Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

International Immunopharmacology 112 (2022) 109224

Contents lists available at ScienceDirect

International Immunopharmacology

journal homepage: www.elsevier.com/locate/intimp

Bioinformatics pipeline unveils genetic variability to synthetic vaccine design for Indian SARS-CoV-2 genomes

Nimisha Ghosh [a,1], Indrajit Saha [b,*,1], Nikhil Sharma [c,1], Suman Nandi [b]

[a] Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India
[b] Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India
[c] Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

ARTICLE INFO

Keywords:
Bioinformatics Pipeline
Clade
Conserved Regions
Non-synonymous signature SNP
SARS-CoV-2
T-cell epitopes

ABSTRACT

In the worrisome scenarios of various waves of SARS-CoV-2 pandemic, a comprehensive bioinformatics pipeline is essential to analyse the virus genomes in order to understand its evolution, thereby identifying mutations as signature SNPs, conserved regions and subsequently to design epitope based synthetic vaccine. We have thus performed multiple sequence alignment of 4996 Indian SARS-CoV-2 genomes as a case study using MAFFT followed by phylogenetic analysis using Nextstrain to identify virus clades. Furthermore, based on the entropy of each genomic coordinate of the aligned sequences, conserved regions are identified. After refinement of the conserved regions, based on its length, one conserved region is identified for which the primers and probes are reported for virus detection. The refined conserved regions are also used to identify T-cell and B-cell epitopes along with their immunogenic and antigenic scores. Such scores are used for selecting the most immunogenic and antigenic epitopes. By executing this pipeline, 40 unique signature SNPs are identified resulting in 23 non-synonymous signature SNPs which provide 28 amino acid changes in protein. On the other hand, 12 conserved regions are selected based on refinement criteria out of which one is selected as the potential target for virus detection. Additionally, 22 MHC-I and 21 MHC-II restricted T-cell epitopes with 10 unique HLA alleles each and 17 B-cell epitopes are obtained for 12 conserved regions. All the results are validated both quantitatively and qualitatively which show that from genetic variability to synthetic vaccine design, the proposed pipeline can be used effectively to combat SARS-CoV-2.

1. Introduction

More than two years ago, SARS-CoV-2 put a massive halt to the freedom of human movement due to its high transmission rates [1]. Early study established the fact that SARS-CoV-2 virus is highly similar to that of the SARS-CoV-1 (95%–100%) [2]. In April 2021, India registered its second sudden surge in official cases with the Delta (B.1.617.2) variant. In late 2021, the third wave hit the country which was led by Omicron. Though, India is pushing towards a very large vaccination drive, concerns over the efficacy of the vaccine for such aggressive mutations are also increasing. Meanwhile, India is not the only country which has witnessed the new mutation strain of the evolving virus, variants in South Africa (501Y.V2) [3], United Kingdom (B.1.1.7) [4], Japan (E484K) [5], Brazil (P.1) [5] are also making their rounds. The latest variant to join the bandwagon is Omicron (B.1.1.529). Although, previously it was suggested that such mutants are not going to affect the effectiveness of the vaccines currently in use, the emergence of Omicron has changed the equation. Moreover, new variants can affect the diagnosing procedure such as primer identification or antibody binding in RT-PCR. Ascoli [6] also suggested that a mutation in the Spike region of SARS-CoV-2 may affect the diagnosing procedure with greatest impact along with the increasing infection rates, transmissibility or even impacting people of younger age.

In the current scenario, it is an important and urgent task to study the frequently occurring mutations within the virus. In this regard, Yuan et al. [7] have analysed 11,183 SARS-CoV-2 genome from around the globe to identify the SNPs and critical SNPs with specific high mutation frequency along with the geographical pattern analysis. Further, they

* Corresponding author.
E-mail address: indrajit@nitttrkol.ac.in (I. Saha).
[1] Equally contributed

https://doi.org/10.1016/j.intimp.2022.109224
Received 24 June 2022; Received in revised form 22 August 2022; Accepted 31 August 2022
Available online 6 September 2022

have found 74 non-synonymous and 43 synonymous mutations. Most importantly they have identified Nucleocapsid (N) as the gene with the highest mutational frequency changes. This directly undermines the claim of Ascoli [6] that Nucleocapsid can be targeted for the diagnosing purposes as N gene undergoes very less mutations or is mostly conserved. Hence, it is important to take a closer look how SARS-CoV-2 is evolving over time. Moreover, Tang et al. [8] have found new developing variations on the receptor binding sites of Spike gene of SARS-CoV-2 in the form of S and L lineages. Here, S and L lineages are defined by two tightly linked SNPs at positions 8,782 (orf1ab:T8517C, synonymous) and 28,144 (ORF8: C251T, S84L) which might affect the virus pathogenesis. Phylogenetic analysis done by Maitra et al. [9] revealed the signature mutations such as C14408T in RdRp along with A23403G change in Spike protein majorly forming A2a clade within 9 Indian sequences. Further, they have also reported a triplet based mutation in N gene 2881–3 GGG/AAC which might affect the miRNAs bindings to original sequences. Genome analysis by Saha et al. [10] for 72 different countries has shown multiple unique mutation points in the form of substitution, deletion, insertion and SNPs in each country, resulting in 7209, 11700, 119 and 53 mutations respectively. Further, they have identified 11 SNPs which are unique to India, the most frequent being T1198K, A97V, T315N and P13L mutation points in NSP3, RdRp, Spike and ORF8. Therefore, it has become more important than ever to constantly monitor the continuous evolving virus in order to take up proper measures to battle the contagious virus. Study conducted by Nagy et al. [11] identified genomic alterations and the association of each mutation and outcome. As a result, they have found 3733 mutation points related to mild outcome in ORF8, NSP6, ORF3a, NSP4 and Nucleocapsid genes whereas the mutations in Spike glycoprotein, RNA polymearse, ORF3a, NSP3, ORF6 and N provided inferior outcome. Also, severe outcomes are associated to the mutations in ORF3a and NSP7 proteins. Thus, mutations are important in the significant genes such as Spike, N etc. and such mutations may even lead to a false diagnosis in RT-PCR testing. Hence, it is also important to extract the conserved regions in a genomic sequence for more effective diagnosis. In this regard, [10] have identified a conserved region in NSP6 gene as a potential target for SARS-CoV-2 detection using RT-PCR.

On the other hand, alteration in the RNA virus can lead to vaccine failures as was noticed in the case of Influenza virus in 2013–14 [12]. Hence, to fight against a highly evolving virus like SARS-CoV-2, it is important to have stable vaccine. In this regard, Ghosh et al. [13] have performed a genome-wide analysis of 10644 SARS-CoV-2 sequences to identify the conserved regions in a virus genome, followed by which they have proposed epitope based vaccine design targeting the T-cell and B-cell epitopes. Another study conducted by Ghosh et al. [14] for identifying the conserved regions specifically focussed on 566-Indian SARS-CoV-2 sequences by considering four different multiple sequence alignment techniques. In both the studies most immunogenic and antigenic epitopes were derived from various coded proteins of the virus which can be targeted for synthetic vaccine design. Alam et al. [15] targeted the Spike glycoprotein to propose non-allergic, highly antigenic and non-mutant synthetic vaccine design targeting Thymus cell (T-cell) and bone marrow. Rahman et al. [16] targeted 3 important genes viz Spike, Membrane and Envelope for multi-epitope-based vaccine design for SARS-CoV-2 with a 90% population coverage. Also, immune simulation suggested a significant increase in primary immune response with increased IgM and secondary immune response with increased IgG1 and IgG2 along with increased proliferation of T-helper cells with increased cytokines. Another study [17] targeted heptad repeats 1 and 2 (HR1 and HR2) in the Spike protein for peptide design using molecular dynamics simulation between the fusion of the viral membrane with the host cell membrane. This eventually limited the spread of the virus in the host cells. Vashi et al. [18] predicted 24 potential epitope fragments of which 20 were on the surface of Spike protein (S protein) and were considered to be helpful for designing potential immunogenic peptide based vaccines.

Motivated by the literature and looking at the sudden surge of SARS-CoV-2 in India, a comprehensive bioinformatics pipeline is proposed in this work to analyse the virus genomes for understanding its evolution for identifying mutations as signature SNPs, conserved regions and subsequently to design epitope based synthetic vaccine. In this regard, we have performed multiple sequence alignment of 4996 Indian SARS-CoV-2 sequences as a case study using MAFFT followed by phylogenetic analysis of the aligned sequences using Nextstrain. As a result, the sequences are found to be distributed in 5 clades, viz 19A, 19B, 20A, 20B and 20C. Thereafter, from the aligned sequences, mutation points as SNPs are identified in each clade. Subsequently, top 10 signature SNPs based on their frequency are identified in each clade resulting in a total of 50 such SNPs. Out of 50 signature SNPs, 40 unique signature SNPs are identified resulting in 23 non-synonymous signature SNPs which gives 28 amino acid changes in protein which are visualised in protein structures as well. Furthermore, the sequence and structural homology-based prediction along with the protein structural stability of the amino acid changes for such SNPs are evaluated using PROVEAN, PolyPhen 2.0 and I-Mutant 2.0 in order to judge the characteristics of the identified clades. As a consequence, A97V in RdRp in 19A, V354L in Nucleocapsid in 19B, Q57H in Nucleocapsid in 20A, R203M in Nucleocapsid in 20B while T85I in NSP2 and Q57H in ORF3a in 20C are the unique amino acid changes which are responsible for defining each clade as they are all deleterious and unstable and are also responsible for decreasing the protein structural stability. Moreover, based on the entropy of each genomic coordinate of the aligned sequences, conserved regions are identified. Conserved regions are such places in genomic sequences for which the corresponding protein sequences remain unchanged. These conserved regions are then filtered based on the criteria that their lengths are greater than or equal to 125nt and their BLAST specificity score is equal to 100% resulting in 12 conserved regions belonging to NSP2, NSP8, NSP10, RdRp, Exon, Spike glycoprotein, ORF3a and ORF7a proteins. Based on its length, one conserved region as potential target is identified in the NSP10 gene for which the primers and probes are reported as well. Such primers and probes can be used for detecting SARS-CoV-2 virus. The 12 conserved regions are also used to identify the T-cell and B-cell epitopes along with their immunogenic and antigenic scores. Using such scores, most immunogenic and antigenic epitopes are selected for the 12 conserved regions thereby identifying 23 MHC-I and 22 MHC-II restricted T-cell epitopes with 10 unique HLA alleles each and 17 B-cell epitopes. Finally, the binding conformation of the MHC-I and MHC-II restricted T-cell epitopes with respect to HLA alleles are shown to judge their relevance. Also, the physico-chemical properties of the epitopes are reported along with structural properties using Ramchandran plots, ERRAT score and Z-Scores. Thus, based on the comprehensive bioinformatics pipeline, the main contributions of this work can be summarised as: (a) phylogenetic analysis in Nextstrain to identify virus clades, (b) identification of SNPs in the aligned sequences, (c) based on frequency, top 10 signature SNPs identification in each virus clade, (d) identification of conserved regions and based on length selecting one such region as potential target for reporting the corresponding primers and probes to detect SARS-CoV-2 and (e) identification of T-cell and B-cell epitopes for peptide based synthetic vaccine design.

2. Material and Methods

In this section, the details of data collection and the preparation are elucidated which is followed by a brief discussion on the pipeline of the workflow that has been considered in this work.

2.1. Data Collection and Preparation

The reference sequence of SARS-CoV-2 virus (NC_045512.2) is

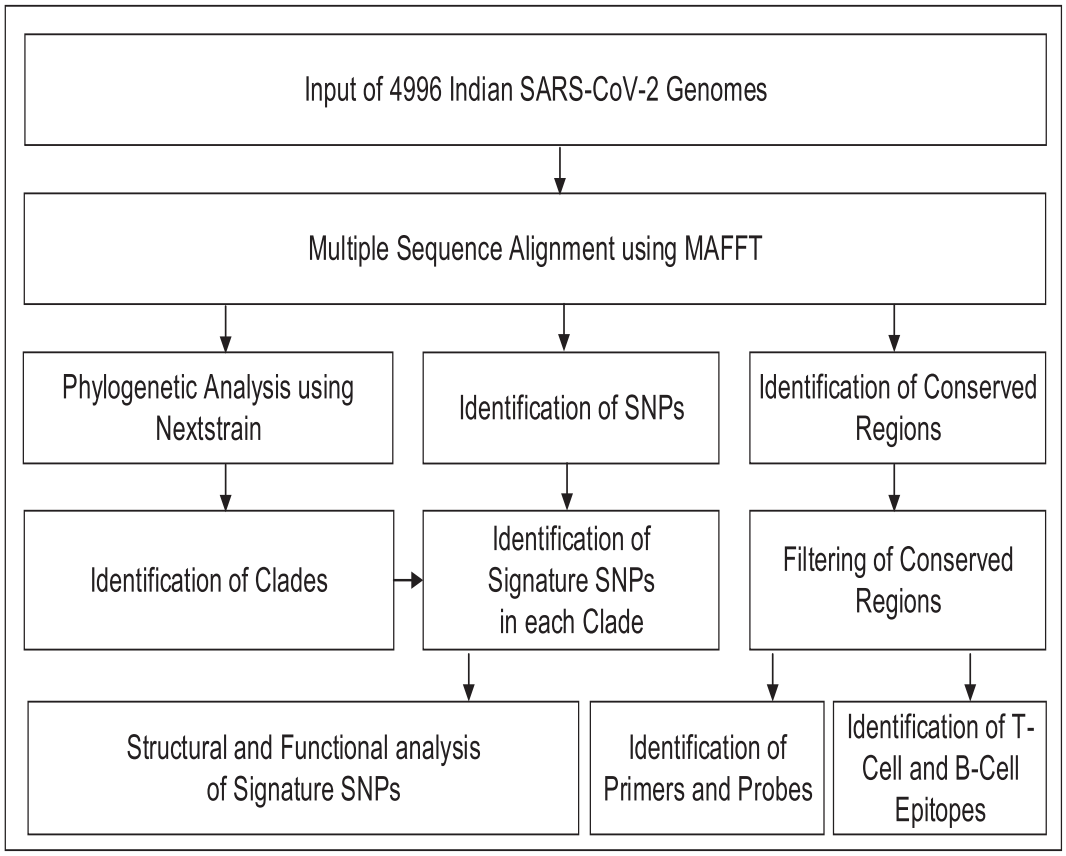


Fig. 1. Pipeline of the work.

collected from National Center for Biotechnology Information (NCBI)[2] while 4996 complete or near complete Indian SARS-CoV-2 genomes are collected from Global Initiative on Sharing All Influenza Data (GISAID)[3] in fasta format. The 4996 SARS-CoV-2 sequences are mostly distributed from January 2020 to January 2021. These sequences are then aligned to find the conserved regions. The coded protein corresponding to each conserved region is extracted as well. Further, to map the protein sequences and changes in the amino acid, protein PDB are collected from Zhang Lab[4] which are then used to model and identify the structural changes. All these analyses are executed on High Performance Computing (HPC) facility of NITTTR, Kolkata while the amino acid changes are checked in MATLAB R2019b. The HPC cluster has a master node with dual Intel Xeon Gold 6130 Processor having 32 Cores, 2.10 GHz, 22 MB L3 Cache and 128 GB DDR4 RAM and 2 GPU and 4 CPU computing nodes with dual Intel Xeon Gold 6152 Processor having 44 Cores, 2.1 GHz, 30 MB L3 Cache and 192 GB DDR4 RAM each, while GPU nodes have NVIDIA Tesla V100 GPU with 16 GB memory each. MSA is performed using the 2 GPU and 4 CPU computing nodes.

2.1.1. Pipeline of the work

The pipeline of this work is provided in Fig. 1. In this work, a comprehensive bioinformatics pipeline is proposed which encompasses identifying mutation points as SNPs, conserved regions and finally design of epitope based synthetic vaccine. To achieve these goals, in the first phase of the pipeline, multiple sequence alignment of 4996 Indian SARS-CoV-2 genomes as a case study using MAFFT [19] is carried out followed by the phylogenetic analyses using Nextstrain [20]. As MAFFT uses fast fourier transform, it outperforms all the other alignment techniques. On the other hand, analysis of the evolution and spread of pathogens is done using Nextstrain by considering phylogenomic and phylogeographic data. The spread and evolution of virus genomes can be visualised at nextstrain.org using auspice. By using this tool, the evolution and geographic distribution of SARS-CoV-2 genomes are visualised by creating the metadata in our High Performance Computing environment. Once the identification of the virus clades are performed using Nextstrain, clade specific aligned sequences are used to identify mutation points as substitutions especially SNPs in each clade. Henceforth, codon table is used to identify the amino acid changes in the virus proteins corresponding to the SNPs. Thereafter, based on their frequency in the virus genome, top 10 signature SNPs are identified in each clade. Please note that the amino acid changes in the SNPs can be either synonymous or non-synonymous. Furthermore, amino acid changes in the non-synonymous SNPs are visualised in the protein structures and they are used to evaluate their functional characteristics as well.

The second phase of the pipeline entails identification of conserved Regions (CnRs) in the aligned sequences using entropy ($\mathscr{N}$) which can be computed as:

$$\mathscr{N} = ln5 + \sum \mathscr{F}_x^y \; [ln(\mathscr{F}_x^y)] \tag{1}$$

where $\mathscr{F}_x^y$ represents the frequency of each residue x occurring at position y and 5 represents the four possible residues as nucleotides plus gap. To identify the conserved regions, a minimum segment length of 15 is considered with maximum average entropy as 0.2 along with a maximum entropy per position of 0.2 as well without any gaps. All these values are taken after following the literature. Thereafter, refinement criteria for the conserved regions are adopted based on the criteria that their lengths are ⩾125nt and their BLAST specificity score as query coverage is equal to 100%. Subsequently, based on its length, a

[2] https://www.ncbi.nlm.nih.gov/nuccore/1798174254
[3] https://www.gisaid.org/
[4] https://zhanglab.ccmb.med.umich.edu/COVID-19/

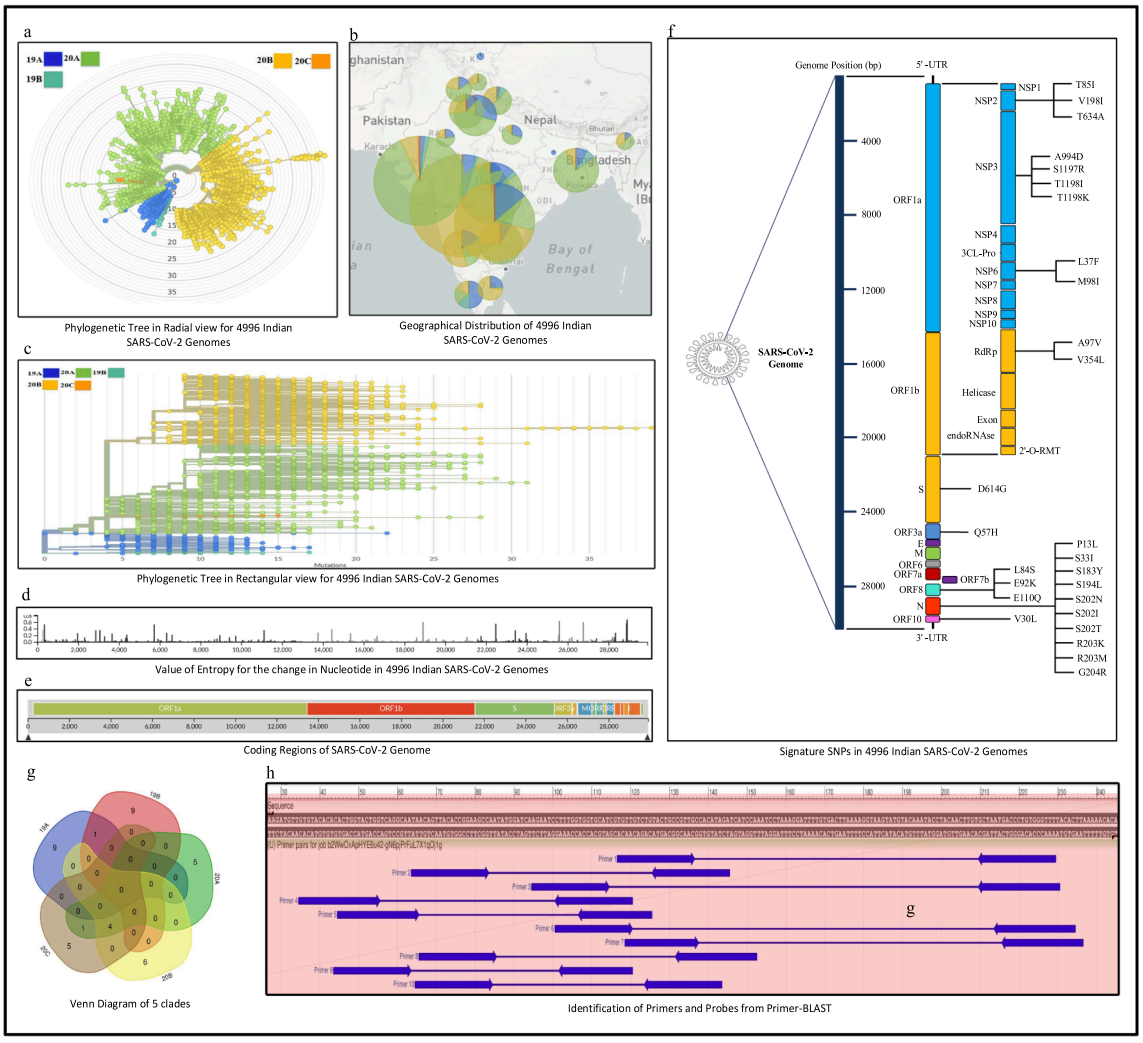


Fig. 2. (a) Phylogenetic Tree in Radial view (b) Geographical Distribution (c) Phylogenetic Tree in Rectangular view (d) Value of Entropy for the change in Nucleotide (e) Coding Regions of SARS-CoV-2 Genome (f) Signature SNPs (g) Venn Diagram of 5 clades and (h) Identification of Primers and Probes using Primer-BLAST.

particular conserved region is considered as potential target which is then used to identify primers and probes using Primer-BLAST[5] for SARS-CoV-2 detection.

In the final phase of the pipeline, T-cell and B-cell epitopes along with their immunogenic and antigenic scores are predicted for the refined CnRs using IEDB[6] and ABCPred[7] respectively. For such MHC-I and MHC-II restricted T-cell epitopes, predictions are carried out using IEDB recommended NetMHCPan EL 4.1[8] and Consensus Approach[9] [21] respectively while ABCPred [22] is used for B-cell epitope prediction. Thereafter, by using these predicted epitopes, antigenic scores are evaluated by VaxiJen 2.0[10] while the validation of the identified T-cell epitopes is carried out by studying their conformational 2D non-covalent structures using LigPlot+ [23]. For the verification of the predicted B-cell epitopes, BepiPred 2.0[11] [24] server is used. Allergen and toxicity properties of the epitopes are evaluated using AllerTop 2.0[12] and ToxinPred[13] respectively. The physico-chemical properties are also evaluated using ToxinPred. Moreover, docking of all the T-cell epitopes are performed using AutoDock Vina [25] and their structural properties are reported using Ramachandran Plot [26], ERRAT score [27] and Verify 3D [28] using SAVES 6.0[14]. Finally, Z-Score evaluation is performed using ProSA [29].

3. Results

3.1. Phylogenetic analysis and Signature SNPs in each clade

To achieve the first step of the bioinformatics pipeline, multiple sequence alignment of 4996 Indian SARS-CoV-2 genomes is performed using MAFFT followed by phylogenetic analysis with the help of Nextstrain. This phylogenetic analysis results in 5 clades viz. 19A, 19B, 20A, 20B and 20C. Thereafter, mutation points as substitutions specifically SNPs are identified in each clade resulting in 708, 161, 3308, 3235 and 47 SNPs for 479, 88, 2486, 1925 and 18 sequences respectively in 19A,

[5] https://www.ncbi.nlm.nih.gov/tools/primer-blast/
[6] https://www.iedb.org/
[7] https://webs.iiitd.edu.in/raghava/abcpred/ABC_submission.html
[8] http://tools.iedb.org/mhci/
[9] http://tools.iedb.org/mhcii/
[10] http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html
[11] http://tools.iedb.org/bcell/
[12] https://www.ddg-pharmfac.net/AllerTOP/
[13] https://webs.iiitd.edu.in/raghava/toxinpred/pep_test.php
[14] https://saves.mbi.ucla.edu/

Table 1
List of Signature SNPs in each clade for 4996 Indian SARS-CoV-2 Genomes.

Clade	Genomic Position	Frequency	Nucleotide Change	Protein Change	Protein Coordinate	Mapped with Coding and Non-Coding Region
19A	11083	425	G>A, G>T	Synonymous, L>F	37	NSP6
	13730	374	C>T	A>V	97	RdRp
	28311	364	C>T	P>L	13	Nucleocapsid
	23929	360	C>T	Synonymous	789	Spike
	6312	359	C>T, C>A	T>I, T>K	1198	NSP3
	19524	111	C>T	Synonymous	495	Exon
	6310	98	C>A, C>T	S>R, Synonymous	1197	NSP3
	1397	77	G>A	V>I	198	NSP2
	29742	77	G>A,G>C, G>T	Not Present	Not Present	3' UTR
	28688	74	T>C	Synonymous	139	Nucleocapsid
19B	28144	87	T>C	L>S	84	ORF8
	8782	86	C>T	Synonymous	76	NSP4
	28878	83	G>A,G>T, G>C	S>N, S>I, S>T	202	Nucleocapsid
	29742	81	G>A,G>C, G>T	Not Present	Not Present	3' UTR
	22468	62	G>T,G>A	Synonymous, Synonymous	302	Spike
	11230	19	G>T	M>I	86	NSP6
	7945	16	C>T	Synonymous	1742	NSP3
	28167	15	G>A	E>K	92	ORF8
	2705	9	A>G	T>A	634	NSP2
	14500	9	G>T	V>L	354	RdRp
20A	23403	2472	A>G	D>G	614	Spike
	241	2458	C>T	Not Present	Not Present	5' UTR
	3037	2455	C>T	Synonymous	106	NSP3
	14408	2377	C>T	P>L	323	RdRp
	26735	1432	C>T	Synonymous	71	Membrane
	18877	1427	C>T	Synonymous	280	Exon
	25563	1418	G>A, G>T, G>C	Synonymous, Q>H, Q>H	57	ORF3a
	28854	1230	C>T	S>L	194	Nucleocapsid
	22444	1191	C>T	Synonymous	294	Spike
	2836	557	C>T	Synonymous	39	NSP3
20B	3037	1923	C>T	Synonymous	106	NSP3
	241	1922	C>T	Not Present	Not Present	5' UTR
	23403	1922	A>G	D>G	614	Spike
	14408	1912	C>T	P>L	323	RdRp
	28881	1868	G>A, G>T	R>K, R>M	203	Nucleocapsid
	28882	1868	G>A	Synonymous	203	Nucleocapsid
	28883	1867	G>A, G>C	G>R, G>R	204	Nucleocapsid
	313	1120	C>T	Synonymous	16	Leader protein
	5700	1106	C>A	A>D	994	NSP3
	4354	281	G>A	Synonymous	545	NSP3
20C	241	18	C>T	Not Present	Not Present	5' UTR
	1059	18	C>T	T>I	85	NSP2
	3037	18	C>T	Synonymous	106	NSP3
	14408	18	C>T	P>L	323	RdRp
	23403	18	A>G	D>G	614	Spike
	25563	18	G>A, G>T, G>C	Synonymous, Q>H, Q>H	57	ORF3a
	16260	9	C>T	Synonymous	8	Helicase
	28821	9	C>A	S>Y	183	Nucleocapsid
	28221	4	G>T, G>C	E>-, E>Q	110	ORF8
	28371	4	G>T	S>I	33	Nucleocapsid

19B, 20A, 20B and 20C. The details of the SNPs are provided in the supplementary Table S1. The resultant phylogenetic trees in radial and rectangular views are shown in Fig. 2(a) and (c) while the clade wise geographical distribution of the 4996 sequences is shown in Fig. 2(b). The clade wise evolution of the sequences for each month of each Indian state is shown in the form of pie charts in supplementary Table S2 while the month wise evolution of such sequences for each clade is reported in supplementary Table S3. The corresponding colour representation for the five major clades and the months are provided in supplementary Figure S1. Moreover, the entropy values for the nucleotide changes and coding regions of the SARS-CoV-2 genome are shown respectively in Fig. 2(d) and (e). It is to be noted that for some sequences, the state name is not mentioned in the GISAID database. Thus, they are aggregated under the state name 'India'.

Once the SNPs are determined for each clade, top 10 SNPs based on their frequency viz. signature SNPs are identified in each clade, thereby resulting in 50 signature SNPs as reported in Table 1 and visualised in Fig. 2(f). In unsupervised learning, feature selection is a very crucial task. In this work, frequency of a SNP is considered to be the feature selection criterion. For example, G11083A and G11083T with a frequency of 425 is the top signature SNP in clade 19A while for 19B, T28144C having frequency of 87 is the top signature SNP. Subsequently, 40 unique SNPs are identified which results in 23 non-synonymous signature SNPs with 28 corresponding amino acid changes. The common signature SNPs in the five clades are visualised using Venn diagram in Fig. 2(g). It is evident from the figure that the clades do not have any common SNPs, thereby confirming the fact that signature SNPs are indeed the defining features of a clade. Moreover, the amino acid changes are visualised in Fig. 3 as well. Please note that 27 amino acid changes are visualised in Fig. 3 as opposed to 28 reported changes; the discarded change is E110* in ORF8 as this amino acid change leads to a stop codon. Also, sequence and structure-based homology prediction of the amino acid changes for the non-synonymous SNPs are reported in Table 2, the details of which are discussed in Discussion section. All the

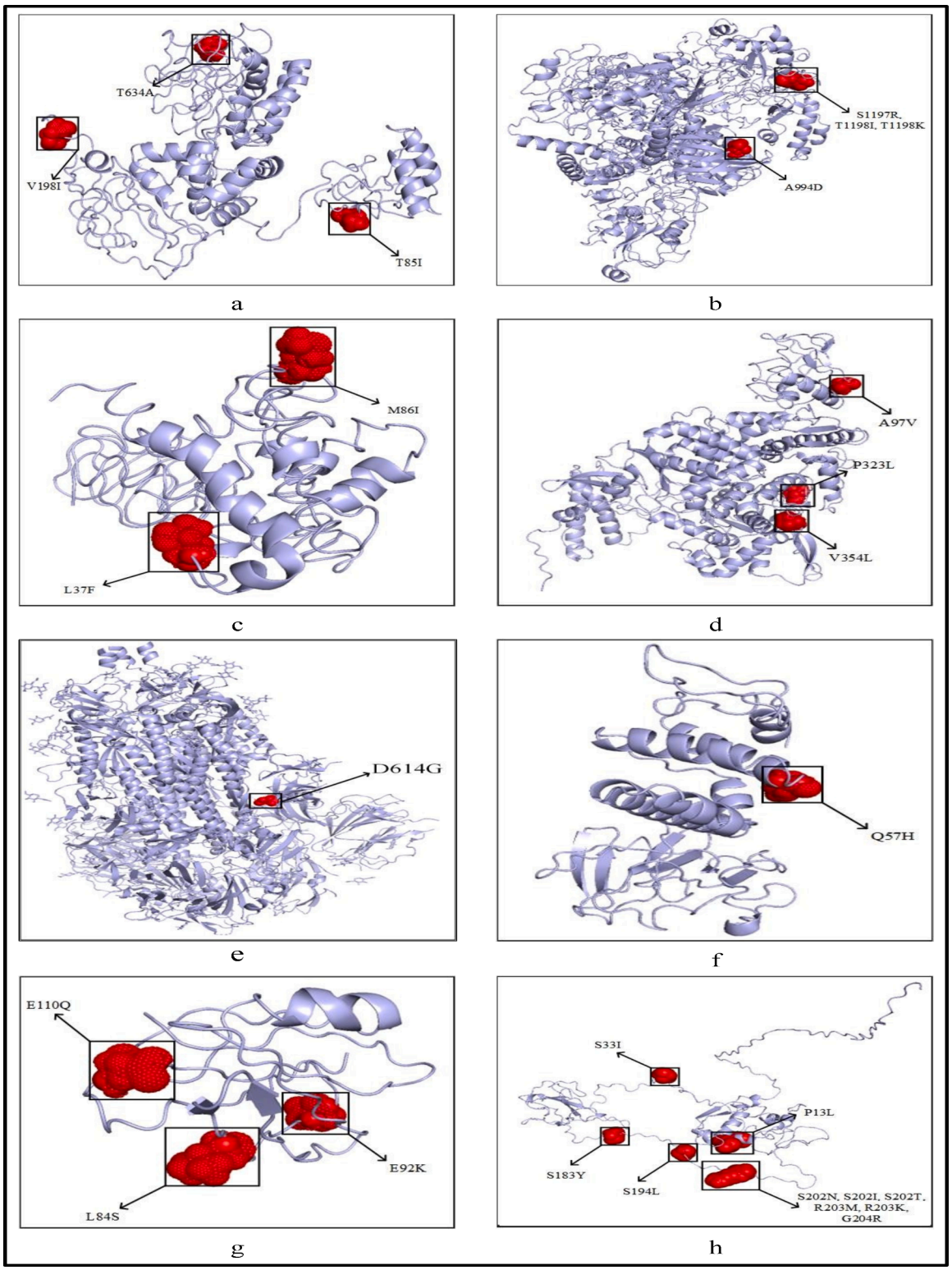


Fig. 3. Highlighted amino acid changes in the protein structures for the non-synonymous signature SNPs of (a) NSP2 (b) NSP3 (c) NSP6 (d) RdRp (e) Spike (f) ORF3a (g) ORF8 and (h) Nucleocapsid.

detailed results are provided in supplementary Table S1.

3.2. Selection of CnRs

For the next phase of this study, we have obtained 473 conserved regions (CnRs) which are then mapped to the 11 coding regions of SARS-CoV-2; ORF1ab, Spike, ORF3a, Envelope, Membrane, ORF6, ORF7a, ORF7b, ORF8, Nucleocapsid and ORF10. For each CnR, the corresponding protein sequence is taken according to the reading frame it is associated with. For example, protein sequence of CnR in Spike region is taken from Frame 2 while that belonging to Envelope and Membrane are taken from Frames 1 and 3 respectively. These 473 conserved regions are then filtered based on the criteria that the length of the CnR should be greater than or equal to 125nt and the their BLAST specificity score as

Table 2
Sequence and structural homology-based prediction for non-synonymous signature SNPs along with their protein structural stability.

Clade	Genomic	Amino residue	Protein	PROVEAN		PolyPhen-2		I-Mutant 2.0	
	Coordinates	Change		Effect	Score	Prediction	Score	Stability	DDG
19A	11083	L37F	NSP6	Neutral	-1.369	Benign	0.027	Decrease	0.05
	13730	**A97V**	**RdRp**	**Deleterious**	**−3.611**	**Probably Damaging**	**0.99**	**Decrease**	**−0.53**
	28311	P13L	Nucleocapsid	Neutral	-1.23	Probably Damaging	1.000	Increase	0.11
	6312	T1198I	NSP3	Neutral	-0.085	Probably Damaging	0.998	Decrease	-0.72
	6312	T1198K	NSP3	Neutral	−0.353	NG	NG	Decrease	-1.37
	6310	S1197R	NSP3	Neutral	-0.835	NG	NG	Decrease	-0.88
	1397	V198I	NSP2	Neutral	0.307	Benign	0.006	Increase	0.18
19B	28144	L84S	ORF8	Neutral	2.333	Benign	0.002	Decrease	-2.87
	28878	S202N	Nucleocapsid	Neutral	-0.404	Probably Damaging	0.994	Decrease	-0.8
	28878	S202I	Nucleocapsid	Deleterious	-3.308	Probably Damaging	0.998	Increase	0.22
	28878	S202T	Nucleocapsid	Neutral	-1.428	Probably Damaging	0.986	Decrease	-0.53
	11230	M86I	NSP6	Neutral	-0.427	Benign	0.025	Decrease	-1.02
	28167	E92K	ORF8	Neutral	-1.5	NG	NG	Decrease	-1.05
	2705	T634A	NSP2	Neutral	-0.004	Benign	0.106	Decrease	-1.13
	14500	**V354L**	**RdRp**	**Deleterious**	**−2.581**	**Probably Damaging**	**0.997**	**Decrease**	**−1.95**
20A	23403	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	-1.94
	14408	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	-0.80
	25563	**Q57H**	**ORF3a**	**Deleterious**	**−3.286**	**Probably Damaging**	**0.966**	**Decrease**	**−1.12**
	28854	S194L	Nucleocapsid	Deleterious	-4.272	Probably Damaging	0.994	Increase	0.45
20B	23403	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	-1.94
	14408	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	-0.80
	28881	R203K	Nucleocapsid	Neutral	-1.604	Probably Damaging	0.969	Decrease	-2.26
	28881	**R203M**	**Nucleocapsid**	**Deleterious**	**−3.305**	**Probably Damaging**	**0.998**	**Decrease**	**−1.52**
	28883	G204R	Nucleocapsid	Neutral	-1.656	Probably Damaging	1	Decrease	0
	5700	A994D	NSP3	Neutral	-1.103	NG	NG	Decrease	-0.78
20C	**1059**	**T85I**	**NSP2**	**Deleterious**	**−4.09**	**Probably Damaging**	**0.998**	**Decrease**	**−1.71**
	14408	P323L	RdRp	Neutral	-0.865	Benign	0.005	Decrease	-0.80
	23403	D614G	Spike	Neutral	0.598	Benign	0.004	Decrease	-1.94
	25563	**Q57H**	**ORF3a**	**Deleterious**	**−3.286**	**Probably Damaging**	**0.966**	**Decrease**	**−1.12**
	28821	S183Y	Nucleocapsid	Deleterious	-2.75	Probably Damaging	0.998	Increase	0
	28221	E110Q	ORF8	Neutral	-0.25	NG	NG	Decrease	-1.13
	28371	S33I	Nucleocapsid	Neutral	-1.372	NG	NG	Increase	0.63

query coverage is equal to 100%. As a result, we have obtained 12 such regions as reported in Table 3. The table also shows the corresponding protein sequences for the conserved regions along with their length, BLAST specificity score, percent of BLAST specificity score as query coverage, coding regions, starting and ending coordinates, length of coding regions and the coded proteins. These CnRs belong to coding regions which code NSP2, NSP8, NSP10, RdRp, Exon, Spike glycoprotein, ORF3a protein and ORF7a protein. The details of all the initial and filtered CnRs are provided in the supplementary as an excel file. Also, based on its length, among these CnRs, one CnR is then chosen as the target for the detection of SARS-CoV-2. Moreover, the protein sequences of these CnRs are used to identify the MHC-I and MHC-II restricted T-cell and B-cell epitopes.

3.3. Identification of Conserved Region as Target and associated Primers and Probes

Among the 12 CnRs identified, the CnR with the largest length of 247nt is considered to be a potential target. This CnR belongs to ORF1ab region, specifically NSP10 gene which is shown in Table 4. With a Nucleotide BLAST score of 457 and BLAST specificity score as query coverage is equal to 100%, the global stability of this CnR as a global target is confirmed. The structure of the NSP10 gene as shown in Table 4 is taken from ZhangLab in the form of a PDB file and the CnR as target is highlighted in red. Using this conserved region, 10 primers and probes are identified from Primer-BLAST and reported in Table 5 and shown in Fig. 2(h). The table reports both the forward and the reverse primers. Moreover, high GC scores (45%-53%) of the identified primers suggest that the identified primers and probes can be used in RT-PCR for SARS-CoV-2 detection in order to correctly diagnose COVID-19 patients. Therefore, the target region of NSP10 gene can be considered as a confirmatory assay. It is to be noted that based on its adhesive properties, Ong et al. [30] have predicted NSP10 as a possible vaccine candidate.

3.4. Identification of T-cell Epitopes

To achieve the final phase of the pipeline, design of epitope based synthetic vaccine is carried out. To predict the epitopes from the 12 CnRs, the corresponding protein sequences are fed to the various tools as inputs. For the prediction of MHC-I restricted T-cell epitopes, IEDB recommended NetMHCPan EL 4.1 [31] is considered targeting 27 unique HLA alleles. For each CnR, this resulted in the selection of 5 best HLA allele binder epitopes based on their immunogenic scores. Thereafter, these best binders are provided as input to VaxiJen 2.0 [32] server for antigenic score prediction [31] with a cut-off score of 0.4. Any epitope beyond this cut-off are considered to be antigenic. Therefore, a total of 60 epitopes, each of length 9–10 mer, are obtained along with their immunogenic and antigenic scores. From each of the 12 CnRs, the most immunogenic and antigenic MHC-I restricted T-cell epitopes are identified resulting in 22 such epitopes and reported in Table 6. With a score of 0.99, the most immunogenic epitopes are SEVGPEHSL, DTDFVNEFY and QEYADVFHLY bounded to HLA-B*40:01, HLA-A*01:01 and HLA-B*44:03 alleles respectively belonging to NSP2 and RdRp coded proteins. On the other hand, with a score of 1.43, HPNPKGFCDL is the most antigenic epitope belonging to NSP10 coded

Table 3
Conserved Regions (CnRs) as derived from 4996 SARS-CoV-2 genomes with associated details

DNA Sequence of Conserved Region (CnR)	Protein Sequence	Length of CnR	BLAST Specificity Score of CnR	% of BLAST Specificity Score as Query Coverage	Coding Region (CR)	Starting Coordinate	Ending Coordinate	Length of Coding Region	Coded Proteins
1282-CACTTGCGAATTTTGTG GCACTGAGAATTTGACTAAAGAAGGT GCCACTACTTGTGGTTACTTAC CCCAAAATGCTGTTGTTAAAATTTATT GTCCAGCATGTCACAATTCAGAAGT AGGACCTGAGCATAGTCTTG-1418	TCEFCGTENLTKEGATTCGY LPQNAVVKIYCPACHNSEVGPEHSL	137	254	100	ORF1ab	266	21552	21287	NSP2
12422-AGAGATGGTTGTGTTC CCTTGAACATAATACCTCTTACAACAGC AGCCAAACTAATGGTTGTCATA CCAGACTATAACACATATAAAAATACGTGTGATGGT ACAACATTTACTTATGCATCAG CATTGTGGGAAAT-12558	RDGCVPLNIIPLTTAAKLMVVI PDYNTYKNTC DGTTFTYASALWE	137	254	100	ORF1ab	266	21552	21287	NSP8
13125-GGGGACAACCAATCA CTAATTGTGTTAAGATGTTGTGTACA CACACTGGTACTGGTCAGGCAATAACAG TTACACCGGAAGCCAATATGG ATCAAGAATCCTTTGGTGGTGCATCGTGTTG TCTGTACTGCCGTTGCCACATAGA TCATCCAAATCCTAAAGGATTTT GTGACTTAAAAGGTAAGTATGTAC AAATACCTACAACTTGTGCTAATGA CCCTGTGGGTTTTACACTTAAAA ACACAGT-13371	GQPITNCVKMLCTHTGTGQAITVTP EANMDQESFGGASCCLYCRCHIDHP NPKGFCDLKGKYVQIPT TCANDPVGFTLKNT	247	457	100	ORF1ab	266	21555	21290	NSP10
14075-TCAATGGTAACTGGTATGATTT CGGTGATTTCATACAAACCACGCC AGGTAGTGGAGTTCCTGTTGTAGATT CTTATTATTCATTGTTAATGCCT ATATTAACCTTGACCAGGGCTTTAACT GCAGAGTCAC-14206	NGNWYDFGDFIQTTPGSGVPVVDS YYSLLMPILTLTRALTAES	132	244	100	ORF1ab	266	21552	21287	RdRp
14221-TTAACAAAGCCTTACATTAAGT GGGATTTGTTAAAATATGACTTCA CGGAAGAGAGGTTAAAACTCTTTG ACCGTTATTTTAAATATTGGGATC AGACATACCACCCAAATTGTG TTAACTGTTTGGATGACAGATGC ATTCTGCATTGTGCAAACTTTAAT GTTTTATTCTCTACAGTGT TCCCA-14406	LTKPYIKWDLLKYDFTEERLK LFDRYFKYWDQTYHPNCVNCLDDRCILHC ANFNVLFSTVFP	186	344	100	ORF1ab	266	21552	21287	RdRp
15607-TTACAACACAGACTTTATGAGT GTCTCTATAGAAATAGAGATGTT GACACAGACTTTGTGAATGAGT TTTACGCATATTTGCGTAAACAT TTCTCAATGATGATACTCTCTGAC GATGCTGTTGTGTGTTT-15737	LQHRLYECLYRNRDVDT DFVNEFYAYLRKHFSM MILSDDAVVC	131	243	100	ORF1ab	266	21552	21287	RdRp
15991-GATGGTACACTTATGATTGAACG GTTCGTGTCTTTAGCTATAGATGCTTAC CCACTTACTAAACATCCTAATC AGGAGTATGCTGATGTCTTTCAT TTGTACTTACAATACATAAGA AAGCTACATGATGAGTTAACAGG	DGTLMIERFVSLAIDAYPLTKH PNQEYADVFHLYLQYIRKLHDE LTGHMLDMYSVMLTNDNTS RYWEPEFY	215	398	100	ORF1ab	266	21552	21287	RdRp

(*continued on next page*)

Table 3 (*continued*)

DNA Sequence of Conserved Region (CnR)	Protein Sequence	Length of CnR	BLAST Specificity Score of CnR	% of BLAST Specificity Score as Query Coverage	Coding Region (CR)	Starting Coordinate	Ending Coordinate	Length of Coding Region	Coded Proteins
ACACATGTTAGACATGTATTCTG TTATGCTTACTAATGATAACACTT CAAGGTATTGGGAACCTG AGTTTTATGA-16205									
18487-ATACCACTTATGTACAAAGG ACTTCCTTGGAATGTAGTGCGTA TAAAGATTGTACAAATGTTAAGTGA CACACTTAAAAATCTCTCT GACAGAGTCGTATTTGTCTTATGGGCACAT GGCTTTGAGTTGACATCTATGAAGTATT TTGTGAAAATAGGACCTGAGCGCA CCTGTTGTCTATGT-18669	IPLMYKGLPWNVVRIKIVQ MLSDTLKNLSDRVVFVLWAHGFELTSM KYFVKIGPERTCCLC	183	339	100	ORF1ab	266	21552	21287	Exon
18980-ACATGGTTGTTAAAGCTGCAT TATTAGCAGACAAATTCCCAGT TCTTCACGACATTGGTAACCCTAAA GCTATTAAGTGTGTACCTCAAGCTGAT GTAGAATGGAAGTTCTATGATGCACAG CCTTGTAGTGACAAAGCTTATAA AATAGAAG-19132	MVVKAALLADKFPVLHDIGNPK AIKCVPQADVEWKFYDAQPC SDKAYKIE	153	283	100	ORF1ab	266	21552	21287	Exon
24490-TTTAAATGATATCCTTTCACGTC TTGACAAAGTTGAGGCTGAAGTGCAAA TTGATAGGTTGATCACAGGCAGACT TCAAAGTTTGCAGACATATGTGAC TCAACAATTAATTAGAGCTGCAGAAATCA GAGC-24621	LNDILSRLDKVEAEVQIDRLITGRLQ SLQTYVTQQLIRAAEIR	132	244	100	Spike	21563	25381	3819	Spike glycoprotein
25913-GCACAACAAGTCCTATTTCTGAACAT GACTACCAGATTGGTGGTTATACTGA AAAATGGGAATCTGGAGTAAAAGACTGTGTT GTATTACACAGTTACTTCACTTCAGACTA TTACCAGCTGTACTCAACTCAATTGAG TACAGACACT-26061	TTSPISEHDYQIGGYTEKWESGVKDCVVLHSYFTSDYYQLYSTQLSTDT	149	276	100	ORF3a	25393	26217	825	ORF3a protein
27394-ATGAAAATTATTCTTTTCTTG GCACTGATAACACTCGCTACTTGT GAGCTTTATCACTACCAAGAGT GTGTTAGAGGTACAACAGTACTTTT AAAAGAACCTTGCTCTTCTGGAA CATACGAGGGCA-27520	MKIILFLALITLATCELYH YQECVRGTTVLLKEPCSSGTYEG	127	235	100	ORF7a	27394	27756	363	ORF7a protein

Table 4
Targeted Conserved Region in SARS-CoV-2 Genome and its corresponding protein sequence in NSP10 which is highlighted by red colour in NSP10 gene.

DNA Sequence of Conserved Region (CnR)	Protein Sequence	NSP10 protein structure with target region
13125-GGGGACAACCAATCACTAATTGTGTTAAGATGTTGTGTACACACTGGTACTG GTCAGGCAATAACAGTTACACCGGAAGCCAATATGGATCAAGAATCCTTTGGTGGTGCATCGTGTTGT CTGTACTGCCGTTGCCACATAGATCATCCAAATCCTAAAGGATTTTGTGACTTAAAAGGTAAGTATGTA CAAATACCTACAACTTGTGCTAATGACCCTGTGGGTTTTACACTTAAAAACACAGT-13371	35-GQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHIDHPNPKGFCDLKGKY VQIPTTCANDPVGFTLKNT-115	

protein and bounded to HLA-B*07:02 allele.

Similarly, MHC-II restricted T-cell epitopes are predicted using IEDB recommended consensus approach targeting a different set of 27 unique HLA alleles resulting in 60 epitopes, each of length 15 mer. Subsequently, the most immunogenic and antigenic MHC-II restricted T-cell epitopes are identified for the 12 CnRs which resulted in 21 such epitopes as reported in Table 6. It is to be noted that a MHC-II restricted T-cell epitope with a low immunogenic score is a better vaccine candidate. Thus, with a score of 0.02, NEFYAYLRKHFSMMI belonging to RdRp coded protein and bounded to HLA-DRB1*11:01 allele is the most immunogenic epitope while the most antigenic epitope is GCVPLNIIPLTTAAK belonging to NSP8 coded protein and bounded to HLA-DRB1*08:02 allele. All the 60 MHC-I and MHC-II restricted T-cell epitopes along with their HLA alleles are provided in the supplementary as an excel file and the corresponding link is provided in Table S1.

3.5. Identification of B-cell Epitopes

Epitope designing consists of both T-cell as well as B-cell epitopes; the latter one is particularly important for antigen production against a virus. In this regard, ABCPred is used for the prediction of B-cell epitopes where a threshold of 0.5 is maintained to consider the epitopes beyond this threshold to be immunogenic. With a cut-off value of 0.4, VaxiJen 2.0 server is used to evaluate the antigenic scores of the epitopes. Thus, we have identified 50 linear B-cell epitopes, each of length 16 mer, for the 12 CnRs, among which 17 are selected to be the most immunogenic and antigenic as shown in Table 6. These epitopes are also verified with the help of BepiPred 2.0 server and their corresponding graphical analysis is shown in supplementary Figure S2 where the red line represents the threshold which is set to 0.35 and the total green and yellow regions indicate a protein sequence. The most immunogenic and antigenic B-cell epitopes as reported in Table 6 are respectively VVKIYCPACHNSEVGP belonging to NSP2 coded protein and PNPKGFCDLKGKYVQI belonging to NSP10 coded protein. Their corresponding graphical representations are provided in supplementary Figure S2 (a) and (c) respectively. All the 50 B-cell epitopes are provided the supplementary as an excel file and the corresponding link is provided in Table S1.

Additionally, in Table 7 we have provided a summarised list of all the epitopes belonging to these 12 CnRs along with their allergic and toxicity characteristics predicted using AllerTOP 2.0[15] and ToxinPred[16] where 12, 6 and 8 allergic MHC-I, MHC-II T-cell and B-cell epitopes are identified respectively while only 1 and 5 epitopes in MHC-I restricted T-cell and B-cell epitopes are found to be toxic. The 3D structures of the epitopes summarised in Table 7 are further highlighted in Fig. 4 using ChimeraX. For better understandability, the identified epitopes are underlined in supplementary Figure S3.

4. Discussion

Since its emergence in Wuhan, China, SARS-CoV-2 has spread very rapidly around the world resulting in a global pandemic. Though the vaccination process has started, the number of COVID affected patients is still quite large. The waves of COVID-19 pandemic are a huge threat to the human population. In this regard, it is important to develop a bioinformatics pipeline in order to conduct in-depth analysis of SARS-CoV-2 genomes in every one or two months for next four to five years to know the evolution, genetic variability, virus strains and conserved regions, thereby to use such information for proper vaccine. Moreover, the mutated variants found in India are also a major concern of the researchers. Thus, identification of virus strains is very essential in today's scenario. Moreover, vaccine is the only ray of hope in this dire situation,

[15] https://www.ddg-pharmfac.net/AllerTOP/
[16] http://crdd.osdd.net/raghava/toxinpred/

Table 5
Details of Primers and Probes of NSP10 gene.

Primer	Primers						
Pair	Type	Sequence (5'->3')	Length	Tm	GC%	Probe Sequence	Probe Length
1	Forward	117-TGTTGTCTGTAC TGCCGTTG-136	20	60.05	50	TGTTGTCTGTACTGCCGTTGCC ACATAGATCATCCAAATCCTAAAGGATTTTGTGAC TTAAAAGGTAAGTATGTACAAATACCTACAACTTG TGCTAATGACCCTGTGGGTTT	113
	Reverse	229-AAACCCACA GGGTCATTAGC-210	20	59.46	50		
2	Forward	64-TAACAGTTACACCGGAAGCC-83	20	59.18	50	TAACAGTTACACCGGAAGCCAATATGGATCAAGA ATCCTTTGGTGGTGCATCGTGTTGTCTGTA CTGCCGTTGCCACATAGA	82
	Reverse	145-TCTATGTG GCAACGGCAGTA-126	20	60.76	50		
3	Forward	95-AGAATCC TTTGGTGGTGCAT-114	20	59.08	45	AGAATCCTTTGGTGGTGCATCGTGTT GTCTGTACTGCCGTTGCCACATAGATCATCCAAAT CCTAAAGGATTTTGTGACTTAAAAGGTAAGTATGT ACAAATACCTACAACTTGTGCTAATGACCCTGTGGGTTTT	136
	Reverse	230-AAAACCCACAGG GTCATTAGC-210	21	60.16	47.62		
4	Forward	35-GTGTACACACAC TGGTACTGG-55	21	59.89	52.38	GTGTACACACACTGGTACTGG TCAGGCAATAACAGTTACACCGGAAGCCAATATG GATCAAGAATCCTTTGGTGGTGCATCGTGTT	86
	Reverse	120-AACACGATGCACC ACCAAAG-101	20	60.97	50		
5	Forward	45-ACTGGTACTGGTCA GGCAATA-65	21	60.16	47.62	ACTGGTACTGGTCAGGCA ATAACAGTTACACCGGAA GCCAATATGGATCAAGAAT CCTTTGGTGGTGCATCGTGTTGTCTG	81
	Reverse	125-CAGACAACACG ATGCACCA-107	19	60	52.63		
6	Forward	101-CTTTGGTGGTG CATCGTGTT-120	20	60.97	50	CTTTGGTGGTGCATCGTGTTGT CTGTACTGCCGTTGCCACATAGATCATCCAAATCC TAAAGGATTTTGTGACTTAAAAGGTAAGTATGTAC AAATACCTACAACTTGTGCTAATGACCCTGTGGGT TTTACAC	134
	Reverse	234-GTGTAAA ACCCACAGGGTCAT-214	21	59.81	47.62		
7	Forward	119-TTGTCTGTACTGCCGTTGC-137	19	60	52.63	TTGTCTGTACTGCCGTTGCCACATAGATCATCCAA ATCCTAAAGGATTTTGTGACTTAAAAGGTAAGTAT GTACAAATACCTACAACTTGTGCTAATGACCCTGTGGGTTTTACACTT	118
	Reverse	236-AAGTGTAAAACCCACAGGGTC-216	21	59.74	47.62		
8	Forward	66-ACAGTTACACCGGAAGCCAA-85	20	61.2	50	ACAGTTACACCGGAAGCCAATATGGATCAAGAAT CCTTTGGTGGTGCATCGTGTTGTCTGTACTGCCGT TGCCACATAGATCATCCA	87
	Reverse	152-TGGATGATCTATGT GGCAACG-132	21	59.81	47.62		
9	Forward	44-CACTGGTACTGGTCAGGCAA-63	20	61.27	55	CACTGGTACTGGTCAGGCAAT AACAGTTACACCG GAAGCCAATATGGATCAAGAATCC TTTGGTGGTG CATCGTGTT	77
	Reverse	120-AACACGATGCACCACCAAA-102	19	59.84	47.37		
10	Forward	65-AACAGTTACACCGGAAGCCA-84	20	61.2	50	AACAGTTACACCGGAAGCCAATATGGATCAAGAATCCTTTGGTGGTG CATCGTGTTGTCTGTACTGCCGTTGCCACATA	79
	Reverse	143-TATGTGGCAACGGCAGTACA-124	20	61.34	50		

thereby making development of peptide based synthetic vaccine viz. epitopes even more necessary. In this regard, we have analysed 4996 Indian SARS-CoV-2 genomes which has resulted in the identification of five clades and subsequently 10 signature SNPs in each clade. Also, based on entropy, conserved regions are identified for the aligned sequences and primers and probes are identified as well for SARS-CoV-2 detection. Furthermore, we have identified T-cell and B-cell epitopes for the development of vaccines.

Structural changes in amino acid residues can often result in changes in the protein translations which is conducive to functional instability of the proteins. In this regard, sequence and structural homology-based prediction of the amino acid changes in the non-synonymous signature SNPs along with their protein stability for the 4996 sequences are reported in Table 2 using PROVEAN (Protein Variation Effect Analyser) [33], PolyPhen-2 (Polymorphism Phenotyping) [34] and I-Mutant 2.0 [35] to judge the characteristics of the identified clades. PROVEAN[17] works with sequence based prediction algorithm while Polyphen-2[18] uses prediction based on sequence, structural and phylogenetic information of a SNP. I-Mutant 2.0[19] uses support vector machine (SVM) for

[17] https://provean.jcvi.org/index.php
[18] http://genetics.bwh.harvard.edu/pph2/
[19] http://folding.biofold.org/i-mutant/i-mutant2.0.html

Table 6
List of most Immunogenic and Antigenic Epitopes for MHC-I, MHC-II restricted T-cell and B-cell Epitopes for 12 CnRs. *I.S.-Immunogenic Score; A.S.-Antigenic Score.

Protein Sequence	Coded Protein	Type	MHC-I restricted T-cell				MHC-II restricted T-cell				B-cell Epitopes		
			Epitopes	Alleles	I.S.*	A.S.*	Epitopes	Alleles	I.S.*	A.S.*	Epitopes	I.S.*	A.S.*
160-TCEFCGTENLTKEGATTCGY LPQNAVVKIYCPACHNSEVGPEHSL-204	NSP2	Immunogenic	SEVGPEHSL	HLA-B*40:01	0.99	0.72	TTCGYLPQNAVVKIY	HLA-DRB5*01:01	4.30	0.04	VVKIYCPACHNSEVGP	0.96	0.66
		Antigenic	NSEVGPEHSL	HLA-B*40:01	0.79	0.82	ATTCGYLPQNAVVKI	HLA-DRB5*01:01	5.20	0.18			
111-RDGCVPLNIIPLTTAAKLMVVIPDY NTYKNTCDGTTFTYASALWE-155		Immunogenic	NTCDGTTFTY	HLA-A*01:01	0.97	-0.03	VPLNIIPLTTAAKLM	HLA-DRB1*08:02	0.25	0.88	MVVIPDYNTYKNTCDG	0.94	0.24
		Antigenic	TTFTYASALW	HLA-B*57:01	0.95	0.40	GCVPLNIIPLTTAAK	HLA-DRB1*08:02	0.27	1.13	VPLNIIPLTTAAKLMV	0.57	0.74
35-GQPITNCVKMLCTHTGTGQAITV TPEANMDQESFGGASCCLYCRCHI DHPNPKGFCDLKGKYVQIPTTCAN DPVGFTLKNT-115	NSP10	Immunogenic	DLKGKYVQI	HLA-B*08:01	0.92	1.38	LKGKYVQIPTTCAND	HLA-DRB1*04:01	0.49	0.63	RCHIDHPNPKGFCDLK	0.93	0.72
		Antigenic	HPNPKGFCDL	HLA-B*07:02	0.69	1.43	DLKGKYVQIPTTCAN	HLA-DRB1*04:01	0.51	0.86	PNPKGFCDLKGKYVQI	0.66	1.55
213-NGNWYDFGDFIQTTPGSGVPVVD SYYSLLMPILTLTRALTAES-255	RdRp	Immunogenic	SLLMPILTL	HLA-A*02:01	0.79	0.21	SYYSLLMPILTLTRA	HLA-DRB1*01:01	0.16	0.55	DFIQTTPGSGVPVVDS	0.93	0.36
		Antigenic	SGVPVVDSY	HLA-B*35:01	0.66	0.59					VDSYYSLLMPILTLTR	0.62	0.47
261-LTKPYIKWDLLKYDFTEERLKLFDR YFKYWDQTYHPNCVNCLDDRCILH CANFNVLFSTVFP-322	RdRp	Immunogenic	KLFDRYFKY	HLA-A*32:01	0.95	-0.05	TEERLKLFDRYFKYW	HLA-DPA1*01:03/ DPB1*02:01	0.76	0.18	YFKYWDQTYHPNCVNC	0.88	0.75
		Antigenic					RLKLFDRYFKYWDQT	HLA-DPA1*01:03/ DPB1*02:01	1.20	0.44			
723-LQHRLYECLYRNRDVDTDFVNEFY AYLRKHFSMMILSDDAVVC-765	RdRp	Immunogenic	DTDFVNEFY	HLA-A*01:01	0.99	0.25	NEFYAYLRKHFSMMI	HLA-DRB1*11:01	0.02	0.23	HRLYECLYRNRDVDTD	0.83	0.23
		Antigenic	YLRKHFSMM	HLA-B*08:01	0.88	0.49	EFYAYLRKHFSMMIL	HLA-DRB1*11:01	0.05	0.39			
851-DGTLMIERFVSLAIDAYPLTKH PNQEYADVFHLYLQYIRKLHDELT GHMLDMYSVMLTNDNTSRYWEPEFY-921	RdRp	Immunogenic	QEYADVFHLY	HLA-B*44:03	0.99	0.27	VFHLYLQYIRKLHDE	HLA-DRB4*01:01	0.37	0.28	GHMLDMYSVMLTNDNT	0.91	0.43
		Antigenic	QEYADVFHL	HLA-B*40:01	0.98	0.36	HMLDMYSVMLTNDNT	HLA-DRB1*04:05	0.42	0.55	HPNQEYADVFHLYLQY	0.77	0.55
150-IPLMYKGLPWNVVRIKIVQMLSDTLKN LSDRVVFVLWAHGFELTSMKYF VKIGPERTCCLC-210	Exon	Immunogenic	NLSDRVVFV	HLA-A*02:03	0.94	0.95	VRIKIVQMLSDTLKN	HLA-DRB4*01:01	0.38	0.29	GFELTSMKYFVKIGPE	0.87	1.17
		Antigenic					PWNVVRIKIVQMLSD	HLA-DRB4*01:01	0.41	0.46			
315-MVVKAALLADKFPVLHDIGNP KAIKCVPQADVEWKFYD AQPCSDKAYKIE-364	Exon	Immunogenic	LLADKFPVL	HLA-A*02:01	0.94	0.08	MVVKAALLADKFPVL	HLA-DPA1*01:03/ DPB1*02:01	1.30	0.40	KCVPQADVEWKFYDAQ	0.80	1.34
		Antigenic	KCVPQADVEW	HLA-B*57:01	0.90	1.09							
977-LNDILSRLDKVEAEVQIDRLITGRLQSLQT YVTQQLIRAAEIR-1019	Spike glycoprotein	Immunogenic	AEVQIDRLI	HLA-B*44:03	0.90	-0.56	VEAEVQIDRLITGRL	HLA-DRB1*03:01	1.10	-0.37	DRLITGRLQSLQTYVT	0.77	-0.36
		Antigenic	RLDKVEAEV	HLA-A*02:01	0.83	0.08	LQTYVTQQLIRAAEI	HLA-DRB4*01:01	2.70	0.02	LNDILSRLDKVEAEVQ	0.51	0.17
175-TTSPISEHDYQIGGYTEKWESGV KDCVVLHSYFTSDYYQLYSTQ LSTDT-223	ORF3a protein	Immunogenic	FTSDYYQLY	HLA-A*01:01	0.98	-0.11	VLHSYFTSDYYQLYS	HLA-DPA1*01:03/ DPB1*04:01	0.17	0.06	TSPISEHDYQIGGYTE	0.93	0.72
		Antigenic	SEHDYQIGGY	HLA-B*44:03	0.91	1.04	HSYFTSDYYQLYSTQ	HLA-DPA1*01:03/ DPB1*04:01	0.33	0.25			
1-MKIILFLALITLATCELYHYQECV RGTTVLLKEPCSSGTYEG-42	ORF7a	Immunogenic	QECVRGTTVL	HLA-B*40:01	0.83	0.60	ILFLALITLATCELY	HLA-DRB1*01:01	0.16	0.19	TCELYHYQECVRGTTV	0.81	0.53
		Antigenic	ILFLALITL	HLA-A*02:01	0.45	0.82							

Table 7
Summary of the most Immunogenic and Antigenic Epitopes along with the Allergic and Toxicity values.

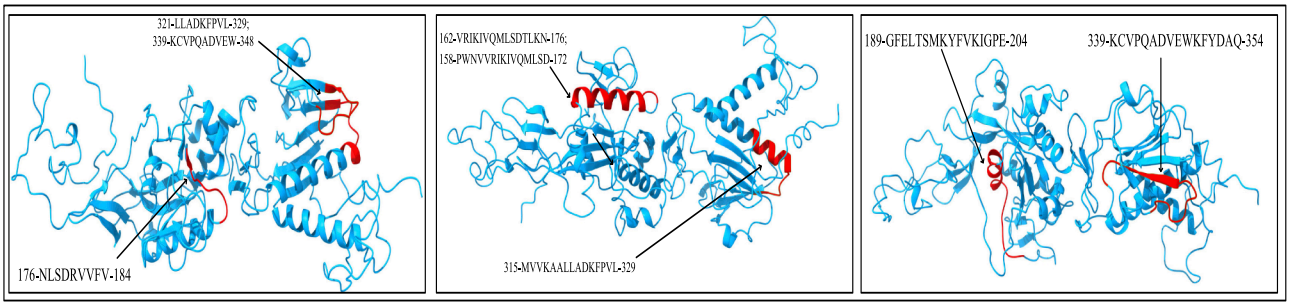
Coded Proteins	MHC-I restricted T-cell Epitopes	Allergic	Toxicity	MHC-II restricted T-cell Epitopes	Allergic	Toxicity	Linear B-cell Epitopes	Allergic	Toxicity
NSP2	SEVGPEHSL	Non-Allergen	Non-Toxin	TTCGYLPQNAVVKIY	Non-Allergen	Non-Toxin	VVKIYCPACHNSEVGP	Allergen	Non-Toxin
	NSEVGPEHSL	Allergen	Non-Toxin	ATTCGYLPQNAVVKI	Non-Allergen	Non-Toxin			
NSP8	NTCDGTTFTY	Allergen	Non-Toxin	VPLNIIPLTTAAKLM	Non-Allergen	Non-Toxin	MVVIPDYNTYKNTCDG	Non-Allergen	Non-Toxin
	TTFTYASALW	Allergen	Non-Toxin	GCVPLNIIPLTTAAK	Non-Allergen	Non-Toxin	VPLNIIPLTTAAKLMV	Non-Allergen	Non-Toxin
NSP10	DLKGKYVQI	Allergen	Non-Toxin	LKGKYVQIPTTCAND	Allergen	Non-Toxin	RCHIDHPNPKGFCDLK	Allergen	Toxin
	HPNPKGFCDL	Allergen	Toxin	DLKGKYVQIPTTCAN	Allergen	Non-Toxin	PNPKGFCDLKGKYVQI	Allergen	Non-Toxin
RdRp	SLLMPILTL	Non-Allergen	Non-Toxin	SYYSLLMPILTLTRA	Non-Allergen	Non-Toxin	DFIQTTPGSGVPVVDS	Non-Allergen	Non-Toxin
	SGVPVVDSY	Allergen	Non-Toxin				VDSYYSLLMPILTLTR	Allergen	Non-Toxin
RdRp	KLFDRYFKY	Non-Allergen	Non-Toxin	TEERLKLFDRYFKYW	Allergen	Non-Toxin	YFKYWDQTYHPNCVNC	Non-Allergen	Toxin
				RLKLFDRYFKYWDQT	Allergen	Non-Toxin			
RdRp	DTDFVNEFY	Allergen	Non-Toxin	NEFYAYLRKHFSMMI	Non-Allergen	Non-Toxin	HRLYECLYRNRDVDTD	Non-Allergen	Toxin
	YLRKHFSMM	Non-Allergen	Non-Toxin	EFYAYLRKHFSMMIL	Non-Allergen	Non-Toxin			
RdRp	QEYADVFHLY	Allergen	Non-Toxin	VFHLYLQYIRKLHDE	Non-Allergen	Non-Toxin	GHMLDMYSVMLTNDNT	Allergen	Non-Toxin
	QEYADVFHL	Allergen	Non-Toxin	HMLDMYSVMLTNDNT	Allergen	Non-Toxin	HPNQEYADVFHLYLQY	Non-Allergen	Toxin
Exon	NLSDRVVFV	Non-Allergen	Non-Toxin	VRIKIVQMLSDTLKN	Non-Allergen	Non-Toxin	GFELTSMKYFVKIGPE	Non-Allergen	Non-Toxin
				PWNVVRIKIVQMLSD	Non-Allergen	Non-Toxin			
Exon	LLADKFPVL	Allergen	Non-Toxin	MVVKAALLADKFPVL	Allergen	Non-Toxin	KCVPQADVEWKFYDAQ	Non-Allergen	Non-Toxin
	KCVPQADVEW	Non-Allergen	Non-Toxin						
Spike glycoprotein	AEVQIDRLI	Non-Allergen	Non-Toxin	VEAEVQIDRLITGRL	Non-Allergen	Non-Toxin	DRLITGRLQSLQTYVT	Non-Allergen	Non-Toxin
	RLDKVEAEV	Allergen	Non-Toxin	LQTYVTQQLIRAAEI	Non-Allergen	Non-Toxin	LNDILSRLDKVEAEVQ	Allergen	Non-Toxin
ORF3a	FTSDYYQLY	Allergen	Non-Toxin	VLHSYFTSDYYQLYS	Non-Allergen	Non-Toxin	TSPISEHDYQIGGYTE	Allergen	Non-Toxin
	SEHDYQIGGY	Non-Allergen	Non-Toxin	HSYFTSDYYQLYSTQ	Non-Allergen	Non-Toxin			
ORF7a	QECVRGTTVL	Non-Allergen	Non-Toxin	ILFLALITLATCELY	Non-Allergen	Non-Toxin	TCELYHYQECVRGTTV	Allergen	Toxin
	ILFLALITL	Non-Allergen	Non-Toxin						

the automatic prediction of protein stability changes for SNPs. PROVEAN and PolyPhen-2 are used to find the deleterious or damaging non-synonymous SNPs. The threshold value of PROVEAN is set to −2.5. If the PROVEAN score of a SNP is *le* this threshold, the corresponding non-synonymous mutation is deleterious. For Polyphen-2, this range is between 0 to 1. If the score is closer to 1, mutations are more confidently considered to be damaging. As reported in Table 2 by considering the consensus of PROVEAN and Polyphen-2, out of the 28 unique amino acid changes, 8 unique changes are deleterious and damaging. Moreover, protein stability is important for considering the functional and structural activity of a protein. Any change in protein stability may cause degradation of proteins. The protein stabilities for the non-synonymous signature SNPs are determined using I-Mutant 2.0. The changes in the protein stability in I-Mutant 2.0 tool is predicted using free energy change values (DDG). A decrease in protein stability is indicated by a zero or a negative value of DDG. Table 2 shows that out of the 8 unique changes, 5 unique changes show a decrease in the stability of the protein structures. As a consequence, A97V in RdRp in 19A, V354L in Nucleocapsid in 19B, Q57H in Nucleocapsid in 20A, R203M in Nucleocapsid in 20B while T85I in NSP2 and Q57H in ORF3a in 20C are the unique amino acid changes which are responsible for defining each clade as they are all deleterious and unstable as well as decrease the stability of the protein structures. All of them are marked in bold in Table 2.

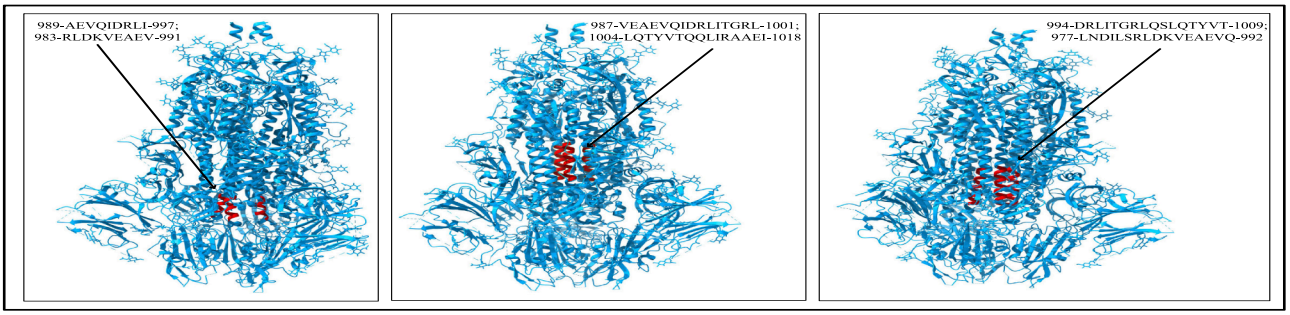
Physico-chemical properties are considered to show the significance of the epitopes as reported in this paper. For each property, the physico-chemical values lie between 0 and 1. The physico-chemical properties for MHC-I, MHC-II restricted T-cells and B-cell epitopes belonging to the 12 CnRs are reported in Supplementary Tables S4, S5 and S6 respectively. As reported in Table S4, MHC-I restricted T-cell epitope SEVGPEHSL has a hydrophobicity value of −0.11, steric hinderance value of 0.52, hydropathicity of −0.64, amphipathicity of 0.44, hydrophilicity of 0.36, net hydrogen of 0.56, charge of −1.5, pI of 4.51 and molecular weight of 954.13. For the other epitopes, their physico-chemical properties are reported in the corresponding tables as well.

For further validation, the conformational 2D non-covalent structures of the MHC-I and MHC-II restricted T-cell epitopes are studied using LigPlot+. Also, it is very important to study their structural characteristics such as binding conformation. Hence, to identify the stable binding interactions, molecular docking of the MHC-I and MHC-II restricted T-cell epitopes are evaluated using AutoDock Vina. For the same, first the 3D structures of the epitopes are prepared with the help of the build structure function of Chimera 1.14 along with the retrieval of the crystal structures of the HLA alleles in PDB format from RCSB Protein Data Bank. To identify the binding energy at the binding groove in the HLA allele, grid search space is set to (60,60,60) with centre of grid at (0,0,0) for X,Y Z coordinate with a spacing parameter of 0.964. The best is selected by higher binding affinity i.e. lowest docking score generated through Autodock Vina. Also, we have used DOE-MBI services such as PROCHECK, ERRAT, Verify3D for Ramchandran Plot, structure quality and 3D structure verification respectively. The results of the docking analysis along with Z-score, respective PDB ID[20], total energy of the 3D complex, van der Waals energy and electric energy of each complex are reported in Tables 8,9 respectively for MHC-I and MHC-II restricted T-cell epitopes. The results of SEVGPEHSL and NEFYAYLRKHFSMMI which are the most immunogenic and HPNPKGFCDL and GCVPLNIIPLTTAAK which are the most antigenic MHC-I and MHC-II restricted T-cell epitopes are shown respectively in Figs. 5–8 while the results of DTDFVNEFY and QEYADVFHLY which are also the most immunogenic MHC-I restricted T-cell epitopes are shown in supplementary Figures S11 and S13 respectively. In these figures, (a) shows the docked complex with epitopes (marked in green) interacting in the HLA pocket where for MHC-I the docking scores are −7.02,

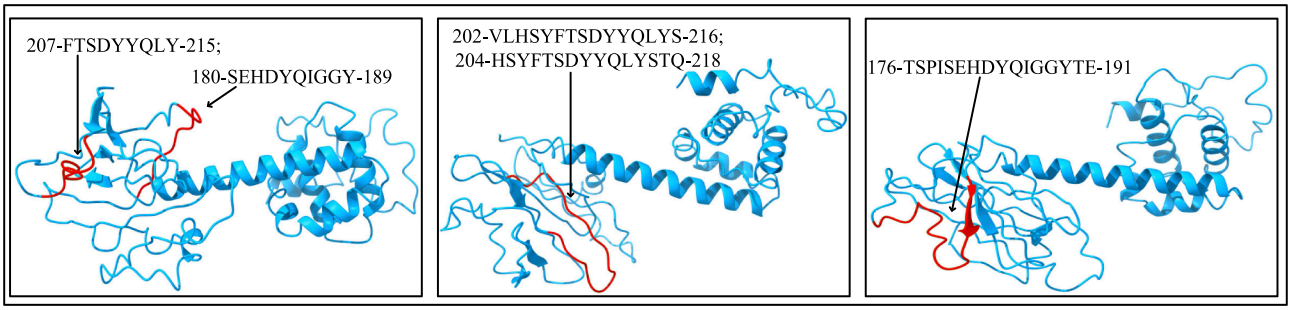
[20] https://www.phla3d.com.br/alleles/index



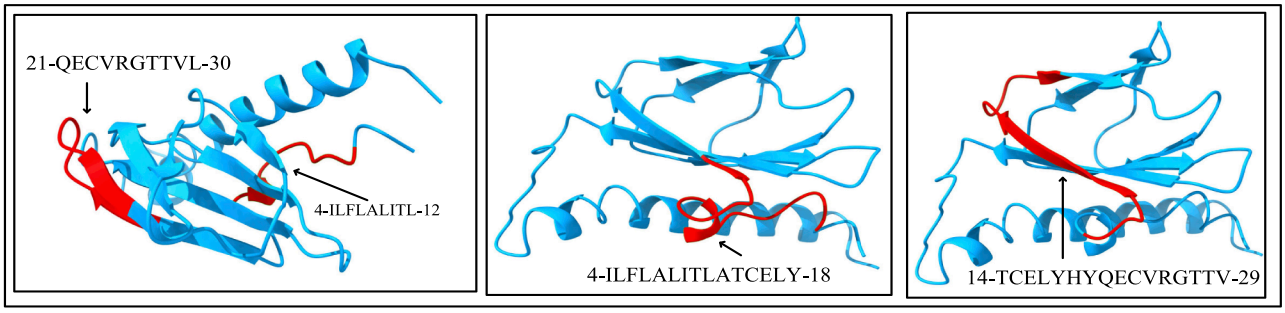
e



f

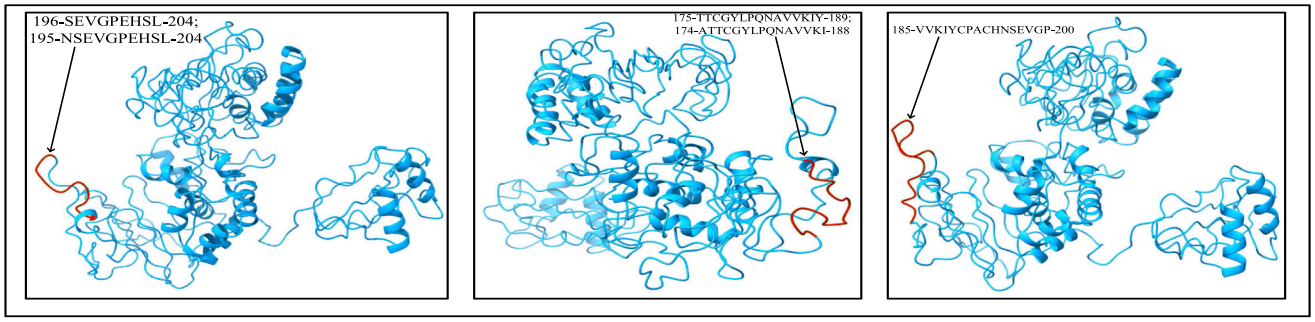


h

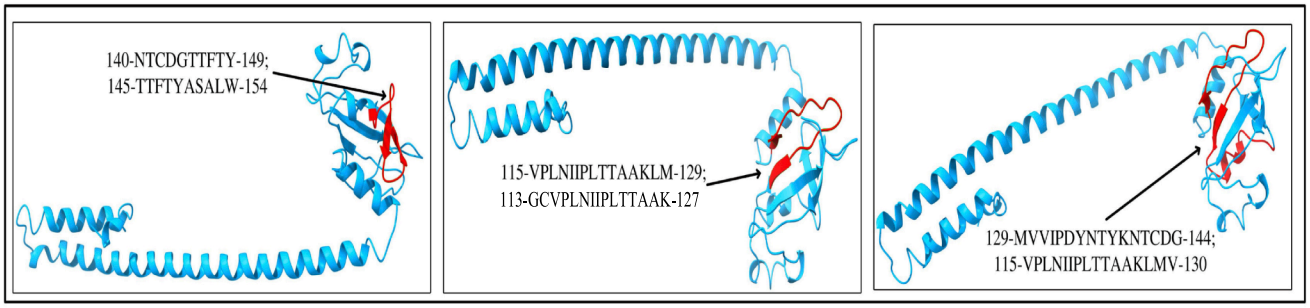


i

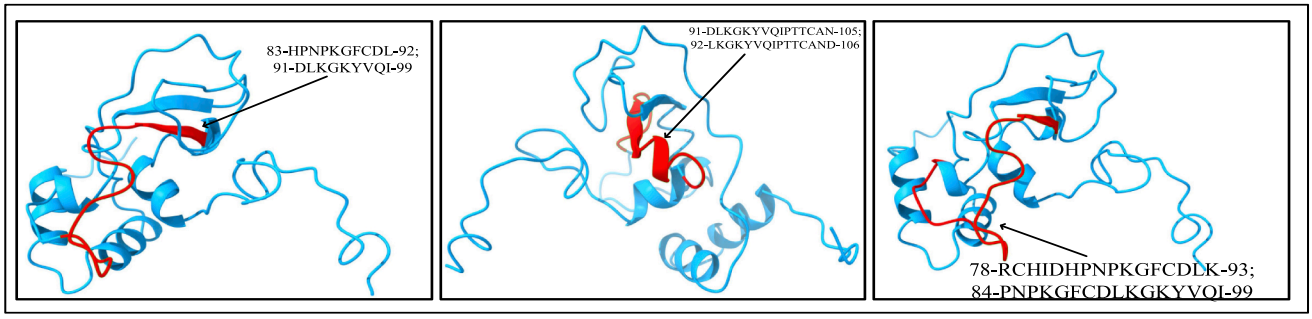
Fig. 4. Modelling of MHC-I, MHC-II restricted T-cell and B-cell epitopes for 12 CnRs belonging to (a) NSP2 (b) NSP8 (c) NSP10 (f) RdRp (f) Exon (g) Spike glycoprotein (h) ORF3a and (i) ORF7a.



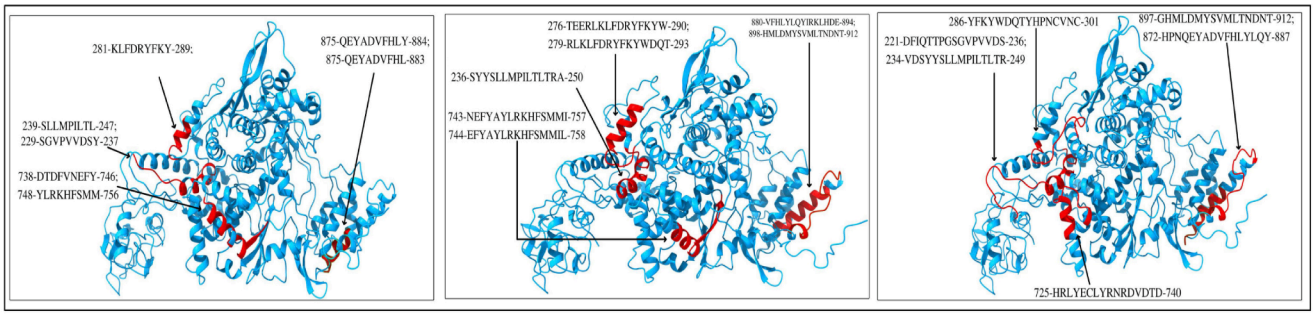
a



b



c



d

Fig. 4. (*continued*).

Table 8
Docking and Z-scores of most Immunogenic and Antigenic MHC-I restricted T-cell epitopes for 12 CnRs.

MHC-I restricted T-cell epitopes	Allele PDB ID	Score from AutoDock Vina	Total Energy	vdW Energy	Electric Energy	ERRAT Score	Z Score
SEVGPEHSL	3LN4:A	-7.02	56.597	4.242	-84.058	92.1127	-8.92
NSEVGPEHSL	3LN4:A	-7.826	62.78	0.135	-71.237	92.1127	-8.92
NTCDGTTFTY	3BO8:A	-7.896	79.478	0.388	-72.211	82.3529	-8.98
TTFTYASALW	3VRI:A	-9.932	131.03	-26.04	-49.8	81.5642	-9.27
DLKGKYVQI	4QRU:A	-8.007	30.829	-7.715	-80.4	80.4469	-9.48
HPNPKGFCDL	4U1H:A	-7.438	51.815	-3.509	-61.083	84.9582	-8.97
SLLMPILTL	3UTQ:A	-8.166	117.669	-10.804	-48.976	83.3333	-9.38
SGVPVVDSY	2CIK:A	-8.074	79.882	-6.491	-77.615	84.0336	-9.28
KLFDRYFKY	5E00:A	-8.323	38.063	0.837	-81.052	85.1955	-8.77
DTDFVNEFY	3BO8:A	-7.786	84.77	-1.521	-75.162	82.3529	-8.98
YLRKHFSMM	4QRU:A	-8.029	40.78	-18.508	-41.459	80.4469	-9.48
QEYADVFHLY	1N2R:A	-8.848	88.793	-9.037	-85.66	85.1955	-8.95
QEYADVFHL	3LN4:A	-7.996	48.824	1.057	-95.906	92.1127	-8.92
NLSDRVVFV	3OX8:A	-7.321	2.558	-17.624	-83.824	82.5843	-9.3
LLADKFPVL	3UTQ:A	-7.845	60.256	-0.423	-73.612	83.3333	-9.38
KCVPQADVEW	3VRI:A	-7.362	44.618	9.799	-82.426	81.5642	-9.27
AEVQIDRLI	1N2R:A	-7.302	-5.739	-14.044	-59.423	85.1955	-8.95
RLDKVEAEV	3UTQ:A	-7.406	-35.156	-10.383	-59.389	83.3333	-9.38
FTSDYYQLY	3BO8:A	-8.007	91.699	-12.984	-63.351	83.3333	-8.98
SEHDYQIGGY	1N2R:A	-9.458	67.521	-29.967	-56.642	85.1955	-8.95
QECVRGTTVL	3LN4:A	-8.409	-0.982	-8.186	-75.82	92.1127	-8.92
ILFLALITL	3UTQ:A	-8.656	123.773	-19.829	-50.913	83.3333	-9.38

Table 9
Docking and Z-scores of most Immunogenic and Antigenic MHC-II restricted T-cell epitopes for 12 CnRs.

MHC-II restricted T-cell epitopes	Allele PDB ID	Score from AutoDock Vina	Total Energy	vdW Energy	Electric Energy	ERRAT Score	Z Score
TTCGYLPQNAVVKIY	1FV1:B	-8.187	51.807	-11.448	-73.616	83.3333	-9.38
ATTCGYLPQNAVVKI	1FV1:B	-7.002	53.457	3.071	-74.542	92.1127	-8.92
VPLNIIPLTTAAKLM	6CPN:B	-7.134	76.07	-0.246	-70.524	82.3529	-8.98
GCVPLNIIPLTTAAK	1X7Q:A	−7.298	117.674	7.064	-70.22	83.7079	−8.91
LKGKYVQIPTTCAND	4MD4:B	-7.168	26.786	18.782	-118.485	84.0336	-9.28
DLKGKYVQIPTTCAN	4MD4:B	-7.598	51.579	-8.601	-62.765	84.0336	-9.28
SYYSLLMPILTLTRA	2G9H:B	-8.185	93.108	-19.626	-34.574	84.0782	-9.21
TEERLKLFDRYFKYW	3WEX:A; 3WEX:B	-8.073	35.351	-8.623	-76.368	83.7079	-8.95
RLKLFDRYFKYWDQT	3WEX:A; 3WEX:B	-8.568	77.593	-17.304	-51.475	88.169	-8.93
NEFYAYLRKHFSMMI	1A6A:B	-8.465	100.048	-14.017	-61.447	87.9552	-9.5
EFYAYLRKHFSMMIL	1A6A:B	-10.032	47.328	-36.397	-46.922	88.4831	-8.97
VFHLYLQYIRKLHDE	1T5W:B	-7.431	33.396	-7.497	-60.178	80.4469	-9.48
HMLDMYSVMLTNDNT	4MD4:B	” -8.019”	88.304	-12.212	-63.943	83.7535	-8.95
VRIKIVQMLSDTLKN	1T5W:B	-6.854	-59.105	37.684	-153.888	77.7465	-9.09
PWNVVRIKIVQMLSD	1T5W:B	-7.877	92.966	-19.085	-38.808	83.3333	-9.38
MVVKAALLADKFPVL	3WEX:A; 3WEX:B	-7.289	7.927	1.388	-98.584	77.7465	-9.09
VEAEVQIDRLITGRL	1A6A:B	-7.845	2.052	-10.221	-87.57	83.7079	-8.95
LQTYVTQQLIRAAEI	1T5W:B	-8.080	24.104	-8.501	-96.551	77.7465	-9.09
VLHSYFTSDYYQLYS	3WEX:A; 3WEX:B	-7.453	40.904	5.179	-116.223	81.5642	-9.27
HSYFTSDYYQLYSTQ	3WEX:A; 3WEX:B	-7.964	107.759	-16.583	-52.05	82.3529	-8.98
ILFLALITLATCELY	2G9H:B	-8.456	39.487	-18.368	-86.629	85.9944	-8.83

−7.786, −8.848 and −7.438 while for MHC-II the scores are −8.465 and −7.298 generated from AutoDock Vina, (b) shows the 2D binding representation between the epitopes and the respective allele pair, (c) shows the ERRAT Score (d) shows the Z-Score where negative scores of −8.92, −8.98, −8.95 and −8.98 for MHC-I and −9.50 and −8.91 for MHC-II represent the stability of the structures of the identified epitopes, (e) represents Ramchandran Plot which has been evaluated using PROCHECK where most favourable region for the residue is shown in the red regions, (f) shows the energy residue plot generated using Verify 3D in Chain A of the docked complex and (g) shows the energy residue plot generated using Verify 3D in Chain B of the docked complex. Similar structural based evaluation are carried out for all the T-cell epitopes of the 12 conserved regions and reported in supplementary figures S4-S42.

It is to be noted that in our previous works [13,14], with a refinement criteria of 60nt, respectively 17 and 23 conserved regions were identified with 30, 24 and 21 and 34, 37 and 29 best immunogenic and antigenic MHC-I and MHC-II T-cell and B-cell epitopes. These experiments were conducted for SARS-CoV-2 sequences till July 2020. As the virus is constantly evolving, a more recent analysis is needed to understand the evolution of the epitopes. Therefore, this work which uses sequences till January 2021 is very relevant in current scenario of constant virus mutation.

5. Conclusion

In the past two years, India has witnessed different surges of COVID-19 cases. Hence, it is important to provide a comprehensive bioinformatics pipeline to understand the virus evolution for identifying the mutation points as SNPs, conserved regions and design potential candidates for vaccine design. In this regard, initially, multiple sequence

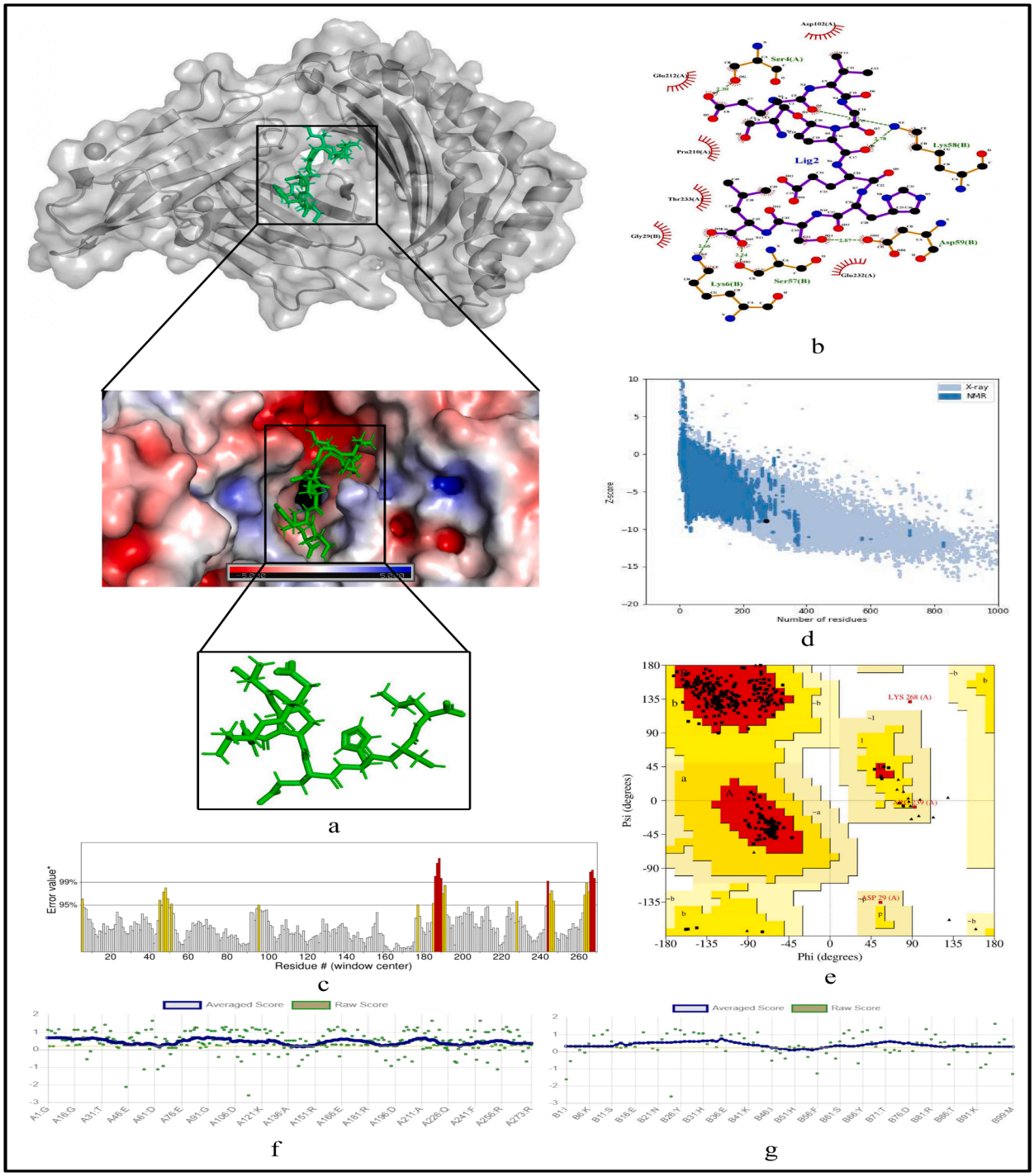


Fig. 5. Structural analysis for the most immunogenic MHC-I restricted T-cell epitope "SEVGPEHSL" in 12 CnRs (a) Docking structure of MHC-I restricted T-cell epitope (b) 2D pose representation between the epitope and HLA allele showing the different non-covalent bonds (c) ERRAT Score (d) Z-Score plot (e) Ramachandran plot of the epitope allele structure showing lower energy sites of the residues in different frames and (f) Verify 3D scores in Chain A of the docked complex (g) Verify 3D scores in Chain B of the docked complex.

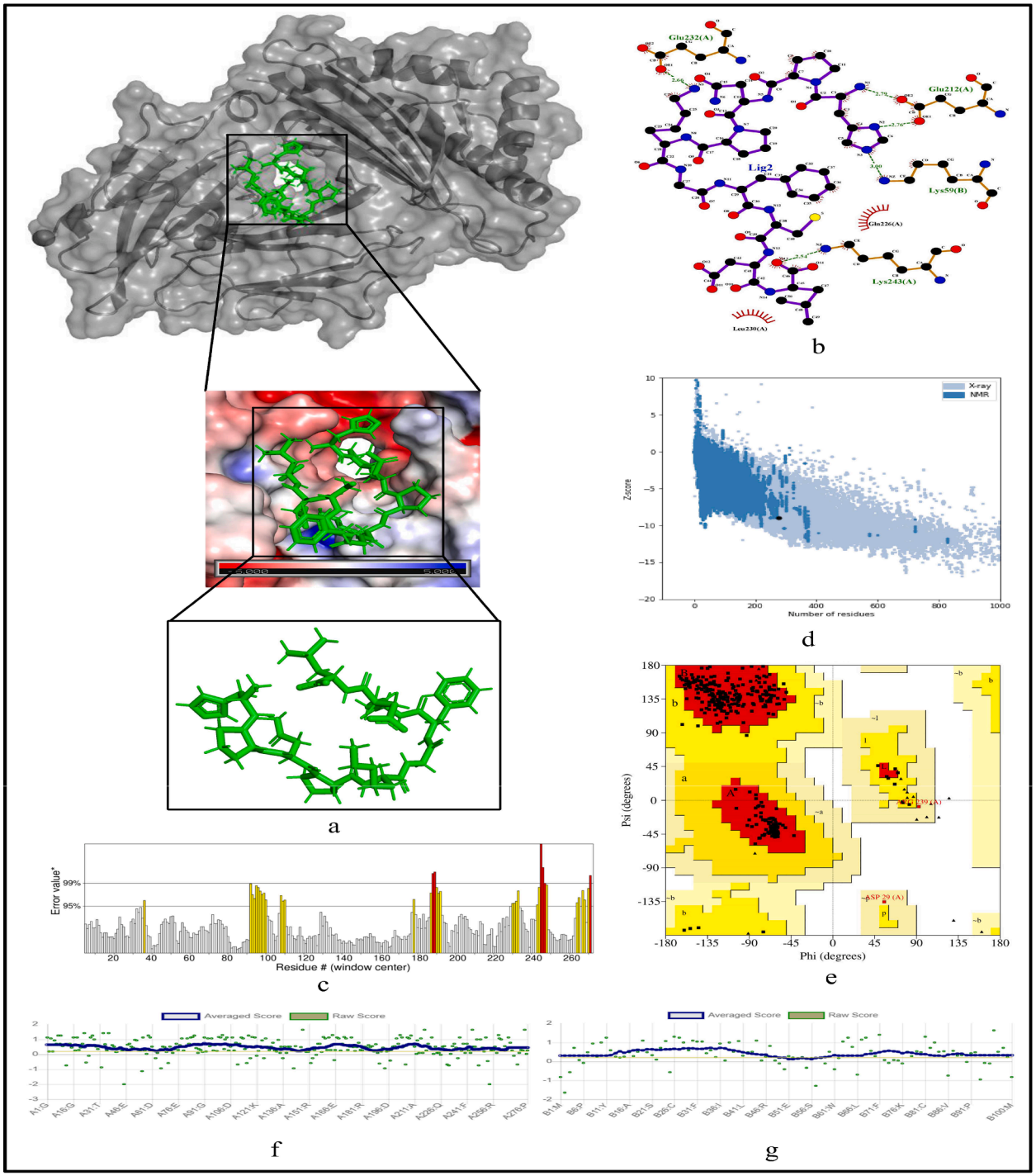


Fig. 6. Structural analysis for the most antigenic MHC-I restricted T-cell epitope "HPNPKGFCDL" in 12 CnRs (a) Docking structure of MHC-I restricted T-cell epitope (b) 2D pose representation between the epitope and HLA allele showing the different non-covalent bonds (c) ERRAT Score (d) Z-Score plot (e) Ramachandran plot of the epitope allele structure showing lower energy sites of the residues in different frames and (f) Verify 3D scores in Chain A of the docked complex (g) Verify 3D scores in Chain B of the docked complex.

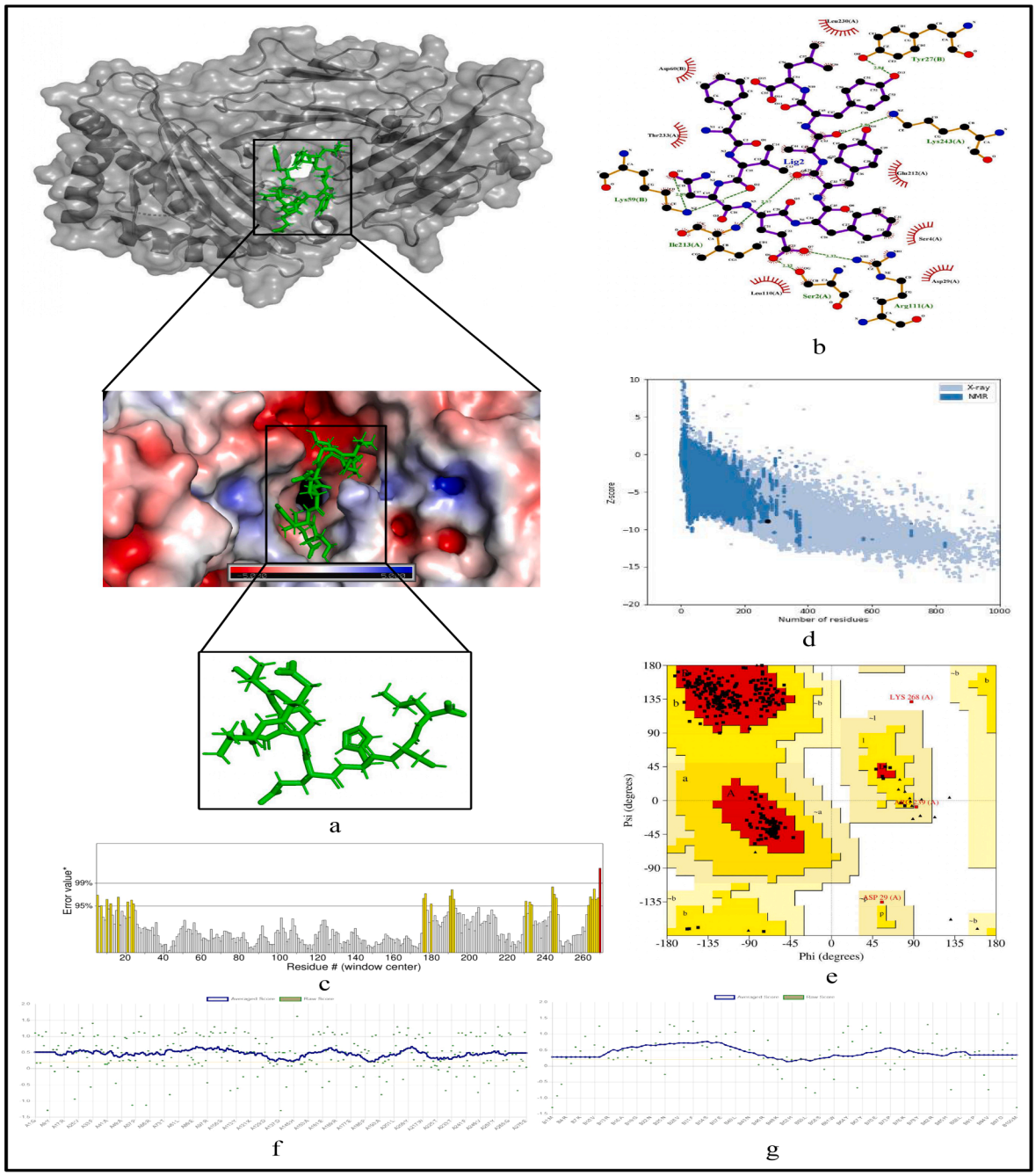


Fig. 7. Structural analysis for the most immunogenic MHC-II restricted T-cell epitope "NEFYAYLRKHFSMMI" in 12 CnRs (a) Docking structure of MHC-I restricted T-cell epitope (b) 2D pose representation between the epitope and HLA allele showing the different non-covalent bonds (c) ERRAT Score (d) Z-Score plot (e) Ramachandran plot of the epitope allele structure showing lower energy sites of the residues in different frames and (f) Verify 3D scores in Chain A of the docked complex (g) Verify 3D scores in Chain B of the docked complex.

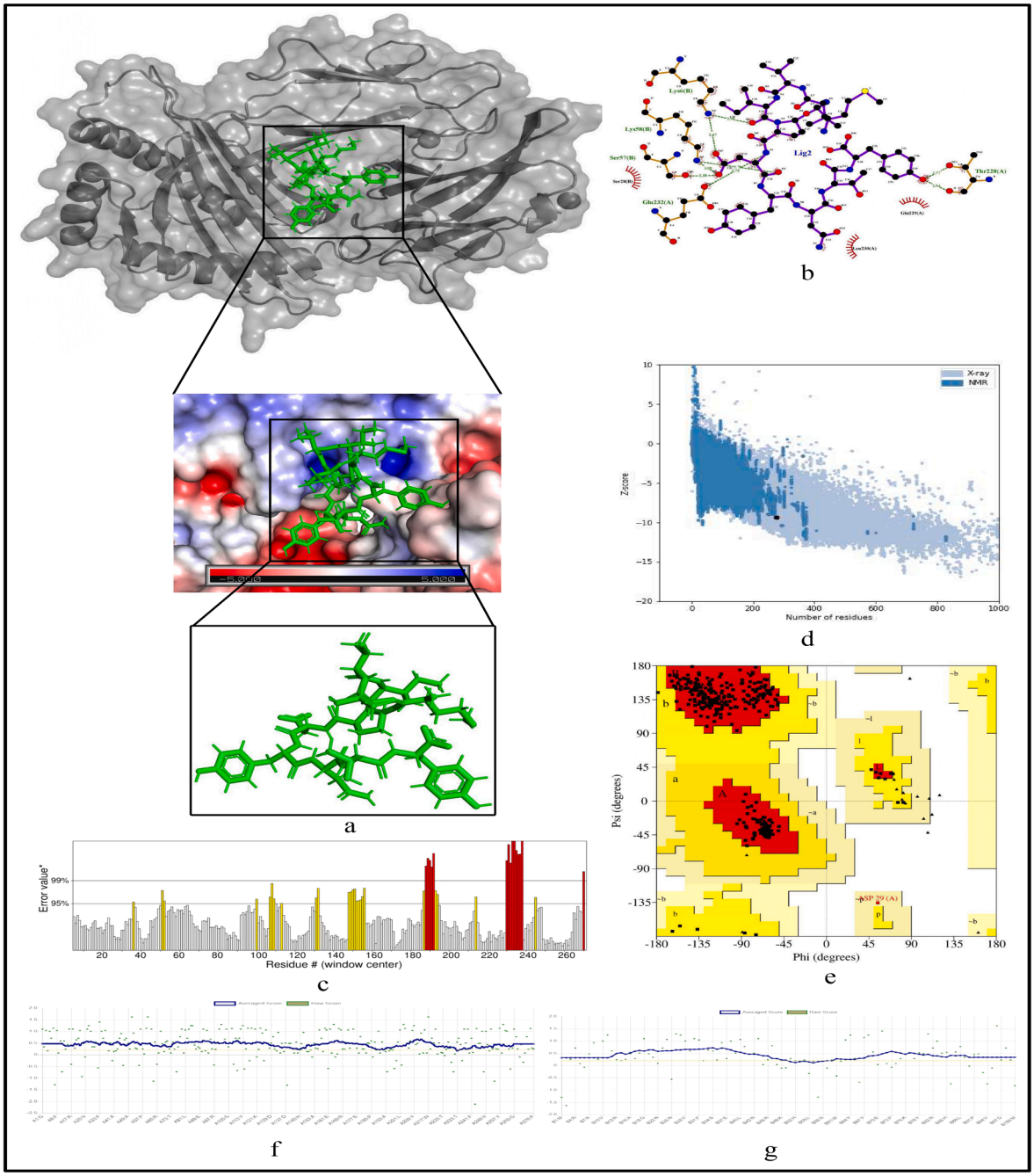


Fig. 8. Structural analysis for the most antigenic MHC-II restricted T-cell epitope "GCVPLNIIPLTTAAK" in 12 CnRs (a) Docking structure of MHC-I restricted T-cell epitope (b) 2D pose representation between the epitope and HLA allele showing the different non-covalent bonds (c) ERRAT Score (d) Z-Score plot (e) Ramachandran plot of the epitope allele structure showing lower energy sites of the residues in different frames and (f) Verify 3D scores in Chain A of the docked complex (g) Verify 3D scores in Chain B of the docked complex.

alignment of 4996 Indian SARS-CoV-2 genomes as a case study are carried out using MAFFT followed by phylogenetic analysis by Nextstrain to identify virus clades, resulting in 5 virus clades; 19A, 19B, 20A, 20B and 20C. Thereafter, mutation points as SNPs are identified in each clade from which top 10 signature SNPs are further identified based on their frequency in each clade. 40 unique signature SNPs are thus identified from the total 50 signature SNPs resulting in 23 non-synonymous signature SNPs which provides 28 amino acid changes in protein. These changes are visualised in their respective protein structure as well. The sequence and structural homology-based prediction of the non-synonymous signature SNPs along with their protein structural stability are evaluated to judge the characteristics of the identified clades. 40 unique signature SNPs are thus identified from the total 50 signature SNPs resulting in 23 non-synonymous signature SNPs which provide 28 amino acid changes in protein. As a consequence, A97V in RdRp in 19A, V354L in Nucleocapsid in 19B, Q57H in Nucleocapsid in 20A, R203M in Nucleocapsid in 20B while T85I in NSP2 and Q57H in ORF3a in 20C are the unique amino acid changes which are responsible for defining each clade as they are all deleterious and unstable as well as they decrease the protein structural stability. Furthermore, based on the entropy of each genomic coordinate of the aligned sequences, 473 conserved regions are identified which are then refined based on the criteria that their lengths are greater than 125nt and their BLAST specificity score as query coverage is equal to 100%. This refinement results in 12 conserved regions belonging to NSP2, NSP8, NSP10, RdRp, Exon, Spike glycoprotein, ORF3a and ORF7a proteins. Based on length, one conserved region belonging to NSP10 gene is considered to be the potential target for which the corresponding primers and probes are reported for SARS-CoV-2 detection. The 12 conserved regions are then used to identify the T-cell and B-cell epitopes along with their immunogenic and antigenic scores. Such scores are then used to select the most immunogenic and antigenic T-cell and B-cell epitopes resulting in 22 MHC-I and 21 MHC-II restricted T-cell epitopes with 10 unique HLA alleles each and 17 B-cell epitopes. Finally, the relevance of these epitopes are validated by showing the binding conformation of the MHC-I and MHC-II restricted T-cell epitopes with respect to HLA alleles. Also, the physico-chemical properties of the epitopes are reported along with the structural properties using Ramchandran plot, ERRAT scores and Z-Scores. Hence, from genetic variability to synthetic pipeline, a comprehensive bioinformatics pipeline is presented in this study to fight against SARS-CoV-2.

6. Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

7. Availability of data and materials

The aligned 4996 Indian SARS-CoV-2 genomes with the reference sequence and the final results of this work are available at 'http://www.nitttrkol.ac.in/indrajit/projects/COVID-Pipeline-5K/". Moreover, the SARS-CoV-2 genomes used in this work are publicly available at GISAID database..

Consent for publication

Not applicable.

Funding

This work has been partially supported by CRG short term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India.

Author contributions

Nimisha Ghosh: Formal analysis; Methodology, Coding; Visualization; Writing - original draft & editing, **Indrajit Saha**: Conceptualization; Data curation; Supervision; Funding acquisition; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Writing - review & editing, **Nikhil Sharma**: Methodology; Visualization; Writing - review & editing, **Suman Nandi**: Conceptualization; Formal analysis; Software; Validation; Visualization; Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

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

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.intimp.2022.109224.

References

[1] P. Zhou, X.L. Yang, X.G. Wang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273, https://doi.org/10.1038/s41586-020-2012-7.
[2] J. Xu, S. Zhao, T. Teng, et al., Systematic comparison of two animal-to-human transmitted human coronaviruses: Sars-cov-2 and sars-cov, Viruses 12 (2020) 244, https://doi.org/10.3390/v12020244.
[3] M. Makoni, South africa responds to new sars-cov-2 variant, The Lancet 397 (2021) 267, https://doi.org/10.1016/S0140-6736(21)00144-6.
[4] J. Tang, O. Toovey, K. Harvey, et al., Introduction of the south african sars-cov-2 variant 501y.v2 into the uk, J. Infect. (2021), https://doi.org/10.1016/j.jinf.2021.01.007, 01.
[5] I. Alam, A. Radovanovic, R. Incitti, et al., Covmt: an interactive sars-cov-2 mutation tracker, with a focus on critical variants, Lancet. Infect. Dis 21 (2021) 602, https://doi.org/10.1016/S1473-3099(21)00078-5.
[6] C. Ascoli, Could mutations of sars-cov-2 suppress diagnostic detection? Nat. Biotechnol. 39 (2021) 1–2, https://doi.org/10.1038/s41587-021-00845-3.
[7] F. Yuan, L. Wang, Y. Fang, et al., Global snp analysis of 11,183 sars-cov-2 strains reveals high genetic diversity, Transboundary and Emerging Diseases (11 2020). doi:10.1111/tbed.13931.
[8] X. Tang, C. Wu, X. Li, et al., On the origin and continuing evolution of sars-cov-2, National Science Review (2020), 03.
[9] A. Maitra, M. Sarkar, H. Raheja, et al., Mutations in sars-cov-2 viral rna identified in eastern india: Possible implications for the ongoing outbreak in india and impact on viral structure and host susceptibility, Journal of Biosciences 45 (12 2020). doi: 10.1007/s12038-020-00046-1.
[10] I. Saha, N. Ghosh, A. Pradhan, et al., Whole genome analysis of more than 10000 sars-cov-2 virus unveils global genetic diversity and target region of nsp6, Briefings in Bioinformatics 22 (2) (2021) 1106–1121, https://doi.org/10.1093/bib/bbab025.
[11] A. Nagy, S. Pongor, B. Győrffy, Different mutations in sars-cov-2 associate with severe and mild outcome, Int. J. Antimicrob. Agents 57 (2020) 106272, https://doi.org/10.1016/j.ijantimicag.2020.106272.
[12] W. Zhu, C. Wang, B.Z. Wang, From variation of influenza viral proteins to vaccine development, International Journal of Molecular Sciences 18 (07 2017). doi: 10.3390/ijms18071554.
[13] N. Ghosh, N. Sharma, I. Saha, Immunogenicity and antigenicity based t-cell and b-cell epitopes identification from conserved regions of 10664 sars-cov-2 genomes, Infection Genetics and Evolution 4 (92) (2021) 104823, https://doi.org/10.1016/j.meegid.2021.104823.
[14] N. Ghosh, N. Sharma, I. Saha, et al., Genome-wide analysis of indian sars-cov-2 genomes to identify t-cell and b-cell epitopes from conserved regions based on immunogenicity and antigenicity, Int. Immunopharmacol. 22 (91) (2020) 107276, https://doi.org/10.1016/j.intimp.2020.107276.
[15] A. Alam, A. Khan, N. Imam, et al., Design of an epitope-based peptide vaccine against the sars-cov-2: A vaccine-informatics approach, Briefings in Bioinformatics 22 (2020) 1–15, https://doi.org/10.1093/bib/bbaa340.
[16] M. Rahman, M. Hoque, R. Islam, et al., Epitope-based chimeric peptide vaccine design against s, m and e proteins of sars-cov-2 1 etiologic agent of global

pandemic covid-19: an in silico approach, PeerJ (2020) e9572, https://doi.org/10.1101/2020.03.30.015164.
[17] R. Ling, Y. Dai, B. Huang, et al., In silico design of antiviral peptides targeting the spike protein of sars-cov-2, Peptides 130 (2020) 170328, https://doi.org/10.1016/j.peptides.2020.170328.
[18] Y. Vashi, V. Jagrit, S. Kumar, Understanding the b and t cells epitopes of spike protein of severe respiratory syndrome coronavirus-2: A computational way to predict the immunogens, Infection, Genetics and Evolution (2020), https://doi.org/10.1101/2020.04.08.013516, 04.
[19] K. Katoh, K. Misawa, K. Kuma, et al., Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform, Nucleic Acids Res. 30 (14) (2002) 3059–3066, https://doi.org/10.1093/nar/gkf436.
[20] J. Hadfield, C. Megill, S. Bell, et al., Nextstrain: real-time tracking of pathogen evolution, Bioinformatics (Oxford, England) 34 (2018), https://doi.org/10.1093/bioinformatics/bty407.
[21] J. Sidney, C. Dow, B. Mothé, et al., A systematic assessment of mhc class ii peptide binding predictions and evaluation of a consensus approach, PLoS computational biology 4 (2008) e1000048, https://doi.org/10.1371/journal.pcbi.1000048.
[22] S. Saha, G. Raghava, Prediction methods for b-cell epitopes, Methods in molecular biology (Clifton, N.J.) 409 (2007) 387–394, https://doi.org/10.1007/978-1-60327-118-9_29.
[23] A.C. Wallace, A.R. Laskowski, J.M. Thornton, Ligplot: a program to generate schematic diagrams of protein-ligand interactions, Protein Engineering, Design and Selection 8 (2) (1995) 127–134, https://doi.org/10.1093/protein/8.2.127.
[24] M. Jespersen, B. Peters, M. Nielsen, et al., Bepipred-2.0: Improving sequence-based b-cell epitope prediction using conformational epitopes, Nucleic acids research 45 (05 2017). doi:10.1093/nar/gkx346.
[25] M.A. Rauf, Ligand docking and binding site analysis with pymol and autodock/vina, International Journal of Basic and Applied Sciences 4 (2015) 168–177, https://doi.org/10.14419/ijbas.v4i2.4123.
[26] A.R. Laskowski, W.M. MacArthur, S.D. Moss, et al., Procheck: a program to check the stereochemical quality of protein structures, J. Appl. Crystallogr. 26 (1993) 283–291.
[27] C. Colovos, T.-O. Yeates, Verification of protein structures: patterns of nonbonded atomic interactions, Protein science 2 (9) (1993) 1511–1519.
[28] D. Eisenberg, R. Luethy, J. Bowie, Verify3d: Assessment of protein models with three-dimensional profiles, Methods in enzymology 277 (1997) 396–404, https://doi.org/10.1016/S0076-6879(97)77022-8.
[29] M. Wiederstein, M. Sippl, Prosa-web: interactive web service for the recognition of errors in three-dimensional structures of proteins, Nucleic acids research 35 (2007) W407–10, https://doi.org/10.1093/nar/gkm290.
[30] E. Ong, M.U. Wong, A. Huffman, et al., Covid-19 coronavirus vaccine design using reverse vaccinology and machine learning, Frontiers in Immunology 11 (2020) 1581, https://doi.org/10.3389/fimmu.2020.01581.
[31] V. Jurtz, S. Paul, M. Andreatta, et al., Netmhcpan-4.0: Improved peptide–mhc class i interaction predictions integrating eluted ligand and peptide binding affinity data, J. Immunol. 199 (2017) ji1700893, https://doi.org/10.4049/jimmunol.1700893.
[32] I. Doytchinova, D. Flower, Vaxijen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. bmc bioinformatics 8:4, BMC bioinformatics 8 (2007) 4, https://doi.org/10.1186/1471-2105-8-4.
[33] Y. Choi, A.P. Chan, Provean web server: a tool to predict the functional effect of amino acid substitutions and indels, Bioinformatics 31 (16) (2015) 2745–2747, https://doi.org/10.1093/bioinformatics/btv195.
[34] I.A. Adzhubei, S. Schmidt, L. Peshkin, et al., A method and server for predicting damaging missense mutations, Nature methods 7 (4) (2010) 248–249, https://doi.org/10.1038/nmeth0410-248.
[35] E. Capriotti, P. Fariselli, R. Casadio, I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure, Nucleic Acid Res. 33 (2005) 306–310, https://doi.org/10.1093/nar/gki375.