Identification and molecular characterization of mutations in Nucleocapsid Phosphoprotein of SARS-CoV-2

Gajendra Kumar Azad Corresp. 1

1 Department of Zoology, Patna University, Patna, Bihar, India

Corresponding Author: Gajendra Kumar Azad
Email address: gkazad@patnauniversity.ac.in

SARS-CoV-2 genome encodes four structural protein that include, Spike glycoprotein, Membrane protein, Envelope protein and Nucleocapsid Phosphoprotein (N-protein). The N-protein interacts with viral genomic RNA and helps in packaging. As the SARS-CoV-2 spread to almost all countries worldwide within 2-3 months; it also acquired mutations in its RNA genome. Therefore, this study was conducted with an aim to identify the variations present in N-protein of SARS-CoV-2. Here, we analysed 4163 reported sequence of N-protein from United States of America (USA) and compared with first reported sequence from Wuhan, China. Our study identified 107 mutations that reside all over the N-protein. Further, we show the high rate of mutations in intrinsically disordered regions (IDRs) of N-protein. Our study show 45% residues of IDR2 harbour mutations. The RNA binding domain (RBD) and dimerization domain of N-protein also have mutations at key residues. We further measured the effect of these mutations on N-protein stability and dynamicity and our data reveals that multiple mutations can cause considerable alterations. Altogether, our data strongly suggests that N-protein is one of the mutational hotspot proteins of SARS-CoV-2 that is changing rapidly and these mutations can potentially interferes with various aspects of N-protein functions including its interaction with RNA, oligomerization and signalling events.

TITLE

Identification and molecular characterization of mutations in Nucleocapsid Phosphoprotein of SARS-CoV-2

AUTHORS

Gajendra Kumar Azad[1#]

[1]Department of Zoology, Patna University, Patna-800005, Bihar (India)

[#]Corresponding Author:

Gajendra Kumar Azad

Email address: gkazad@patnauniversity.ac.in

Keywords: COVID-19; SARS-CoV-2; Mutations; Nucleocapsid Phosphoprotein (N-protein); Infectious diseases; USA

ABSTRACT

SARS-CoV-2 genome encodes four structural protein that include, Spike glycoprotein, Membrane protein, Envelope protein and Nucleocapsid Phosphoprotein (N-protein). The N-protein interacts with viral genomic RNA and helps in packaging. As the SARS-CoV-2 spread to almost all countries worldwide within 2-3 months; it also acquired mutations in its RNA genome. Therefore, this study was conducted with an aim to identify the variations present in N-protein of SARS-CoV-2. Here, we analysed 4163 reported sequence of N-protein from United States of America (USA) and compared with first reported sequence from Wuhan, China. Our study identified 107 mutations that reside all over the N-protein. Further, we show the high rate of mutations in intrinsically disordered regions (IDRs) of N-protein. Our study show 45% residues of IDR2 harbour mutations. The RNA binding domain (RBD) and dimerization domain of N-protein also have mutations at key residues. We further measured the effect of these mutations on N-protein stability and dynamicity and our data reveals that multiple mutations can cause considerable alterations. Altogether, our data strongly suggests that N-protein is one of the mutational hotspot proteins of SARS-CoV-2 that is changing rapidly and these mutations can potentially interferes with various aspects of N-protein functions including its interaction with RNA, oligomerization and signalling events.

INTRODUCTION

In the late December, 2019, Wuhan, the Hubei province of China, reported a surge in hospitalisation due to pneumonia like symptoms (Zhu et al., 2020). The causative agent was identified as a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that shares close similarity with earlier known SARS-CoV (Chen et al., 2020). The SARS-CoV-2 is highly contagious that lead to its rapid spread worldwide, and in March 2020, the World Health Organization (WHO) declared the outbreak a pandemic. The disease caused by SARS-CoV-2 has been named as coronavirus disease 19 (COVID-19) that exhibits mild to severe respiratory distress in the infected individuals. As of 28th June, 2020 the COVID-19 has affected all countries worldwide with close to 10 million reported cases and 0.5 million confirmed deaths. Further, the epidemiological studies revealed that the mortality rate from COVID-19 is significantly higher among individuals over 60 years of age with weak immunity (Liu et al., 2020).

The SARS-CoV-2 has positive sense, single stranded RNA genome of approximately 29.8 kb (Wu et al., 2020b). The majority of viral genome encodes non-structural proteins that are proteolytically processed from a single Orf1ab polypeptide. SARS-CoV-2 genome also encode four structural proteins, including the Spike glycoprotein (S), Membrane protein (M), Envelope protein (E) and Nucleocapsid Phosphoprotein (N) (Wu et al., 2020a). The S, M and E proteins are located in the lipid bilayer of the virus and contribute to the formation of viral envelope; however, the N-protein contributes to the viral genomic RNA packaging and remains embedded in the central core of the virion. N-protein binds with viral genomic RNA and forms helical structure to maintain the structural integrity of RNA genome (Chang et al., 2014). This is one of the most abundant structural proteins encoded by the SARS-CoV-2 genome. The SARS-CoV-2 N-protein resembles N-protein from other RNA viruses, known to modulate host intracellular machinery and also involved in the regulation of virus life cycle (McBride, van Zyl & Fielding, 2014). Evidence show that N-protein is recruited to the Replication-Transcription Complexes (RTC) via Nsp3 and plays a crucial role in coronaviral life cycle (Cong et al., 2019). The abrogation of this interaction impairs the stimulation of genomic RNA and viral mRNA transcription in vivo and in vitro. Furthermore, the N-protein interactions with M promotes completion of viral assembly by stabilizing N protein-RNA complex, inside the internal virion (Astuti & Ysrafil, 2020).

The crystal structure of N-protein revealed two distinct domains at N and C terminus (Kang et al., 2020). The domain present towards the N terminus is also known and RNA binding domain (RBD). The C terminal side harbours dimerization domain which interacts with other N-protein to make dimer. Apart from these two domains there are three intrinsically disordered regions (IDRs) at N and C terminal ends as well as between the RBD and dimerization domain of N-protein. Since, this protein plays critical role in packaging of SARS-CoV-2 RNA genome, the mutations in N-protein or interfering its function can lead to diverse outcome on viral life cycle (Rabi Ann Musah, 2005; Chenavas et al., 2013).

Moreover, the study of N-protein is also important because of its unique immunological properties. For instance, earlier study with SARS N-protein has shown that this protein is a potential candidate for vaccine development because it can induce a strong immunological response (Liu et al., 2006). A recent study revealed that the B and T cell epitopes of N protein of SARS-CoV-2 shows close resemblance with that of SARS-CoV indicating that immune targeting of these identical epitopes may offer protection against this virus (Ahmed, Quadeer & McKay, 2020). Moreover, the sera of COVID-19 patients contains abundant amount of IgA, IgM and IgG antibodies against N-protein antigen demonstrating the importance of this antigen in host

immunity and diagnostics (Shang et al., 2005; Zeng et al., 2020). Therefore, the N-protein is one of the candidate target molecule that needs to be properly studied to understand its role in virus pathogenesis, vaccine development and pharmacological implications. Here, we compared the N-protein sequences obtained from USA with first reported sequence from China to identify the variations present between them. We have identified 107 mutations and their impact on N-protein structure and function are discussed.

MATERIALS AND METHODS

Sequence retrieval from NCBI-virus-database

The NCBI-virus-database stores the deposited sequences of SARS-CoV-2 which is updated regularly as the new sequences are reported. As of 23rd June 2020, 4163 SARS-CoV-2 sequences of N-protein were deposited from USA. We downloaded these sequences and used them for analysis in this study. The first reported N-protein sequence from Wuhan was used as reference sequence or wild type sequence (Wu et al., 2020b). The protein accession identification number of reference sequence used in this study is YP_009724397 and rest of the 4163 IDs (reported from USA) are mentioned in supplementary table 1.

Multiple sequence alignment by Clustal-Omega program

To identify the mutations present in the SARS-CoV-2 N-protein reported from USA, we did multiple sequence alignments and compared them with the first reported N-protein sequence (YP_009724397) from Wuhan, China as described earlier (Azad, 2020). The multiple sequence alignment was performed using Clustal Omega tool (Madeira et al., 2019).

Calculation of free energy and vibrational entropy between wild type and mutant N-proteins

In order to measure the impact of mutations identified in this study on the structural dynamicity and stability of N-protein, we calculated the differences in free energy (ΔΔG) and vibrational entropy (ΔΔSvib) ENCoM between wild type and mutants as described earlier (Chand, Banerjee & Azad, 2020a). This analysis was performed by DynaMut program (Rodrigues, Pires & Ascher, 2018). To perform DynaMut protein modelling we used RCSB protein ID: 6VYO (Kang et al., 2020) for RBD molecular modelling and RCSB protein ID: 6WJI for dimerization domain molecular modelling of N-protein. DynaMut also provide the visual representation of fluctuation in protein structure. The blue colour represents gain in rigidity and red colour represents gain in flexibility upon mutation.

Generation of weblogo to show conservation of N-protein sequences

The weblogo was generated using a webserver as described earlier (Crooks et al., 2004). The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position (Crooks et al., 2004).

RESULTS

Identification of mutations in IDR1, IDR2 and IDR3 of N-protein

The crystal structure of N-protein of SARS-CoV-2 has been recently solved (Kang et al., 2020), the structural details show it is comprised of three distinct regions; the N terminal domain (contains RNA binding domain), C terminal domain (contains dimerization domain) and IDRs as shown in figure 1. There are three IDRs in N-protein; IDR1 (at the N terminal end), IDR2 (between RBD and CTD) and IDR3 (at the C terminal end). IDR2 is also referred as linker region (LKR) because it connects RBD and dimerization domain of N-protein. In order to identify the variations present in N-protein of SARS-CoV-2 reported from the USA, we performed multiple sequence alignments. Here, we used Clustal Omega program to align 4163 N-protein polypeptide sequences from USA and compared them with the first reported sequence from Wuhan, China.

Our analysis identified eighteen mutations in IDR1 (Table1). The IDR1 is present from 1-43 residues towards the N terminal end of N-protein. These eighteen mutations correspond to approximately 40% (18 out of 43) of the residues of IDR1. Among these the most frequently mutated residues are Gly and Arg (both are mutated at four positions) and Pro residue is mutated at three different positions in IDR1 (Table 1).

Similar analysis with IDR2 identified thirty six mutations which correspond to approximately 45% of residues of IDR2 (Table 2). The IDR2 is present from 181-256 residues of the N-protein and connects RBD and dimerization domains. The most frequently mutated residue in IDR2 was found to be Ser, it is mutated at twelve positions. Further, the Ala, Gly and Arg residues are mutated at five positions, respectively.

Similarly, we identified fifteen mutations in IDR3 (Table 3). The IDR3 is present from 365-419 residues towards the C terminal end of N-protein. Most notable mutations are Thr and Ala residues are mutated at three positions and Pro, Asp, and Gln are mutated at two positions, respectively (Table 3). Altogether, we identified sixty nine mutations in intrinsically disordered regions IDR1, IDR2 and IDR3 of N-protein.

Identification of mutations in RBD and dimerization domain of N-protein

The RBD of N-protein starts from 44th residue till 180th residue. We mapped the mutation in this region of N-protein and our analysis revealed presence of twenty two mutations (Table 4). These twenty two mutations also correspond to approximately 16% of the residues of RBD. Our mutational analysis shows the most frequently mutated residues are Pro and Ala at five positions and Asp at three positions as shown in table 4.

Similar analysis with the dimerization domain of N-protein revealed that it harbours sixteen mutations (Table 5). The dimerization domain of N-protein starts from 257th residue till 364th residue. Our mutational analysis shows Thr is mutated at four positions and Asp at three positions. Further, only 14 % residues are mutated in this domain which is least among all other regions of the N-protein identified here. Altogether, we identified thirty eight mutations in RBD and dimerization domain of N-protein. We have highlighted the location of amino acids in the representative crystal structure of N-protein that are mutated in RBD (Figure 1B) and dimerization domain (Figure 1C)

Subsequently, we also calculated the frequency of each mutation identified in this study. The table 6 shows the top ten mutants arranged in descending order of their respective frequencies. The R203K mutation is having the highest frequency of 4.9% followed by G204R with 4.7%. Further, we generated weblogo of the 4163 polypeptide sequences of N-protein to observe their amino acid conservation as shown in figure 1D. Altogether, we have identified 107 mutations in N-protein that resides in its IDRs and RBD and dimerization domain.

Mutations causes alteration in dynamic stability of N-protein

In order to understand the effect of mutations on the stability of the protein we calculated the differences in free energy (ΔΔG) between wild type and mutants. We performed this analysis using DynaMut program. The positive ΔΔG corresponds to increase in stability while negative ΔΔG corresponds to decrease in stability. We performed this analysis with all of the mutations that reside in RBD and dimerization domain of N-protein. The IDRs do not have proper 3D structure therefore; this analysis is not accurate for those regions. Our data revealed the noticeable increase or decrease in free energy in various mutations as shown in table 6. The top five positive and negative ΔΔG values are highlighted in table 6. The maximum increase in ΔΔG was observed for T271I (1.184 kcal/mol) and the highest negative ΔΔG was obtained for I292T (-1.952 kcal/mol), both of these mutations reside in dimerization domain of N-protein.

We also measured the change in vibrational entropy energy ($\Delta\Delta S_{Vib}$ENCoM) between the wild type and the mutants present in RBD and dimerization domain of N-protein (Table 7).Vibration

entropy contributes to the configurational-entropy of the proteins (Goethe, Fita & Rubi, 2015). The negative $\Delta\Delta S_{Vib}$ENCoM of mutant N-protein corresponds to the increase in rigidification and positive $\Delta\Delta S_{Vib}$ENCoM corresponds to gain in flexibility of the protein structure. The maximum positive $\Delta\Delta S_{Vib}$ENCoM was obtained for P364L (0.256 kcal.mol-1.K-1) and negative $\Delta\Delta S_{Vib}$ENCoM was obtained for G284E (-0.844 kcal.mol-1.K-1). The variation in vibrational entropy between wild type and mutant can also be visualised as shown in figure 2. The blue colour corresponds to rigidification in protein structure and red colour corresponds to gain in flexibility upon mutation. The top three positive and negative $\Delta\Delta S_{Vib}$ENCoM are shown in figure 2 (A-F). Altogether, the data obtained from $\Delta\Delta G$ and $\Delta\Delta S_{Vib}$ENCoM strongly suggests that the mutations identified in this study can influence N-protein stability and dynamicity.

Intramolecular interactions are altered due to mutations in N-protein

Next, we sought to closely analyse the changes in the intramolecular interactions in some of the mutants that exhibited significant alterations in $\Delta\Delta G$. We compared the intramolecular interaction for T271I ($\Delta\Delta G$: 1.184 kcal/mol) and I292T ($\Delta\Delta G$: -1.952 kcal/mol) as these two mutants showed maximum variations among thirty eight mutants present in RBD and dimerization domain of N-protein (Table 4 and 5). Our data clearly showed the variations in the interactions mediated by wild type and mutant residues in the pocket, where these amino acids resides as shown in figure 3A-B (T271I) , and 3C-D (I292T). Altogether, our data strongly suggests that the mutants identified in our study are affecting the dynamic stability as well as intramolecular interactions in the N-protein.

DISCUSSIONS

SARS-CoV-2 is an RNA virus, a causative agent of COVID-19. This virus spread worldwide within a span of few months and during its spread it also acquired mutations. Several recent studies reported the appearance of mutations in SARS-CoV-2 proteins (Korber et al., 2020; Pachetti et al., 2020; Chand, Banerjee & Azad, 2020b). This study was performed with an aim to identify mutations in N-protein which is one of the main structural proteins of SARS-CoV-2. Here, we analysed 4163 sequences of N-protein from USA and identified 107 mutations upon comparison from first reported sequences of the same protein from Wuhan, China. We also observed around 64% (69 out of 107) of these mutations reside in the IDRs of N-protein. Among IDRs, the IDR2 harbours 36 mutations that correspond to the most number of mutations observed in a single distinct region of the N-protein.

Earlier studies demonstrated that Ser and Arg-rich linker region (IDR2) plays indispensible role in intracellular signalling events primarily by phosphorylation at Ser residues (Wootton, Rowland & Yoo, 2002; McBride, van Zyl & Fielding, 2014). The wild type LKR/ IDR2 contains sixteen Ser residues, and our study revealed that out of those, twelve serine residues are mutated (table 2). Therefore, we can safely assume that these mutations of Ser residues might contribute to alteration of phosphorylation dependent signalling. A recent study shows that S197, S202, R203 and G204 are important sites of phosphorylation by Aurora kinase A/B, GSK-3 as well as for its interactions with 14-3-3 protein (Tung & Limtung, 2020). Surprisingly, our study report mutation in all of these four residues suggesting that these mutant might have altered phosphorylation signaling. We have also observed that R203 and G204 is the most frequently mutated residue of N-protein (Table 6). Similar observations were also reported from other locations (Franco-Munoz et al., 2020). Furthermore, two recent independent studies revealed that SARS-CoV-2 is capable of suppressing the type-I IFN innate immune pathway possibly due to the role of N-protein in signalling events (Blanco-Melo et al., 2020; Zhou et al., 2020a) which can potentially alter the virulence of SARS-CoV-2.

We also measured ΔΔG and $\Delta\Delta S_{Vib}$ENCoM for the mutants that reside in the RBD and dimerization domain of N-protein. The four mutants that exhibited highest values for ΔΔG and $\Delta\Delta S_{Vib}$ENCoM identified in our study are T271I, I292T, G284E and P364L. Since, all of them are in the dimerization domain; therefore, it is possible that these mutations might lead to alteration in the dimerization potential of N-protein. The structural study of N-protein (C terminal domain) has revealed that residue 247-279 are essential for RNA binding (Zhou et al., 2020b) which harbours seven mutations (T247A, K249R, S250F, A252S, S255A, V270L T271I). The occurrence of these mutations in C terminal domain could possibly affect its interaction with RNA that might translate into viral RNA packaging and stability. Furthermore, the N-protein is also proposed as a candidate for vaccine development because it is known to elicit strong immunological response in SARS-CoV infected patients (Lin et al., 2003). A recent study shows that several B cell epitope of SARS-CoV were identical with SARS-CoV-2 (Ahmed, Quadeer & McKay, 2020). This study revealed that one of the most important B and T cell epitope lies between residues 305-340 of N-protein; however, our study identified multiple mutations including, P309L, M322I, S327L, T329M, T334I, D340G, D340N in that stretch. Therefore, it is possible that due to these mutations the properties of epitope might change that can affect host immunological response. Another mutation, P344S mutation has been implicated to decrease the protein stability (Khan et al., 2020). Hence, the development of

vaccines that target SARS-COV-2 N-protein must consider the mutations that occur in various populations and locations.

Evidences indicate that the N-protein of coronaviruses functions as an RNA chaperones (Zúñiga et al., 2007, 2010) and also contributes to packaging and maintenance of the RNA genome. It is also involved in RNA metabolism because N-protein interaction assays have shown the core stress granule components G3BP1 and G3BP2 are its interacting partners (Gordon et al., 2020). This interaction can either enhance stress granule induction or inhibit stress granule formation by sequestering G3BP1/G3BP2 (Hou et al., 2017). Hence, the drugs that can either inhibit the interactions of RNA with N-protein or interfere with dimerization of N-protein can be a potential antiviral candidates (Lo et al., 2013). One such drug is Nucleozin and its derivatives that targets ribonucleoprotein formation in influenza virus by interfering N-protein oligomerization (Gerritz et al., 2011). Furthermore, a recent study was conducted to identify inhibitors of SARS-CoV-2 N-protein, identified various promising candidate drugs including Conivaptan, Ergotamine, Venetoclax and Rifapentine (Onat Kadioglu, 2020). These candidate drugs interact with the residues that are either mutated (residue 154, 155, 156, 166) or are in the close vicinity of the mutations (residue 67, 81, 163, 169) identified in our study. Furthermore, bioinformatics analysis predicted Dihydroergotamine , Rifabutin and Nystatin as a potential candidate drugs (Onat Kadioglu, 2020) that interacts with a stretch of residues (from residues 150 to 160) of N-protein. Surprisingly, our study revealed that this stretch harbour four mutations (151, 152, 154 and 156), which can potentially alter the interactions of these drugs with N-protein. Altogether, the mutation revealed in this study can interfere with various aspects of N-protein functions that include oligomerization, interaction with RNA and interference in N-protein mediated signalling events.

CONCLUSIONS

In this study we identified 107 mutations in N-protein of SARS-CoV-2 reported from USA. Further, we demonstrate these mutations can potentially alter dynamic stability of N-protein. Altogether, the data presented here, warrants further investigations to understand its impact on SARS-CoV-2 phenotype and drugs that target N-protein.

ACKNOWLDEGEMENTS

We would like to acknowledge the Department of Zoology, Patna University, Patna, Bihar (India) for providing infrastructural support for this study.

FIGURE AND TABLE LEGENDS

Figure 1: The schematic structure of Nucleocapsid Phosphoprotein (N-protein) of SARS-CoV-2. The N-protein comprising of 419 residues is shown. The RNA binding domain, dimerization domain, intrinsically disordered regions including IRD1, IRD2, and IRD3 are labelled. B-C) Cartoon representation of crystal structure of the RNA binding domain and dimerization domain of N-protein. The stick shows the location of residues that are mutated in the respective domains. The structural representations are made using Autodock software. D) The weblogo diagram showing the conservation status of polypeptide sequence of N-protein. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position

Figure 2: Visual representation of Δ Vibrational Entropy Energy between Wild-Type and Mutant N-protein. The amino acids residues are colored according to the vibrational entropy change as a consequence of mutation of N-protein. **BLUE** represents a rigidification of the structure and **RED** a gain in flexibility. (A-C) represents the top three mutants that show rigidification in structure upon mutation. (D-F) represents the top three mutants that show gain in flexibility upon mutation. Each panel also shows the mutation and the location of the residues.

Figure 3: Visual representation of interatomic interactions contributed by T271I and I292T of N-protein. Both of these mutants showed maximum positive and negative ΔΔG among mutants present in RBD and dimerization domain of N-protein. (A-B) represents threonine to isoleucine substitution at 271st position; (C-D) represents isoleucine to threonine substitution at 292nd position. Wild-type and mutant residues are represented in light-green color. The interactions made by wild type and mutant residues are highlighted in each panel. The polar interactions are depicted in red dotted line, hydrophobic interaction in green and weak hydrogen bonds in orange.

Table 1: The table show the location and details of mutations identified in IDR1 of N-protein.

Table 2: The table show the location and details of mutations identified in IDR2.

Table 3: The table show the location and details of mutations identified in IDR3.

Table 4: The table show the location and details of mutations identified in RBD of N-protein.

Table 5: The table show the location and details of mutations identified in dimerization domain of N-protein.

Table 6: The frequency of top ten mutations observed in our study

Table 7: The table show the ΔΔG and ΔΔSvib ENCoM of the mutants present in RBD and dimerization domain of N-protein. DynaMut program was used to calculate both parameters. The top five positive and negative ΔΔG values are highlighted in bold digits. The top three positive and negative ΔΔSvib ENCoM values are highlighted in bold digits.

REFERENCES

Ahmed SF, Quadeer AA, McKay MR. 2020. Preliminary identification of potential vaccine targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses*. DOI: 10.3390/v12030254.

Astuti I, Ysrafil. 2020. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*. DOI: 10.1016/j.dsx.2020.04.020.

Azad GK. 2020. Identification of novel mutations in the methyltransferase complex (Nsp10-Nsp16) of SARS-CoV-2. *Biochemistry and Biophysics Reports*. DOI: 10.1016/j.bbrep.2020.100833.

Blanco-Melo D, Nilsson-Payant BE, Liu WC, Uhl S, Hoagland D, Møller R, Jordan TX, Oishi K, Panis M, Sachs D, Wang TT, Schwartz RE, Lim JK, Albrecht RA, tenOever BR. 2020. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell*. DOI: 10.1016/j.cell.2020.04.026.

Chand GB, Banerjee A, Azad GK. 2020a. Identification of twenty-five mutations in surface glycoprotein (Spike) of SARS-CoV-2 among Indian isolates and their impact on protein dynamics. *Gene Reports* 21. DOI: 10.1016/j.genrep.2020.100891.

Chand GB, Banerjee A, Azad GK. 2020b. Identification of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure. *PeerJ* 8:e9492. DOI: 10.7717/peerj.9492.

Chang CK, Hou MH, Chang CF, Hsiao CD, Huang TH. 2014. The SARS coronavirus nucleocapsid protein - Forms and functions. *Antiviral Research*. DOI: 10.1016/j.antiviral.2013.12.009.

Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, Zhang L. 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*. DOI: 10.1016/S0140-6736(20)30211-7.

Chenavas S, Cre´pin T, Delmas B, Ruigrok RWH, Slama-Schwok A. 2013. Influenza virus nucleoprotein: Structure, RNA binding, oligomerization and antiviral drug target. *Future Microbiology*. DOI: 10.2217/fmb.13.128.

Cong Y, Ulasli M, Schepers H, Mauthe M, V'kovski P, Kriegenburg F, Thiel V, de Haan CAM, Reggiori F. 2019. Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle. *Journal of Virology*. DOI: 10.1128/jvi.01925-19.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. *Genome Research*. DOI: 10.1101/gr.849004.

Franco-Munoz C, Alvarez-Diaz DA, Laiton-Donato K, Wiesner M, Escandon P, Usme-Ciro JA, Franco-Sierra ND, Florez-Sanchez AC, Gomez-Rangel S, Calderon LDR, Ramirez JB, Baez EO, Walteros DM, Martinez MLO, Mercado-Reyes M. 2020. Substitutions in Spike and Nucleocapsid proteins of SARS-CoV-2 circulating in Colombia. *medRxiv*. DOI: 10.1101/2020.06.02.20120782.

Gerritz SW, Cianci C, Kim S, Pearce BC, Deminie C, Discotto L, McAuliffe B, Minassian BF, Shi S, Zhu S, Zhai W, Pendri A, Li G, Poss MA, Edavettal S, McDonnell PA, Lewis HA, Maskos K, Mor̈tl M, Kiefersauer R, Steinbacher S, Baldwin ET, Metzler W, Bryson J, Healy MD, Philip T, Zoeckler M, Schartman R, Sinz M, Leyva-Grado VH, Hoffmann HH, Langley DR, Meanwell NA, Krystal M. 2011. Inhibition of influenza virus replication via small molecules that induce the formation of higher-order nucleoprotein oligomers. *Proceedings of the National Academy of Sciences of the United States of America*. DOI: 10.1073/pnas.1107906108.

Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj V V., Guo JZ, Swaney DL, Tummino TA, Huettenhain R, Kaake RM, Richards AL, Tutuncuoglu B, Foussard H, Batra J, Haas K, Modak M, Kim M, Haas P, Polacco BJ, Braberg H, Fabius JM, Eckhardt M, Soucheray M, Bennett MJ, Cakir M, McGregor MJ, Li Q, Meyer B, Roesch F, Vallet T, Mac Kain A, Miorin L, Moreno E, Naing ZZC, Zhou Y, Peng S, Shi Y, Zhang Z, Shen W, Kirby IT, Melnyk JE, Chorba JS, Lou K, Dai SA, Barrio-Hernandez I, Memon D, Hernandez-Armenta C, Lyu J, Mathy CJP, Perica T, Pilla KB, Ganesan SJ, Saltzberg DJ, Rakesh R, Liu X, Rosenthal SB, Calviello L, Venkataramanan S, Liboy-Lugo J, Lin Y,

Huang XP, Liu YF, Wankowicz SA, Bohn M, Safari M, Ugur FS, Koh C, Savar NS, Tran QD, Shengjuler D, Fletcher SJ, O'Neal MC, Cai Y, Chang JCJ, Broadhurst DJ, Klippsten S, Sharp PP, Wenzell NA, Kuzuoglu D, Wang HY, Trenker R, Young JM, Cavero DA, Hiatt J, Roth TL, Rathore U, Subramanian A, Noack J, Hubert M, Stroud RM, Frankel AD, Rosenberg OS, Verba KA, Agard DA, Ott M, Emerman M, Jura N, von Zastrow M, Verdin E, Ashworth A, Schwartz O, d'Enfert C, Mukherjee S, Jacobson M, Malik HS, Fujimori DG, Ideker T, Craik CS, Floor SN, Fraser JS, Gross JD, Sali A, Roth BL, Ruggero D, Taunton J, Kortemme T, Beltrao P, Vignuzzi M, García-Sastre A, Shokat KM, Shoichet BK, Krogan NJ. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. DOI: 10.1038/s41586-020-2286-9.

Hou S, Kumar A, Xu Z, Airo AM, Stryapunina I, Wong CP, Branton W, Tchesnokov E, Götte M, Power C, Hobman TC. 2017. Zika Virus Hijacks Stress Granule Proteins and Modulates the Host Stress Response. *Journal of Virology*. DOI: 10.1128/jvi.00474-17.

Kang S, Yang M, Hong Z, Zhang L, Huang Z, Chen X, He S, Zhou Z, Zhou Z, Chen Q, Yan Y, Zhang C, Shan H, Chen S. 2020. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharmaceutica Sinica B*. DOI: 10.1016/j.apsb.2020.04.009.

Khan MI, Khan ZA, Baig MH, Ahmad I, Farouk A-E, Song YG, Dong J-J. 2020. Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An in silico insight. *PloS one* 15:e0238344. DOI: 10.1371/journal.pone.0238344.

Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, Foley B, Giorgi EE, Bhattacharya T, Parker MD, Partridge DG, Evans CM, Silva T de, LaBranche CC, Montefiori DC, Group SC-19 G. 2020. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*. DOI: 10.1101/2020.04.29.069054.

Lin Y, Shen X, Yang RF, Li YX, Ji YY, He YY, Shi MD, Lu W, Shi TL, Wang J, Wang HX, Jiang HL, Shen JH, Xie YH, Wang Y, Pei G, Shen BF, Wu JR, Sun B. 2003. Identification of an epitope of SARS-coronavirus nucleocapsid protein. *Cell research*. DOI: 10.1038/sj.cr.7290158.

Liu K, Chen Y, Lin R, Han K. 2020. Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. *Journal of Infection*. DOI: 10.1016/j.jinf.2020.03.005.

Liu SJ, Leng CH, Lien SP, Chi HY, Huang CY, Lin CL, Lian WC, Chen CJ, Hsieh SL, Chong P. 2006. Immunological characterizations of the nucleocapsid protein based SARS vaccine

candidates. *Vaccine*. DOI: 10.1016/j.vaccine.2006.01.058.

Lo YS, Lin SY, Wang SM, Wang CT, Chiu YL, Huang TH, Hou MH. 2013. Oligomerization of the carboxyl terminal domain of the human coronavirus 229E nucleocapsid protein. *FEBS Letters*. DOI: 10.1016/j.febslet.2012.11.016.

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research*. DOI: 10.1093/nar/gkz268.

McBride R, van Zyl M, Fielding BC. 2014. The coronavirus nucleocapsid is a multifunctional protein. *Viruses*. DOI: 10.3390/v6082991.

Onat Kadioglu MSHJGTE. 2020. Identification of novel compounds against three targets of SARS CoV2 coronavirus by combined virtual screening and supervised machine learning . *Bull World Health Organ*. DOI: 10.2471/BLT.20.251561.

Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, Zella D, Ippodrino R. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine*. DOI: 10.1186/s12967-020-02344-6.

Rabi Ann Musah. 2005. The HIV-1 Nucleocapsid Zinc Finger Protein as a Target of Antiretroviral Therapy. *Current Topics in Medicinal Chemistry*. DOI: 10.2174/1568026043387331.

Rodrigues CHM, Pires DEV, Ascher DB. 2018. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research*. DOI: 10.1093/nar/gky300.

Shang B, Wang XY, Yuan JW, Vabret A, Wu XD, Yang RF, Tian L, Ji YY, Deubel V, Sun B. 2005. Characterization and application of monoclonal antibodies against N protein of SARS-coronavirus. *Biochemical and Biophysical Research Communications*. DOI: 10.1016/j.bbrc.2005.08.032.

Tung HYL, Limtung P. 2020. Mutations in the phosphorylation sites of SARS-CoV-2 encoded nucleocapsid protein and structure model of sequestration by protein 14-3-3. *Biochemical and Biophysical Research Communications*. DOI: 10.1016/j.bbrc.2020.08.024.

Wootton SK, Rowland RRR, Yoo D. 2002. Phosphorylation of the Porcine Reproductive and Respiratory Syndrome Virus Nucleocapsid Protein. *Journal of Virology*. DOI: 10.1128/jvi.76.20.10569-10576.2002.

Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J, Sheng J, Quan L, Xia Z, Tan W, Cheng G, Jiang T. 2020a. Genome Composition and Divergence of

475 the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host and Microbe*. DOI:
476 10.1016/j.chom.2020.02.001.
477 Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML,
478 Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020b. A
479 new coronavirus associated with human respiratory disease in China. *Nature*. DOI:
480 10.1038/s41586-020-2008-3.
481 Zeng W, Liu G, Ma H, Zhao D, Yang Y, Liu M, Mohammed A, Zhao C, Yang Y, Xie J, Ding C,
482 Ma X, Weng J, Gao Y, He H, Jin T. 2020. Biochemical characterization of SARS-CoV-2
483 nucleocapsid protein. *Biochemical and Biophysical Research Communications*. DOI:
484 10.1016/j.bbrc.2020.04.136.
485 Zhou Z, Ren L, Zhang L, Zhong J, Xiao Y, Jia Z, Guo L, Yang J, Wang C, Jiang S, Yang D,
486 Zhang G, Li H, Chen F, Xu Y, Chen M, Gao Z, Yang J, Dong J, Liu B, Zhang X, Wang W,
487 He K, Jin Q, Li M, Wang J. 2020a. Heightened Innate Immune Responses in the
488 Respiratory Tract of COVID-19 Patients. *Cell Host and Microbe*. DOI:
489 10.1016/j.chom.2020.04.017.
490 Zhou R, Zeng R, von Brunn A, Lei J. 2020b. Structural characterization of the C-terminal
491 domain of SARS-CoV-2 nucleocapsid protein. *Molecular Biomedicine* 1:2. DOI:
492 10.1186/s43556-020-00001-4.
493 Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F,
494 Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. 2020. A novel coronavirus from patients
495 with pneumonia in China, 2019. *New England Journal of Medicine*. DOI:
496 10.1056/NEJMoa2001017.
497 Zúñiga S, Cruz JLG, Sola I, Mateos-Gómez PA, Palacio L, Enjuanes L. 2010. Coronavirus
498 Nucleocapsid Protein Facilitates Template Switching and Is Required for Efficient
499 Transcription. *Journal of Virology*. DOI: 10.1128/jvi.02011-09.
500 Zúñiga S, Sola I, Moreno JL, Sabella P, Plana-Durán J, Enjuanes L. 2007. Coronavirus
501 nucleocapsid protein is an RNA chaperone. *Virology*. DOI: 10.1016/j.virol.2006.07.046.
502

Table 1(on next page)

IDR1 Mutations

The table show the location and details of mutations identified in IDR1 of N-protein.

1 Table 1:

S. No.	Wild type residue	Position of mutation	Mutated residue
1	Asp	3	Tyr
2	Asn	4	Asp
3	Pro	6	Thr
4	Gln	9	His
5	Pro	13	Leu
6	Arg	14	His
7	Gly	18	Cys
8	Gly	19	Arg
9	Pro	20	Leu
10	Asp	22	Tyr
11	Ser	23	Thr
12	Gly	30	Ala
13	Glu	31	Asp
14	Arg	32	Leu
15	Gly	34	Leu
16	Ala	35	Thr
17	Arg	36	Leu
18	Arg	40	Cys
19	Arg	40	Leu

2

Table 2(on next page)

IDR2 mutations

The table show the location and details of mutations identified in IDR2.

1 Table 2:

S. No.	Wild type residue	Position of mutation	Mutated residue
1	Ser	183	Tyr
2	Arg	185	Cys
3	Arg	185	Leu
4	Ser	187	Leu
5	Ser	188	Leu
6	Ser	190	Ile
7	Arg	191	Leu
8	Asn	192	Ser
9	Ser	193	Ile
10	Ser	194	Leu
11	Arg	195	Ile
12	Ser	197	Leu
13	Pro	199	Ser
14	Ser	202	Asn
15	Arg	203	Lys
16	Arg	203	Met
17	Gly	204	Arg
18	Thr	205	Ile
19	Ala	208	Gly
20	Arg	209	Lys
21	Arg	209	Thr
22	Ala	211	Ser
23	Gly	212	Cys
24	Asn	213	Tyr
25	Gly	215	Ser
26	Ala	218	Val
27	Ala	220	Thr
28	Gln	229	His
29	Ser	232	Arg
30	Ser	232	Thr
31	Met	234	Ile
32	Ser	235	Pro
33	Ser	235	Phe
34	Gly	236	Val
35	Gly	238	Cys
36	Gly	243	Cys
37	Thr	247	Ala
38	Lys	249	Arg
39	Ser	250	Phe

40	Ala	252	Ser
41	Ser	255	Ala

2

Table 3(on next page)

IDR3 mutations

The table show the location and details of mutations identified in IDR3.

1 Table 3:

S. No.	Wild type residue	Position of mutation	Mutated residue
1	Pro	365	Ser
2	Pro	365	Leu
3	Asp	377	Tyr
4	Asp	377	Gly
5	Thr	379	Ile
6	Gln	380	His
7	Ala	381	Val
8	Pro	383	Ser
9	Pro	383	Leu
10	Gln	386	Lys
11	Gln	386	His
12	Thr	391	Ile
13	Thr	393	Ile
14	Ala	397	Ser
15	Ala	398	Val
16	Asp	399	Glu
17	Ser	413	Ile
18	Ser	416	Leu

2

3

4

Table 4(on next page)

RBD mutations

The table show the location and details of mutations identified in RBD of N-protein

1 Table 4:

S. No.	Wild type residue	Position of mutation	Mutated residue
1	Pro	46	Ser
2	Glu	62	Val
3	Pro	67	Ser
4	Asp	81	Tyr
5	Ala	90	Ser
6	Ala	119	Ser
7	Pro	122	Leu
8	Ala	125	Thr
9	Asp	128	Tyr
10	Asn	140	Thr
11	Pro	142	Ser
12	Asp	144	Tyr
13	Asp	144	His
14	Ile	146	Phe
15	Pro	151	Leu
16	Ala	152	Ser
17	Asn	154	Tyr
18	Ala	156	Ser
19	Gln	163	Arg
20	Thr	166	Ile
21	Lys	169	Arg
22	Ser	180	Ile

2

Table 5(on next page)

Dimerization domain mutations

The table show the location and details of mutations identified in dimerization domain of N-protein.

1 Table 5:

S. No.	Wild type residue	Position of mutation	Mutated residue
1	Val	270	Leu
2	Thr	271	Ile
3	Gly	284	Glu
4	Gln	289	His
5	Ile	292	Thr
6	Gln	294	Leu
7	Asp	297	Val
8	Pro	309	Leu
9	Met	322	Ile
10	Ser	327	Leu
11	Thr	329	Met
12	Thr	334	Ile
13	Asp	340	Gly
14	Asp	340	Asn
15	Asp	348	Tyr
16	Thr	362	Ile
17	Pro	364	Leu

2

Table 6(on next page)

Frequency of N-protein mutations

The frequency of top 10 mutations observed in this study

1 Table 6:

Mutation	Number of samples that harbour the mutation	% frequency
R203K	207	4.97357
G204R	196	4.709274
E62V	39	0.937049
A208G	24	0.576646
S183Y	20	0.480538
S194L	18	0.432484
T362I	16	0.384431
T205I	15	0.360404
P13L	11	0.264296
R185C	10	0.240269

2

Table 7 (on next page)

ΔΔG and ΔΔSvib ENCoM calculations

The table show the ΔΔG and ΔΔSvib ENCoM of the mutants present in RBD and dimerization domain of N-protein. DynaMut programe was used to calculate both parameters. The top five positive and negative ΔΔG values are highlighted in bold digits. The top three positive and negative ΔΔSvib ENCoM values are highlighted in bold digits.

1 Table 7:

S. No.	Mutant	PDB ID	ΔΔG (kcal/mol)	ΔΔSvibENCoM (kcal.mol-1.K-1)
1	E62V	6VYO	0.105	0.091
2	P67S	6VYO	**-0.486**	0.16
3	D81Y	6VYO	0.454	-0.425
4	A90S	6VYO	0.274	0.043
5	A119S	6VYO	0.073	-0.069
6	P122L	6VYO	**-0.166**	-0.049
7	A125T	6VYO	**-0.565**	-0.022
8	D128Y	6VYO	**0.846**	-0.236
9	N140T	6VYO	0.318	-0.177
10	P142S	6VYO	0.26	-0.17
11	D144Y	6VYO	0.291	-0.293
12	D144H	6VYO	-0.036	0.06
13	I146F	6VYO	0.708	**-0.837**
14	P151L	6VYO	**0.771**	-0.14
15	A152S	6VYO	0.298	-0.051
16	N154Y	6VYO	-0.096	-0.063
17	A156S	6VYO	0.428	-0.256
18	Q163R	6VYO	-0.092	-0.017
19	T166I	6VYO	0.194	-0.055
20	K169R	6VYO	0.231	0.077
21	V270L	6WJI	0.679	-0.194
22	T271I	6WJI	**1.184**	-0.472
23	G284E	6WJI	0.553	**-0.844**
24	Q289H	6WJI	0.18	**0.181**
25	I292T	6WJI	**-1.952**	**0.186**
26	Q294L	6WJI	0.447	-0.078
27	D297V	6WJI	-0.113	-0.072
28	P309L	6WJI	**0.887**	**-0.524**
29	M322I	6WJI	**-0.348**	0.045
30	S327L	6WJI	**0.894**	-0.259
31	T329M	6WJI	0.569	-0.189
32	T334I	6WJI	0.236	-0.115
33	D340G	6WJI	0.398	-0.114
34	D340N	6WJI	0.194	-0.088
35	D348Y	6WJI	0.136	-0.121
36	T362I	6WJI	0.396	0.047
37	P364L	6WJI	-0.061	**0.256**

2

Figure 1

The schematic structure of Nucleocapsid Phosphoprotein (N protein) of SARS-CoV-2.

The schematic structure of Nucleocapsid Phosphoprotein (N-protein) of SARS-CoV-2. The N-protein comprising of 419 residues is shown. The RNA binding domain, dimerization domain, intrinsically disordered regions including IRD1, IRD2, and IRD3 are labelled. B-C) Cartoon representation of crystal structure of the RNA binding domain and dimerization domain of N-protein. The stick shows the location of residues that are mutated in the respective domains. The structural representations are made using Autodock software. D) The weblogo diagram showing the conservation status of polypeptide sequence of N-protein. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino acid at that position

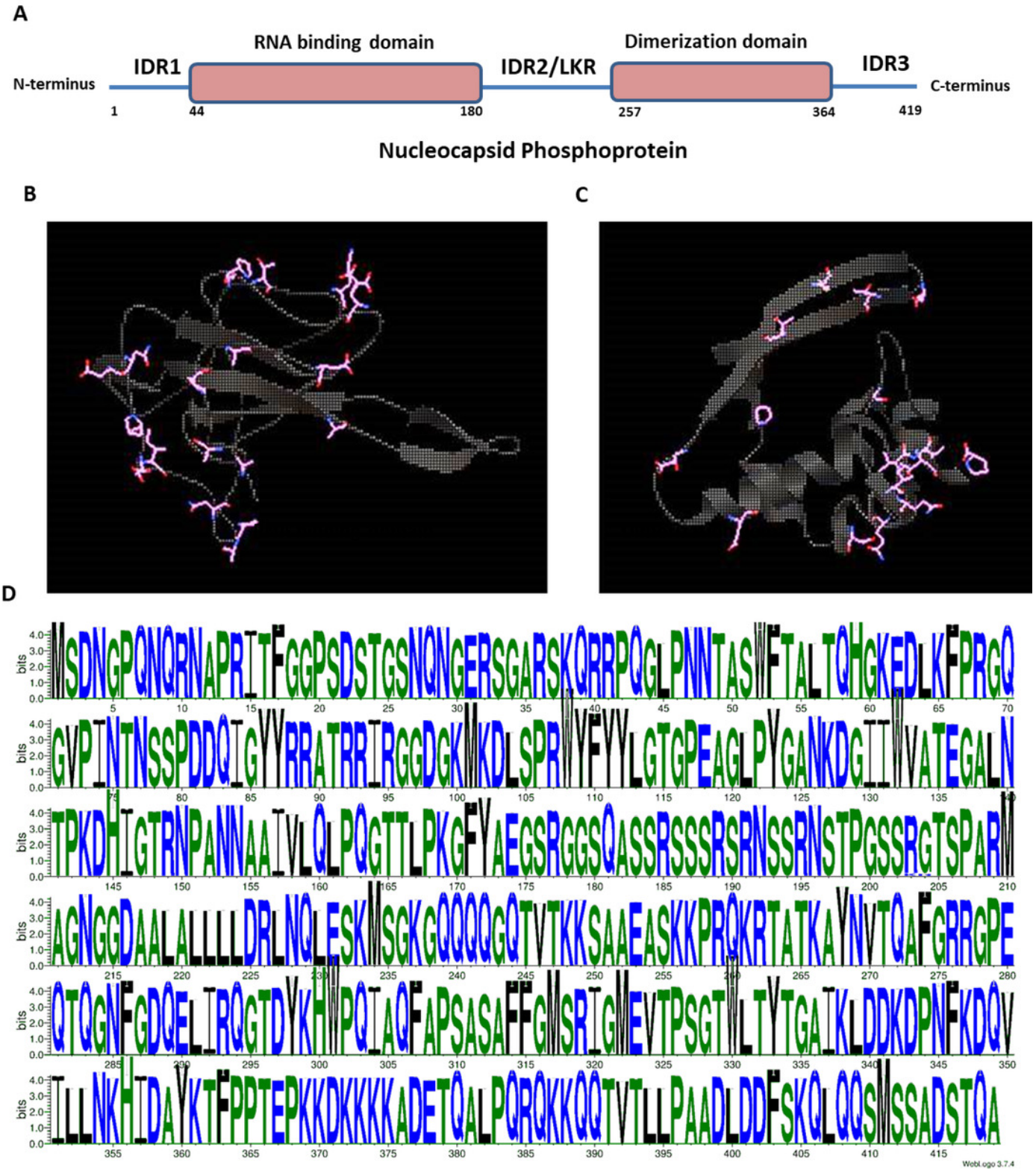

A
RNA binding domain
Dimerization domain
IDR1
IDR2/LKR
IDR3
N-terminus
C-terminus
1
44
180
257
364
419
Nucleocapsid Phosphoprotein
B
C
D
bits
WebLogo 3.7.4

Figure 2

Visual representation of Δ Vibrational Entropy Energy between Wild-Type and Mutant N protein.

The amino acids residues are colored according to the vibrational entropy change as a consequence of mutation of N-protein.**BLUE**represents a rigidification of the structure and **RED**a gain in flexibility. (A-C) represents the top three mutants that show rigidification in structure upon mutation. (D-F) represents the top three mutants that show gain in flexibility upon mutation. Each panel also shows the mutation and the location of the residues.

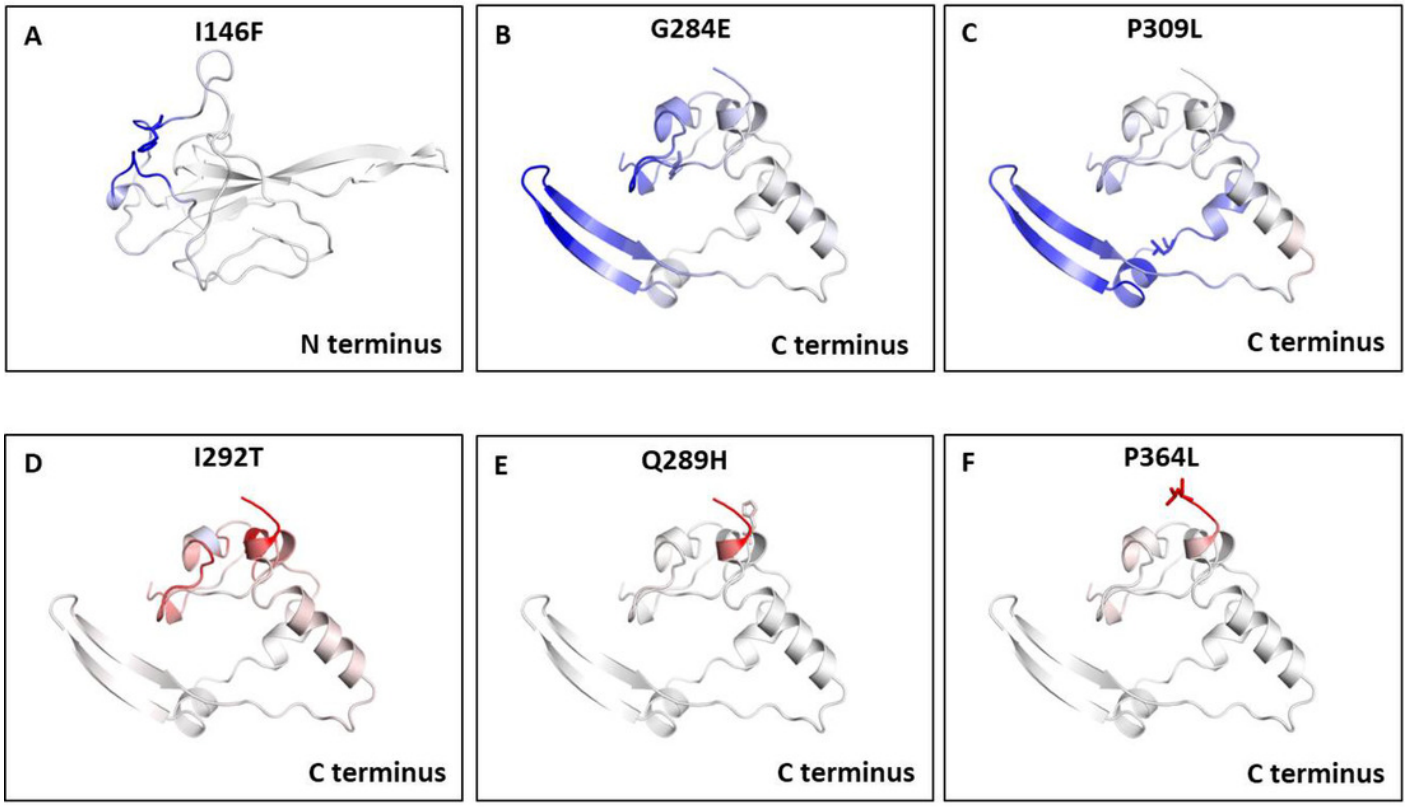
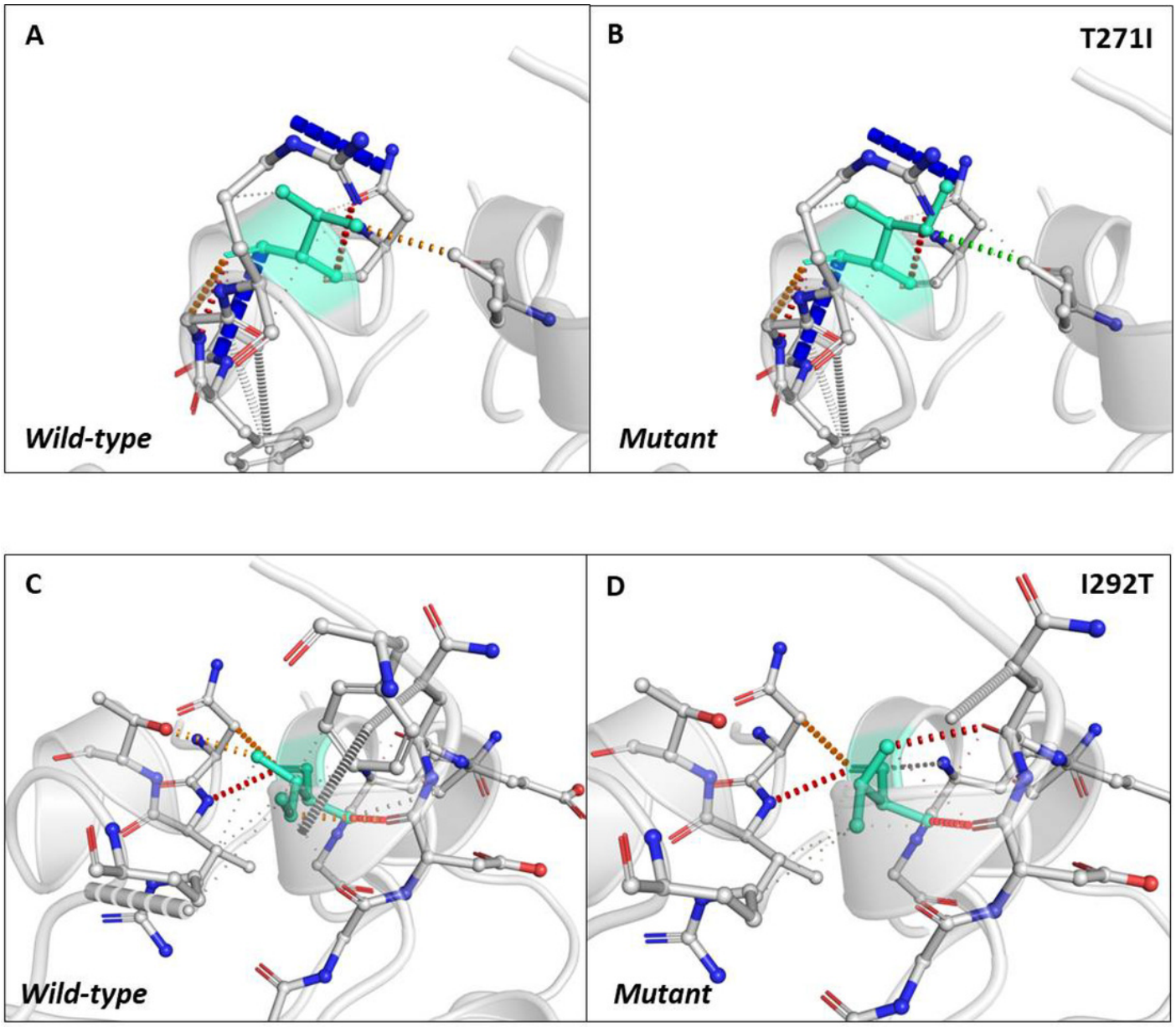


Figure 3

Analysis of interatomic interactions.

Visual representation of interatomic interactions contributed by T271I and I292T of N-protein. Both of these mutants showed maximum positive and negative ΔΔG among mutants present in RBD and dimerization domain of N-protein. (A-B) represents threonine to isoleucine substitution at 271st position; (C-D) represents isoleucine to threonine substitution at 292nd position. Wild-type and mutant residues are represented in light-greencolor. The interactions made by wild type and mutant residues are highlighted in each panel. The polar interactions are depicted in red dotted line, hydrophobic interaction in green and weak hydrogen bonds in orange.


A
Wild-type
B
T271I
Mutant
C
Wild-type
D
I292T
Mutant