SARS-CoV-2 transcriptome analysis and molecular cataloguing of immunodominant epitopes for multi-epitope based vaccine design

Sandeep Kumar Kushwaha[1*], Veerbhan Kesarwani[1,2], Samraggi Choudhury[1], Sonu Gandhi[1], Shailesh Sharma[1]

[1]DBT-National Institute of Animal Biotechnology (NIAB), Hyderabad-500032, Telangana, India.

[2]Center of Bioinformatics, University of Allahabad, Prayagraj-211002, Uttar Pradesh, India.

Email: sandeep@niab.org, veerbhan8586@gmail.com samraggi.1995@gmail.com, gandhi@niab.org.in shailesh.sharma@niab.org.in

*Corresponding author

Abstract

SARS-CoV-2 is a single-stranded RNA virus that has caused more than 0.29 million deaths worldwide as of May 2020, and influence of COVID-19 pandemic is increasing continuously in the absence of approved vaccine and drug. Moreover, very limited information is available about SARS-CoV-2 expressed regions and immune responses. In this paper an effort has been made, to facilitate vaccine development by proposing multiple epitopes as potential vaccine candidates by utilising SARS-CoV-2 transcriptome data. Here, publicly available RNA-seq data of SARS-CoV-2 infection in NHBE and A549 human cell lines were used to construct SARS-CoV-2 transcriptome to understand disease pathogenesis and immune responses. In the first step, epitope prediction, MHC class I and II gene identification for epitopes, population coverage, antigenicity, immunogenicity, conservation and cross-reactivity analysis with host antigens were performed by using SARS-CoV-2 transcriptome, and in the second step, structural compatibility of identified T-and B-cell epitopes were evaluated with MHC molecules and B-cell receptors through molecular docking studies. Quantification of MHC gene expression was also performed that indicated high variation in allele types and expression level of MHC genes with respect to cell lines. In A549 cell line, HLA-A*30:01:01:01 and HLA-B*44:03:01:01 were highly expressed, whereas 92 variants of HLA-A*24 genes such as HLA-A*24:02:01:01, HLA-A*24:286, HLA-A*24:479Q, HLA-A*24:02:134 and HLA-A*24:02:116 were highly expressed in NHBE cell lines. Prevalence of HLA-A*24 alleles was suggested as risk factors for H1N1 infection, and associated with type-1 diabetes. HLA-C*03:03, linked with male infertility factors was also highly expressed in SARS-CoV-2 infected NHBE cell lines. Finally, three potential T-cell and five B-cell epitopes were selected for molecular docking studies with twenty-two MHC molecules and two B-cell receptors respectively. The results of *in silico* analysis indicated that proposed epitopes have high potential to recognize immune response of SARS-CoV-2 infection. This study will facilitate *in vitro* and *in vivo* vaccine related research studies.

Keywords: SARS-CoV-2, COVID-19, T- and B-cell, Peptide, Epitopes, Vaccine

1. Introduction

Corona viruses are a group of related viruses that are responsible for causing diseases ranging from the common cold to severe diseases like Middle East respiratory syndrome (MERS), SARS-CoV-2 in mammals and birds. SARS-CoV-2 is a positive sense single-stranded RNA virus belonging to the family *Coronaviridae* and subgenus *Sarbecovirus*. It is responsible for the widespread global pandemic causing an upper respiratory tract infection of humans [1]. SARS-CoV-2 virion ranges from approximately 50-200 nm in diameter [2]. SARS-CoV-2 is made up of four structural proteins known as the S (spike), E (envelope), M (membrane) and N (nucleocapsid) proteins. The nucleocapsid protein contains the viral RNA and the spike, membrane, envelope make up the viral envelope. The spike protein is responsible for the viral attachment with angiotensin-converting enzyme 2 (ACE2) receptors and facilitates entry into the host cells [3]. The ACE2 receptors are present in the goblet (secretory) cells of ciliated cells in the nose, back of the throat, lungs, gut, heart muscles and kidney which facilitates the hand to mouth transmission route. The viral RNA is released in the nasal cells when the transmembrane serine protease 2 (TMPRSS2) splits the spike proteins, and enters inside the cell, the viral genetic material replicates into millions. Seroconversion of SARS-CoV-2 took place within four days of infection and was found in most patients by day 14 and persistent specific IgG and antibody production was reported even after 2 years of infection [4]. Whereas limited serological details of SARS-CoV-2 are available at the moment, it is reported that a patient showed the presence of IgM after 9 days of infection, and later production of IgG after 2 weeks [5]. In an *in vitro* plaque testing with patient sera, it was confirmed that it is able to neutralize SARS-CoV highlighting the successful mounting of humoral response [6]. The current evidences have shown that Th1 immune response can be successful for controlling SARS-CoV and may work for SARS-CoV-2 as well, since the epitopes overlap for both, the T-epitopes can be identified and will be valuable for designing the cross-reactive vaccines.

Epitopes are the antigenic regions of an antigen, causing an immune response which is identified by antibodies generated from T-and B-cells. T-cell epitopes present on the cell surface binds to the major histocompatibility complex (MHC) molecules. The MHC I molecules presents peptides of 8-11 amino acids in length, which are $CD8^+$ T-cell epitopes. In contrast, MHC II molecules present longer peptides of 13-17 amino acids in length, which are $CD4^+$ T-cell epitopes. The epitope-based vaccine development offers prospective advantages over the whole protein approach because the immune response against highly reserved epitopes over a widespread population can be used for the treatment of highly variable pathogens [7, 8]. Various successful studies were reported for the epitope-based vaccine design against West Nile virus [9], dengue virus [10], chikungunya virus [11], shigellosis [12] etc. COVID-19 first cases were observed in Wuhan, China in December 2019, which seems to be the origin of SARS-CoV-2 virus. As of April 2020, there are more than 3.04 million confirmed cases, with 211 thousand deaths globally. Vaccines and commercial detection kits are mostly in the developmental stages to combat this viral infection, and currently, chloroquine and hydroxychloroquine drugs are being

used for treatment, but there is no approved drug or vaccine triggering the immune response in the body against SARS-CoV-2 in the market.

In the present study, an integrated bioinformatics approach was used to identify expressed T- and B-cell epitopes from RNA-seq data of SARS-CoV-2 infection in normal human bronchial epithelial (NHBE) and human adeno carcinomic alveolar basal epithelial (A549). To the best of our knowledge, no previous study has been reported a list of expressed T-and B-cell epitopes for multi-epitope based vaccine development. The specific objectives of this research study were: (1) SARS-CoV-2 transcriptome construction and annotation to explore expressed region of SARS-CoV-2 genome, (2) identification of potential T-and B-cell epitopes by using SARS-CoV-2 transcriptome, (3) modelling and docking studies to explore structural compatibility of epitopes with MHC complexes and B-cell receptors, and (4) gene expression of MHC class I and II genes by using RNA-seq data.

2. Material and Method

2.1. SARS-CoV-2 data retrieval, processing and transcriptome assembly

Due to the recent outbreak of SARS-CoV-2, several countries were started to generate molecular resources to understand pandemic caused by SARS-CoV-2. We used publicly available transcriptome data (PRJNA615032) of SARS-CoV-2 infection in A549 and NHBE cell lines [13]. All the available data were download from the sequence read archive of NCBI database and fastq-dump program of SRAtoolkit [14] was used to extract fastq reads. Quality assessment and control of RNA-seq data was performed through the FastQC version 0.11.5 [15], MultiQC version 1.8 [16] and trimmomatic version 0.39 software [17]. All high-quality reads were mapped over SARS-CoV-2 isolate Wuhan-Hu-1(MN908947.3) by using HISAT2 version 2.1.0 on default parameters [18]. Samtools version 1.1.0 [19] and Bedtools version 2.26.0 [20] were used to extract all the mapped read from each sample and extracted reads were used to construct *de novo* assemblies by using the Trinity assembler version 2.5.1 [21]. TransDecoder program [22] was used to generate protein sequence from assembled transcriptome. Kallisto, a pseudo aligner for bulk RNA-seq data alignment, was used for expression quantification [23].

2.2. T-Cell epitope prediction from SARS-CoV-2 transcriptome sequences

T-cell epitopes are short peptide fragments of infectious agents such as viruses and bacteria, which has potential to induce specific immune responses and can be used as a key molecular resource for epitope-based vaccine design. NetCTL 1.2 program [24] was used to predict cytotoxic T-lymphocyte (CTL) epitopes from protein sequences, translated from assembled SARS-CoV-2 transcriptome. NetCTL 1.2 has 12 super types i.e. A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, and B62, and combined score of proteasomal C terminal cleavage for $CD8^+$ T-cell epitopes, MHC class I binding, and TAP transport efficiency was considered for epitope identification at threshold value 1.00. To explore antigenic potential of identified peptides, VaxiJen v2.0 [25] server was used at threshold value 0.4, and IEDB program (http://tools.iedb.org/immunogenicity/) was used to identify immunogenicity score for each identified epitopes. Combination of antigenicity and immunogenicity was used to select highly

antigenic and immunogenic peptide for further analysis. To explore MHC genes and alleles, all the identified epitopes were analysed through IEDB program mhci (http://tools.iedb.org/mhci/download/) and mhcii (http://tools.iedb.org/mhcii/download/). Peptide length of predicted epitopes 9.0 and inhibitory concentration (IC50) value less than or equal to 200nM were selected as parameters for the identification of MHC class I and II binding genes and alleles. Conservation level of selected epitopes were calculated from IEDB program conservancy (http://tools.iedb.org/conservancy/) by using transcriptome sequences, and publically available SARS-CoV-2 protein sequences. IEDB population coverage tool (http://tools.iedb.org/population/) was used to analyze population coverage through predicted MHC alleles of epitopes [26, 27]. Peptide toxicity prediction was performed through the ToxinPred web server [28]. Cross-reactivity with host antigenic proteins might leads to adverse immune responses. Therefore, selected epitopes were checked for similarities with the human proteome sequences (Homo sapiens: GRCh38) through standalone NCBI BLAST similarity search tool.

2.3. B-cell epitopes prediction from SARS-CoV-2 transcriptome sequences

Sequence and structure based approaches were used to identify B-cell epitopes. In the sequence based approach, VaxiJen server was used to identify most antigenic proteins from translated transcriptome, and BepiPred-2.0 program [29] was used to identify B-cell epitopes from the identified antigenic proteins. IEDB conformational B-cell prediction tool ElliPro (http://tools.iedb.org/ellipro/) was used to predict epitopes based on protein structure with the parameters PI (protrusion index) value 0.8 as minimum score, and 7 Å as maximum distance [30]. Protein sequences of SARS-CoV-2 transcriptome were showed strong sequence similarity with modelled 3D structure of SARS-CoV-2 genome at Zhang lab. Hence, we downloaded 24 structure of SARS-CoV-2 genome from Zhang lab (https://zhanglab.ccmb.med.umich.edu/) for structure-based epitope prediction. Epitopes identified by both the approaches were evaluated for toxicity, antigenicity, and immunogenicity same as done for T-cell epitopes. Cross-reactivity of selected epitopes were checked with human proteome sequences.

2.4. Molecular docking studies

Selected epitopes were used for structural compatibility and interaction analysis with available MHC class I and II genes, and B-cell receptor structures. Protein structure of MHC class genes, and B-cell receptors were retrieved from RCSB Protein Data Bank (PDB) (https://www.rcsb.org/) in PDB format. To identify the interactions between predicted T- and B- cell epitopes and protein receptors, the molecular docking studies were performed using AutoDockTools, AutoDock Vina and CABS-dock server [31-33]. Structural compatibility and interaction prediction between peptide and receptor in real world is even more complex and challenging. Therefore, initial screening of peptides for receptors were performed through Autodock, whereas CABSdocks, considered full flexibility of peptide and small fluctuation in receptor backbone, was used for final binding and compatibility analysis. Protein complex visualization and hydrogen bonds were calculated through UCSF chimera [34] and LIGPLOT software package [35], and also ensured the number of genuine hydrogen bonds through cavity prediction by using D3Packets server [36].

3. Results

T- and B-cells are the working horses of the adaptive immune system which are capable to produce immunological protection against specific pathogens. The function of T- and B cells are based on the recognition of antigens through specialized receptors and recognized antigenic regions are known as epitopes. Therefore, identification of epitopes for pathogens is crucial for the understanding of disease etiology, disease diagnostics, and epitope-based vaccine development. In this study, an effort was made to identify expressed T- and B-cell epitopes from RNA-seq data of SARS-CoV-2 infection in human NHBE and A549 cell lines. To achieve this goal, various bioinformatics approaches, tools and software were used as summarized in figure-1.

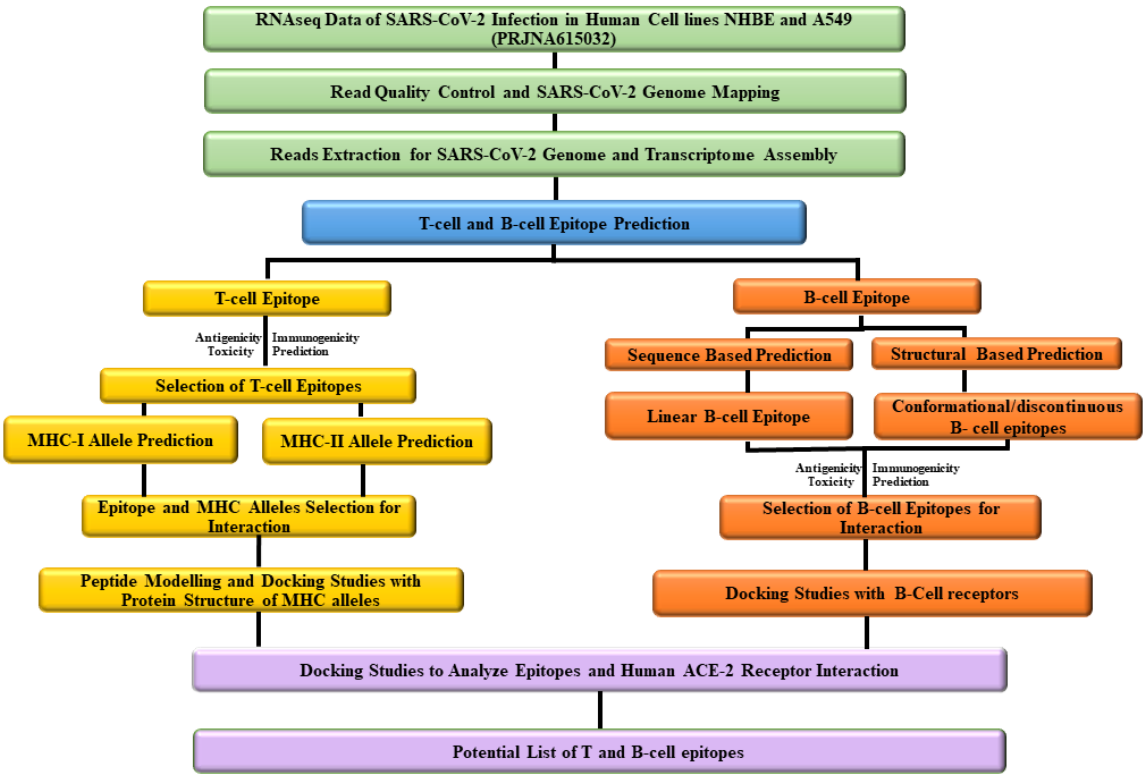

Figure-1. Schematic representation of used approach for transcriptome assembly from RNA-seq data of SARS-CoV-2 infection in human cell lines, epitope identification, and molecular docking studies.

3.1 SARS-CoV-2 transcriptome assembly and annotation

De novo SARS-CoV-2 transcriptome was constructed by using publicly available transcriptome data from SRA project (PRJNA615032). Transcriptome data was generated to the study of SARS-CoV-2 infection in human cell lines NHBE and A549 [13]. Extracted reads were trimmed by removing adapter and low quality sequences by using trimmomatic-0.39. Reads with a length of less than 20 bps were also removed from dataset [37]. In order to develop SARS-CoV-2 transcriptome assembly, raw reads were aligned to the SARS-CoV-2 genome (Accession number: MN908947.3, Wuhan-Hu-1 isolate) by

using the RNA-seq alignment tool HISAT2 on default conditions. After read mapping, Samtools and Bedtools were used to extract mapped RNA reads on SARS-CoV-2 genome. Total, 87,716 reads were extracted from all the samples. Detail description of experiment, sample name, description, and number of mapped reads per sample over SARS-CoV-2 genome were given in supplementary material file1 (Table - S1). All the extracted RNA-seq reads were used to construct *de novo* SARS-CoV-2 transcriptome through Trinity software. In total, 54,814 bases were assembled into 27 transcripts with median contig length 650 bps, N50 value of 10,677 bps and approximate average transcript length of 2030 bps. The generated transcriptome assembly was clustered at 90 % sequence identity through CD-HIT software that produced 27 non-redundant transcripts, the same number of non-redundant transcripts showed the sequence variability among assembled transcripts. Non-redundant transcripts were translated into 44 protein sequences through TransDecoder program. Protein sequences were annotated against Uniport databases by using BLAST similarity search at evalue threshold 1e-10 (Table - 1).

Table-1 SARS-CoV-2 transcriptome annotation along with expression values. H: Helicase, M: Membrane protein, N: Nucleoprotein, NendoU: Uridylate-specific endoribonuclease, NSP: Non-structural protein, ORF: open reading frame, SG: Surface glycoprotein, RdRp: RNA-dependent RNA polymerase, TPM: Transcript per million.

Transcripts ids	Functional Class	Uniport Annotation	Expression (TPM)
TRINITY_DN10_c0_g1_i1_p1	H	ORF1ab polyprotein Tax=BtRs-BetaCoV/YN2013 TaxID=1503303	3661.09
TRINITY_DN0_c0_g1_i1_p3	M	Membrane protein Tax=Bat SARS-like coronavirus TaxID=1508227	91746.3
TRINITY_DN0_c0_g1_i2_p5	M	Membrane protein Tax=Bat SARS-like coronavirus TaxID=1508227	1716.75
TRINITY_DN0_c0_g1_i4_p5	M	Membrane protein Tax=Bat SARS-like coronavirus TaxID=1508227	3188.07
TRINITY_DN0_c0_g1_i6_p5	M	Membrane protein Tax=Bat SARS-like coronavirus TaxID=1508227	12432
TRINITY_DN0_c0_g1_i1_p1	N	Nucleoprotein Tax=Bat SARS-like coronavirus TaxID=1508227	91746.3
TRINITY_DN0_c0_g1_i2_p3	N	Nucleoprotein Tax=Bat SARS-like coronavirus TaxID=1508227	1716.75
TRINITY_DN0_c0_g1_i4_p3	N	Nucleoprotein Tax=Bat SARS-like coronavirus TaxID=1508227	3188.07
TRINITY_DN0_c0_g1_i6_p3	N	Nucleoprotein Tax=Bat SARS-like coronavirus TaxID=1508227	12432
TRINITY_DN0_c0_g1_i7_p1	N	Nucleoprotein Tax=Bat SARS-like coronavirus TaxID=1508227	838657
TRINITY_DN0_c0_g1_i2_p2	NendoU	Non-structural polyprotein 1ab Tax=Bat SARS-like coronavirus	1716.75
TRINITY_DN0_c0_g1_i4_p2	NendoU	Non-structural polyprotein 1ab Tax=Bat SARS-like coronavirus	3188.07
TRINITY_DN0_c0_g1_i6_p2	NendoU	Non-structural polyprotein 1ab Tax=Bat SARS-like coronavirus	12432
TRINITY_DN0_c0_g1_i3_p1	NSP1	Non-structural polyprotein 1ab Tax=Betacoronavirus TaxID=694002	10488.1
TRINITY_DN0_c0_g1_i5_p1	NSP1	Non-structural polyprotein 1ab Tax=Betacoronavirus TaxID=694002	3765.1
TRINITY_DN20_c0_g1_i1_p1	NSP2	Non-structural polyprotein 1ab Tax=Betacoronavirus TaxID=694002	1786.89
TRINITY_DN4_c0_g1_i1_p1	NSP2	Non-structural polyprotein 1ab Tax=Bat SARS-like coronavirus	3107.15
TRINITY_DN13_c0_g1_i1_p1	NSP3	Non-structural polyprotein 1ab Tax=Betacoronavirus TaxID=694002	1009.35
TRINITY_DN17_c0_g1_i1_p1	NSP3	Non-structural polyprotein 1ab Tax=Bat SARS-like coronavirus	1416.36
TRINITY_DN19_c0_g1_i1_p1	NSP3	Non-structural polyprotein 1ab Tax=Betacoronavirus TaxID=694002	1759.06
TRINITY_DN2_c0_g1_i1_p1	NSP3	Non-structural polyprotein 1ab Tax=Bat SARS-like coronavirus	1602.68
TRINITY_DN5_c0_g1_i1_p1	NSP3	Non-structural polyprotein 1ab Tax=Betacoronavirus TaxID=694002	553.586
TRINITY_DN6_c0_g1_i1_p1	NSP3	Non-structural polyprotein 1ab Tax=Betacoronavirus TaxID=694002	1813.9
TRINITY_DN7_c0_g1_i1_p1	NSP6	Non-structural polyprotein 1ab Tax=Betacoronavirus TaxID=694002	697.846

TRINITY_DN1_c0_g1_i1_p1	NSP8	Non-structural polyprotein 1ab Tax=Bat SARS-like coronavirus	1941.65
TRINITY_DN0_c0_g1_i1_p2	ORF3a	Uncharacterized protein Tax=Human SARS coronavirus TaxID=694009	91746.3
TRINITY_DN0_c0_g1_i2_p4	ORF3a	Uncharacterized protein Tax=Human SARS coronavirus TaxID=694009	1716.75
TRINITY_DN0_c0_g1_i4_p4	ORF3a	Uncharacterized protein Tax=Human SARS coronavirus TaxID=694009	3188.07
TRINITY_DN0_c0_g1_i6_p4	ORF3a	Uncharacterized protein Tax=Human SARS coronavirus TaxID=694009	12432
TRINITY_DN0_c0_g1_i1_p4	ORF7a	SARS_X4 domain-containing protein Tax=Bat SARS-like coronavirus	91746.3
TRINITY_DN0_c0_g1_i2_p6	ORF7a	SARS_X4 domain-containing protein Tax=Bat SARS-like coronavirus	1716.75
TRINITY_DN0_c0_g1_i4_p6	ORF7a	SARS_X4 domain-containing protein Tax=Bat SARS-like coronavirus	3188.07
TRINITY_DN0_c0_g1_i6_p6	ORF7a	SARS_X4 domain-containing protein Tax=Bat SARS-like coronavirus	12432
TRINITY_DN0_c0_g1_i1_p5	ORF8	Uncharacterized protein, Bat SARS-like coronavirus TaxID=1508227	91746.3
TRINITY_DN0_c0_g1_i2_p7	ORF8	Uncharacterized protein, Bat SARS-like coronavirus TaxID=1508227	1716.75
TRINITY_DN0_c0_g1_i4_p7	ORF8	Uncharacterized protein, Bat SARS-like coronavirus TaxID=1508227	3188.07
TRINITY_DN0_c0_g1_i6_p7	ORF8	Uncharacterized protein, Bat SARS-like coronavirus TaxID=1508227	12432
TRINITY_DN15_c0_g1_i1_p1	Proteinase	UPI0001D192D5 related cluster, TaxID= RepID=UPI0001D192D5	1706.65
TRINITY_DN9_c0_g1_i1_p1	Proteinase	UPI000181CE36 related cluster, TaxID= RepID=UPI000181CE36	811.289
TRINITY_DN11_c0_g1_i1_p1	RdRp	ORF1ab polyprotein n=1 Tax=BtRf-BetaCoV/JL2012 TaxID=1503299	1168.03
TRINITY_DN12_c0_g1_i1_p1	RdRp	RNA-dependent RNA polymerase Tax=Human SARS coronavirus	1833.88
TRINITY_DN0_c0_g1_i2_p1	SG	Spike protein n=1 Tax=Bat SARS-like coronavirus TaxID=1508227	1716.75
TRINITY_DN0_c0_g1_i4_p1	SG	Spike protein n=1 Tax=Bat SARS-like coronavirus TaxID=1508227	3188.07
TRINITY_DN0_c0_g1_i6_p1	SG	Spike protein n=1 Tax=Bat SARS-like coronavirus TaxID=1508227	12432

3.2. T-cell epitopes identification of SARS-CoV-2 transcriptome

T-cell epitopes are presented by MHC class I and II that are recognized by two distinct subsets of T-cells, $CD8^+$ and $CD4^+$ T-cells, respectively. NetCTL 1.2 program was used for the prediction T-cell epitopes, and 1144 epitopes were selected at combined prediction threshold value1.0 for 12 super type categories i.e. A1 (330), A24(314), A26(242), A2(247), A3(328), B27(175), B39(263), B44(133), B58(284), B62(473), B7(157), and B8(193). The predicted T-cell epitopes were further evaluated for antigenicity by VaxiJen server and immunogenicity by IEDB prediction tools, and 598 and 625 epitopes were shown antigenicity and immunogenicity potential respectively. Finally, 598 antigenicity and immunogenicity T-cell epitopes were selected for further study. To determine epitopes potential to elicit effective immune response, 598 selected epitopes were used to explore interacting MHC alleles through IEDB program. Peptide length nine and the IC50 value =< 200nM were selected as parameters for the identification of MHC class I binding gene and alleles. MHC class I alleles such as HLA-A, HLA-B, and HLA-C were recognized with parameter human as MHC source species and IEBD recommended method to predict a distinct set of MHC class I alleles for all selected 598 epitopes. In MHC-I allele analysis, HLA-A type alleles were found as more frequently occurring alleles (HLA-A*01:01, HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*03:01, HLA-A*11:01, HLA-A*23:01, HLA-A*24:02, HLA-A*26:01, HLA-A*30:01, HLA-A*30:02, HLA-A*31:01, HLA-A*32:01, HLA-A*33:01, HLA-A*68:01, HLA-A*68:02) than HLA-B type alleles (HLA-B*07:02, HLA-B*08:01, HLA-B*15:01, HLA-B*35:01, HLA-B*40:01, HLA-B*44:02, HLA-B*44:03, HLA-B*53:01, HLA-B*57:01). In our analysis, HLA-C class genes were not found for any epitopes. ToxinPred was used to

explore toxicity of predicted epitopes. Finally, 40 $CD8^+$ T-cell epitopes (Table-2) were selected as high immunogenic, antigenic, non-toxic epitopes with good binding affinity to MHC class I alleles. Further, these epitopes were explored for SARS-CoV-2 proteome for functional characterization.

Table - 2 List of potential T-cell epitopes for MHC class I. Immscore: Immunogencity score, Antiscore: Antigencity score, IC: inhibitory constant, NT: Non-toxic, H: Helicase, M: Membrane protein, N: Nucleoprotein, NendoU: Uridylate-specific endoribonuclease, NSP: Non-structural protein, ORF: open reading frame, SG: Surface glycoprotein, RdRp: RNA-dependent RNA polymerase, 2'-O-MT: 2'-O-methyltransferase, ExoN: Guanine-N7 methyltransferase

Epitopes	Protein	Immscore	Antiscore	Toxicity	IC50	MHCI
DIADTTDAV	SG	0.151	1.0904	NT	13.83	HLA-A*68:02
EQWNLVIGF	M	0.226	1.3869	NT	170.94	HLA-B*15:01
ETSWQTGDF	NSP2	0.134	1.314	NT	151.38	HLA-A*26:01
FEHIVYGDF	NendoU	0.223	1.1633	NT	142.3	HLA-B*40:01
FELEDFIPM	NendoU	0.335	1.2669	NT	60.81; 8.64; 9.78	HLA-A*02:06;HLA-B*35:01; HLA-B*40:01
FLFLTWICL	M	0.354	1.4835	NT	138.72; 32.26	HLA-A*02:01;HLA-A*02:06
FLHVTYVPA	SG	0.115	1.3346	NT	152.58; 65.16; 7.14	HLA-A*02:01;HLA-A*02:03; HLA-A*02:06;
FTIGTVTLK	ORF3a	0.180	2.0317	NT	23.28; 4.9	HLA-A*11:01; HLA-A*68:01;
FVKRVDWTI	ExoN	0.253	1.9477	NT	67.59	HLA-A*02:06;
HFAIGLALY	H	0.196	1.4046	NT	63.58	HLA-A*30:02;
HSIGFDYVY	ExoN	0.233	1.0882	NT	124.39; 12.47; 151.12; 27.64; 34.81; 41.91; 56.72	HLA-A*26:01; HLA-A*30:02; HLA-A*68:01; HLA-B*15:01; HLA-B*35:01; HLA-B*57:01; HLA-B*58:01;
IFWRNTNPI	2'-O-MT	0.142	1.1927	NT	187.29; 61.88	HLA-A*23:01; HLA-A*32:01
ILGTVSWNL	NSP3	0.118	1.3875	NT	23.63; 95.09	HLA-A*02:01; HLA-A*02:03
ILMTARTVY	NSP6	0.126	1.097	NT	14.92; 88.57	HLA-A*30:02; HLA-B*15:01
IQYIDIGNY	ORF8	0.304	2.096	NT	15.96; 59.22	HLA-A*30:02; HLA-B*15:01
KLSYGIATV	H	0.157	1.0767	NT	17.61; 3.28; 6.77	HLA-A*02:01; HLA-A*02:03; HLA-A*02:06
KSHNIALIW	NSP3	0.239	1.2831	NT	15.69; 3.68; 6.85	HLA-A*32:01; HLA-B*57:01; HLA-B*58:01;
KSVNITFEL	NSP3	0.330	2.1377	NT	15.38; 39.62; 49.16	HLA-A*02:06; HLA-A*32:01; HLA-B*58:01
LAAVYRINW	M	0.208	1.4322	NT	127.84; 23.1; 64.87	HLA-B*53:01; HLA-B*57:01; HLA-B*58:01
LEPEYFNSV	H	0.102	1.067	NT	30.81	HLA-A*02:06;

LPVNVAFEL	NendoU	0.241	1.2581	NT	14.41; 44.35; 83.6	HLA-B*07:02; HLA-B*35:01; HLA-B*53:01
LSPRWYFYY	N	0.357	1.2832	NT	48.64; 74.89	HLA-A*01:01;HLA-A*30:02;
LSYGIATVR	H	0.256	1.696	NT	14.11; 24.87	HLA-A*31:01; HLA-A*68:01
LTAVVIPTK	NSP3	0.233	1.1167	NT	17.58; 183.23; 37.55	HLA-A*11:01;HLA-A*30:01; HLA-A*68:01;
LVSDIDITF	NSP3	0.254	1.783	NT	10.85; 152.54; 184.23; 185.25;	HLA-A*02:06; HLA-B*15:01; HLA-B*35:01; HLA-B*53:01;
NVAFNVVNK	NendoU	0.194	1.1634	NT	10.54; 73.52	HLA-A*11:01; HLA-A*68:01;
NYVFTGYRV	H	0.228	1.0902	NT	103.91;	HLA-A*23:01;
QQWGFTGNL	ExoN	0.281	1.0003	NT	86.89;	HLA-A*02:06;
QYIKWPWYI	SG	0.216	1.4177	NT	13.22; 6.13	HLA-A*23:01; HLA-A*24:02;
RELHLSWEV	H	0.108	2.2601	NT	38.44; 47.51	HLA-A*02:06; HLA-B*40:01
SLENVAFNV	NendoU	0.198	1.0488	NT	152.28; 38.77; 70.84	HLA-A*02:01; HLA-A*02:03; HLA-A*02:06;
TLNDFNLVA	Proteinase	0.143	1.4845	NT	41.06; 82.34	HLA-A*02:01; HLA-A*02:03;
TSFGPLVRK	RdRp	0.116	1.7142	NT	15.92; 20.71; 9.98	HLA-A*03:01; HLA-A*11:01; HLA-A*68:01
VFITLCFTL	ORF7a	0.142	1.249	NT	117.58; 60.42	HLA-A*23:01; HLA-A*24:02;
VLSDRELHL	H	0.123	1.5809	NT	107.63; 166.19	HLA-A*02:01; HLA-A*02:03;
VVFLHVTYV	SG	0.128	1.5122	NT	13.02; 21.97; 36.56; 51.51	HLA-A*02:01; HLA-A*02:03; HLA-A*02:06; HLA-A*68:02
VVNARLRAK	H	0.144	1.933	NT	168.01; 48.82; 71.14	HLA-A*03:01; HLA-A*11:01; HLA-A*31:01
WLIVGVALL	ORF3a	0.183	1.2686	NT	121.13; 16.77; 59.19	HLA-A*02:01; HLA-A*02:03; HLA-A*02:06
WPWYIWLGF	SG	0.417	1.4953	NT	193.4; 42.3; 8.91	HLA-B*07:02; HLA-B*35:01; HLA-B*53:01;
YIDIGNYTV	ORF8	0.188	1.3128	NT	10.59; 22.57	HLA-A*02:01; HLA-A*02:06

Similarly, MHC class II gene and allele's prediction was performed through IEDB analysis resources by using the same parameters as for MHC class I except the selection of SMM method for the prediction of a distinct set of MHC class II alleles. Total, 4072 epitopes were identified from protein sequences with good binding affinity to MHC class II alleles. To select MHC class II alleles and epitopes, we decided to take those MHC class II alleles and epitopes which have MHC class I epitopes as a core sequence. Total, 34 MHC class II epitope sequences (15-mer) were selected by using previously

selected 40 (9-mer) antigenic and immunogenic epitope sequences as core sequences (Supplementary file1: Table -S2). Among all MHC-II alleles, HLA-DPA1*01:03/DPB1*04:01, HLA-DPA1*02:01/DPB1*14:01, HLA-DRB3*02:02, HLA-DRB1*01:01, HLA-DRB1*07:01, HLA-DPA1*01:03/DPB1*02:01, and HLA-DRB5*01:01were the most abundant alleles. Among all, various 9-mer epitopes such as LSPRWYFYY, KSVNITFEL, IQYIDIGNY, EQWNLVIGF, DIADTTDAV, TSFGPLVRK and RELHLSWEV were also have core sequence among 15-mer MHC class II alleles epitopes. But IC_{50} value of these epitopes were more than 200nM. After conservation analysis, twelve most antigenic and immunogenic MHC class II epitopes (APHGVVFLHVTYVPA, FLHVTYVPAQEKNFT, GVVFLHVTYVPAQEK, HGVVFLHVTYVPAQE, PHGVVFLHVTYVPAQ, QSAPHGVVFLHVTYV, QYIKWPWYIWLGFIA, SAPHGVVFLHVTYVP, VFLHVTYVPAQEKNF, VVFLHVTYVPAQEKN, and YIKWPWYIWLGFIAG) were selected which contains 9-mer core sequences of four epitopes from previously selected 40 epitopes. Peptide sequences of 9-mer epitopes, $CD4^+$ T-cell epitopes sequence and MHC class II alleles were given in Table-3.

Table - 3: List of potential non-toxic, conserved epitopes for MHC class II along with their core 9-mer epitopes. Anno: Annotation, Immscore: Immunogencity score, Antiscore: Antigencity score, SG: Surface glycoprotein, *: overlapping epitopes

Anno	Peptide	Immscore	Antiscore	MHC-II	Epitopes
SG	QYIKWPWYI* WPWYIWLGF*	0.41673	1.4953	HLA-DPA1*01:03/DPB1*02:01 HLA-DPA1*01:03/DPB1*04:01	QYIKWPWYIWLGFIA YIKWPWYIWLGFIAG
SG	VVFLHVTYV	0.1278	1.5122	HLA-DRB1*01:01; HLA-DRB1*04:05; HLA-DRB1*07:01; HLA-DRB3*02:02;	QSAPHGVVFLHVTYV SAPHGVVFLHVTYVP APHGVVFLHVTYVPA PHGVVFLHVTYVPAQ
SG	FLHVTYVPA	0.11472	1.3346	HLA-DPA1*01:03/DPB1*02:01; HLA-DPA1*01:03/DPB1*04:01; HLA-DPA1*02:01/DPB1*14:01; HLA-DRB1*01:01; HLA-DRB1*04:05; HLA-DRB1*07:01; HLA-DRB3*02:02;	APHGVVFLHVTYVPA PHGVVFLHVTYVPAQ HGVVFLHVTYVPAQE GVVFLHVTYVPAQEK VVFLHVTYVPAQEKN VFLHVTYVPAQEKNF FLHVTYVPAQEKNFT

3.3. Population coverage analysis

MHC molecules can form complexes with millions of epitopes which are reflecting the polymorphic nature of MHC genes. If MHC polymorphism occurs in peptide-binding region, binding specificity of MHC molecules will be changed. MHC variability has evolutionary advantage to identify variety of pathogens. But genetic variability among MHC alleles are also a major obstacles in the development of peptide-based vaccines. Therefore, population coverage is an important criterion to design a generalized an effective vaccine[38].

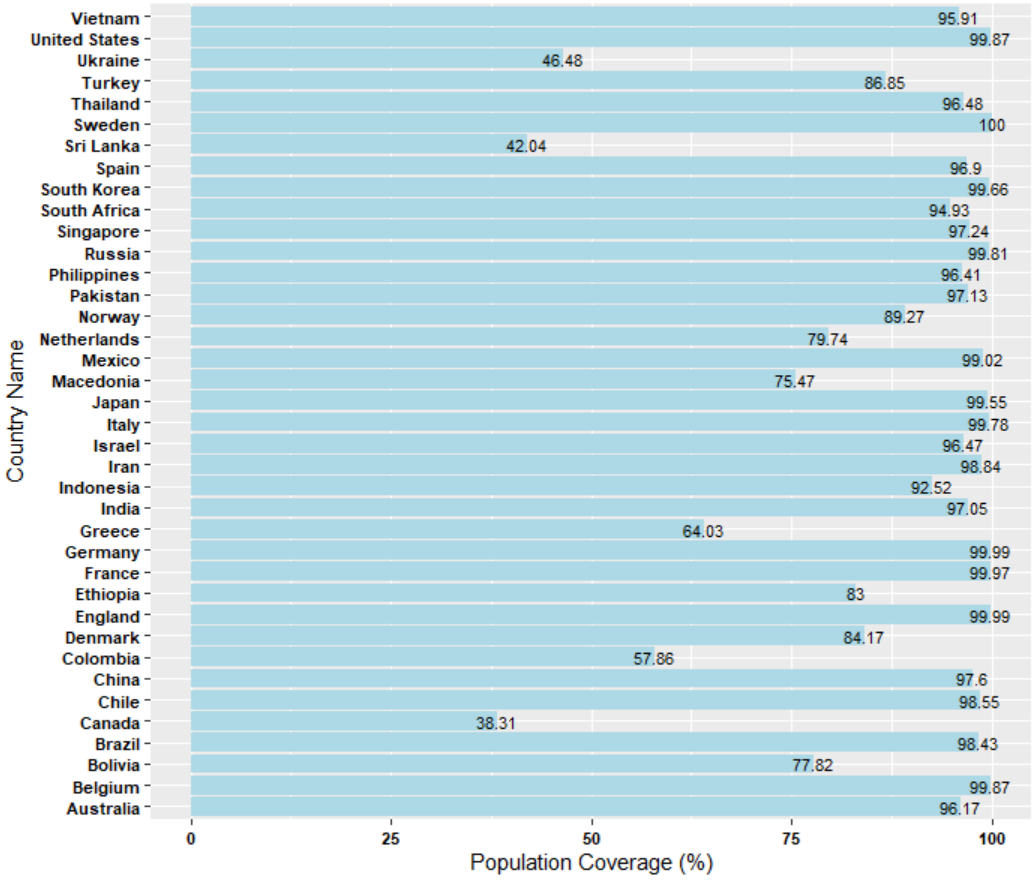

Figure -2. Global population coverage through identified epitopes.

In population coverage analysis, MHC class I allele's of 40 epitopes and MHC class II allele's of 33 epitopes were used, and a significant population coverage was found for different geographic regions around the world (Figure-2). MHC class alleles of selected epitopes were covered approximately 90% of the world population. Highest population coverage was found for Sweden (100%) which was closely followed by England, Germany, France, Belgium, United States, Russia, Italy, South Korea, Japan, Mexico, Iran, Chile, Brazil, China, Singapore, Pakistan, India, Spain, Thailand, Israel, Philippines, Australia, and Vietnam with a population coverage of 99.99%, 99.99%, 99.97%, 99.87%, 99.87%, 99.81%, 99.78%, 99.66%, 99.55%, 99.02%, 98.84%, 98.55%, 98.43%, 97.6%, 97.24%, 97.13%, 97.05%, 96.9%, 96.48%, 96.47%, 96.41%, 96.17%, and 95.91% respectively. The lowest population coverage were found for Canada (38.31%), Srilanka (42.04) and Ukrain (46.48). United States and Europe has highest number of COVID-19 cases [39]. Hence, the population coverage prediction is essential for vaccine design. Population of ethnic groups were also significantly covered (Supplementary file1: Table -S3), and average coverage for ethnic group across the world is around 93%.

3.4. B-cell epitope identification of SARS-CoV-2 transcriptome.

B-cell epitope is a precise region of the antigenic protein that is detected by B-cell receptors (BCR) through membrane-bound immunoglobulins. Once B-cell activated, it secretes soluble forms of the

immunoglobulins to neutralize antigenic proteins. Thus, B-cell epitope and B-cell receptor information is essential for epitope-based vaccine design. Prediction of B-cell epitopes was performed by using protein sequences of the assembled transcriptome and SARS-CoV-2 protein structures. Total, 330 B-cells epitopes were predicted through BepiPred-2.0 program by using protein sequences, whereas 77 B-cell epitopes were predicted through the IEDB conformational tool ElliPro by using 24 SARS-CoV-2 protein structure, download from Zhang lab. B-cell epitopes prediction from the protein structure is highly useful information for epitope-based vaccine design and development [40, 41]. Therefore, a separate analysis was performed for B-cell epitopes identified from protein structures, and 19 epitopes were identified after toxicity (non-toxic), immunogenicity and antigenicity analysis (Supplementary file1: Table-S4). In order to explore most suitable B-cell epitopes, epitopes generated by both the approaches were combined, and total 130 unique B-cell epitopes were identified from length range 9 to 15. 73 B-cell epitopes were predicted as antigenic epitopes through VaxiJen v2.0 webserver and 53 B-cell epitopes were identified as immunogenic. Total, 16 non-toxic B-cell epitopes were identified with immunogenicity score and antigenicity score more than 0.1 and 0.4 respectively.

Table - 4: List of B-cell epitopes identified form protein sequences of assembled transcriptome and modelled protein structure of various coding protein of SAR-CoV-2 genome. Immscore: Immunogencity score, Antiscore: Antigencity score, NT: Non-toxic, H: Helicase, M: Membrane protein, N: Nucleoprotein, NendoU: Uridylate-specific endoribonuclease, NSP: Non-structural protein, ORF: open reading frame, SG: Surface glycoprotein, RdRp: RNA-dependent RNA polymerase, 2'-O-MT: 2'-O-methyltransferase, ExoN: Guanine-N7 methyltransferase

B-cell Epitopes	Anno	Len	Immscore	Antiscore	Toxicity
SEQLDFIDTKRGV	NSP2	13	0.13632	1.7773	NT
KGTLEPEYF	H	9	0.17084	1.3504	NT
HCGETSWQTGDFV	NSP2	13	0.24817	1.1314	NT
KTVGELGDVRE	NSP3	11	0.25316	0.9231	NT
LTGTGVLTESNK	SG	12	0.10111	0.8122	NT
TGVVGEGSEGLN	NSP2	12	0.17371	0.7539	NT
QTTETAHSC	H	9	0.11862	0.7078	NT
MEVTPSGTWLT	N	11	0.13056	0.5982	NT
SDARTAPHG	NSP1	9	0.14895	0.5706	NT
LKATEETFK	H	9	0.34467	0.5278	NT
NENGTITDA	SG	9	0.2408	0.5257	NT
KGHFDGQQGEVPVS	NendoU	14	0.14142	0.5183	NT
LQAGNATEVPANS	NSP10	13	0.2729	0.4491	NT
VQIPTTCANDPVGFT	NSP10	15	0.2336	0.4488	NT

Cross-reactivity analysis of epitopes were performed against human proteome sequences through BLAST similarity search, and found that two B-cell epitopes (DNNFCGPDGYPLE,

NQDLNGNWYD) showed significant similarity with human proteins ENSP00000390696.1 and ENSP00000263390.3 respectively (Supplementary file1: Table-S5). Finally, 14 B-cell epitope were found for further analysis (Table - 4). On the basis of immunological parameters, the five B-cell epitopes, NSP2 (SEQLDFIDTKRGV, HCGETSWQTGDFV), Helicase (KGTLEPEYF), NSP3 Papain-like (KTVGELGDVRE), and Surface glycoprotein (LTGTGVLTESNK) were selected from Table - 4 for molecular docking studies of B-cell receptors.

3.5. Molecular docking analysis

Cellular immunity gets activated when MHC molecules binds to intracellular and extracellular proteins displayed on the cell surface. Structural analysis of epitopes and MHC molecules can improve our knowledge about T-cell based mechanism to reduce disease burden. To explore structural compatibility between T-cell epitopes and MHC complexes, molecular docking studies were performed to analyze binding affinities between MHC complexes and T-cell epitopes. Twenty-two MHC proteins were explored with selected three T-cell epitopes, and the best interaction was identified with the highest binding affinity. The compatible structural model of epitopes (WPWYIWLGF, VVFLHVTYV, and FLHVTYVPA) and the MHC molecules were retaining a binding affinity range from of -136.54 to -7.12 kcal/mol. Detail description of molecular interaction analysis of peptide VVFLHVTYV and identified protein structure of MHC genes were given in Table - 5. Detail docking descriptions of other two epitopes with MHC molecule were given in supplementary material file1 (Table - S6).

Table -5. Molecular interaction and docking analysis of identified antigenic and immunogenic T-cell epitope (VVFLHVTYV) and selected protein structures of MHC alleles. TPM (Transcript per million)

MHC Alleles (PDB Code)	Expression (TPM)	Protein Chain	Interaction Energy	Total Energy	Hydrogen Bond (Peptide - Receptor)
HLA-DRB3*02:02 (2Q6W)	NoExp	B	-96.73	-1560.1	THR7-ASP152, LEU4-SER120
HLA-A*02:03 (3OX8)	46.1625	A	-63.72	-1937.21	THR7-ASP77, THR7-TYR84
HLA-DQA1*05:01/DQB1*03:01 (4D8P)	NoExp	A	-69.82	-890.23	VAL1-SER8, TYR8-THR93, VAL1-VAL6, PHE3-SER8, TYR8-THR83, TYR8-ASP142
HLA-DRB1*04:01 (5JLZ)	NoExp	B	-57.25	-1377.11	THR7-GLU187, HIS5-GLU187, PHE3-VAL101
HLA-DRB1*01:01 (5V4N)	NoExp	C	-63.35	-1430.47	VAL1-HIS360, THR7-GLU411
HLA-A*23:01 (5WWJ)	78.8768	C	-95.26	-1611.15	TYR8-CYS264, TYR8-CYS388, VAL9-TYR327, VAL9-ASN331
HLA-A*24:02 (5XOV)	2597.88	A	-82.36	-2186.76	THR90-HIS5, HIS191-HIS5, THR190-THR7, THR190-THR7
HLA-A*11:01 (6ID4)	NoExp	A	-7.12	-1452.3	LEU4-TYR27, LEU4-ARG6, TYR8-SER4, SER4-VAL9
HLA-B*44:02 (3DX6)	NoExp	A	-83.81	-2384.29	THR7-ASP114,VAL1-SER167, VAL1-GLU55
HLA-DPA1*01:03/DPB1*02:01 (3LQZ)	NoExp	A	-75.51	-3195.72	HIS5-GLU134, HIS5-HIS144
HLA-A*02:06 (3OXR)	NoExp	C	-47.82	-2993.06	VAL1-HGLN54, PHE3-GLN54, HIS5-GLU55, HIS5-TRP51

HLA-B*57:01 (3X11)	NoExp	A	-31.02	-2183.55	TYR8-GLN155, VAL9-TYR99, VAL1-SER116, PHE3-SER8
HLA-A*68:02 (4HWZ)	61.582	A	-56.06	-2054.65	PHE3-ARG144, VAL1-ASP77, VAL9-THR73, HIS5-THR73
HLA-B*44:03 (4JQX)	38.0344	A	-39.35	-1709.95	THR7-ASN77, VAL9-SER69, TYR6-GLN89
HLA-B*35:01 (4PRA)	53.433	A	-22.74	-1780.49	VAL1-ARG6, VAL9-SER116, THR7-ASP144, TYR8-TYR74
HLA-A*02:01(4U6Y)	27.0879	A	-29.1	-1507.37	VAL1-GLU63, VAL1-TYR99, HIS5-HIS114, THR7-HIS144
HLA-DRB1*11:01(5NI9)	NoExp	B	-90.52	-1566.14	VAL6-TYR30, TYR8-TYR48
HLA-B*58:01 (5VWH)	NoExp	A	-76.23	-2288.87	THR7-ARG14, HIS5-ARG21
HLA-A*30:02 (6J1V)	NoExp	A	-80.19	-2377.92	VAL6-ARG202
HLA-A*30:01 (6J1W)	NoExp	A	-46.93	-1869.26	PHE3-ARG273, HIS5-THR271
HLA-A*03:01 (6O9B)	23.4843	A	-62.78	-2263.52	TYR8-TYR-123, TYR8-ARG114, HIS5-ARG114
HLA-A*68:01 (6PBH)	61.582	A	-88.24	-2076.8	TYR6-TYR99, ALA9-THR143

MHC class I and II gene expression analysis was performed by using RNA-seq data of cell lines through pseudo aligner Kallisto[23]. Expression values of expressed MHC class alleles were given in supplementary material file1 (Table - S7). HLA-A*24:02 allele was highly expressed among most frequent occurring MHC alleles, but highest interaction of peptide (VVFLHVTYV) was shown with HLA-DRB3*02:02 allele with two hydrogen bonds (THR7-ASP152, LEU4-SER120). The epitope position in protein structure and the binding interactions had shown in Figure - 3. We were also ensured genuine hydrogen bonds interaction with receptors through cavity prediction (Supplementary material file2)

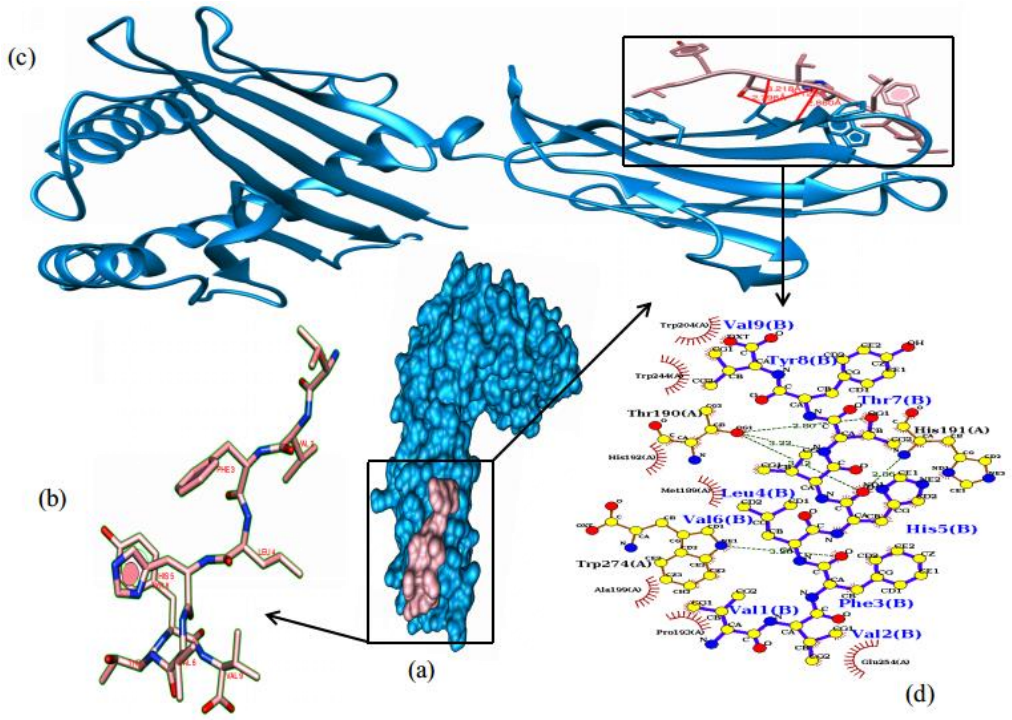

Figure-3. Molecular interaction of peptide (VVFLHVTYV) and protein structure (5XOV) of HLA allele (HLA-A*24:02). a and c) cartoon structure of peptide and class-I HLA allele. b) interaction of peptide and MHC-II alleles (Hydrogen bonds in red colour). d) molecular level description of interaction through LIGPlot software (Hydrogen bonds in green).

Similarly, five most antigenic and immunogenic B-cell epitopes were selected for molecular docking studies with two B-cell receptors (5DRW and 1K1F) through CABSDocks server. 5DRW protein structure is a crystal structure of BCR Fab fragment from subset of chronic lymphocytic leukaemia whereas 1K1F is dimer structure of Bcr-Abl oncoprotein. 1K1F protein structure provided a base to design an inhibitor to disrupt Bcr-Abl oligomerization. Moreover, 1K1F was without Fab region unlike to 5DRW [42]. In our analysis, most of the peptides were shown higher binding affinities with 5DRW than 1K1F. Detail description of docking studies of B-cell receptor and peptides were given in Table -6.

Table -6: Molecular docking and interaction analysis of B-cell epitope and B-cell protein receptors.

Peptide	**PDB Code**	**Protein Chain**	**Interaction Energy**	**Total Energy**	**Hydrogen Bonds (Peptide - Receptor)**
SEQLDFIDTKRGV	5DRW	B	-94.17	-1656.88	THR9-PRO49, THR9-SER48, THR9-ARG51, GLU7-LYS55
	1K1F	A	-25.53	-362.98	THR9-ARG50, TRP7-ARG51, GLU4ARG51,CYS2-ASN39
HCGETSWQTGDFV	5DRW	B	-85.9	-1579.94	SER6-GLN205, THR9-GLN205
	1K1F	A	-29.14	-366.08	GLN8-ARG22
KGTLEPEYF	5DRW	B	-91.21	-1767.94	THR3-SER20, GLU7-SER10
	1K1F	A	-35.76	-629.66	GLU5-GLN51
KTVGELGDVRE	5DRW	B	-37.47	-1490.12	THR2-PRO125, VAL3-SER127, GLU11-ASN144, GLY7-SER168, GLY7-SER182
	1K1F	A	-54.36	-686.76	THR5-ARG25
LTGTGVLTESNK	5DRW	B	-40.46	-1235.01	THR4-TYR37, THR4-ASN39, SER10-GLN43
	1K1F	A	-110.1	-652.47	GLU9-SER18, GLU9-SER41

For protein structure 5DRW, interaction and total energy of peptide and B-cell receptor were varied from -94.17 to -37.47 kcal/mol and -1656.88 to -1235.01 kcal/mol respectively. Whereas interaction and total energy for 1K1F were varied from -110.1 to -25 kcal/mol and -686.76 to -362.98 kcal/mol respectively.

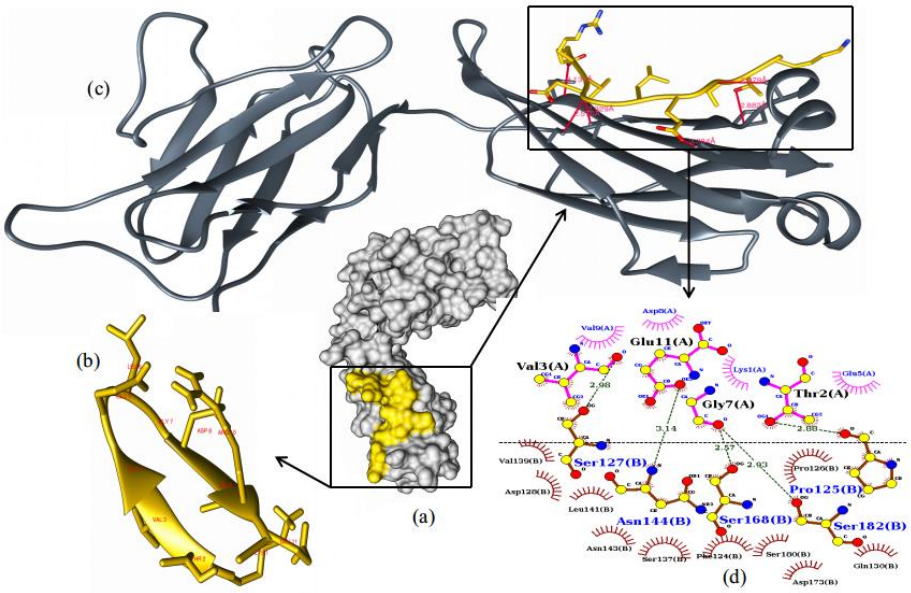

Figure – 4. Molecular interaction of peptide (KTVGELGDVRE) and protein structure of Fab region of B-cell receptor (5DRW). a and c) cartoon structure of peptide and B-cell receptor. b) interaction of peptide and B-cell receptor (Hydrogen bonds in red colour). d) molecular level description of interaction through LIGPlot software (Hydrogen bonds in green colour)

4. Discussion

SARS-CoV-2 virus has infected more than four million people worldwide, and contagious nature of virus has imposed the biggest challenge of COVID-19 treatment and prevention. Therefore, vaccine design, development and production against COVID-19 diseases is an urgent requirement to protect people from the rising viral attacks. In practice, whole process of vaccine development takes several years to be completed [43]. But, integration of immunological understanding, high throughput genomics technologies, and bioinformatics tools and techniques can help us to design effective and safe vaccines in a short duration. In human research studies, vaccines were shown variable length of protection period such as chikungunya, rift valley fever virus, and measles, are about 30 years, 12 years, and 65 years, respectively [44]. In order to develop epitope-based vaccine, the surface glycoprotein is the primary focus because it is involved in the interaction between virus and human cell receptor, and contribute significant role in pathogenesis. But, other viral elements are also important to cause disease [45]. Information of expressed regions of viral genome is very important to identify potential vaccine candidates. In the present study, SARS-CoV-2 transcriptome was used for the molecular cataloguing of immunodominant epitopes, and result of performed analysis is summarized in Figure-5.

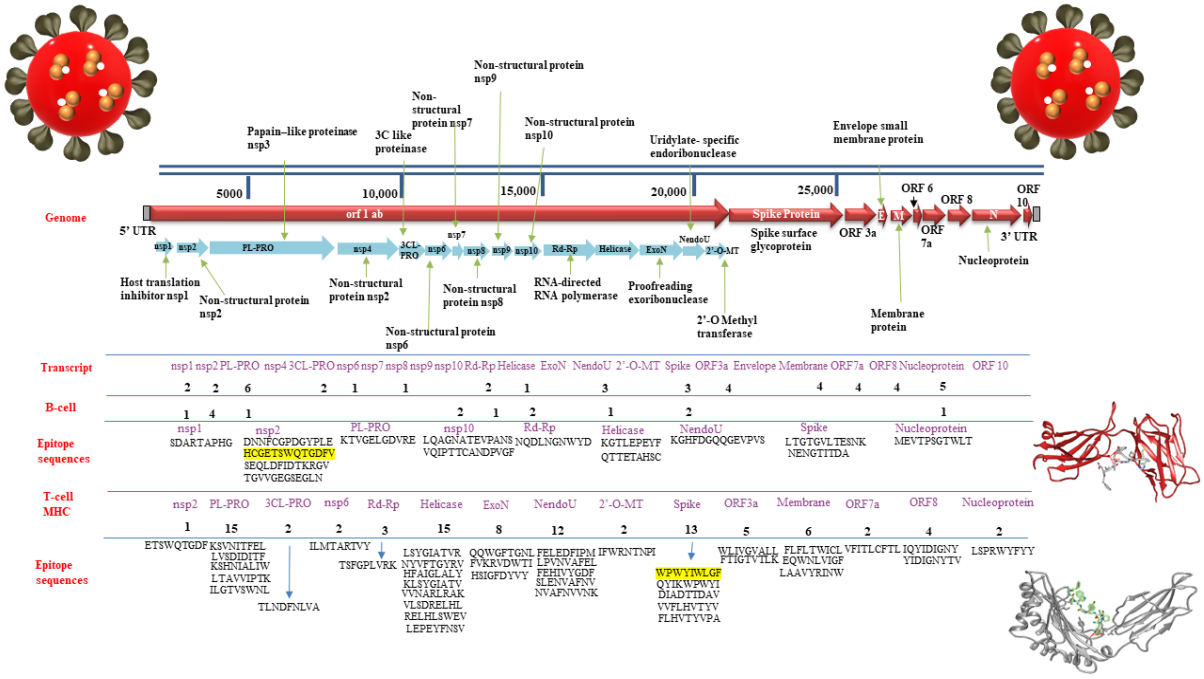

Figure -5. Overview of performed study: genome-wide transcriptome analysis of SARS-CoV-2, T- and B-cell epitopes, and docked complexes. Docked structures of B-cell receptor and MHC alleles were given at right-side and their T-and B-cell epitopes were highlighted in yellow colour.

In this study, we used RNA-seq data of SARS-CoV-2 infection in human cell lines NHBE and A549. We extracted the RNA-seq reads for SARS-CoV-2 genome and constructed a transcriptome to explore expressed regions of SARS-CoV-2 genome. In assembled transcripts, non-structural polyprotein 1ab related transcripts were found as most abundant transcripts, whereas nucleoprotein, membrane protein, SARS_X4 domain-containing protein, spike protein were also found in significant numbers. The largest region of SARS-CoV-2 genome (ORF1ab) was expressed various smaller proteins such as non-structural proteins (NSP1, NSP2, NSP3, NSP4, NSP6 and NSP8), helicase, uridylate-specific endoribonuclease, RNA-dependent RNA polymerase. Expression of polyprotein 1ab protein subcomponent (Table - 1) is clearly reflecting about disease initiation and progression such as expression of NSP1 protein was indicating the inhibition of host translation machinery by making NSP1-40S ribosome complex which can cause an endonucleolytic cleavage near the 5' UTR of host mRNAs for degradation. NSP1 also facilitates viral gene expression in infected cells by suppressing host gene expression [46]. NSP2 is another expressed transcripts that may have a role in the alteration of host cell survival signalling pathway by interacting with host prohibitin (PHB) and prohibitin-2(PHB2)[47]. NSP3 is known as papain-like proteinase, which is responsible for N-terminus cleavage of the replicase polyprotein and involved in the assembly of virally-induced cytoplasmic double-membrane vesicles together with NSP4, necessary for viral replication. NSP3 is an important viral molecular factor to suppress innate immune induction of type I interferon by blocking the phosphorylation, dimerization and subsequent nuclear translocation of host interferon regulatory transcription factor (IRF3), and also suppress NF-kappa-B signalling of host [48, 49]. NSP6 initiate

induction of autophagosomes from host reticulum endoplasmic. But later, it reduces the expansion of phagosomes to stop delivery of viral components to lysosomes [50]. NSP8, together with NSP7 forms hexadecamer act as a primase to participate in a viral replication. Expression of helicase is required for the RNA and DNA duplex-unwinding activities [51].

Antigenicity and immunogenicity are the important parameters of epitopes selection. Antigenicity of epitopes represents the ability to bind or interact with B-cell or T-cell receptors. T-cell receptors recognize amino acid sequences of epitopes, when it binds with MHC molecules whereas in B-cell, B-cell receptor interact with these epitopes. To identify antigenic epitopes, all the epitopes were selected at a threshold antigenicity score greater than 1 (Table - 2). In our antigenicity analysis, epitopes KSVNITFEL (2.138), IQYIDIGNY (2.096), RELHLSWEV (2.260), and FTIGTVTLK (2.032) were containing high antigenicity. In comparison of antigenicity, immunogenic features of epitopes triggers the innate immune response, and later induces adaptive immune response. Antigenic epitopes may or may not have immunogenicity. But, all immunogenic epitopes will have antigenic potential [52]. In order to select best immunodominant epitopes, antigenicity (>=1) and immunogenicity (>=0.1), both criteria were used (Table-2) i.e. WPWYIWLGF (0.417), LSPRWYFYY (0.357), FLFLTWICL (0.354), FELEDFIPM (0.335), KSVNITFEL (0.330), IQYIDIGNY (0.304), QQWGFTGNL (0.281), LSYGIATVR (0.256), LVSDIDITF (0.254), and FVKRVDWTI (0.253). Epitopes with higher antigenicity and immunogenicity scores will have a higher probability of binding with T-and B-receptors to elicit an effective immune response. MHC proteins helps to distinguish cell own proteins from foreign proteins such as viruses and bacteria. Thus, the binding affinity of these epitopes to MHC protein is another very important criteria. MHC class I and II genes and alleles were predicted with lower IC_{50} values (IC_{50} =<200) to ensure a higher affinity of epitopes binding with MHC class I proteins. When MHC class I molecules binds to epitopes, immune system recognizes these epitopes as a foreign peptide, and the infected cell presents itself as an antigen-presenting cell for self-destruction. For better sensitivity, $CD8^+$ T-cell epitopes should be generated from both structural and non-structural proteins because both types of proteins will be processed by infected cells in the cytoplasm, whereas structural proteins are of interest for $CD4^+$ T-cell epitopes, as it might provide help to cognate interaction [53-55]. In this study, we selected $CD4^+$ T-cell epitopes on the basis of $CD8^+$ T-cell epitope core sequences to find out the best T-cell epitopes which might provide an immune response for both kinds of MHC classes. Twelve 15-mer MHC class II epitopes were selected which have core sequences of four 9-mer epitopes of MHC class I, and all four $CD8^+$ cell epitopes belong to surface glycoprotein (Table - 3).

Diverse repertoire of MHC molecules with the binding ability to a wide range of epitopes, and genetic variability among SARS-CoV clade antigens are the major scientific challenges to develop generalized vaccine. In order to address these two challenges, we performed conservation analysis of identified epitopes through IEDB resources and NCBI Blast, and epitope conservation were ensured among known sequences. IC_{50} threshold 500nM or lower values were suggested to selects strong binding affinity between MHC class protein and epitopes [56]. To ensure MHC class allele specificity, lower

IC_{50} (=<200nM) values were considered for the selection of MHC alleles and epitopes. Genetic diversity of MHC molecules across the various ethnic groups worldwide is another the major limitation such as different MHC class alleles form different geographical region might be presented by a particular set of epitopes only. To understand demographic coverage of epitopes, population coverage analysis was performed through selected T-cell epitopes, and analysis revealed that approximately 89.44% and 93% average coverage can be achieved for world population and population of ethnic groups respectively (Figure-2, Supplementary file1: Table -S3).

In various B-cell research studies, particular antigen induces distinct class or subclass of antibodies such as schistosomiasis and filariasis induced a mixed response of IgE and IgG [57, 58]. In order to select distinct class of epitopes, sequence and structure-based dual approaches were used to identify B-cell epitope by using BepiPred-2.0 and ElliPro programs. 14 non-toxic, non cross-reactive, antigenic, and immunogenic B-cell epitopes were identified of different length (Table - 4). Predicted epitopes may or may not be a key feature of proteins because prediction methods, ignored epitope and receptor interaction, may be predicted only putative epitopes, which might lead to produce an antibody of no use. The real epitopes cannot be identified without considering the structural compatibility of complex formation [41]. Therefore, it became important to determine the structural coordinates of peptide-binding pockets to identify motifs for peptide binding. The specificities of different MHC class alleles and B-cell receptor's phenotypes can be used to predict the recognition patterns of epitopes derived from antigens. The molecular docking approach was used to validate the interaction of three T-cell epitopes to most frequently occurring twenty-two MHC allele's structures (Table - 5). Similarly, the top five B-cell epitopes were used to explore peptide interaction with two different kind of B-cell receptor proteins (5DRW and 1K1F). First protein structure, 5DRW was considered to evaluate binding affinity of peptides to BCR antibody light chain, whereas 1K1F was a Bcr-Abl oncoprotein and formed a tetramer through oligomerization. Monomer of 1K1F protein provided a basis to design an inhibitors to disrupt Bcr-Abl oligomerization[42]. Therefore, 1K1F was considered as control to compare the peptide binding affinity to B-cell receptor Fab region binding affinity (Table -6). All peptides showed higher binding affinity to 5DRW than 1K1F except peptide KTVGELGDVRE (Supplementary file1: Table - S8). Docked complex of 1K1F and peptides was contained good interaction and total energy (-37.47, and -1490.12) and six hydrogen bonds (THR2-PRO125, VAL3-SER127, GLU11-ASN144, GLY7-SER168, GLY7-SER182). Hydrogen bond visualization of 1K1F protein and peptide were given in Figure-4.

The MHC genes and regions are one of the highly studied parts of human genome because it is highly associated with different diseases, immune responses and natural targets for molecular evolution, and very well characterized at functional levels [59]. Due to biomedical interest, MHC class gene expression profiling was performed through SARS-CoV-2 RNA-seq data, and HLA-A variants (HLA-A*01, HLA-A*02, HLA-A*03, HLA-A*11, HLA-A*24, HLA-A*25, HLA-A*26, HLA-A*30, HLA-A*31, HLA-A*32, HLA-A*33, HLA-A*34) were more expressed than HLA-B (HLA-B*08, HLA-B*15,HLA-

B*18,HLA-B*37,HLA-B*44,HLA-B*45) and HLA-C variants (HLA-C*03, HLA-C*04, HLA-C*06, HLA-C*07, HLA-C*12, HLA-C*16). In MHC class II analysis, HLA-DMA*01, HLA-DMB*01, HLA-DOB*01, HLA-DPB1*08, HLA-DPB1*75, HLA-DQA1*01, HLA-DQB1*06, HLA-DRA*01 and HLA-DRB1*13 gene variants were highly expressed to present exogenous antigens to $CD4^+$ T cells. Name of expressed MHC class genes and alleles, expression count (TPM) of alleles with respective to human cells were given in supplementary material file1 (Table -S8). Total, 92 gene variants of HLA-A*24 were expressed. Interestingly, very low or no expression for HLA-A*24 gene variants were observed for A549 cell line. In literature, prevalence of HLA-A*24 alleles was suggested as risk factors for severe H1N1 infection [60], and HLA-A*24:02 alleles were also reported to increase diabetes-associated risk together with HLA-B*39:01 gene [61]. A HLA-C allele, HLA-C*03:03, linked male infertility was also highly expressed in SARS-CoV-2 infected NHBE cell lines. The performed study on semen quality was reported that presence of HLA-C*03:03 allele was increased two fold in human papillomavirus virus infected individuals [62]. Three genetic variant of HLA-B*08:01 genes, myasthenia gravis autoimmune disease characterized by muscle weakness and abnormal fatigability were also highly expressed in NHBE cell lines [63]. In present, a combination of malaria and AIDS drugs are in use for the treatment of COVID-19. So, it would be interesting to explore malaria and HIV associated MHC class allele's in SARS-CoV-2 transcriptome. Therefore, all the MHC class genes were analyzed and filtered, and a list of expressed MHC class alleles for malaria and HIV were generated by using IEDB resources along with expression counts for human cell lines (Supplementary file1: Table -S9). HLA-A*02:01:131, HLA-A*02:01:160 and HLA-A*03:01:01:02N were expressed for malaria and HIV in NHBE cell lines. SARS related two HLA genes (HLA-A*23:01:03, HLA-A*23:01:31) were expressed, and 34 HLA-A*24:02 were expressed for HIV gag polyprotein in NHBE cell lines. In our analysis, HLA-A*23:01 HLA-A*24:02 and HLA-A*02:01 were predicted for proposed T-cell epitopes, and molecular interaction studies were also performed between T-cell epitopes and revealed MHC alleles.

5. Conclusion

This study has high scientific relevance to understand immune responses of SARS-CoV-2 in the scarcity of experimental resources. Performed study has provided the extremely useful information about the expressed region of SARS-CoV-2 genome, potential T- and B- cell epitopes, molecular interaction of identified epitopes to receptors through various bioinformatics approaches, and gene expression of MHC class I and II genes. Expressed regions of SARS-CoV-2 genome and putative expressed targets of human immune response will facilitate vaccine related research studies. Proposed epitopes are possessing T- and B-cell selectivity, nontoxicity, higher population coverage, and significant interaction with MHC class I and II genes, and B-cell receptors. However, the presented list of T-and B-cell epitopes is an outcome of computational analysis. But, all the epitopes were identified from transcriptome data of SARS-CoV-2 infection in human cell-lines. Therefore, these epitopes have high

potential to reflect SARS-CoV-2 immune response, and become vaccine candidates after experimental validation.

References

1. **Surveillance case definitions for human infection with novel coronavirus (nCoV): interim guidance v1, January 2020** [https://apps.who.int/iris/handle/10665/330376]
2. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y *et al*: **Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study**. *The Lancet* 2020, **395**(10223):507-513.
3. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, Wang Q, Xu Y, Li M, Li X *et al*: **Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods**. *Acta Pharmaceutica Sinica B* 2020.
4. Stadlbauer D, Amanat F, Chromikova V, Jiang K, Strohmeier S, Arunkumar GA, Tan J, Bhavsar D, Capuano C, Kirkpatrick E *et al*: **SARS-CoV-2 Seroconversion in Humans: A Detailed Protocol for a Serological Assay, Antigen Production, and Test Setup**. *Current Protocols in Microbiology* 2020, **57**(1):e100.
5. Liu W, Fontanet A, Zhang P-H, Zhan L, Xin Z-T, Baril L, Tang F, Lv H, Cao W-C: **Two-year prospective study of the humoral immune response of patients with severe acute respiratory syndrome**. *J Infect Dis* 2006, **193**(6):792-795.
6. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L *et al*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin**. *Nature* 2020, **579**(7798):270-273.
7. Sette A, Fikes J: **Epitope-based vaccines: an update on epitope identification, vaccine design and delivery**. *Current opinion in immunology* 2003, **15**(4):461-470.
8. Gras S, Burrows SR, Turner SJ, Sewell AK, McCluskey J, Rossjohn J: **A structural voyage toward an understanding of the MHC-I-restricted immune response: lessons learned and much to be learned**. *Immunological reviews* 2012, **250**(1):61-81.
9. Larsen MV, Lelic A, Parsons R, Nielsen M, Hoof I, Lamberth K, Loeb MB, Buus S, Bramson J, Lund O: **Identification of CD8+ T cell epitopes in the West Nile virus polyprotein by reverse-immunology using NetCTL**. *PLoS One* 2010, **5**(9):e12697-e12697.
10. Chakraborty S, Chakravorty R, Ahmed M, Rahman A, Waise TM, Hassan F, Rahman M, Shamsuzzaman S: **A computational approach for identification of epitopes in dengue virus envelope protein: a step towards designing a universal dengue vaccine targeting endemic regions**. *In silico biology* 2010, **10**(5-6):235-246.
11. Hasan MA, Khan MA, Datta A, Mazumder MH, Hossain MU: **A comprehensive immunoinformatics and target site study revealed the corner-stone toward Chikungunya virus treatment**. *Molecular immunology* 2015, **65**(1):189-204.

12. Oany AR, Pervin T, Mia M, Hossain M, Shahnaij M, Mahmud S, Kibria KMK: **Vaccinomics Approach for Designing Potential Peptide Vaccine by Targeting Shigella spp. Serine Protease Autotransporter Subfamily Protein SigA**. *Journal of Immunology Research* 2017, **2017**:6412353.
13. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, Møller R, Panis M, Sachs D, Albrecht RA, tenOever BR: **SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems**. *bioRxiv* 2020:2020.2003.2024.004655.
14. SRAToolkit: https://ncbi.github.io/sra-tools/install_config.html
15. FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
16. MultiQC: https://multiqc.info/
17. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
18. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nature methods* 2015, **12**(4):357-360.
19. HISAT2: http://www.htslib.org/
20. Bedtools: https://bedtools.readthedocs.io/en/latest/
21. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nature protocols* 2013, **8**(8):1494-1512.
22. TransDecoder: https://github.com/TransDecoder/TransDecoder/wiki
23. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification**. *Nature Biotechnology* 2016, **34**(5):525-527.
24. Larsen MV, Lelic A, Parsons R, Nielsen M, Hoof I, Lamberth K, Loeb MB, Buus S, Bramson J, Lund O: **Identification of CD8+ T Cell Epitopes in the West Nile Virus Polyprotein by Reverse-Immunology Using NetCTL**. *PLoS One* 2010, **5**(9):e12697.
25. Flower DR, Macdonald IK, Ramakrishnan K, Davies MN, Doytchinova IA: **Computer aided selection of candidate vaccine antigens**. *Immunome Res* 2010, **6 Suppl 2**(Suppl 2):S1-S1.
26. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B: **The Immune Epitope Database (IEDB): 2018 update**. *Nucleic Acids Res* 2019, **47**(D1):D339-D343.
27. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M: **NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W509-W512.
28. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GP: **In silico approach for predicting toxicity of peptides and proteins**. *PLoS One* 2013, **8**(9):e73957.

29. Jespersen MC, Peters B, Nielsen M, Marcatili P: **BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes**. *Nucleic Acids Res* 2017, **45**(W1):W24-W29.
30. Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, Peters B: **ElliPro: a new structure-based tool for the prediction of antibody epitopes**. *BMC bioinformatics* 2008, **9**:514.
31. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ: **AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility**. *Journal of computational chemistry* 2009, **30**(16):2785-2791.
32. Trott O, Olson AJ: **AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading**. *Journal of computational chemistry* 2010, **31**(2):455-461.
33. Kurcinski M, Jamroz M, Blaszczyk M, Kolinski A, Kmiecik S: **CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site**. *Nucleic Acids Res* 2015, **43**(W1):W419-W424.
34. Goddard TD, Huang CC, Ferrin TE: **Software Extensions to UCSF Chimera for Interactive Visualization of Large Molecular Assemblies**. *Structure* 2005, **13**(3):473-482.
35. Wallace AC, Laskowski RA, Thornton JM: **LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions**. *Protein engineering* 1995, **8**(2):127-134.
36. Chen Z, Zhang X, Peng C, Wang J, Xu Z, Chen K, Shi J, Zhu W: **D3Pockets: A Method and Web Server for Systematic Analysis of Protein Pocket Dynamics**. *Journal of chemical information and modeling* 2019, **59**(8):3353-3358.
37. Chauhan P, Hansson B, Kraaijeveld K, de Knijff P, Svensson EI, Wellenreuther M: **De novo transcriptome of Ischnura elegans provides insights into sensory biology, colour and vision genes**. *BMC Genomics* 2014, **15**(1):808.
38. Bui H-H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A: **Predicting population coverage of T-cell epitope-based diagnostics and vaccines**. In: *BMC bioinformatics.* vol. 7; 2006: 153.
39. WHO: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200502-covid-19-sitrep-103.pdf?sfvrsn=d95e76d8_4
40. Darnell SJ, Riese M, Edlunde EG, Blattner FR: **Enhancing B-Cell Epitope Predictions by Integrating Protein Sequence and Structural Bioinformatics**. *Biophysical Journal* 2014, **106**(2):207a.
41. Potocnakova L, Bhide M, Pulzova LB: **An Introduction to B-Cell Epitope Mapping and In Silico Epitope Prediction**. *Journal of immunology research* 2016, **2016**:6760830-6760830.
42. Zhao X, Ghaffari S, Lodish H, Malashkevich VN, Kim PS: **Structure of the Bcr-Abl oncoprotein oligomerization domain**. *Nature Structural Biology* 2002, **9**(2):117-120.

43. Singh K, Mehta S: **The clinical development process for a novel preventive vaccine: An overview**. *J Postgrad Med* 2016, **62**(1):4-11.
44. Tesh RB, Vasconcelos PFC: **CHAPTER 72 - Sandfly Fever, Oropouche Fever, and Other Bunyavirus Infections**. In: *Tropical Infectious Diseases: Principles, Pathogens and Practice (Third Edition)*. Edited by Guerrant RL, Walker DH, Weller PF. Edinburgh: W.B. Saunders; 2011: 481-482.
45. Tilston-Lunel NL, Acrani GO, Randall RE, Elliott RM: **Generation of Recombinant Oropouche Viruses Lacking the Nonstructural Protein NSm or NSs**. *Journal of Virology* 2016, **90**(5):2616.
46. Tanaka T, Kamitani W, DeDiego ML, Enjuanes L, Matsuura Y: **Severe acute respiratory syndrome coronavirus nsp1 facilitates efficient propagation in cells through a specific translational shutoff of host mRNA**. *J Virol* 2012, **86**(20):11128-11137.
47. Graham RL, Sims AC, Brockway SM, Baric RS, Denison MR: **The nsp2 Replicase Proteins of Murine Hepatitis Virus and Severe Acute Respiratory Syndrome Coronavirus Are Dispensable for Viral Replication**. *Journal of Virology* 2005, **79**(21):13399.
48. Kikkert M: **Innate Immune Evasion by Human Respiratory RNA Viruses**. *Journal of Innate Immunity* 2020, **12**(1):4-20.
49. Gao Y, Goonawardane N, Ward J, Tuplin A, Harris M: **Multiple roles of the non-structural protein 3 (nsP3) alphavirus unique domain (AUD) during Chikungunya virus genome replication and transcription**. *PLoS Pathog* 2019, **15**(1):e1007239-e1007239.
50. Cottam EM, Whelband MC, Wileman T: **Coronavirus NSP6 restricts autophagosome expansion**. *Autophagy* 2014, **10**(8):1426-1441.
51. te Velthuis AJW, van den Worm SHE, Snijder EJ: **The SARS-coronavirus nsp7+nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension**. *Nucleic Acids Res* 2012, **40**(4):1737-1747.
52. Ilinskaya AN, Dobrovolskaia MA: **Understanding the immunogenicity and antigenicity of nanomaterials: Past, present and future**. *Toxicol Appl Pharmacol* 2016, **299**:70-77.
53. Tian Y, Grifoni A, Sette A, Weiskopf D: **Human T Cell Response to Dengue Virus Infection**. *Front Immunol* 2019, **10**:2125-2125.
54. Sette A, Moutaftsi M, Moyron-Quiroz J, McCausland MM, Davies DH, Johnston RJ, Peters B, Rafii-El-Idrissi Benhnia M, Hoffmann J, Su H-P *et al*: **Selective CD4+ T cell help for antibody responses to a large viral pathogen: deterministic linkage of specificities**. *Immunity* 2008, **28**(6):847-858.
55. Parvizpour S, Pourseif MM, Razmara J, Rafi MA, Omidi Y: **Epitope-based vaccine design: a comprehensive overview of bioinformatics approaches**. *Drug Discovery Today* 2020.

56. Zhao W, Sher X: **Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes**. *PLoS Comput Biol* 2018, **14**(11):e1006457-e1006457.

57. Hagan P, Blumenthal UJ, Dunn D, Simpson AJG, Wilkins HA: **Human IgE, IgG4 and resistance to reinfection with Schistosoma haematobium**. *Nature* 1991, **349**(6306):243-245.

58. Ottesen EA, Skvaril F, Tripathy SP, Poindexter RW, Hussain R: **Prominence of IgG4 in the IgG antibody response to human filariasis**. *The Journal of Immunology* 1985, **134**(4):2707.

59. Trowsdale J, Knight JC: **Major Histocompatibility Complex Genomics and Human Disease**. *Annual Review of Genomics and Human Genetics* 2013, **14**(1):301-323.

60. Hertz T, Oshansky CM, Roddam PL, DeVincenzo JP, Caniza MA, Jojic N, Mallal S, Phillips E, James I, Halloran ME *et al*: **HLA targeting efficiency correlates with human T-cell response magnitude and with mortality from influenza A infection**. *Proceedings of the National Academy of Sciences* 2013, **110**(33):13492-13497.

61. Mikk M-L, Heikkinen T, El-Amir MI, Kiviniemi M, Laine A-P, Härkönen T, Veijola R, Toppari J, Knip M, Ilonen J *et al*: **The association of the HLA-A*24:02, B*39:01 and B*39:06 alleles with type 1 diabetes is restricted to specific HLA-DR/DQ haplotypes in Finns**. *HLA* 2017, **89**(4):215-224.

62. Marques PI, Gonçalves JC, Monteiro C, Cavadas B, Nagirnaja L, Barros N, Barros A, Carvalho F, Lopes AM, Seixas S: **Semen quality is affected by HLA class I alleles together with sexually transmitted diseases**. *Andrology* 2019, **7**(6):867-877.

63. Varade J, Wang N, Lim CK, Zhang T, Zhang Y, Liu X, Piehl F, Matell R, Cao H, Xu X *et al*: **Novel genetic loci associated HLA-B*08:01 positive myasthenia gravis**. *Journal of Autoimmunity* 2018, **88**:43-49.

Author's contributions

SKK: conceived the study, planned the data processing and analysis, and written first draft of manuscript.

VB and SS: performed the docking studies.

SC and SG: contributed in biological data interpretation, manuscript writing, and generation of figures and tables.

Final manuscript is read and approved by all the authors.

Acknowledgment

We are grateful to the Director NIAB for providing support to carry out the study. Bioinformatics facility, NIAB is acknowledged for the bioinformatics data analysis. We would also thankful to Aakash Chawade and Pallavi Chauhan for useful discussion and suggestions.

Competing interest

Authors have declared that they have no competing interest.

Supplementary information

Supplementary file 1(Table S1-S9): Meta data for used transcriptome, detail description of T- and B-cell epitopes, detail description docking studies and gene expression analysis of MHC-alleles

Supplementary file 2: Verification of genuine hydrogen bonds involved in the interaction through cavity prediction.

Data availability

All the used RNA-seq data is available at NCBI SRA under the project accession number PRJNA615032.