 doi:10.20944/preprints202005.0413.v1

Characterization of nucleocapsid (N) protein from novel coronavirus SARS-CoV-2

Abhishek Kumar[1, 2, *] Alisha Parveen[3], Narendra Kumar[4], Sneha Bairy[5], Vibha Kaushik[6], Chetan Chandola[1], Jyoti Sharma[1, 2], Phulwanti Sharma[6], Akhil Agarwal[7], Akhilesh Pandey[2, 8, 9], Pankaj Goyal[6, *] and Muniasamy Neerathilingam[1, 2, *]

[1]Institute of Bioinformatics, International Technology Park, Bangalore, 560066 India
[2]Manipal Academy of Higher Education (MAHE), Manipal 576104, Karnataka, India
[3] Medical Research Center, Medical Faculty of Mannheim, University of Heidelberg, Mannheim, Germany;
[4]Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan - 173 234, Himachal Pradesh, India
[5]Institute for Stem Cell Science and Regenerative Medicine, GKVK - Post, Bellary Rd, Bengaluru, Karnataka 560065, India
[6]Department of Biotechnology, School of Life Sciences, Central University of Rajasthan, Bandarsindri, Kishangarh, Rajasthan 305817 India
[7]Department of Microbiology, School of Life Sciences, Central University of Rajasthan, Bandarsindri, Kishangarh, Rajasthan 305817 India
[8]Center for Molecular Medicine, National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore, India
[9]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States 7Center for Individualized Medicine, Mayo Clinic, Rochester, MN, United States

*To whom correspondence may be addressed:
AK - abhishek@ibioinformatics.org; Tel.: +91 80-2841-6140
PG - pankaj_bio@curaj.ac.in
MN - muniasamy@ibioinformatics.org; Tel.: +91 80-2841-6140

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Abstract

Severe acute respiratory syndrome novel coronavirus 2 (SARS-CoV-2) has caused the global pandemic as COVID-19, which is the most notorious global public health crisis in the last 100 years. SARS-CoV-2 is composed of four structural proteins and several non-structured proteins. The multi-facet nucleocapsid (N) protein is the major component of structural proteins of CoVs, However, there are no dedicated genomic, sequences and structural analyses focusing on potential roles of N protein. Hence, there is an urgent requirement of a detailed study on N protein of SARS-CoV-2. Herein, we are presenting a comprehensive study on N protein from SARS-CoV-2.

We have identified seven motifs conserved in the three major domains namely N-terminal domain, linker regions and the C-terminal domains. Out of seven motifs, six motifs are conserved across different members of *coronaviridae*, while motif4 is specific for SARS CoVs with potential amyloidogenic properties. Additionally, we report this protein has large patches of disordered regions flanking with these seven motifs. These motifs are hubs of epitopes with 67 experimentally verified epitopes from related viruses. We report the presence of three nuclear localization signals (NLS1-NLS3 mapped to 36-41, 256-26, and 363-389 residues, respectively) and two nuclear export signals (NES1-NLS2 from 151-161 and 217-230 residues, respectively) in the N protein of SARS-CoV-2. These deciphered two Q-patches as Q-patch1 and Q-patch2, mapped in the regions of 266-306, and 361-418 residues, which potentially help in the aggregation of the viral proteins along with ^{219}LALLLLDR^{226} patch. Additionally, we have identified 14 antiviral drugs potentially binding to seven motifs of N-proteins using docking-based drug discovery methods.

Keywords: SARS-CoV-2; nucleocapsid (N); genomics; coronavirus; Wuhan; Pandemic;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

1. Introduction

A novel coronavirus (CoV) named as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, initially called as 2019-nCoV) pushed the entire world on a halt with outbreak of coronavirus disease 2019 (COVID-19). This disease has rapidly become a new emerging human disease and a global pandemic as SARS-CoV-2 has already infected more than 4,731,458 people across globe and over 316,169 deaths due to COVID-19 as Situation Report–120 of World Health Organization (WHO, Webpage https://www.who.int/emergencies/diseases/novel-coronavirus-2019 on 19th May 2020).

This suggested that the rate of mortality is 6.7%, which is higher than initial estimation of about 2%. Several countries have locked down their cities, states and their international borders, which has led almost no international travels and ultimately resulted in a massive slowdown in the global economy and businesses.

Coronaviruses (CoV) are members of coronaviridae (NCBI Taxonomy ID: 69399) and these are enveloped positive-sense, single-stranded RNA viruses. These members of coronaviridae are evolutionary grouped into four genera with Greek number prefixes as alpha (α-CoV), beta (β-CoV), gamma (γ-CoV) and delta-coronaviruses (δ-CoV). CoVs are known to cause infections of the mammalian respiratory and gastrointestinal tracts, but the infection mechanism of these viruses is not fully understood. The current virus SARS-CoV-2 is seventh CoV, which caused infections to humans and the other four infections by members of β-CoV genera were SARS, Middle East respiratory syndrome (MERS), HCoV-OC43 and HCoV-HKU1, while two α-CoV infections were reported as HCoV-NL63 and HCoV-229E. One examining carefully, it is clear that CoV have tendencies to break species boundaries during infecting humans, they use intermediate hosts such as bats, ant-eaters and camels [1,2]. Hence, it is cautioned that CoV-based infections will come back every now and then with our exposure to intermediate hosts most probably from a wild animal species [2].

Although known that CoV clearly causes infections multiple times and every time a new infection comes, we tend to control its impact by containing spread of the virus by lockdown and using other viral drugs. It is the high time that we focus on CoV using an insight out approach focusing on the long term solutions for the current disease – COVID19 and also how we can counter any other future outbreak of the CoV. Hence, it is an urgent requirement that several computational methods have to come up for characterization of various components of CoV genetics and biology. This will be unravelling future therapeutic targets against SARS-CoV-2 and related viruses.

We have an advantage now that using next-generation sequencing methods several genomes of CoV are available and massive global attempts are made to make more and more genomic data to be available soon. Recently available genome of SARS-CoV-2 is 29.3 kb in size (NCBI Accession ID ASM985889v3) harbouring four essential structural proteins, including spike (S) glycoprotein, small envelope (E) protein, matrix (M) protein, and nucleocapsid (N) protein [3]. The nucleocapsid (N) protein of a coronavirus is a multifunctional protein that plays a crucial role in virus assembly and in its RNA transcription. The N protein is crucial in the formation of helical ribonucleoproteins during packaging the RNA genome, regulating viral RNA synthesis during replication, transcription. This protein is also capable of regulating infected host cells and their cellular mechanisms. The primary functions of N protein are binding to the viral RNA genome, and packing them into a long

 doi:10.20944/preprints202005.0413.v1

helical nucleocapsid structure or ribonucleoprotein (RNP) complex. The N protein possesses highly immunogenic properties and it is a highly expressed protein during infection, capable of inducing protective immune responses against SARS-CoV and SARS-CoV-2.

Herein, we have carried on a comprehensive characterization of nucleocapsid N protein of SARS-CoV-2 from sequence, phylogenetic and structural perspectives for deciphering potential therapeutic targets and epitome inventories. We have seven motifs with different roles either *coronaviridae*-specific (motif2) or SARS-CoVs-specific (motif4), two glutamine-rich Q-patches and potential roles in generating higher aggregation propensity in the human cells.
We also explored a library of FDA-approved drugs against seven motifs of N protein for their roles as anti-SARS-CoV-2 drugs and usages in drug repurposing and majority of drugs are HIV

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

2. Materials and Methods

2.1. Scanning nucleocaspid (N) proteins of several coronavirus genomes

We detected putative nucleocapsid (N) proteins from different viruses using BLASTP [4] with an E-value < $1e^{-10}$ with nucleocapsid (N) protein (Genbank Accession ID YP_009724397.2) as the query after setting up the local BLAST database. We performed annotation of these sequences using OMICSBOX [5] and CELLO2GO [6].

2.2. Genomic analyses of nucleocapsid (N) protein

We utilized the genome of Wuhan SARS-CoV-2 isolate (originating from the Wuhan seafood market, Wuhan, China [3]) with Genbank accession ID NC_045512.2 and carried out the comparisons with four Indian isolates as Indian SARS-CoV-2 samples. Indian genomic data was collected from GISAID (Global Initiative on Sharing All Influenza Data, Web: https://www.gisaid.org/) with accession details as EPI_ISL_413523 (hCoV-19/India/1-31/2020), EPI_ISL_426179 (hCoV-19/India/c31/2020), EPI_ISL_426414 (hCoV-19/India/GBRC1/2020). and EPI_ISL_426415 (CoV-19/India/GBRC1s/2020). We visualized genomic organization flanking nucleocapsid (N) of SARS-CoV-2 using SNAPGENE (web: https://www.snapgene.com/). In Addition, it was compared with other CoV using UCSC genome browser [7] after aligning 43 CoV strains (derived from bats) with Multiz aligner [8].

2.3. Protein sequence and strucural analyses of nucleocapsid (N) protein from selected CoV

We aligned nucleocapsid (N) proteins from selected CoVs using MUSCLE alignment suite [9] and resulting alignments were visualized using either ESPrint3.0 tool [10] and/or JALVIEW [11]. Homology model of full length N protein (accession ID - QHD43423.pdb) was taken from I-TASSER [12]. This protein sequence alignment was constructed using Muscle alignment tool and this protein sequence alignment is visualized along with secondary structural elements using ESPrint3.0 tool [10]. We predicted nuclear localization and nuclear export locations using and NetNES1.1 [13], respectively and these signals are manually curated further. We analysed disordered regions using PREDICTPROTEIN and DOSOPRET3.0 [14]. We visualized the motifs in the homology model using PyMol (www.pymol.org) and YASARA (www.yasara.org).

2.5. Construction of sequence logo

We constructed sequence logos of selected nucleocapsid (N) motif using WEBLOGO3.0 [15], as described previously [16,17].

2.6. Phylogenetics analyses

We have carried out phylogenetic analyses of nucleocapsid (N) protein from selected CoV with Neighbor-joining (NJ) method [18] using bootstrap value of 500 under using MEGA-X [19].

2.7. Mapping of experimentally validated epitopes using IEDB database

We mapped experimentally validated immune epitopes, flanking identified 7 motifs of N protein with cut-off of minimum 70% sequence identity from Immune Epitope Database and Analysis Resource (IEDB, Web: http://www.iedb.org/), derived by BLASTP [4].

2.8. Prediction of aggregation propensity for nucleocapsid (N) protein

For predicting aggregation propensity of nucleocapsid (N) protein, we have supplied N protein sequence (YP_009724397.2) to four different tools namely TANGO [20],

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

AGGRESCAN [21], FoldAmyloid [22], and Amylpred [23]. For TANGO following parameters were used pH, temperature, ionic strength and protein concentration by setting to values of 7.4, 310 K, 0.1 M and 0.1 M, respectively and the protein was analyzed in its native form without any N- or C-terminal protection. The data were then compared with those of a well-known amyloidogenic protein, TAR-DNA binding protein 43 (TDP-43, accession no. NP_031401.1), which is involved in the amytropic lateral sclerosis (ALS) [24].

2.9. Docking analyses focusing on seven motifs

We downloaded FDA approved drugs in SDF format from Drugbank [25]. For our docking analyses, we have unutilized the protein model of full length N protein from I-TASSER) website (accession ID - QHD43423.pdb). In this model, we mapped the seven conserved motifs. We docked A total of 1908 drugs to different motifs using Autodock vina (3), after energy minimization with help of ff14SB amber force field in the UCSF Chimera Software (4). We performed the virtual screening using VSvina custom scripts (https://github.com/narekum/VSvina). After docking, we carefully examined the binding modes of best affinity binders for each of the binding motifs. Resulting data was visualized using PyMol (www.pymol.org) and YASARA (www.yasara.org).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

3. Results

3.1. Overview of genomic location of SARS-CoV-2 N protein

Recently the genome of SARS-CoV-2 has become available by Wu *et al.* [3] with Genbank accession ID NC_045512.2. This genome of SARS-CoV-2 is 29,903 in the size and we compared the genome sizes for four Indian SARS-CoV-2 strains and we found that assembled genome size is same for EPI_ISL_426414 (hCoV-19/India/GBRC1/2020) and EPI_ISL_426415 (CoV-19/India/GBRC1s/2020), while it is slighter shorter as 29,851 and 29,860 nucleotides for EPI_ISL_413523 (hCoV-19/India/1-31/2020) and EPI_ISL_426179 (hCoV-19/India/c31/2020), respectively (**Fig. 1**).

The SARS-CoV-2 genome possesses 14 open-reading frames (ORFs) encoding 24 proteins [3]. The ORF N is localized in the 3′-terminal end of the genome in a 1280 bp region from 28.274 bp to 29,533 bp flanked by ORF8 (366 bp) on the one side while ORF10 (117 bp) is on the other side (**Fig. 1**). The ORF N is also mapped into genomes of four Indian SARS-CoV-2 strains at equivalent locations as 28.254-29,510 bp, 28.261-29,517 bp, 28.274-29,530 bp, and 28.261-29,517 bp for four Indian samples as EPI_ISL_413523 (hCoV-19/India/1-31/2020), EPI_ISL_426179 (hCoV-19/India/c31/2020), EPI_ISL_426414 (hCoV-19/India/GBRC1/2020) and EPI_ISL_426415 (CoV-19/India/GBRC1s/2020), respectively (**Fig. 1**). This genomic architecture is conserved in several CoVs (**Fig. 1**) as illustrated on the genomic scales of the 43 bat-CoVs genomes (**Suppl. Table 1**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

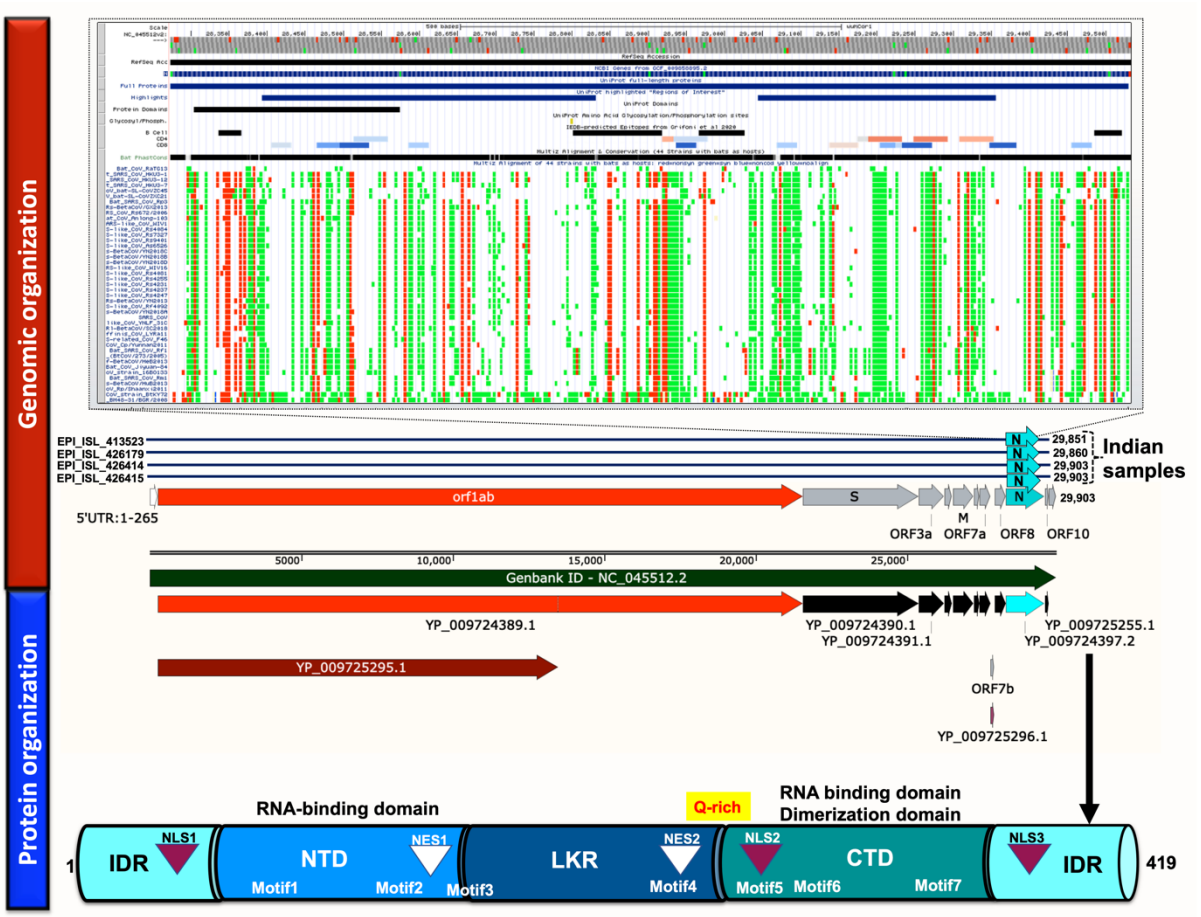

Fig 1. Overview of genomic and protein architectures of N protein from SARS-CoV-2, depicting location of seven motifs identified in this study. Four Indian SARS-CoV-2 genomes was compared with the genome of Wuhan SARS-CoV-2 isolate (Genbank accession ID NC_045512.2) by Wu *et al.* [3].
Indian SARS-CoV-2 samples were
EPI_ISL_413523 (hCoV-19/India/1-31/2020),
EPI_ISL_426179 (hCoV-19/India/c31/2020),
EPI_ISL_426414 (hCoV-19/India/GBRC1/2020) and
EPI_ISL_426415 (CoV-19/India/GBRC1s/2020)
CTD - C-terminal dimerization domain, IDR - intrinsically disordered region, NES – Nuclear Export signal, NLS – Nuclear localization signal and NTD - N-terminal RNA-binding domain

3.2. Protein architectures and sequence of N protein from SARS-CoV-2

We aligned Nucleocapsid (N) from representative viruses as listed and a re-annotation summary is provided in **Table 1**. Nucleocapsid (N) protein of SARS-CoV-2 is 419 amino acid long with a molecular mass of 45.6 kDa. This N protein has two major intrinsically disordered regions (IDRs) in the N- and C-terminal ends ranging from 1 to 41 and 366 to 419 residues (**Figs. 1-2**).

There are three conserved domains in protein architectures of coronavirus N protein namely an N-terminal RNA-binding domain (NTD), a C-terminal dimerization domain (CTD), and intrinsically disordered Ser-Arg (SR)-rich linker region. We have built the model structure of the full length N protein from SARS-CoV-2, which is composed of eight α-helices namely helix α1 to helix α8, ten η-helices designed as helix η1 to helix η10 and nine β-sheets as β-sheet β1-to β-sheet β9 (**Figs 2-3**).

Nucleocapsid (N) protein of SARS-CoV-2 is the close homolog of protein N of BtCoV-RatG12 and SARS CoV (SARS-CoV) with 99% and 90%sequence identities and 99% and 94% sequence similarities, respectively (**Fig. 2**). During

 doi:10.20944/preprints202005.0413.v1

our sequence comparisons, we found that protein N from SARS CoVs have higher sequence identities and similarities ranged from 89% to 99% identities and 93% to 94% similarities, whereas N protein from other viruses namely PHEV-1 (Genbank accession ID. ARC95215.1) and PHEV-2 (AAY68302.1) from porcine hemagglutinating encephalomyelitis virus and MHV-1 (AAA76578.1) and MHV-2 (BAJ04701.1) from murine hepatitis virus show 30-31% identities and 47% similarities, respectively.

We also examined the N protein from four Indian SARS-CoV-2 genomes versus N protein from SARS-CoV-2 Wuhan isolates. We found that sequence identity and sequence similarity are in ranges of 99-100% for N protein deduced from these four genomes versus that of SARS-CoV-2 Wuhan isolates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

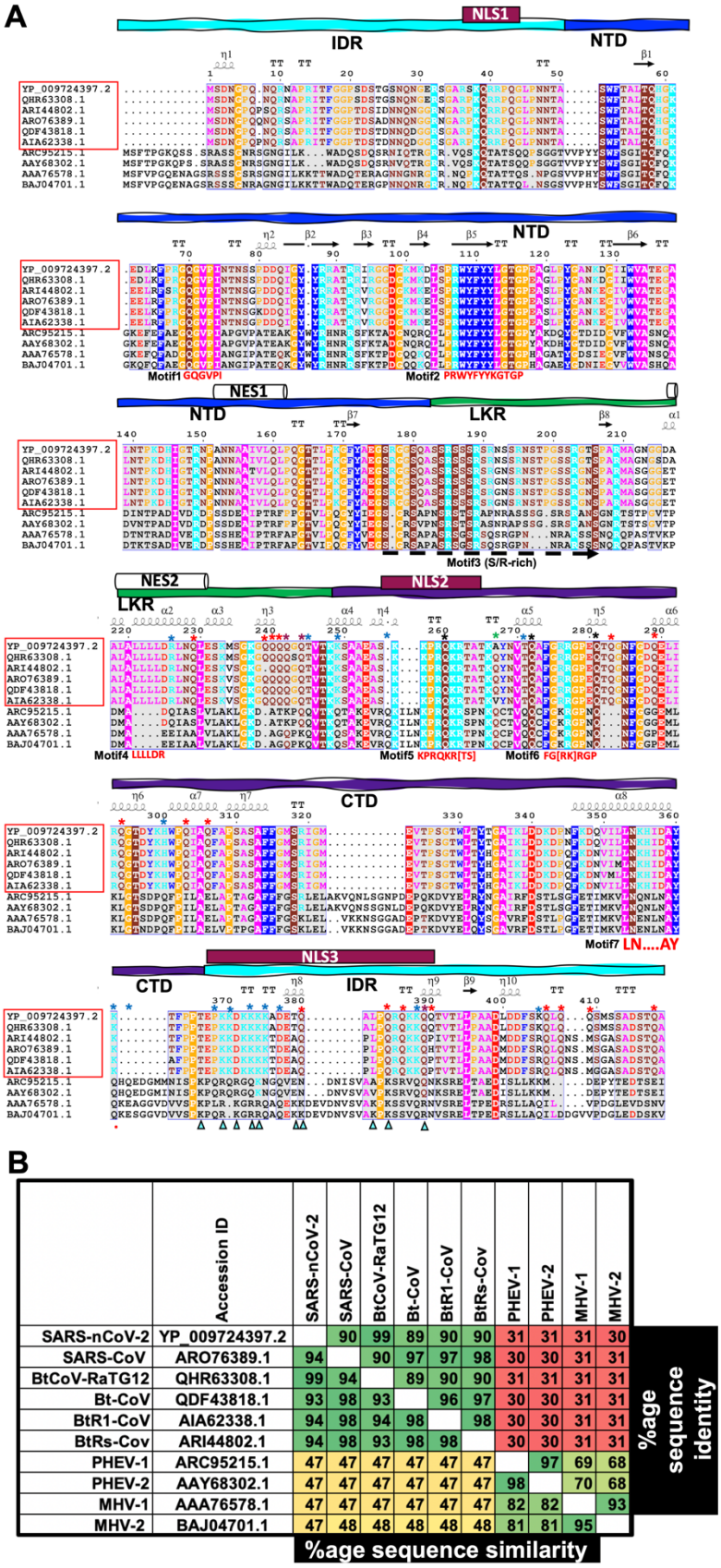

Fig 2. Protein sequence alignment of N protein illustrates secondary structural elements and positions of different motifs.

A. The full-length N protein of SARS-CoV-2 has different secondary structural elements namely 8 α-helices, 10 η-helices & 9 β-sheets, mapped on top of the alignment, derived from I-TASSER [12] model (PDB ID - QHD43416.pdb). This protein sequence alignment was constructed using MUSCLE suite [9] and this protein sequence alignment is visualized along with secondary structural elements using ESPrint3.0 tool [10]. Amino acids are coloured based on physicochemical properties as following cyan – HKR, red – DE, maroon – STNQ, pink – AVLIM. CTD - C-terminal dimerization domain, IDR - intrinsically disordered region, NES – Nuclear Export signal, NLS – Nuclear localization signal and NTD - N-terminal RNA-binding domain; cyan triangle – alternative K or R for NLS3
*-position of glutamines as
red * - glutamine present only in SARS-CoVs
green * - glutamine absent in SARS-CoV-2
blue * - glutamine present only in other viruses
black * - glutamine present in all viruses
maroon * - glutamine present in SARS-CoVs and some other viruses

B. Sequence identity and similarity scores depicts that CoVs (top 6) have higher sequence identities and similarities (green shades) than other viruses (bottom 4) - porcine hemagglutinating encephalomyelitis virus (like PHEV-1 (Genbank accession ID. ARC95215.1) and PHEV-2 (AAY68302.1) and murine hepatitis virus (MHV-1, AAA76578.1 and MHV-2, BAJ04701.1)

This is clearly evident that there is grouping into two these classes of viruses (**Fig. 1**) as (a) first six N protein sequences from different CoVs (marked in red box, **Fig. 1**)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

and (b) lower four N protein sequences with lower sequence identities (**Fig. 1**). Upon examining carefully, we have identified seven motifs present in several CoVs mapped into different regions of the protein N namely motif1-motif7 (**Figs. 2-3**) as listed in **table 2**.

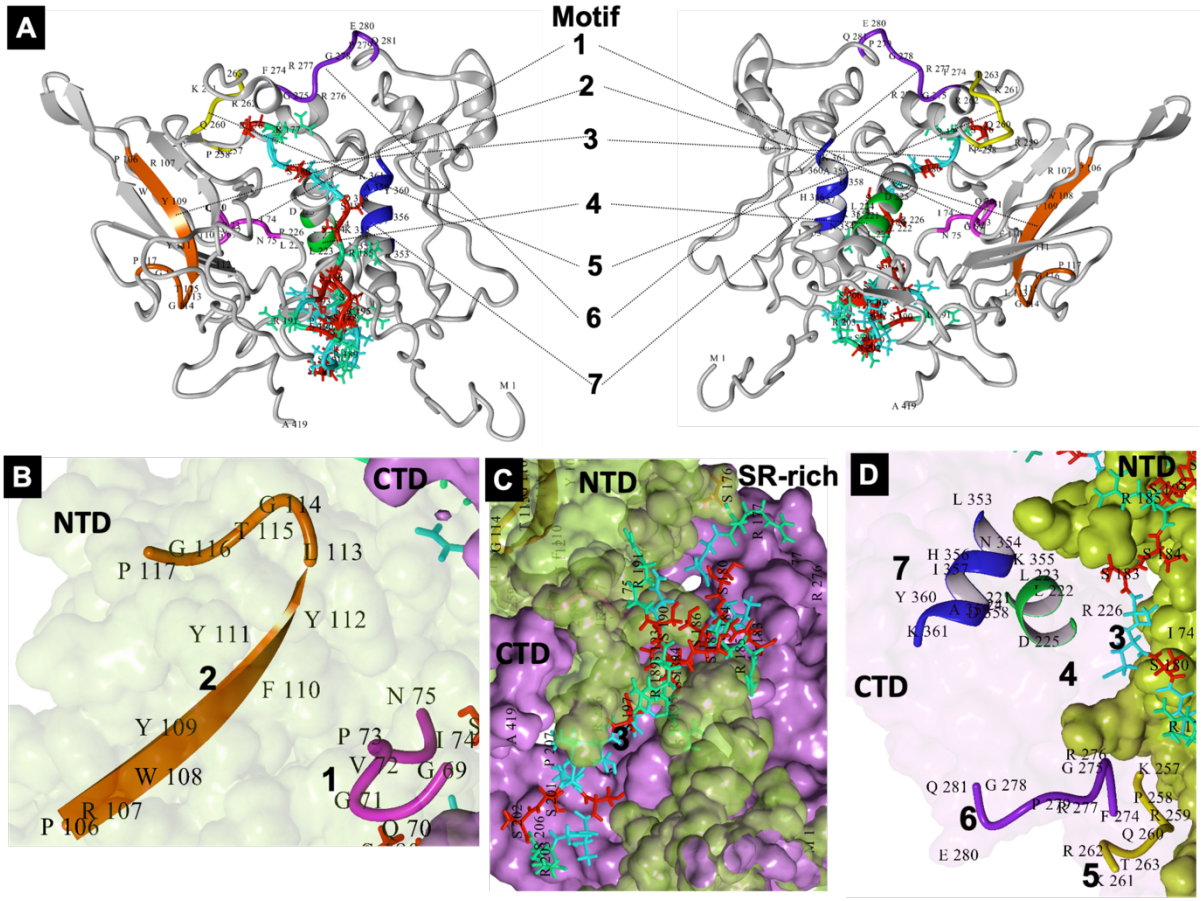

Fig 3. Structural modelling and visualization of the full-length N protein of SARS-CoV-2 depicting locations of 7 identified motifs of N protein across secondary structural elements. Secondary structural elements are 8 α-helices, 10 η-helices & 9 β-sheets.

A. Front & back view illustrating motifs 1-7.
B. Location of motif1 (^{69}GQGVPI^{75}) and motif2 (^{106}PRWYFYYLGTGP^{117}) in the N-terminal RNA-binding domain (NTD).
C. Location of motif3 (S/R rich region) in the turn between β-sheet β4 and Π-helix Π1, mapped into at the end of NTD and the linker region (LKR).
D. Illustration of motifs4 to motif7 mapped into the C-terminal dimerization domain (CTD) located in the turn between the Π-helix Π4 and the α-helix α5, the turn between the α-helix α5 and the Π-helix Π5 and, at the end of the α-helix α8, respectively.

First two motifs are present in the NTD region as motif1 and motif2 in the amino acid positions at 69-75 and 106-117 as ^{69}GQGVPI^{75} and ^{106}PRWYFYYLGTGP^{117} and these two motifs are mapped in the turn between β-sheets β1-β2 and at the β-sheet β5, respectively (**Figs. 2-3**). The third motif is a large stretch of S/R-rich region, starting at the of the NTD region and surpassing over the linker region (LKR) from 176 to 207 residue positions (according to YP_009724397.2 numbering). This motif is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

present in the LKR region connecting the NTD to CTD domains as visualised on the structural model, in the turn between the β-sheet β4 and the Π-helix Π1 (**Figs 1-2**). The fourth motif is six-residues long as ^{221}LLLLDR^{226} in the LKR region, residing in the turn in between the β-sheet β4 and the Π-helix Π1.

Remaining three motifs are localized in the C-terminal domain (CTD) in 257-263, 274-^{53}LN....AY.^{361}, which is localized at the end of the α-helix α8.

3.3. Motif2 of N protein is coronaviridae-specific and it is potential hub of with various epitope design

Motif2 is 12 residues long as ^{106}PRWYFYYLGTGP^{117} which is mapped at the β-sheet β5, (**Figs. 1-2**). This motif is 281 and 353-361 as motif5-motif7, respectively. Motif5 is seven amino acid long as ^{257}KPRQKR[ST]^{263} with either serine or threonine at the 7th position, which is structurally mapped in the tern connecting the Π-helix Π4 and the α-helix α5, where motif6 is closely mapped on the next tern connecting the α-helix α5 and the Π-helix Π5 and it is six residues long as ^{274}FG[KR]RGP^{281}. The seventh motif is [3] present in various members of *coronaviridae* including various coronaviruses as evident from phylogenetic tree (**Fig. 4**), where other viruses are marked in blue boxes like murine hepatitis viruses and porcine hemagglutinating encephalomyelitis. This motif is 100% identical at this location and it hinted us to examine its potential immunological roles like epitope formation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

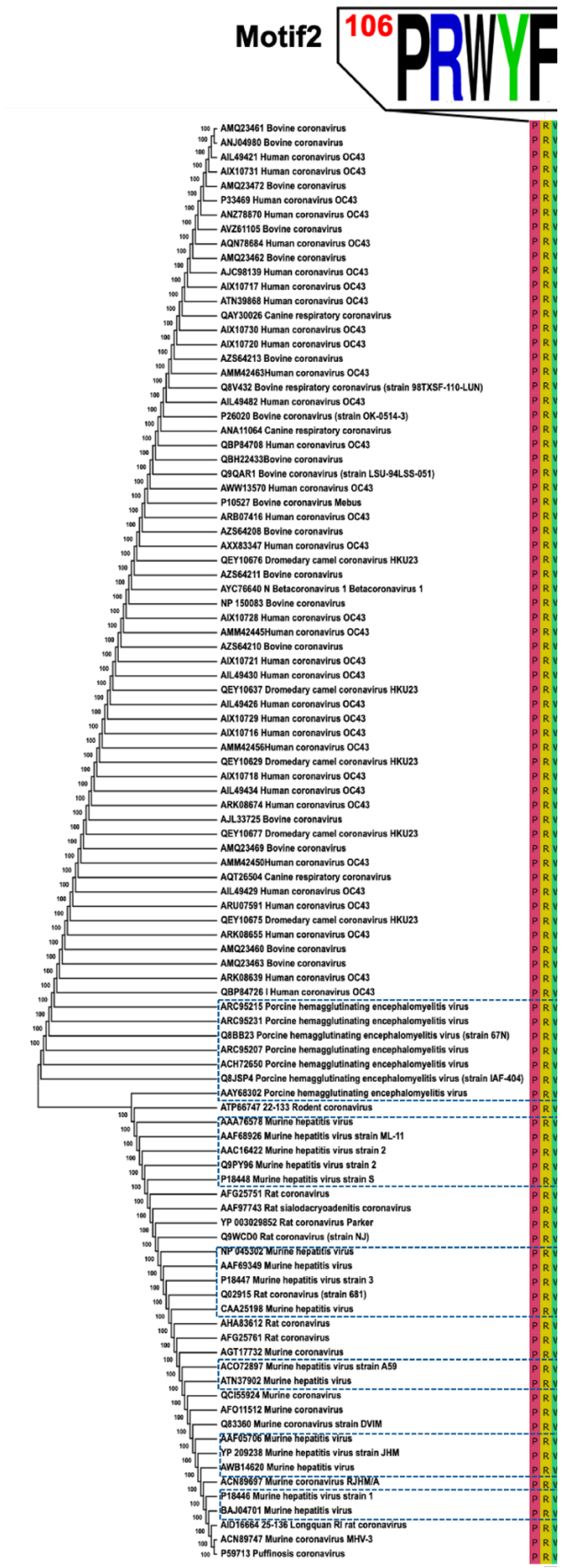

Fig 4. Phylogenetic analysis reveals that motif2 (^{106}PRWYFYYLGTGP^{117}) of nucleocapsid N is highly conserved across different viruses including various CoV.

Interestingly, we found that this motif is residing to a central location, of which 9-10 residues flanking in both directions, have be experimentally validated for immune

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

epitopes in various experiments for different SARS-CoVs using T-Cell and MHC arrays in human and mice (BALB/c). We have summarized 19 epitopes (epitope1 to epitope19) using motif2, either fully or partially matching to these epitopes in other SARS-CoVs and also in related viruses like feline infectious peritonitis virus (strain KU-2) and murine hepatitis virus strain JHM (**Table 3**). The presence of 19 epitope is this region of N protein for CoVs and total conservation of this motif, together hints that motif2 is a potentially the hub of the epitope designing for SARS-CoV-2, which is supported by various experimental validation in closely related viruses like different strains of SARS-CoV as summarized in **Table 3.**

3.4. Motif4 is SARS-CoV-specific with amyloidogenic properties

Motif4 is clearly an insertion in the linker (LKR) region of the nucleocapsid protein of the SARS-CoV-2 and related CoVs (**Fig. 2)** but not present in other viruses. We examined this motif in various members of coronaviridae and we confirmed that this leucine-rich six-residue motif (^{221}LLLLDR^{226}). This motif is present in various SARS-CoVs (**Fig. 5A**) in the turn between the β-sheet β4 and the Π-helix Π1 in the linker (LKR) region (**Fig. 3**). This motif is the central element of a potential nuclear export signal (NES, **Fig. 5B**) .

Using four tools for amyloidogenic propensity prediction, we identified eight amyloidogenic stretches, present throughout the N protein sequence (**Table 4**). Interestingly, two of the eight predicted stretches; ^{108}WYFYYL^{113} and ^{219}LALLLLDR^{226} (extension of motif4) of the protein possess the highest aggregation score for all the tools.

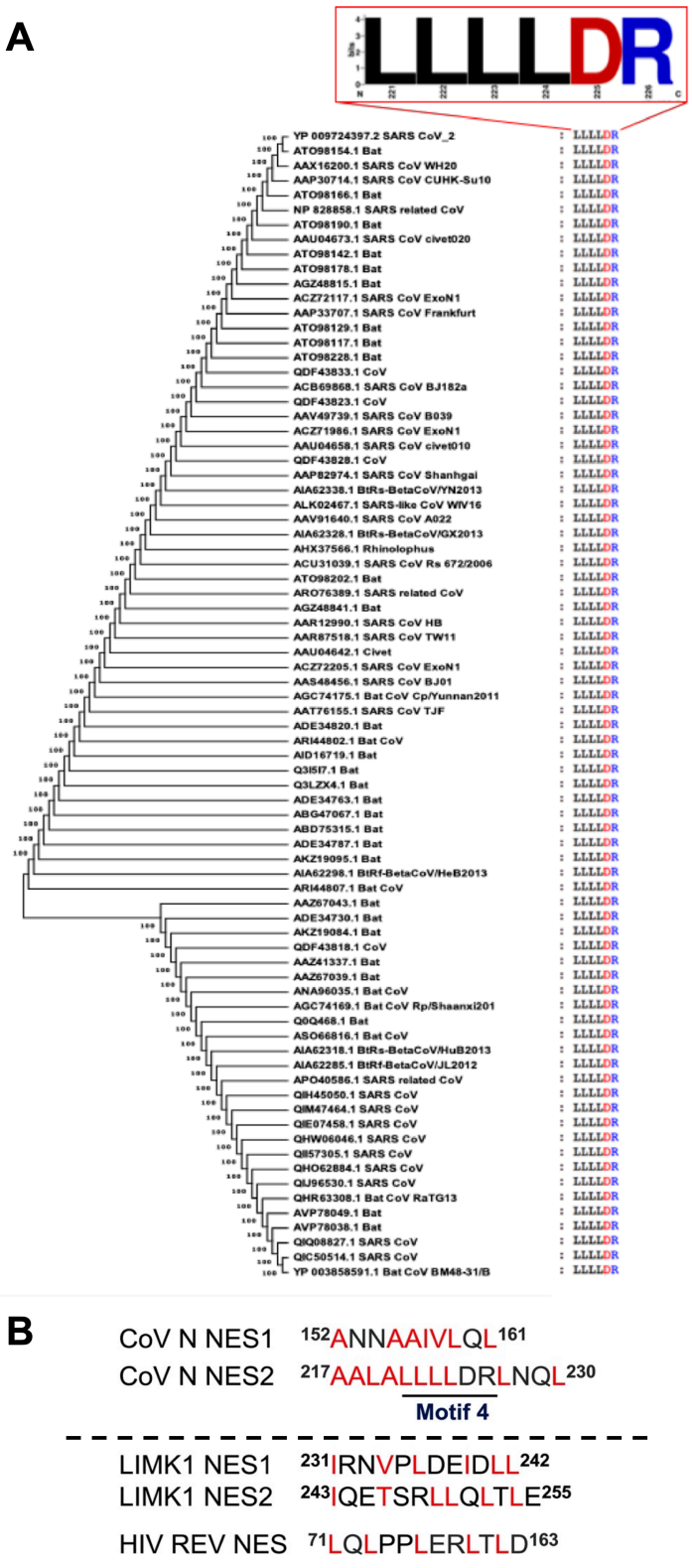

Fig 5. Motif4 (^{221}LLLLDR^{226}) is SARS-CoVs-specific with harbouring a nuclear export signal (NES)

A. Phylogenetic analysis reveals that motif4 is present in various SARS-CoVs.

B. Motif4 is in the central fragment of nuclear export signal 2 (NES2) of SARS-CoV N protein. This protein has two NES as NES1 and NES2 mapped to 152–161 and 217–230 residues respectively and these two NESs were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

compared with well-known examples of NES from human LIMK1 and HIV REV. Hydrophobic residues are shown in red colour.

Notably, the sequence alignment of representatives of SARS-CoV family and other viruses showed that the first region remains conserved among the related virus while the motif4 has specifically got inserted in the SARS-CoV family. These data suggest that the nucleocapsid protein has a high aggregation propensity in the host cells. The specific insertion of motif4 (harbouring leucine-rich region) might enhance the aggregation propensity of this protein compared to that of the other viruses and thus might form amyloid-like structures.

Additionally, it is interesting to note that two amyloidogenic stretches ^{350}VILLN^{354} and ^{392}VTLLP^{396} in SARS-nCoV-2 N protein (**Fig. 2 & Table 4**) bears sequence homology with ^{75}VLVVL^{79} from the ORF8b. Shi et al. previously demonstrated that the VLVVL motif confers the aggregation ability to ORF8b protein [26]. Thus, we might speculate that these stretches in N protein might confer the ability for aggregation in SARS-nCoV-2, however it needs to be experimentally verified. It is known that intracellular protein aggregation contributes to the pathogenesis of a variety of diseases and is both propagated by and contributes to inflammation [27,28]. Recently, ORF8b protein in SARS-CoV has shown aggregation that leads to cytotoxicity in epithelial cells, and this cytotoxicity can be partially rescued by preventing the aggregation of this protein. The aggregation of ORF8b protein induces endoplasmic reticulum (ER) stress, lysosomal damage, and the activation of transcription factor EB (TFEB)[26].. Since N protein in SARS-CoV-2 shows amyloidogenic stretches, it would be interesting to explore the possibility of formation of aggregates and the role in inflammation and cytotoxicity in epithelial cells. Overall, all these amyloidogenic stretches (**Table 4**) require experimental validations for their detailed roles in aggregation, inflammation, cytotoxicity, autophagy and cellular apoptosis.

3.6. Potential roles of other motifs of SARS-CoV-2 N protein

We have evaluated epitome profiles of other motifs using from IEDB and we found 48 epitopes flanking these 6 motifs (**Table 5**). The first motif (^{69}GQGVPI^{75}) mapped to the turn between β-sheets β1-β2 β-sheet β5, respectively (**Figs. 2-3**) has eight experimentally verified epitopes and seven of these are reported for SARS-CoV and one for murine hepatitis virus strain JHM (**Table 5**). The S/R-rich motif3 in the LKR has a total of ten epitopes experimentally verified for different SARS-CoV strains (**Table 5**). The six-residues long motif4 (^{221}LLLLDR^{226}) has 9 epitopes flanking in closely related SARS-CoV strains. Similarly, there are 5, 3, and 13 validated epitome regions motif5, motif6 and motif7, respectively (**Table 5**). Taken together 67 epitopes of seven motifs (**Tables 3, 5**), 94% are deduced from SARS-CoV strains.

3.7. Deciphering nuclear localization signal and nuclear export signals in the N protein

There are three nuclear localization signals (NLS) in the N protein of SARS-CoV-2 (**Fig. 2**), namely NLS1-NLS3, mapped on one each to the N-terminal IDR ranged from 36-41, to the CTD from 256-262 and to the N-terminal IDR from 363-389. NLS2 is mapped to motif5. NLS3 is the longest NLS and it is conserved for SARS-CoVs but alternative lysine (K) and arginine (R) are present in the same region for other viruses (marked by cyan triangles in **Fig 2**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

This hints that NLS3 is common to viral N proteins with some variations.

The nuclear export of proteins is mainly governed by CRM1 (chromosome region maintenance 1 protein) or XPO1 (exportin-1)[29]. These recognize the protein nuclear export signal (NES) in the cargo proteins [29]. NESs are hydrophobic rich (preferably leucine) regions of 8-15 amino acids long sequences [29]. A comparison of NES containing proteins [30] with SARS-CoV-2 N protein, we identified two NES with the NES1 is mapped in the NTD from 152-161 whereas the second NES (NES2) mapped in the LKR ranged from 217-230 (**Figs. 2** and **5B**). Interestingly, the NES2 was found in SARS-CoV-specific motif4 (**Fig. 5B**). These data suggest that motif4 might enhance the cytoplasmic localization of N protein from SARS-CoV and have a SARS-CoV-specific cytoplasmic function.

Overall, N protein of SARS-CoV-2 is a unique protein harbouring three NLS and two NES signals. This potentially allows SARS-CoV-2 to use it as system of dynamically transporting

3.8. Identification of two glutamine-rich patches as Q-patches in the N protein

As hinted by disordered region prediction (**Fig. S2**) for a glutamine-rich stretch present in the N protein of SARS-CoVs as ^{239}QQQQGQ^{244} as depicted in protein sequence alignment (**Fig. 2**). However, the manual inspection of the protein alignment, we found two large patches of glutamine rich regions, which we named as Q-patches (marked by * in **Fig. 2**). These two Q-patches – Q-patch1 and Q-patch2 are present in 266-306 and 361-418 residue ranges, respectively (**Table 6**). The Q-patch1 has a total of 20 glutamines (marked by * in **Fig. 2**) mapped with eight glutamines are SARS-CoV-specific as Q229, Q239, Q240, Q241, Q289, Q294, Q303 and Q306 (red stars in **Fig. 2**), whereas six glutamines are only found in other viruses as R227Q, T245Q, K249Q, del255Q, T271Q, and H300Q (blue stars in **Fig. 2**). Additionally, three glutamines are conserved in all viruses as Q260, Q272 and Q281 (black stars in **Fig. 2**), whereas Q242 and Q244 as conserved in all SARS-CoVs along with some other viruses (maroon stars in **Fig. 2**). Interestingly, SARS-CoV-2 lost one conserved glutamine and it is replaced by alanine as A267Q (green star in **Fig. 2**). Q-patch2 is mapped from the end of the motif7 at the position 361 to almost end of the C-terminal end at the position 418 and this patch harbours 16 glutamines (marked by * in **Fig. 2**) Out of 16 glutamines, with seven glutamines are SARS-CoV2 -specific as Q380, Q384, Q386, Q406, Q408, Q409 and Q418 (red stars in **Fig. 2**), whereas eight are present only in other viruses, like K361Q, P368Q, K370Q, K373Q, K375Q, D377Q, K388Q, and K405Q (blue stars in **Fig. 2**) and one glutamine is present in all SARS-CoVs along with some other viruses as Q389 (maroon star in **Fig. 2**).

All in all, two patches of glutamine-rich regions are deciphered during our sequence analyses of N proteins SARS-CoV-2 and related viruses.

3.9. Small molecule docking at the conserved motifs of SARS-CoV-2 N protein

Previously, structure-based computer modelling leveraged the drug discovery process for the development of antiviral drugs of several viruses including hepatitis C virus (HCV[31]), hepatitis delta virus [32], Ebola virus [33] and Zika virus [34]. For the identification of potential drug targets against the seven conserved motifs of N protein, we employed the screening of the FDA approved drugs using molecular docking method.

As the protein pockets of similar shapes can bind to diverse drugs with different chemical properties [35]. We identified four drugs namely, adapalene, naldemedine, dihydroergotamine and midostaurin binding to the seven conserved motifs of N protein (**Supplementary Table 2**). SARS-CoV-specific motif4 of N protein showed the highest affinity (-11.9

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Kcal/Mol) to dihydroergotamine (**Fig. 6**). Adapalene exhibited the binding affinity towards motif1 and motif2 of N protein (**Fig. 6**). We further identified 14 antiviral top antiviral FDA-approved drugs with affinity score lower than -8.0 Kcal/Mol, which can potentially target motifs of the nucleocapsid protein of SARS-CoV-2 (**Table 7**). This list has ten drugs approved against HIV as abacavir, darunavir, delavirdine, dolutegravir, elvitegravir, indinavir, nelfinavir raltegravir, rilpivirine, and tipranavir (**Table 7**). In addition, we deduced three drugs, which are used against HCVs as dasabuvir, boceprevir and sofosbuvir, and trifluridine (**Table 7**) as a drug against herpes simplex viruses (HSV). All these drugs belongs to six drug classes namely, integrase inhibitors, non-nucleoside reverse transcriptase inhibitors (NNRTIs), nucleoside reverse transcriptase inhibitors (NRTIs), 5-substituted 2-deoxyuridines, HCV NS5B inhibitor + HCV NS5A inhibitor and protease inhibitors [36].

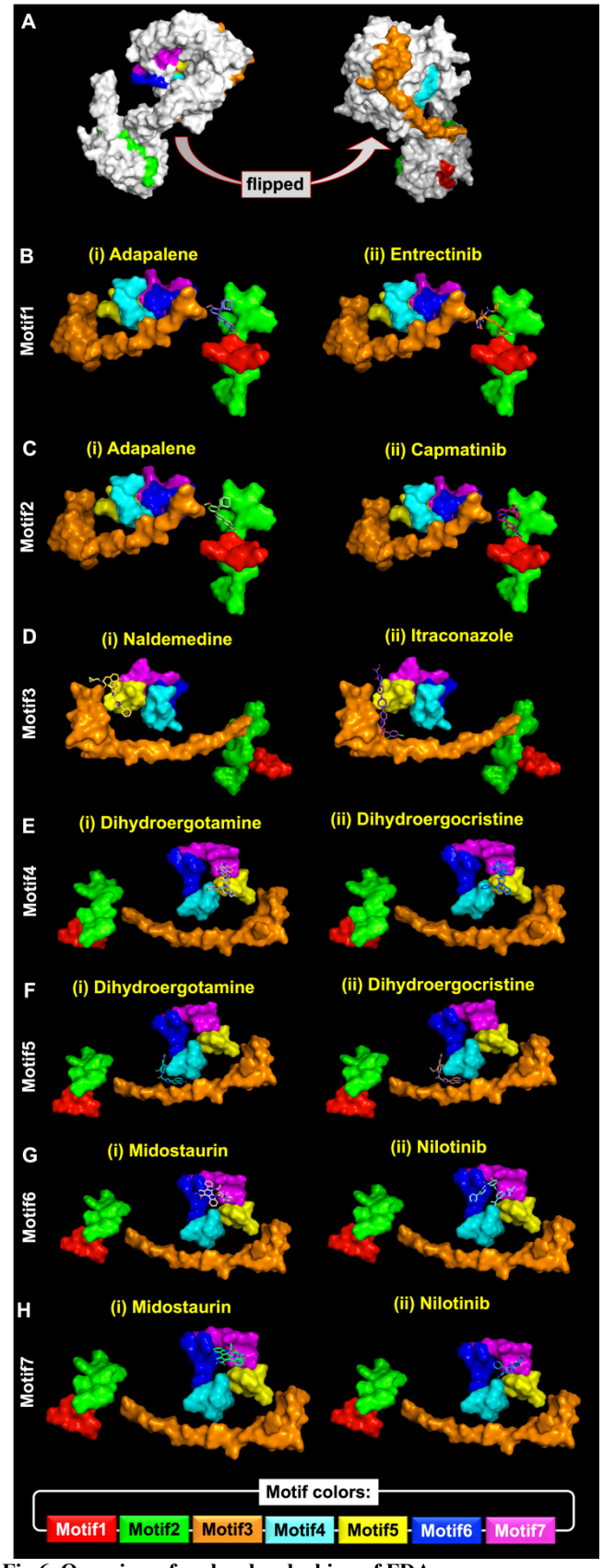

Fig 6. Overview of molecular docking of FDA approved drugs against seven identified motifs of SARS-CoV-2 N protein.

 doi:10.20944/preprints202005.0413.v1

Integrase inhibitors is an important class of drug as it targets virus integrase to inhibit the integration of viral DNA into human chromosomes. NNRTI binds directly to virus reverse transcriptase and inhibits DNA synthesis. Mechanism of each drug action in listed in the **Table 7**.

When we compared motif-wise antiviral drug targets, we found that motif2 can be targeted by 13 different drugs, whereas motif1 and motif2 by four each, motif5 by 3 drugs and motif6 and motif7 can be targets of 2 antiviral drugs each (**Table 7**). These hint for repurposing of a set of antiviral drugs are possible against COVID-19 and it requires further investigation in this direction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

4. Discussion

COVID-19 caused by SARS-CoV-2, is a major pandemic in 100 years. It has challenged our global movements by locking down the human population at homes for the hunt of safety against this virus. SARS-CoV-2 is the seventh CoVs that infected humans via with an intermediate host of zoonotic origins. It is well known that CoVs e.g., SARS, MERS and the recent outbreak of SARS-CoV-2 pandemic has caused great loss of life as well as economy time and again. Various strategies are being used to develop drugs against cell surface viral proteins [37]. Neutralizing antibodies bind to the surface proteins on viruses to prevent entry to the host cells, but the frequent mutation in the coat proteins can abolish the antibody-mediated immunity [37]. As an alternative, N protein can be used an antigen for early diagnosis and development of vaccines for many viruses due to its conserved gene sequence. In particular, we focus on identifying motifs in N protein in SARS-CoV-2 which are conserved across various CoVs.

N protein is mapped in the 3' end of SARS-CoV-2 genome from 28.274 bp to 29,533 bp with total ORF size of 1280 bp and is maintained in Indian strains like EPI_ISL_426414, EPI_ISL_426415, EPI_ISL_413523 and EPI_ISL_426179 and also in 43 bat-CoVs genomes (**Fig. 1**). Using a set representative sequences, we have illustrated seven motifs present in several CoVs mapped into different region of the protein N namely motif1-motif7 (**Figs. 1-3, Table 2**) The 12 amino acid long motif2 (^{106}PRWYFYYLGTGP^{117}) is conserved in all representative sequences and hence it is *coronaviridae*-specific (**Fig. 4**), whereas motif4 is clearly SARS-CoVs-specific (**Fig. 5A**).

Utilizing IEDB we have generated the largest epitome inventory of any SARS-CoV with 67 epitopes of seven motifs and 63 are deduced from SARS-CoV strains (**Tables 3, 5**). Given that N proteins of SARS-CoV-2 and SARS CoV are nearly identical with 99% sequence identities and 94% sequence similarities (**Fig. 2**).Being nearly identical proteins, this epitome inventory is suitable for testing against SARS-CoV-2.

We have identified three nuclear localization signals (NLSs) and two leucine-rich nuclear export signals (NESs) in the N proteins of different SARS-CoVs (**Figs 1-2 and 5**).

Previously, it has been shown that the N protein was localized mainly in cytoplasm SARS-CoV infected Vero E6 cells [38] suggesting a strong NES may be present in this protein. Timani *et al*. showed that the region (^{220}LALLLLDRLNRL^{231}) of the N protein in SARS-CoV was a functional NES [39]. Interestingly, the EGFP tagged NTD also showed cytoplasmic localization of the protein [39] and authors hinted for the presence of an additional NES in the NTD of the protein. In this study, we have identified NES1 in the NTD region, conserved in several SARS-CoVs (Fig. 2).

These nuclear localization and export signals are conserved in the different SARS-CoVs (**Fig. 2**). It collaborates that the N protein is involved in the dynamic nuclear–cytoplasmic trafficking. This trafficking of proteins controls many cellular processes, including gene expression, signal transduction, cell differentiation, and immune response. Utilizing of CRM1 using two NES by SARS-CoVs reflects that CoVs are capable of mimicking and exploiting for the conserved and constitute mechanism for cellular protein. There are some well-established viral examples of nuclear exports, exploited by matrix M1 protein of influenza A virus [40, 41], E7 oncoprotein of human papillomavirus [40] and REV protein of HIV-1 [41]. Presence of both

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

NLS and NES is also known for two viral proteins, VP1 and VP3 from the chicken anemia virus and using these signals and these proteins regulate the VP2 shuttling in cells [42,43].
Altogether, it is clear that with exploitation of these NLSs and NESs from the N protein become capable of shuttling between the cytoplasm and the nucleus, during the SARS CoV life cycle and plays an important role in the SARS CoV replication, assembly, and SARS CoV budding.

We have identified two glutamate-rich patches in N protein of SARS-CoV2 as Q-patch1 (20 Qs) and Q-patch2 (16 Qs). Q-patch1 also contents GKGQQQQGQ, which is conserved in N proteins in 1727 genomes of Bat CoVs and SARS-CoVs [44]. Repeats of Q and Q-rich regions are very well described in eukaryotic and viral genomes, which interfere with autophagy by causing viral proteins aggregation [44]. There are several examples of Q-rich regions assisting in controlling virus replication like bovine leukemia virus infection [45].
For RNA viruses like CoVs, these long patches of Q-rich may play instrumental roles in genome replications and environmental sensing [46]. Along with other RNA viruses including MERS and SARS-CoVs, Q-rich patches are required dsRNA folding domains near the 5′ end of these genomes as reported for SARS-CoVs [47] and Flavivirus [48]. Additionally, there are some other possibilities in different viruses as like dsRNA intermediate formation, roles in the interferon response and viral interfere with the OAS/RNaseL system as discussed recently [44].
Such longer patches of Q-rich regions are not explained previously in RNA viruses specifically *coronaviridae*. It is clear that coupled by extended motif4 (as ^{219}LALLLLDR^{226}, **Table 4**) and these two Q-rich patches, it is clear that N protein of SARS-CoV-2 has very high aggregation propensity in the host cells. Presence of Q-rich patches are known to have interference with autophagy [49] and their RNA binding abilities are known in several model organisms [44]. The Q-rich regions are often associated with human diseases like Huntington's disease and Huntington's disease-like 2 [44] and other neurological disorders [49]. In these diseases, The Q-rich patches targeting are potentially provide therapeutics solutions against these diseases [50]. During this study, we have found sequence stretches of
N protein, which are capable of creating higher aggregation propensity in the human cells. Hence, we recommend targeting these patches such as extended motif4, Q-patch1 and Q-patch2. Additional stretches are also critical in aggregation formation (**Table 4**). However, a careful mutational studies will be required to full-proof these findings. There are some studies on other viruses which have reflected that N proteins are potentially forming the aggregations like rhabdovirus uses N proteins to form aggregates in Cho cells and bacteria [51]. This study further demonstrated that osmolytes and a chaperone like phosphoprotein (P) maintain the N protein in a correctly folded form; however, authors did not report about the sequence patches in the N protein that might facilitate this aggregation [51]. The presence of amyloid patches in N protein might be one of the contributing reasons for this aggregation. All-in-all, amyloidogenic stretches in the nucleocapsid protein of SARS-nCoV-2 may lead to aggregation which may induce an overwhelming inflammatory stress in lung epithelial cells leading to cytotoxicity. During the pandemic of COVID-19 the biggest rush is towards the repurposing existing drugs because drugs of known safety profiles are keys to rapid deployments after initial success at level of the computational docking experiments [52]. Hence, we have also performed this strategy and explored potential antiviral drugs targeted against identified seven N protein motifs of SARS-CoV-2 and we

 doi:10.20944/preprints202005.0413.v1

have identified 14 antiviral drugs from drugbanks (**Table 7**).

Overall, we have characterized N protein of SARS-CoV-2 during COVID-19 pandemic using genomic data, protein sequence, structure, epitopes and docking experiment to provide quick suggestions for further studies. We are sure this work will be setting of further platforms for the characterization of N proteins from different viruses, specifically CoVs.

References:

[1] P.C. Woo, S.K. Lau, C.S. Lam, C.C. Lau, A.K. Tsang, J.H. Lau, R. Bai, J.L. Teng, C.C. Tsang, M. Wang, B.J. Zheng, K.H. Chan, K.Y. Yuen, Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus, J Virol 86 (2012) 3995-4008. 10.1128/JVI.06540-11.
[2] Y.Z. Zhang, E.C. Holmes, A Genomic Perspective on the Origin and Emergence of SARS-CoV-2, Cell (2020). 10.1016/j.cell.2020.03.035.
[3] A. Wu, Y. Peng, B. Huang, X. Ding, X. Wang, P. Niu, J. Meng, Z. Zhu, Z. Zhang, J. Wang, J. Sheng, L. Quan, Z. Xia, W. Tan, G. Cheng, T. Jiang, Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China, Cell Host Microbe 27 (2020) 325-328. 10.1016/j.chom.2020.02.001.
[4] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res 25 (1997) 3389-3402.
[5] A. Conesa, S. Gotz, Blast2GO: A comprehensive suite for functional analysis in plant genomics, Int J Plant Genomics 2008 (2008) 619832. 10.1155/2008/619832.
[6] C.S. Yu, C.W. Cheng, W.C. Su, K.C. Chang, S.W. Huang, J.K. Hwang, C.H. Lu, CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation, PLoS One 9 (2014) e99368. 10.1371/journal.pone.0099368.
[7] C.M. Lee, G.P. Barber, J. Casper, H. Clawson, M. Diekhans, J.N. Gonzalez, A.S. Hinrichs, B.T. Lee, L.R. Nassar, C.C. Powell, B.J. Raney, K.R. Rosenbloom, D. Schmelter, M.L. Speir, A.S. Zweig, D. Haussler, M. Haeussler, R.M. Kuhn, W.J. Kent, UCSC Genome Browser enters 20th year, Nucleic Acids Res 48 (2020) D756-D761. 10.1093/nar/gkz1012.
[8] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, D. Haussler, W. Miller, Aligning multiple genomic sequences with the threaded blockset aligner, Genome Res 14 (2004) 708-715. 10.1101/gr.1933104.
[9] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (2004) 113. 10.1186/1471-2105-5-113.
[10] X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server, Nucleic Acids Res 42 (2014) W320-324. 10.1093/nar/gku316.
[11] A.M. Waterhouse, J.B. Procter, D.M. Martin, M. Clamp, G.J. Barton, Jalview Version 2--a multiple sequence alignment editor and analysis workbench, Bioinformatics 25 (2009) 1189-1191. 10.1093/bioinformatics/btp033.
[12] J. Yang, Y. Zhang, I-TASSER server: new development for protein structure and function predictions, Nucleic Acids Res 43 (2015) W174-181. 10.1093/nar/gkv342.
[13] T. la Cour, L. Kiemer, A. Molgaard, R. Gupta, K. Skriver, S. Brunak, Analysis and prediction of leucine-rich nuclear export signals, Protein Eng Des Sel 17 (2004) 527-536. 10.1093/protein/gzh062.
[14] D.T. Jones, D. Cozzetto, DISOPRED3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics 31 (2015) 857-863. 10.1093/bioinformatics/btu744.
[15] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res 14 (2004) 1188-1190. 10.1101/gr.849004.
[16] A. Parveen, R. Kumar, R. Tandon, S. Khurana, C. Goswami, A. Kumar, Mutational hotspots of HSP47 and its

potential role in cancer and bone-disorders, Genomics 112 (2020) 552-566. 10.1016/j.ygeno.2019.04.007.
[17] A. Kumar, A. Bhandari, S.J. Sarde, C. Goswami, Genetic variants and evolutionary analyses of heparin cofactor II, Immunobiology 219 (2014) 713-728. 10.1016/j.imbio.2014.05.003.
[18] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol Biol Evol 4 (1987) 406-425. 10.1093/oxfordjournals.molbev.a040454.
[19] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms, Mol Biol Evol 35 (2018) 1547-1549. 10.1093/molbev/msy096.
[20] R. Linding, J. Schymkowitz, F. Rousseau, F. Diella, L. Serrano, A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins, Journal of molecular biology 342 (2004) 345-353. 10.1016/j.jmb.2004.06.088.
[21] O. Conchillo-Sole, N.S. de Groot, F.X. Aviles, J. Vendrell, X. Daura, S. Ventura, AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides, BMC bioinformatics 8 (2007) 65. 10.1186/1471-2105-8-65.
[22] S.O. Garbuzynskiy, M.Y. Lobanov, O.V. Galzitskaya, FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence, Bioinformatics (Oxford, England) 26 (2010) 326-332. 10.1093/bioinformatics/btp691.
[23] K.K. Frousios, V.A. Iconomidou, C.M. Karletidi, S.J. Hamodrakas, Amyloidogenic determinants are usually not buried, BMC structural biology 9 (2009) 44. 10.1186/1472-6807-9-44.
[24] F. Chiti, C.M. Dobson, Protein misfolding, functional amyloid, and human disease, Annual review of biochemistry 75 (2006) 333-366. 10.1146/annurev.biochem.75.101304.123901.
[25] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res 46 (2018) D1074-D1082. 10.1093/nar/gkx1037.
[26] C.S. Shi, N.R. Nabar, N.N. Huang, J.H. Kehrl, SARS-Coronavirus Open Reading Frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes, Cell Death Discov 5 (2019) 101. 10.1038/s41420-019-0181-7.
[27] A. Currais, W. Fischer, P. Maher, D. Schubert, Intraneuronal protein aggregation as a trigger for inflammation and neurodegeneration in the aging brain, FASEB J 31 (2017) 5-10. 10.1096/fj.201601184.
[28] X.D. Liu, S. Ko, Y. Xu, E.A. Fattah, Q. Xiang, C. Jagannath, T. Ishii, M. Komatsu, N.T. Eissa, Transient aggregation of ubiquitinated proteins is a cytosolic unfolded protein response to inflammation and endoplasmic reticulum stress, J Biol Chem 287 (2012) 19687-19698. 10.1074/jbc.M112.350934.
[29] H.Y. Fung, S.C. Fu, Y.M. Chook, Nuclear export receptor CRM1 recognizes diverse conformations in nuclear export signals, Elife 6 (2017). 10.7554/eLife.23961.
[30] P. Goyal, D. Pandey, A. Behring, W. Siess, Inhibition of nuclear import of LIMK2 in endothelial cells by protein kinase C-dependent phosphorylation at Ser-283, J Biol Chem 280 (2005) 27569-27577. 10.1074/jbc.M504448200.
[31] J.P. Yang, D. Zhou, F. Wong-Staal, Screening of small-molecule compounds as inhibitors of HCV entry, Methods Mol Biol 510 (2009) 295-304. 10.1007/978-1-59745-394-3_22.
[32] S. Singh, S.K. Gupta, A. Nischal, S. Khattri, R. Nath, K.K. Pant, P.K. Seth,

Identification and characterization of novel small-molecule inhibitors against hepatitis delta virus replication by using docking strategies, Hepat Mon 11 (2011) 803-809. 10.5812/kowsar.1735143X.737.
[33] J. Schuler, M.L. Hudson, D. Schwartz, R. Samudrala, A Systematic Review of Computational Drug Discovery, Development, and Repurposing for Ebola Virus Disease Treatment, Molecules 22 (2017). 10.3390/molecules22101777.
[34] L. Wang, R. Liang, Y. Gao, Y. Li, X. Deng, R. Xiang, Y. Zhang, T. Ying, S. Jiang, F. Yu, Development of Small-Molecule Inhibitors Against Zika Virus Infection, Front Microbiol 10 (2019) 2725. 10.3389/fmicb.2019.02725.
[35] M. Gao, J. Skolnick, A comprehensive survey of small-molecule binding pockets in proteins, PLoS Comput Biol 9 (2013) e1003302. 10.1371/journal.pcbi.1003302.
[36] E. De Clercq, G. Li, Approved Antiviral Drugs over the Past 50 Years, Clin Microbiol Rev 29 (2016) 695-747. 10.1128/CMR.00102-15.
[37] S. Belouzard, J.K. Millet, B.N. Licitra, G.R. Whittaker, Mechanisms of coronavirus cell entry mediated by the viral spike protein, Viruses 4 (2012) 1011-1033. 10.3390/v4061011.
[38] M.S. Chang, Y.T. Lu, S.T. Ho, C.C. Wu, T.Y. Wei, C.J. Chen, Y.T. Hsu, P.C. Chu, C.H. Chen, J.M. Chu, Y.L. Jan, C.C. Hung, C.C. Fan, Y.C. Yang, Antibody detection of SARS-CoV spike and nucleocapsid protein, Biochem Biophys Res Commun 314 (2004) 931-936. 10.1016/j.bbrc.2003.12.195.
[39] K.A. Timani, Q. Liao, L. Ye, Y. Zeng, J. Liu, Y. Zheng, L. Ye, X. Yang, K. Lingbao, J. Gao, Y. Zhu, Nuclear/nucleolar localization properties of C-terminal nucleocapsid protein of SARS coronavirus, Virus Res 114 (2005) 23-34. 10.1016/j.virusres.2005.05.007.
[40] Z. Onder, V. Chang, J. Moroianu, Nuclear export of cutaneous HPV8 E7 oncoprotein is mediated by a leucine-rich nuclear export signal via a CRM1 pathway, Virology 474 (2015) 28-33. 10.1016/j.virol.2014.10.012.
[41] R.T. Behrens, M. Aligeti, G.M. Pocock, C.A. Higgins, N.M. Sherer, Nuclear Export Signal Masking Regulates HIV-1 Rev Trafficking and Viral RNA Nuclear Export, J Virol 91 (2017). 10.1128/JVI.02107-16.
[42] C. Feng, Y. Liang, J.G. Teodoro, The Role of Apoptin in Chicken Anemia Virus Replication, Pathogens 9 (2020). 10.3390/pathogens9040294.
[43] J.H. Cheng, G.H. Lai, Y.Y. Lien, F.C. Sun, S.L. Hsu, P.C. Chuang, M.S. Lee, Identification of nuclear localization signal and nuclear export signal of VP1 from the chicken anemia virus and effects on VP2 shuttling in cells, Virol J 16 (2019) 45. 10.1186/s12985-019-1153-5.
[44] C.H. Schein, Polyglutamine Repeats in Viruses, Mol Neurobiol 56 (2019) 3664-3675. 10.1007/s12035-018-1269-4.
[45] M. Reichert, Proteome analysis of sheep B lymphocytes in the course of bovine leukemia virus-induced leukemia, Exp Biol Med (Maywood) 242 (2017) 1363-1375. 10.1177/1535370217705864.
[46] L. Atanesyan, V. Gunther, B. Dichtl, O. Georgiev, W. Schaffner, Polyglutamine tracts as modulators of transcriptional activation from yeast to mammals, Biol Chem 393 (2012) 63-70. 10.1515/BC-2011-252.
[47] R. Madhugiri, N. Karl, D. Petersen, K. Lamkiewicz, M. Fricke, U. Wend, R. Scheuer, M. Marz, J. Ziebuhr, Structural and functional conservation of cis-acting RNA elements in coronavirus 5'-terminal genome regions, Virology 517 (2018) 44-55. 10.1016/j.virol.2017.11.025.
[48] M.A. Garcia-Blanco, S.G. Vasudevan, S.S. Bradrick, C. Nicchitta, Flavivirus RNA transactions from viral entry to genome replication, Antiviral Res 134 (2016) 244-249. 10.1016/j.antiviral.2016.09.010.
[49] A. Ashkenazi, C.F. Bento, T. Ricketts, M. Vicinanza, F. Siddiqi, M. Pavel, F. Squitieri, M.C. Hardenberg, S. Imarisio, F.M. Menzies, D.C. Rubinsztein,

Polyglutamine tracts regulate beclin 1-dependent autophagy, Nature 545 (2017) 108-111. 10.1038/nature22078.
[50] A.M. Monteys, S.A. Ebanks, M.S. Keiser, B.L. Davidson, CRISPR/Cas9 Editing of the Mutant Huntingtin Allele In Vitro and In Vivo, Mol Ther 25 (2017) 12-23. 10.1016/j.ymthe.2016.11.010.
[51] A. Majumder, S. Basak, T. Raha, S.P. Chowdhury, D. Chattopadhyay, S. Roy, Effect of osmolytes and chaperone-like action of P-protein on folding of nucleocapsid protein of Chandipura virus, J Biol Chem 276 (2001) 30948-30955. 10.1074/jbc.M011705200.
[52] J.M. Parks, J.C. Smith, How to Discover Antiviral Drugs Quickly, N Engl J Med (2020). 10.1056/NEJMcibr2007042.

 doi:10.20944/preprints202005.0413.v1

Author Contributions

Conceived and designed the experiments: AK, and MN

Supervision: AK, PG and MN

Analyzed the data: AK, AP, NK, SB, VK, CC, JS, PS, AA, AP, PG and MN
Contributed to the writing of the manuscript: AK, AP, NK, JS, AP, PG and MN
All authors have approved final version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgements: AK is a recipient of Ramalingaswami Re-Retry Faculty Fellowship (Grant; BT/RLF/Re-entry/38/2017) from Department of Biotechnology (DBT), Government of India (GOI). JS is a recipient of DBT-BioCARe Women Scientist award from Department of Biotechnology (DBT), Government of India (GOI). AK also thanks BioBam for their supports to COVID-19 Research Project in India with OMICSBOX. Authors also acknowledge GISAID for providing access of genomic data used in this study.

 doi:10.20944/preprints202005.0413.v1

Table legends

Table 1. Summary of re-annotation of selected viral nucleocapsid (N) proteins used as representative sequences in this study

Table 2. Overview of 7 motifs deduced from sequence analyses of nucleocapsid (N) proteins from CoVs and related viruses.

Table 3. Summary of experimentally validated epitopes flanking N protein motifs, except motif2. This data is derived from Immune Epitope Database and Analysis Resource (IEDB, Web: http://www.iedb.org/)

Table 4. Predicted amyloidogenic stretches in the nucleocapsid protein of SARS-CoV-2

Table 5. Overview of experimentally measured immune epitopes, mapped to the motif2 of N protein and flanking regions. This data is derived from IEDB, (Web: http://www.iedb.org/).

Table 6. Summary of two extended patches of glutamine (Q) residues as Q-patches.

Table 7. List of top antiviral FDA-approved drugs (affinity lower than -8 Kcal/Mol), which can potentially target identified motifs of the nucleocapsid protein of SARS-CoV-2.

Supplementary file 1 – it contains entire supplementary data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

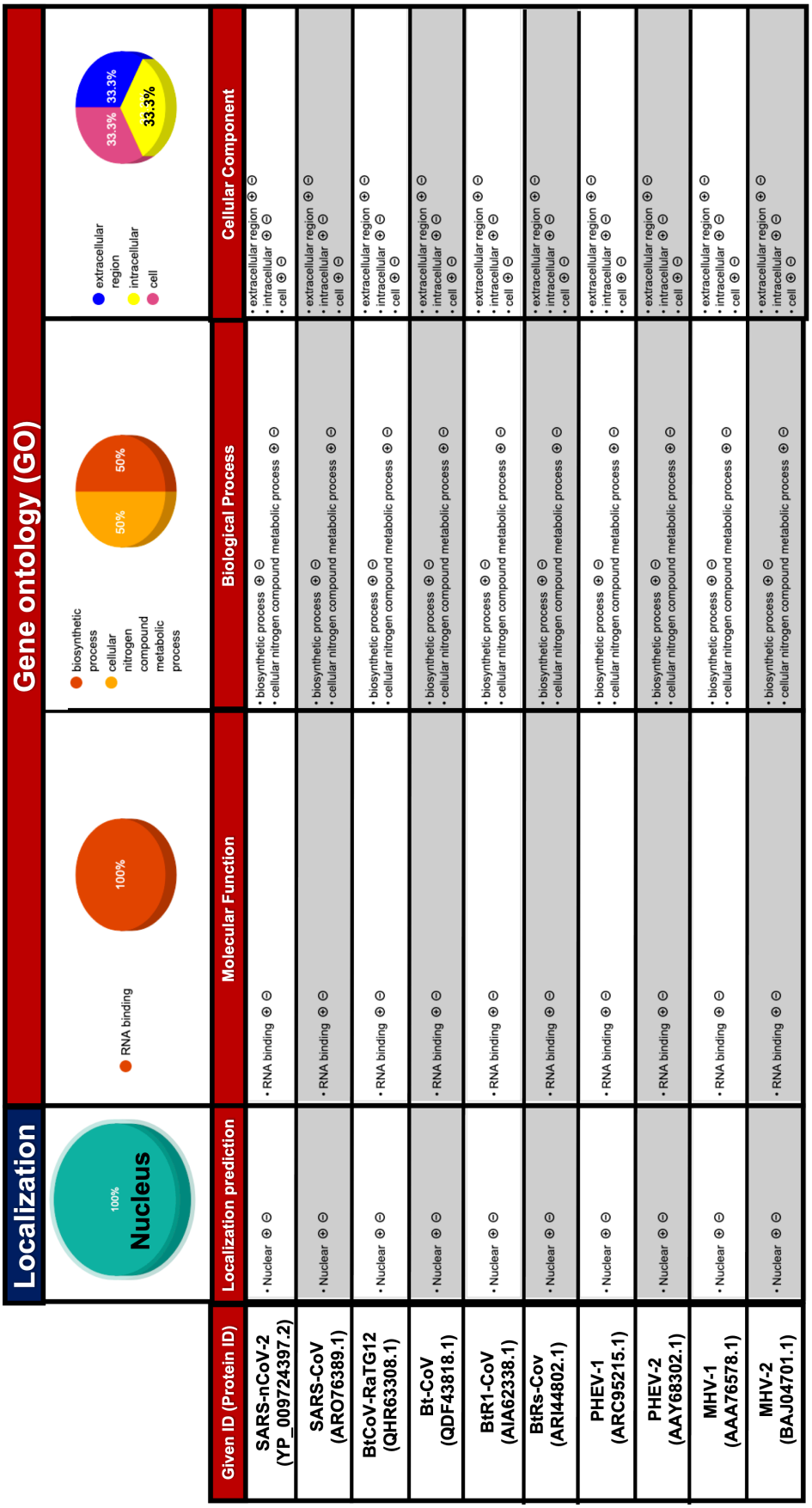

Figure S1. Summary of GO annotation derived from CELLO2GO [6] for representative Protein N from SARS-CoVs and other viruses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

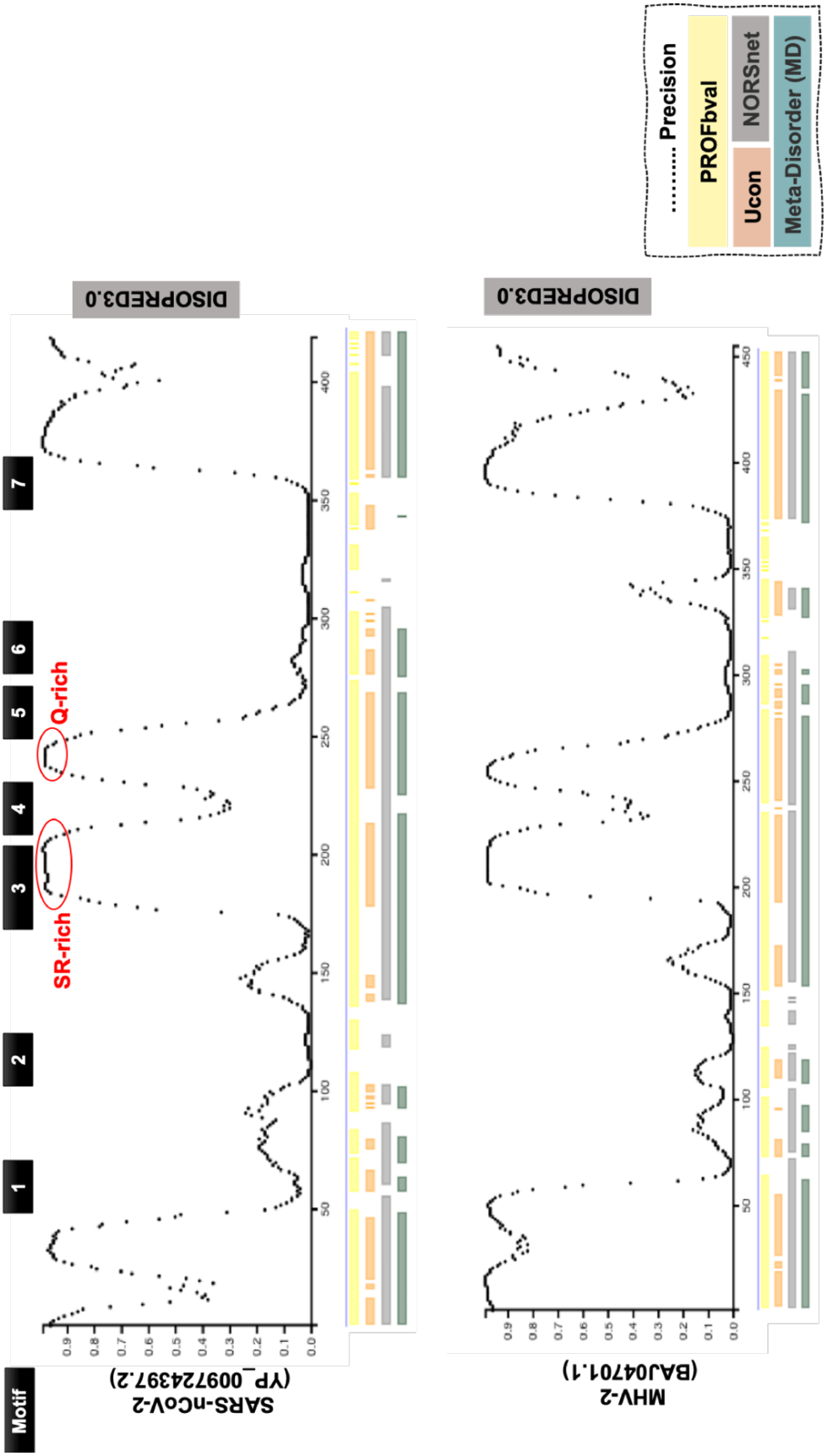

Figure S2. Overview of disordered regions in Protein N from SARS-CoV-2 and murine hepatitis virus (MHV-2).

 doi:10.20944/preprints202005.0413.v1

Table S1. Summary of Genomes of Bat CoV, used in this study.

Table S2. Summary of all FDA-approved drugs docked against seven motifs of the nucleocapsid protein of SARS-CoV-2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Table 1. Summary of re-annotation of selected viral Nucleocapsid (N) proteins used as representative sequences in this study

Name Given	Genbank ID	Virus species	Protein Length	e-Value	Mean Similarities	#GO	GO IDs	GO Names	InterPro IDs	InterPro GO IDs & Name
SARS-nCoV-2	YP_009724397.2	Severe acute respiratory syndrome coronavirus 2	419	0	97.06	4	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus	IPR001218 (PFAM); IPR001218 (PIRSF); 9x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037195 (SUPERFAMILY); IPR037179 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid
SARS-CoV	ARO76389.1	Severe acute respiratory syndrome-related coronavirus	422	0	99.03	5	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177; C:GO:0044220	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus; C:host cell perinuclear region of cytoplasm	IPR001218 (PIRSF); IPR001218 (PFAM); 9x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037179 (SUPERFAMILY); IPR037195 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid
BtCoV-RaTG12	QHR63308.1	Bat coronavirus RaTG13	419	0	98.93				IPR001218 (PFAM); IPR001218 (PIRSF); 9x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037179 (SUPERFAMILY); IPR037195 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid
Bt-CoV	QDF43818.1	Coronavirus BtRl-BetaCoV/SC2018	421	0	99.12	4	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus	IPR001218 (PIRSF); IPR001218 (PFAM); 10x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037195 (SUPERFAMILY); IPR037179 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid
BtR1-CoV	AIA62338.1	BtRs-BetaCoV/YN2013	422	0	99.05	5	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177; C:GO:0044220	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus; C:host cell perinuclear region of cytoplasm	IPR001218 (PFAM); IPR001218 (PIRSF); 9x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037179 (SUPERFAMILY); IPR037195 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid
BtRs-Cov	ARI44802.1	Bat coronavirus	421	0	98.26	5	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177; C:GO:0044220	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus;	IPR001218 (PIRSF); IPR001218 (PFAM); 9x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037179	C:GO:0019013 C:viral nucleocapsid

 doi:10.20944/preprints202005.0413.v1

								C:host cell perinuclear region of cytoplasm	(SUPERFAMILY); IPR037195 (SUPERFAMILY)	
PHEV-1	ARC95215.1	Porcine hemagglutinating encephalomyelitis virus	449	0	98.39	4	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus	IPR001218 (PIRSF); IPR001218 (PFAM); 8x mobidb-lite (MOBIDB_LITE); (IPR001218 (HAMAP); IPR037195 (SUPERFAMILY); IPR037179 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid
PHEV-2	AAY68302.1	Porcine hemagglutinating encephalomyelitis virus	449	0	97.56	4	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus	IPR001218 (PFAM); IPR001218 (PIRSF); 8x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037179 (SUPERFAMILY); IPR037195 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid
MHV-1	AAA76578.1	Murine hepatitis virus	451	0	96.34	4	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus	IPR001218 (PIRSF); IPR001218 (PFAM); 7x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037195 (SUPERFAMILY); IPR037179 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid
MHV-2	BAJ04701.1	Murine hepatitis virus	455	0	96.81	4	F:GO:0003723; C:GO:0019013; C:GO:0044172; C:GO:0044177	F:RNA binding; C:viral nucleocapsid; C:host cell endoplasmic reticulum-Golgi intermediate compartment; C:host cell Golgi apparatus	IPR001218 (PFAM); IPR001218 (PIRSF); 7x mobidb-lite (MOBIDB_LITE); IPR001218 (HAMAP); IPR037195 (SUPERFAMILY); IPR037179 (SUPERFAMILY)	C:GO:0019013 C:viral nucleocapsid

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Table 2. Overview of 7 motifs deduced from sequence analyses of Nucleocapsid (N) protein and related viruses

Motifs	Sequence motifs*, #, $, %	Residue positions*	Location in protein organization	Secondary Structural Locations	Comments
Motif1	^{69}GQGVPI^{75}	69-75	N-terminal RNA-binding domain (NTD)	Loop between β-sheets β1-β2	
Motif2	^{106}PRWYFYYLGTGP^{117}	106-117	N-terminal RNA-binding domain (NTD)	β-sheet β3	Largest and it can Be target for various viruses
Motif3	^{176}SR[4]^{207}	176-207	At the end of NTD and Linker region (LKR)	Loop between β-sheet β4-helix Π1	
Motif4	^{221}LLLLDR^{226}	221-226	LKR	Loop between β-sheet β4-helix Π1	Specific for CoVs, potentially increasing aggregation propensity
Motif5	^{257} KPRQKR[ST]^{263}	257-263	C-terminal dimerization domain (CTD)	Loop between helices α2-α3	
Motif6	^{274}FG[KR]RGP^{281}	274-281	CTD	Loop between helices α3-α4	
Motif7	^{353}LN....AY.^{361}	353-361	CTD	Spanning β-sheets β5-β6	

*Amino acid numbering of Nucleocapsid (N) protein of SARS-CoV-2 is followed
- "[4]" indicates presence of SR minimum two times and maximum of seven times
$ - "[XY]" indicates presence of any of two given residues in the bracket
% - "." Indicates presence of any 20 amino acids

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Table 3. Overview of experimentally measured immune epitopes, mapped to the motif2 of N protein and flanking regions. This data is derived from Immune Epitope Database and Analysis Resource (IEDB, Web: http://www.iedb.org/).

Epitope Name	Epitope ID	Alignment with motif2	Starting Position	Ending Position	Antigen Accession	Species	Experimental evidences
Epitope1	13278	**------ELSPRWYFY----------**	104	112	P59595.1	SARS-CoV Tor2	MHC HLA-A*26:01 (Positive-High), HLA-A*29:02 (Positive-High), and HLA-A*30:02 (Positive-Intermediate).
Epitope2	13279	**------ELSPRWYFYY----------**	104	113	P59595.1	SARS-CoV Tor2	MHC HLA-A*01:01 (Positive-Intermediate), HLA-A*26:01 (Positive-High), HLA-A*29:02 (Positive-High), and HLA-A*30:02 (Positive-Intermediate).
Epitope3	20633	**----------FYYLGTGPEASLPYG---**	100	114	P59595.1	SARS-CoV	ELISPOT IFNg release - Positive - immunized mice (BALB/c)
Epitope4	18487	**--GKMKELSPRWYFYYL----------**	111	125	P59595.1	SARS-CoV	T Cell Assays (Mus musculus C57BL/6 - H2-b class II) a) 3H-thymidine proliferation – Positive b) Antibody help - Positive
Epitope5	38698	**-----KELSPRWYFY----------**	123	131	P03417.1	Murine hepatitis virus strain JHM	MHC HLA-B*40:02 (Positive-Intermediate), HLA-B*44:02 (Positive-High), HLA-B*44:03 (Positive-High), and HLA-B*45:01 (Positive-Low).
Epitope6	30515	**---KMKELSPRWYFYYLG----------**	103	112	P59595.1	SARS-CoV Tor2	ELISA - IFNg release - immunized mice (BALB/c) - Positive ELISPOT - IFNg release – Human – HLA class II - Positive ICS - IFNg release – Human – HLA class II - Positive
Epitope7	32340	**--------LPRWYFYYL----------**	101	115	AAP13445.1	SARS-CoV Urbani	T Cell Assays using IFNg release is positive in mouse, HLA class II in human.
Epitope8	39576	**-------LSPRWYFYY----------**	105	113	P59595.1	SARS-CoV Tor2	MHC HLA-A*01:01 (Positive-High), HLA-A*11:01 (Positive-Low), HLA-A*23:01(Positive-Low), HLA-A*24:02(Positive-Low), HLA-A*29:02 (Positive-High), HLA-A*30:02(Positive-Intermediate) and HLA-A*31:01 **(Positive-Low)**
Epitope9	39577	**-------LSPRWYFYYLGTGPEASL------**	105	122	AAP49024.1	SARS-CoV Frankfurt 1	ELISPOT - IFNg release – Human – Positive
Epitope10	49278	**---------PRWYFYYLGTGPEAS-------**	107	121	P59595.1	SARS coronavirus	New Zealand White rabbits (Oryctolagus cuniculus)-western blot qualitative binding (Positive)
Epitope11	60242	**--------SPRWYFYYL----------**	106	114	P59595.1	SARS-CoV Tor2	MHC HLA-B*07:02 (Positive-High), HLA-B*51:01 (Positive-Low), HLA-B*53:01(Positive-Low), and HLA-B*54:01 (Positive-Low)
Epitope12	60243	**--------SPRWYFYYLGTGPEA--------**	106	120	AAP13445.1	SARS-CoV Urbani	Multiple experiments in immunized mice (BALB/c) ELISPOT IFNg release (Positive), IL-10 release (Positive), IL-2 release - MHC H2-d class II (Positive), and IL-4 release - MHC H2-d class II (Positive)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Epitope13	74683	**----------YLGTGPEASL------**	113	122	AAP13445.1	SARS-CoV Urbani	MHC HLA-A*02:01 (Positive-High) - two studies
Epitope14	74684	**----------YLGTGPEASLPYGANK**	113	128	P59595.1	SARS-CoV BJ01	ELISA qualitative binding (Positive) ELISPOT IFNg release (Positive).
Epitope15	125060	**GDGKMKELSPRWYFY----------**	98	112	AAP49024.1	SARS-CoV Frankfurt 1	Multiple experiments in immunized mice (BALB/c) ELISPOT IFNg release (Positive), IL-10 release (Positive), IL-2 release - MHC H2-d class II (Positive), and IL-4 release - MHC H2-d class II (Positive)
Epitope16	125158	**----MKELSPRWYFYYLGT----------**	102	116	AAP49024.1	SARS-CoV Frankfurt 1	Multiple experiments in immunized mice (BALB/c) ELISPOT IFNg release (Positive), IL-10 release (Positive), IL-2 release - MHC H2-d class II (Positive), and IL-4 release - MHC H2-d class II (Positive)
Epitope17	125322	**----------YFYYLGTGPEASL------**	110	122	AAP49024.1	SARS-CoV Frankfurt 1	Multiple experiments in immunized mice (BALB/c) ELISPOT IFNg release (Positive), IL-10 release (Positive), IL-2 release - MHC H2-d class II (Positive), and IL-4 release - MHC H2-d class II (Positive)
Epitope18	142171	**--GQRKELPERWFFYFLGTGPH---------**	81	100	AAB47502.1	Feline infectious peritonitis virus (strain KU-2)	ELISPOT IFNg release (Positive) - MHC H2 class I, ELISPOT IFNg release (Positive), and ELISPOT IFNg release (Positive-Low)
Epitope19	151508	**----------YLGTGPYAGA------**	130	139	CAA25497.1	Murine hepatitis virus strain JHM	Mice (C57BL/6) - western blot qualitative binding (Positive)
	Motif2	**---------PRWYFYYLGTGP----------**	106	117			

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Table 4. Predicted amyloidogenic stretches in the Nucleocapsid protein of SARS-CoV-2

S.No.	Sequence of the Stretches	Secondary Structure (PsiPred)	TANGO	AGGRESCAN	AmylPred	FoldAmyloid
1.	^{51}SWFTA^{55}	^{51}CCCCC^{55}	9.82	0.543	Yes	22.92
2.	**^{108}WYFYYL^{113}**	**^{108}EEEEEC^{113}**	**88.92**	**3.968**	**Yes**	**24.93**
3.	^{130}IIWVA^{134}	^{130}EEEEE^{134}	53.27	2.034	Yes	23.45
4.	^{156}AIVLQ^{160}	^{156}CCCCC^{160}	16.64	1.127	Yes	22.70
5.	**^{219}LALLLLDR^{226}**	**^{219}HHHHHHHH^{226}**	**62.81**	**2.454**	**Yes**	**23.11**
6.	^{313}AFFGM^{317}	^{313}HHHCC^{317}	0.59	3.65	Yes	22.43
7.	^{329}TWLTY^{333}	^{329}EEEEE^{333}	7.16	2.431	Yes	22.65
8.	^{350}VILLN^{354}	^{350}HHHHH^{354}	16.32	1.109	Yes	22.85
10.	^{392}VTLLP^{396}	^{392}CCCCCC^{396}				22.378

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Table 5. Summary of experimentally validated epitopes flanking N protein motifs, except motif2.

Epitope ID	Description (with motif underlined)	Start Position	End Position	Antigen Accession	Organism	MHC Restriction	Isotype	Assay Type	Effector Origin
motif1	^{69}GQGVPI^{75}								
17385	FPRGQGVPI	67	75	ABI96968.1	SARS-CoV				
20544	GKEELRFPRGQGVPI	61	75	P59595.1	SARS-CoV	H2-dII		EL	DEV
20545	GKEELRFPRGQGVPINTNSG	61	80	P59595.1	SARS-CoV				
23917	HGKEELRFPRGQGVPI	60	75	P59595.1	SARS-CoV BJ01				
39087	LRFPRGQGVPINTNS	65	79	P59595.1	SARS-CoV			WB	
53758	RFPRGQGVPINTNSGPDDQI	66	85	P59595.1	SARS-CoV BJ01				
125079	GQGVPINTNSGPDDQ	70	84	AAP49024.1	SARS-CoV Frankfurt 1	H2-dII		EL	STRE
151185	FQFAQGQGVPIA	81	92	CAA25497.1	MHV		IgG	WB	
Motif3	**SRGGSQASSRSSSRSRNSSRNSTPGSSRGTS**								
956	AEGSRGGSQA	174	183	P59595.1	SARS-CoV Tor2				
19650	GFYAEGSRGGSQASS	171	185	AAP13445.1	SARS-CoV Urbani			EL	**STRE**
22481	GSRGGSQASSRSSSR	176	190	AAP13445.1	SARS-CoV Urbani				
48067	PKGFYAEGSRGGSQASSR	169	186	P59595.1	SARS-CoV BJ01				
51485	QLPQGTTLPKGFYAEGSRGGSQ	161	182	P59595.1	SARS-CoV				
60379	SQASSRSS	181	188	P59595.1	SARS-CoV				
60380	SQASSRSSSR	181	190	P59595.1	SARS-CoV Tor2				
60669	SRGGSQASSRSSSRSR	177	192	P59595.1	SARS-CoV BJ01				
60749	SRNSTPGSSRGNSPARMA	195	212	P59595.1	SARS-CoV BJ01				
78665	GNSRNSTPGS	193	202	AAP33707.1	SARS-CoV Frankfurt 1		IgG	WB	
motif4	**^{221}LLLLDR^{226}**								
2431	ALALLLLDR	219	227	P59595.1	SARS-CoV				
2433	ALALLLLDRLNQLESKV	219	235	P59595.1	SARS-CoV				

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

19443	GETALALLLLDRLNQ	216	230	ABI96968.1	SARS-CoV				
34851	LALLLLDRL	220	228	P59595.1	SARS-CoV				
37515	LLLLDRLNQL	222	231	P59595.1	SARS-CoV				
41120	MASGGGETALALLLLDRL	211	228	P59595.1	SARS-CoV				
41121	MASGGGETALALLLLDRLNQ	211	230	P59595.1	SARS-CoV				
41122	MASGGGETALALLLLDRLNQLESKV	211	235	AAP13445.1	SARS-CoV				
193498	LLLLDRLNQ	222	230	NP_828858.1	SARS-CoV				
motif5	**^{257}KPRQKR[ST]^{263}**								
31329	KILNKPRQKRTPNK	266	279	P03417.1	MHEV				
31692	KKSAAEASKKPRQKRTA	249	265	P59595.1	SARS-CoV BJ01				
31693	KKSAAEASKKPRQKRTATKQYNVTQ	249	273	ABI96968.1	SARS-CoV				
58868	SKKPRQKRTATKQYNV	256	271	P59595.1	SARS-CoV BJ01				
92294	AGQPKQVTKQSAKEVRQKILNKPRQKRTP	249	277	P03417.1	MHV				
motif6	**^{274}FG[KR]RGP^{281}**								
64655	TKQYNVTQAFGRRGP	266	280	AAP13445.1	SARS-CoV Urbani				
71468	VTQAFGRRGPEQTQGNFGDQ	271	290	P59595.1	SARS-CoV				
75235	YNVTQAFGRRGPEQTQGNF	269	287	P59595.1	SARS-CoV BJ01				
motif7	**^{353}LN....AY.^{361}**								
7808	DDKDPQFKDNVILLNKHIDA	341	360	P59595.1	SARS-CoV				
27182	ILLNKHIDA	352	360	P59595.1	SARS-CoV	HLA-A*02:01		E	Cell Line / Clone
27183	ILLNKHIDAYKTFPP	352	366	P59595.1	SARS-CoV				
31115	KHIDAYKTFPPTEPK	356	370	AAP13445.1	SARS-CoV Urbani				
31116	KHIDAYKTFPPTEPKKDKKK	356	375	P59595.1	SARS-CoV				
37611	LLNKHIDAYKTFPPTEPK	353	370	P59595.1	SARS-CoV				
38249	LNKHIDAYKTFPPTEPK	354	370	P59595.1	SARS-CoV BJ01				
44501	NKHIDAYKTFPPTEP	355	369	P59595.1	SARS-CoV Tor2				

 doi:10.20944/preprints202005.0413.v1

48995	PQFKDNVILLNKHIDAYK	345	362	AAP13445.1	SARS-CoV				
50781	QFKDNVILLNKHIDAYK	346	362	P59595.1	SARS-CoV BJ01				
69035	VILLNKHIDAYKTFP	351	365	AAP13445.1	SARS-CoV Urbani	H2-dII, H2-kII, H2-bII, H2-IAd, HLA-DRB1*03:01, HLA-DRB1*15:01		E, I	STRE, DEV
125100	ILLNKHID	352	359	P59595.1	SARS-CoV				
985589	LLNKHIDAYKTFP	353	365	P59595.1	SARS-CoV	H2-dII, H2-kII, H2-bII, H2-IAd, HLA-DRB1*03:01, HLA-DRB1*15:01		E, I	STRE, DEV

MHV - Murine hepatitis virus strain JHM; SARS-CoV - Severe acute respiratory syndrome-related coronavirus; STRE- Short Term Restimulated; DEV - Direct Ex Vivo; EL – ELISPOT; WB-western blot; I – ICS; H2-dII - H2-d class II; H2-kII - H2-k class II; H2-bII - H2-b class II

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Table 6. Summary of two extended patches of glutamate (Q) residues as Q-patches.

Types	Numbers of Q residues	Positions of glutamate (Q)[$, #, *]	Marked in Fig. 2
Q-patch1 266-306			
Present only in SARS-CoVs	8	Q229, Q239, Q240, Q241, Q289, Q294, Q303, Q306	by red stars
Absent in SARS-CoV-2	1	A267Q	by green star
Present only in other viruses	6	R227Q, T245Q, K249Q, del255Q, T271Q, H300Q	by blue stars
Present in all viruses	3	Q260, Q272, Q281	by black stars
Present in SARS-CoVs and some other viruses	2	Q242, Q244	by maroon stars
Total Q in Q-patch1	**20**		
Q-patch2 361-418			
Present only in SARS-CoV	7	Q380, Q384, Q386, Q406, Q408, Q409, Q418	by red stars
Present only in other viruses	8	K361Q, P368Q, K370Q, K373Q, K375Q, D377Q, K388Q, K405Q	by blue stars
Present in SARS-CoVs and some other viruses	1	Q389	by maroon star
Total Q in Q-patch2	**16**		

$Numbering of N protein of SARS-CoV-2 was used.
#R227Q means at the position R277 in SARS-CoV-2, other viruses have Q as a substitution.
*present in some species only

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

Table 5. List of top antiviral FDA-approved drugs (affinity lower than -8 Kcal/Mol), which can potentially target motifs of the nucleocapsid protein of SARS-CoV-2.

MOTIF	DRUGBANK ID	MODE	AFFINITY (KCAL/MOL)	DRUG NAME	DRUG CLASS	BRAND NAME	APPROVED CLINICAL USE	MECHANISM(S) OF DRUG ACTION
MOTIF1	DB08930	1	-8.7	Dolutegravir	Integrase inhibitors	Tivicay®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosomes
MOTIF1	DB08864	1	-8.2	Rilpivirine	Non-Nucleoside Reverse Transcriptase Inhibitors	Edurant®	HIV-1	Binds directly to HIV RT and inhibits DNA synthesis
MOTIF1	DB06817	1	-8.2	Raltegravir	Integrase inhibitors	Isentress®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosome
MOTIF1	DB00432	1	-8	Trifluridine	5-Substituted 2'-deoxyuridines	Viroptic®	HSV	Inhibits HSV DNA replication
MOTIF2	DB09183	1	-8.3	Dasabuvir	HCV NS5B inhibitor + HCV NS5A inhibitor	Viekira Pak®	HCV	Inhibits activities of HCV NS5A and NS5B polymerase
MOTIF2	DB06817	1	-8.3	Raltegravir	Integrase inhibitors	Isentress®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosome
MOTIF2	DB08930	1	-8.2	Dolutegravir	Integrase inhibitors	Tivicay®	HIV	Targets HIV integrase to inhibit the integration of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

								viral DNA into human chromosome
MOTIF2	DB08864	1	-8.2	Rilpivirine	Non-Nucleoside Reverse Transcriptase Inhibitors	Edurant®	HIV-1	Binds directly to HIV RT and inhibits DNA synthesis
MOTIF4	DB08930	1	-9.7	Dolutegravir	Integrase inhibitors	Tivicay®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosome
MOTIF4	DB00932	1	-9.6	Tipranavir	Protease inhibitors	Aptivus®	HIV-1	Blocks the active site of HIV protease to prevent cleavage of viral precursor proteins
MOTIF4	DB00705	1	-9.5	Delavirdine	Non-Nucleoside Reverse Transcriptase Inhibitors	Rescriptor®	HIV-1	Binds directly to HIV RT and inhibits DNA synthesis
MOTIF4	DB06817	1	-9.4	Raltegravir	Integrase inhibitors	Isentress®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosome
MOTIF4	DB08864	1	-8.9	Rilpivirine	Non-Nucleoside Reverse Transcriptase Inhibitors	Edurant®	HIV-1	Binds directly to HIV RT and inhibits DNA synthesis
MOTIF4	DB00224	1	-8.8	Indinavir	Protease inhibitors	Crixivan®	HIV	Blocks the active site of HIV protease to prevent cleavage of viral precursor proteins
MOTIF4	DB09183	1	-8.8	Dasabuvir	HCV NS5B inhibitor + HCV NS5A inhibitor	Viekira Pak®	HCV	Inhibits activities of HCV NS5A and NS5B polymerase
MOTIF4	DB08873	1	-8.5	Boceprevir	Protease inhibitors	Victrelis®	HCV	NA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2020 doi:10.20944/preprints202005.0413.v1

MOTIF4	DB00220	1	-8.5	Nelfinavir	Protease inhibitors	Viracept®	HIV	Blocks the active site of HIV protease to prevent cleavage of viral precursor proteins
MOTIF4	DB09101	1	-8.3	Elvitegravir	Integrase inhibitors	Vitekta ®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosome
MOTIF4	DB01264	1	-8.3	Darunavir	Protease inhibitors	Prezista®	HIV	Blocks the active site of HIV protease to prevent cleavage of viral precursor proteins
MOTIF4	DB08934	1	-8	Sofosbuvir	HCV NS5B inhibitor + HCV NS5A inhibitor	Sovaldi®	HCV	Inhibits activities of HCV NS5A and NS5B polymerase
MOTIF4	DB01048	1	-8	Abacavir	Nucleoside Reverse Transcriptase Inhibitors	Ziagen®	HIV	Targets HIV RT and competes with dGTP to inhibit DNA synthesis
MOTIF5	DB09183	1	-8.7	Dasabuvir	HCV NS5B inhibitor + HCV NS5A inhibitor	Viekira Pak®	HCV	Inhibits activities of HCV NS5A and NS5B polymerase
MOTIF5	DB06817	1	-8.3	Raltegravir	Integrase inhibitors	Isentress®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosome
MOTIF5	DB08930	1	-8.1	Dolutegravir	Integrase inhibitors	Tivicay®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosome
MOTIF6	DB00932	1	-8.7	Tipranavir	Protease inhibitors	Aptivus®	HIV-1	Blocks the active site of HIV protease to prevent cleavage of viral precursor proteins
MOTIF6	DB06817	1	-8.4	Raltegravir	Integrase inhibitors	Isentress®	HIV	Targets HIV integrase to inhibit the integration of

 doi:10.20944/preprints202005.0413.v1

								viral DNA into human chromosome
MOTIF7	DB00932	1	-8.7	Tipranavir	Protease inhibitors	Aptivus®	HIV-1	Blocks the active site of HIV protease to prevent cleavage of viral precursor proteins
MOTIF7	DB06817	1	-8.4	Raltegravir	Integrase inhibitors	Isentress®	HIV	Targets HIV integrase to inhibit the integration of viral DNA into human chromosome