gamma distribution (GTR+G). NSP12: nonstructural protein 12; RdRP: RNA-dependent RNA polymerase.

[PNG File, 4475 KB - bioinform_v1i1e20735_app3.png]

Multimedia Appendix 4
Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of the NSP2 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the general Time Reversible model (GTR). NSP3: nonstructural protein 3.

[PNG File, 4291 KB - bioinform_v1i1e20735_app4.png]

Multimedia Appendix 5
Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of the NSP3 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the Tamura-3 model having invariant sites (T92+I). NSP6: nonstructural protein 6.

[PNG File, 4466 KB - bioinform_v1i1e20735_app5.png]

Multimedia Appendix 6
Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of NSP4 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the Tamura-3 model having invariant sites (T92+I). NSP6: nonstructural protein 6.

[PNG File, 4466 KB - bioinform_v1i1e20735_app6.png]
Molecular phylogenetic analysis by maximum likelihood method. Phylogenetic dendrogram based on nucleotide sequences of NSP7 gene of SARS-CoV-2 strains circulating in India during early 2020, with other known strains of respective genotype. The representative Indian strains have been marked with a solid circle. Scale - bar was set at 0.1 nucleotide substitutions per site. Bootstrap values of less than 70% are not shown. The best fit model, which was used for constructing the phylogenetic dendrogram, was the Tamura-3 model with gamma distribution (T92+G). NSP7: nonstructural protein 8.


Abbreviations

CFSSP: Chou and Fasman Secondary Structure Prediction
GISAID: Global Initiative on Sharing All Influenza Data
MEGAX: Molecular Evolutionary Genetics Analysis, version X
MERS-CoV: Middle East respiratory syndrome–related coronavirus
N: nucleocapsid
NSP: nonstructural protein
ORF: open reading frame
RdRP: RNA-dependent RNA polymerase
S: spike
SARS: severe acute respiratory syndrome
SNP: single nucleotide polymorphism
5′ UTR: five prime untranslated region
Molecular Docking Using Chimera and Autodock Vina Software for Nonbioinformaticians

Sania Safdar Butt¹, MS; Yasmin Badshah¹, MPhil; Maria Shabbir¹, PhD; Mehak Rafiq², PhD

¹Atta Ur Rahman School of Applied Biosciences, National University of Sciences and Technology, Islamabad, Pakistan
²Research Centre for Modelling and Simulation, National University of Sciences and Technology, Islamabad, Pakistan

Corresponding Author:
Mehak Rafiq, PhD
Research Centre for Modelling and Simulation
National University of Sciences and Technology
H-12
Islamabad
Pakistan
Phone: 92 519085733
Email: mehak@rcms.nust.edu.pk

Abstract

In the field of drug discovery, many methods of molecular modeling have been employed to study complex biological and chemical systems. Experimental strategies are integrated with computational approaches for the identification, characterization, and development of novel drugs and compounds. In modern drug designing, molecular docking is an approach that explores the confirmation of a ligand within the binding site of a macromolecule. To date, many software and tools for docking have been employed. AutoDock Vina (in UCSF [University of California, San Francisco] Chimera) is one of the computationally fastest and most accurate software employed in docking. In this paper, a sequential demonstration of molecular docking of the ligand fisetin with the target protein Akt has been provided, using AutoDock Vina in UCSF Chimera 1.12. The first step involves target protein ID retrieval from the protein database, the second step involves visualization of the protein structure in UCSF Chimera, the third step involves preparation of the target protein for docking, the fourth step involves preparation of the ligand for docking, the fifth step involves docking of the ligand and the target protein as Mol.2 files in Chimera by using AutoDock Vina, and the final step involves interpretation and analysis of the docking results. By following the guidelines and steps outlined in this paper, researchers with no previous background in bioinformatics research can perform computational docking in an easier and more user-friendly manner.

(Keywords: virtual screening; molecular docking; Autodock Vina; UCSF Chimera; nonbioinformaticians; binding affinity; protein database; computer-aided docking, free offline docking; drug discovery)

Introduction

In the modern era of pharmaceutical research, many methods of molecular modeling have been employed to study complex chemical and biological systems in a variety of programs of drug discovery. It is very important to integrate experimental strategies into computational approaches in the identification, characterization, and development of novel and propitious compounds. Molecular docking is an approach used extensively in modern drug designing and development; it explores the conformations of ligands within the macromolecular target binding site, providing an estimation of receptor-ligand binding free energy for all different conformations. Small molecular compounds (ligands) are docked into the binding site of the receptor, following which the binding affinity of the complex is estimated. This constitutes a significant part of the structure-based drug design process. For a thorough understanding and estimation of the ligand/protein complex, the ability to visualize the binding interactions and geometries by using a fast and accurate protocol for docking is required [1].

To date, a variety of algorithms for docking are available, which can lead to a better understanding of the benefits and drawbacks of these methods. However, most of the free tools rely on the knowledge of the command-line interface. For biologists, this is a laborious process and hence they avoid it. The proper
estimation of each method can lead to the development of plausible strategies and the origination of reproducible and relevant results.

AutoDock and AutoDock Vina (The Scripps Research Institute) are some of the most widely used, free, open-source tools for molecular docking simulations [2]. AutoDock is a collection of command-line programs that can be employed to predict binding conformations of a small flexible ligand to a macromolecular target whose structure is known. This technique combines the rapid grid-based method used for energy evaluation with conformation searching and simulated annealing.

AutoDock 4 was used for molecular docking previously. The new AutoDock Vina has a more accurate binding algorithm that can speed up the rate by approximately 2 orders of magnitude as compared to AutoDock 4. In addition, AutoDock 4 has significantly improved predictions of binding mode, assessed by the training tests employed in the AutoDock 4. By the use of multithreading on the multicore machines, faster processing can be achieved from parallelism. AutoDock Vina clusters the results for the user in a transparent way and automatically calculates the grid maps.

The UCSF (University of California, San Francisco) Chimera software is used for visualization as well as analysis of density maps, 3D microscopy, molecular structures, and the associated data [3]. The challenges in the scope, size, and types of data used with the experimental cutting-edge methods are addressed by this software. It provides advanced options for high-quality rendering (reliable calculations of the molecular surface, interactive ambient occlusion, etc) and provides professional approaches to the design and distribution of the software. Chimera is a freely available software for noncommercial use and shows advances particularly in its performance, extensibility, visualization, and usability.

Chimera is segmented into major components: a core that has a role in providing visualization and basic services and extensions that have a higher-level functionality. Two major extensions of Chimera are very important: the first one is the multiscale, which can visualize the molecular assemblies of large-scale components such as the viral coats, and the second one is collaborative interface, which allows sharing of the chimera session interactively, despite being at separate locales. The other extensions of chimera include the Multaligln Viewer, which shows multiple sequence alignments and the associated structures, the Movie that replays the trajectories of molecular dynamics, the Volume Viewer that is responsible for displaying and analyzing the volumetric data, and ViewDock that screens the docked ligand orientations. Chimera is available for all operating systems. It can be freely used by academic and nonprofit users.

For the purpose of this protocol, Akt and flavonoid fisetin are used. Protein kinase B, also known as Akt, is a serine/threonine-specific protein that regulates cell growth and survival [4]. In various cancers, the PI3K or Akt signaling cascade is upregulated and linked with enhanced progression and proliferation of cancer cells. Akt is an important part of signaling cascades for cell endurance and growth throughout the progression and proliferation in cancer. It controls the cell cycle, growth, and survival by indirectly altering cyclin D1 levels and directly activating inhibitors of cyclin-dependent kinases (WAF1/p21 and KIP1/p27) [5].

The plant-derived flavonoid named fisetin present in various edible natural sources is reported to possess antiproliferative potential [6]. Invasion, proliferation, and metastatic growth are inhibited significantly by the use of various concentrations of fisetin, especially in lung cancer. Current research has reported that the PI3K/Akt cascade is a direct target of fisetin in human cells, which is a hallmark for growth and survival [7].

The tools employed in Chimera are robust, simple, and interactive, and the computations involved take a few seconds. The major benefit of Chimera is that it integrates a large collection of interactive methods. These tools also play a role in the preparation of input and examination of results from more specialized, complex, and noninteractive algorithmic analysis software. Both the interactive and the noninteractive analyses are beneficial.

**Methods**

**Requirements for Docking**

Docking requires the following: (1) Windows 7, 8, or 10 or Mac operating system and Linux, and (2) UCSF Chimera 1.12.

**Instructions**

The stepwise instructions for docking are provided below:

**Retrieval**

Retrieve the required target protein structure from the major database Protein Data Bank (PDB) [8,9] as a PDB file.

**Use of UCSF Chimera for Docking the Target Protein**

UCSF Chimera is an extensible program that is meant mainly for visualization and analysis of the molecular structures. However, in this paper, we are operating Autodock Vina in Chimera for docking purposes.

1. Click on the file and fetch by ID, as shown in Figure 1.
2. Input the PDB ID of the protein (Akt: 3QKK). Figure 2 presents a screenshot of how to obtain the protein structure through PDB ID in Chimera. Any protein can be fetched by inserting the PDB ID of the protein.
3. When the protein is fetched, its structure is downloaded through the website; hence, a working internet connection is required, or the PDB file can be downloaded beforehand and simply be opened thorough File > open. Figure 3 displays the Akt structure retrieved in UCSF Chimera.
4. Create a working directory for the docking project that is convenient to access, such as Users/Desktop/Docking/. Start saving all your prepared files there, for example, save 3QKK as Akt.pdb.
Figure 1. Screenshot of the process to fetch/deliver a protein structure in Chimera.
Figure 2. Screenshot of the retrieved protein structure of Akt from the RCSB protein database.
Preparing the Target Protein for Docking

1. To easily define the active site, the already present inhibitor needs to be identified. To do so, select the inhibitor by click on Select > Residue > SMH (nonstandard residue), as seen in Figure 4. In this screenshot, Akt bears an HOH group and SMH residues as nonstandard residues. Due to the selection, SMH appears to be highlighted in green.
2. After selecting the nonstandard (inhibitor) residues, the residues must be accorded a color. To distinguish the chosen residue from the rest of the protein (Figure 5), change the color by clicking on Actions > Color > red (any color of your choice).
3. The protein needs to be optimized for docking. Click on Tools > Structure Editing > Dock Prep (Figure 6). The required dock prep tools are all available within Chimera. These dock prep tools are available in the structure editing file menu option.
4. In the dock prep box, select all options except “Delete non-complexed ions” and click OK (Figure 7).
5. Add hydrogen to the proteins by selecting the appropriate following options and click OK (Figure 8). We allow the program to make the best choice according to the model by selecting the abovementioned options.
Figure 8. Screenshot of adding hydrogen atoms to the protein.

6. Assign charges to the protein by clicking on the Gasteiger charges (Figure 9) and click OK.
7. Select the net charges (Figure 10) and click OK. For most proteins, the net charges equal to zero.
Preparing the Ligand for Docking

Similar to the process of obtaining the protein, drugs with Pubchem compound ID (CID) can be fetched through the software with a working internet connection.

1. Click on Structure Editing > Build Structure > PubChem CID or you can even insert the simplified molecular-input line-entry system (SMILES) of the novel compound being used. Figure 11 shows how to fetch ligands from PubChem using its ID.
2. Enter the PubChem CID and click apply.
3. The ligand needs to be optimized as the protein was optimized. Click on Tools > Structure Editing > Dock Prep, and repeat the same steps followed for preparing the protein. These steps include removing solvents, adding hydrogens, and determining the charge. Figure 12 shows an overview of the dock prep for the ligand.
4. The ligand Fisetin is saved as prep_fisetin.mol2 file in the working directory earlier created (Figure 13).
Figure 11. Screenshot of fetching the ligand compound fisetin through its PubChem ID in Chimera.
Figure 12. Screenshot of preparing the ligand for docking.
**Docking**

The following steps outline the process for docking:

1. Click on Tools > Surface or Binding Analysis > Autodock Vina (Figure 14).
2. We will set up the grid box values on the active site; this is usually where the previous inhibitor was present. In case an inhibitor is absent or the active site is relatively unknown, the size of the box and the location of the amino acids are determined by reading the literature (Figure 15). For the purpose of this protocol, we will use the active site that already had an inhibitor attached to it.
3. Browse the output file and save as Akt Fisetin.pdbqt in the same directory.

4. Delete the inhibitor molecule attached to the original 3D structure. Thereafter, select Actions > Atoms and Bonds > Delete (Figure 16). The removal of the inhibitor is important to easily visualize the docking results. The 3QKK PDB needs to be saved again as prep Ald.Akt.PDB.
Figure 16. Screenshot of deletion of the inhibitor that is bound to the protein.

5. Choose the receptor as the protein (preped_Akt) from the drop-down menu and the ligand as prep_fisetin.mol2. It is important to set the right receptor and ligand. In the receptor and ligand options, change everything to TRUE (Figure 17).
6. Select the Opal Web service or enter the local path where the installed version of Autodock Vina is placed and click on OK (Figure 18).
**Results**

**Outcome of Docking**

After the successful run of Autodock Vina, the following dialogue box will appear with the solution. Figure 19 portrays the final step of Docking, that is, outcome/results of docking, which are score, root-mean-square deviation (RMSD) lower bound, and RMSD upper bound.
Visualizing the Docking

Visualization of docking can be done as follows:

1. To see the hydrogen bonding between the receptor and the ligands using the result dialogue box (Figure 20), select H Bonds > Add Count to the Entire Receptor.
2. This opens an H-bond parameter dialogue box (Figure 21). Select Intermodel to visualize bonding between receptor and ligand. Different parameters can be adjusted to better picture the bonding. The table showing all the information on hydrogen bonds and RMSD is presented at the end of the docking session (Figure 22).

Figure 20. Screenshot of visualizing hydrogen bonding between the receptor and ligand.
Figure 21. Screenshot of the H-bond Parameters dialogue box.
3. To be able to retrieve the docking session later at any stage, it can be saved by selecting File > Save Session as > An Akt Fisetin Docking (name the session).

Discussion

Computationally fast and accurate docking of a ligand with a target protein can be performed using AutoDock Vina in Chimera. This protocol will help researchers who are not able to use Autodock and Autodock Vina due to its command-line interface and do not have access to high-end software such as Gold Suite and Molecular Operating Environment to perform computational docking easily. The use of Chimera with AutoDock Vina has not been demonstrated before, and due to the ease of the graphical user interface of Chimera, it can be a go-to tool for someone who is just starting to learn bioinformatics.

Conflicts of Interest

None declared.

References


Abbreviations

CID: compound ID
PDB: Protein Data Bank
RMSD: root-mean-square deviation
SMILES: simplified molecular-input line-entry system
UCSF: University of California, San Francisco

©Sania Safdar Butt, Yasmin Badshah, Maria Shabbir, Mehak Rafiq. Originally published in JMIR Research Protocols (http://www.researchprotocols.org), 19.06.2020. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Research Protocols, is properly cited. The complete bibliographic information, a link to the original publication on http://bioinform.jmir.org, as well as this copyright and license information must be included.