Title

Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein

Journal

bioRxiv : the preprint server for biology, 1(06-18)

Authors

Ye, Qiaozhen
West, Alan MV
Silletti, Steve
et al.

Publication Date

2020-05-17

DOI

10.1101/2020.05.17.100685

Peer reviewed

Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein

Qiaozhen Ye[1], Alan M.V. West[1], Steve Silletti[2], Kevin D. Corbett[1,2,3]*

[1]Department of Cellular & Molecular Medicine, University of California San Diego, La Jolla, CA

[2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA

[3]Ludwig Institute for Cancer Research, San Diego Branch, La Jolla, CA

*Correspondence should be addressed to Kevin D. Corbett:

9500 Gilman Drive, #3206
La Jolla, CA 92093
(858) 534-7267
kcorbett@ucsd.edu

Abstract

The COVID-2019 pandemic is the most severe acute public health threat of the twenty-first century. To properly address this crisis with both robust testing and novel treatments, we require a deep understanding of the life cycle of the causative agent, the SARS-CoV-2 coronavirus. Here, we examine the architecture and self-assembly properties of the SARS-CoV-2 nucleocapsid (N) protein, which binds viral RNA and assembles into a filament that is packaged into new virions. We determined a 1.4 Å resolution crystal structure of this protein's N2b domain, revealing a compact, intertwined dimer very similar to that of related coronaviruses SARS-CoV and MERS-CoV. Using size exclusion chromatography and multi-angle light scattering, we find that this domain forms a dimer in solution, and that addition of the C-terminal spacer B/N3 domain mediates tetramer formation. Using hydrogen-deuterium exchange mass spectrometry, we find evidence that at least part of this putatively disordered domain is structured, potentially forming an α-helix that either self-associates or docks against the N2b domain to mediate tetramer formation. Finally, we map the locations of over 4,400 individual amino acid substitutions in the N protein from ~17,000 SARS-CoV-2 genome sequences, and find that they are strongly clustered in the protein's N2a linker domain. The nearly 300 substitutions identified within the N1b and N2b domains cluster away from their functional RNA binding and dimerization interfaces. Overall, this work reveals the architecture and self-assembly properties of a key protein in the SARS-CoV-2 life cycle. As the N protein is a common target of patient antibodies, this work will also benefit ongoing efforts to develop robust and specific serological tests, and could also benefit the analysis of patient-derived antibodies.

Significance Statement

After infecting a cell, the SARS-CoV-2 coronavirus replicates its genome and packages it into new virus particles. The virus's nucleocapsid (N) protein is critical for this process, binding the genomic RNA and assembling into large filaments before being packaged into new virions. Here, we sought to understand the structure of the SARS-CoV-2 N protein and how it forms large assemblies. We combined x-ray crystallography and two biochemical methods to show that the protein's N2b and N3 domains mediate self-association into dimers and tetramers, respectively. We have therefore outlined the first two steps of a probable three-step self-assembly mechanism of the SARS-CoV-2 nucleocapsid. This work will benefit ongoing efforts to develop robust viral tests and to develop new drugs targeting key steps in the viral life cycle.

Introduction

SARS-CoV-2 (1, 2) is the third coronavirus to cross from an animal reservoir to infect humans in the 21st century, after SARS (severe acute respiratory syndrome coronavirus) (3, 4) and MERS (Middle-East respiratory syndrome coronavirus) (5). Isolation and sequencing of SARS-CoV-2 was reported in January 2020, and the virus was found to be highly related to SARS and share a probable origin in bats (2, 6). Since its emergence in December 2019 in Wuhan, China, the virus has infected over 4.6 million people and caused over 312,000 deaths as of May 17, 2020 (https://coronavirus.jhu.edu). The high infectivity of SARS-CoV-2 and the worldwide spread of this ongoing outbreak highlight the urgent need for public health measures and therapeutics to limit new infections. Moreover, the severity of the atypical pneumonia caused by SARS-CoV-2 (COVID-2019), often requiring multi-week hospital stays and the use of invasive ventilators (7–9), highlights the need for therapeutics to lessen the severity of individual infections.

Current therapeutic strategies against SARS-CoV-2 target major points in the life-cycle of the virus. The antiviral Remdesivir, first developed against Ebola virus (10, 11), inhibits the viral RNA-dependent RNA polymerases of a range of coronaviruses including SARS-CoV-2 (12–14) and has shown promise against SARS-CoV-2 in small-scale trials in both primates and humans (15, 16). Another target is the viral protease (Mpro/3CLpro), which is required to process viral polyproteins into their active forms (17). Finally, the transmembrane spike (S) glycoprotein mediates binding to host cells through the angiotensin converting enzyme 2 (ACE2) and transmembrane protease, serine 2 (TMPRSS2) proteins, and mediates fusion of the viral and host cell membranes (18–21). As the most prominent surface component of the virus, the spike protein is the major target of antibodies in patients, and is the focus of several current efforts at SARS-CoV-2 vaccine development. Initial trials using antibody-containing plasma of convalescent COVID-19 patients has also shown promise in lessening the severity of the disease (22).

While the above efforts target viral entry, RNA synthesis, and protein processing, there has so far been less emphasis on other steps in the viral life cycle. One critical step in coronavirus replication is the assembly of the viral genomic RNA and nucleocapsid (N) protein into a ribonucleoprotein (RNP) complex, which in betacoronaviruses like SARS-CoV-2 forms a helical filament structure that is packaged into virions through interactions with the membrane-spanning membrane (M) protein (23–26). Despite its location within the viral particle rather than on its surface, patients infected with SARS-CoV-2 show higher and earlier antibody responses to the nucleocapsid protein than the surface spike protein (27, 28). As such, a better understanding of the SARS-CoV-2 N protein's structure, and structural differences between it and N proteins of related coronaviruses including SARS-CoV, may aid the development of sensitive and specific immunological tests.

Coronavirus N proteins possess a shared domain structure with an N-terminal RNA-binding domain and a C-terminal domain responsible for dimerization. The assembly of N protein dimers into higher-order helical filaments is not well understood, but likely involves cooperative interactions between the dimerization domain and other regions of the protein, plus the bound RNA (29–37). Here, we present a high-resolution structure of the SARS-CoV-2 N dimerization domain, revealing an intertwined dimer similar to that of related betacoronaviruses. We also

analyze the self-assembly properties of the SARS-CoV-2 N protein, and show that higher-order assembly requires both the dimerization domain and the extended, disordered C-terminus of the protein. Together with other work revealing the structure and RNA-binding properties of the nucleocapsid N-terminal domain, these results lay the groundwork for a comprehensive understanding of SARS-CoV-2 nucleocapsid assembly and architecture.

Results

Structure of the SARS-CoV-2 Nucleocapsid dimerization domain

Betacoronavirus Nucleocapsid (N) proteins share a common overall domain structure, with ordered RNA-binding (N1b) and dimerization (N2b) domains separated by short regions with high predicted disorder (N1a, N2a, and spacer B/N3; **Figure 1A**). Self-association of the full-length SARS-CoV N protein and the isolated C-terminal region (domains N2b plus spacer B/N3; residues 210-422) was first demonstrated by yeast two-hybrid analysis (29), and the purified full-length protein was shown to self-associate into predominantly dimers in solution (30). The structures of the N2b domain of SARS-CoV and several related coronaviruses confirmed the obligate homodimeric structure of this domain (31–37), and other work showed that the region C-terminal to this domain mediates further self-association into tetramer, hexamer, and potentially higher oligomeric forms (38–40). Other studies have suggested that the protein's N-terminal region, including the RNA-binding N1b domain, can also self-associate (41, 42), highlighting the possibility that assembly of full-length N into helical filaments is mediated by cooperative interactions among several interfaces.

To characterize the structure and self-assembly properties of the SARS-CoV-2 nucleocapsid, we first cloned and purified the protein's N2b dimerization domain (N_{2b}; residues 247-364) (43, 44). We crystallized and determined two high-resolution crystal structures of N_{2b}; a 1.45 Å resolution structure of His_6-tagged N_{2b} at pH 8.5, and a 1.42 Å resolution structure of N_{2b} after His_6-tag cleavage, at pH 4.5 (see **Methods** and **Table S1**). These structures reveal a compact, tightly intertwined dimer with a central four-stranded β-sheet comprising the bulk of the dimer interface (**Figure 1B**). This interface is composed of two β-strands and a short α-helix from each protomer that extend toward the opposite protomer and pack against its hydrophobic core. The asymmetric units of both structures encompass two N_{2b} dimers, giving four crystallographically independent views of the N_{2b} dimer. These four dimers differ only slightly, showing overall Cα r.m.s.d values of 0.15-0.19 Å and with most variation arising from loop regions (**Figure S1**). Our structures also overlay closely with another recently-deposited structure of the SARS-CoV-2 N2b domain, which was determined to 2.05 Å resolution at pH 7.5 (PDB ID 6WJI; Center for Structural Genomics of Infectious Diseases (CSGID), unpublished). Including this structure, there are now seven independent crystallographic views of the SARS-CoV-2 N2b domain dimer (14 total protomers) in three crystal forms at pH 4.5, 7.5, and 8.5, all of which overlay with an overall Cα r.m.s.d of 0.15-0.31 Å (**Figure S1**). We examined the packing of N_{2b} dimers in the three different crystal forms, but did not identify any consistent dimer-dimer interactions that would suggest independent higher-order assembly of this domain.

The structure of N_{2b} closely resembles that of related coronaviruses, including SARS-CoV, Infectious Bronchitis Virus (IBV), MERS-CoV, and HCoV-NL63 (31–36). The structure is particularly similar to that of SARS-CoV, with which the N2b domain shares 96% sequence identity; only five residues differ between these proteins' N2b domains (SARS-CoV Gln268 → SARS-CoV-2 A267, D291→E290, H335→Thr334, Gln346→Asn345, and Asn350→Gln349), and the structures are correspondingly similar with an overall Cα r.m.s.d of 0.44 Å across the N_{2b} dimer (**Figure 1C**).

N protein variation in SARS-CoV-2 patient samples

Since the first genome sequence of SARS-CoV-2 was reported in January 2020 (2, 6), nearly 28,000 full genomic sequences have been deposited in public databases (as of May 17, 2020). To examine the variability of the N protein in these sequences, we downloaded a comprehensive list of reported mutations within the SARS-CoV-2 N gene in a set of 16,975 genome sequences from the China National Center for Bioinformation, 2019 Novel Coronavirus Resource. Among these sequences, there are 4,454 instances of amino acid substitutions in the N protein (**Figure 2A**). These variants are strongly clustered in the N2a linker domain, particularly in the serine/arginine-rich subdomain (SR in **Figure 2A**). The most common substitutions are R203K and G204R, which occur together as the result of a common trinucleotide substitution in genomic positions 28881-28883, from GGG to AAC (~1,500 of the 16,975 sequences in our dataset; **Figure S2**). While positions 203 and 204 accounted for nearly two-thirds of the total individual amino acid substitutions in this dataset, the N2a region shows a strong enrichment of mutations even when these positions are not considered (**Figure 2A**). One sequence also showed an in-frame deletion of residues 208-210. In contrast to the enrichment of missense mutations in the N2a domain, synonymous mutations were distributed equally throughout the protein (not shown). Thus, these data suggest that the N2a domain is uniquely tolerant of mutations, in keeping with its likely structural role as a disordered linker between the RNA-binding N1b domain and the N2b dimerization domain.

While the majority of N protein mutations are in the N2a domain, we nonetheless identified 152 instances of amino acid variants in the RNA-binding N1b domain, and 135 instances in the N2b domain. We mapped these onto high-resolution structures of both domains (**Figure 2B-C**). Two high-resolution crystal structures of the SARS-CoV-2 N1b domain have been determined (PDB ID 6M3M and 6VYO) (45), and a recent NMR study determined a solution structure of the domain and defined its likely RNA binding surface (**Figure 2B**; left panel) (46). In keeping with its functional importance, the identified RNA binding surface shows no mutations in this dataset (**Figure 2B**; middle panel). In the N2b domain, most mutations occur on surface residues, particularly in loop regions, while the functionally-important dimer interface is largely invariant (**Figure 2C**).

Finally, the 16,975 SARS-CoV-2 genome sequences contain seven sequences with nonsense/premature stop codons in the N protein. These mutations are all located in the spacer B/N3 region C-terminal to the N2b domain, at positions 372-418 (**Figure 2A**). An eighth sequence showed an in-frame deletion of residues 389-392 (not shown).

Self-association of the SARS-CoV-2 N protein

Our structures of the SARS-CoV-2 N protein N2b domain reveal that, as in related coronaviruses, this domain mediates homodimer formation. We next investigated the molecular basis for higher-order self-assembly of the SARS-CoV-2 nucleocapsid. Prior work with the Murine Hepatitis Virus (MHV) N protein showed that this protein's C-terminal N3 domain (residues 409-454 of 454) can, on its own, incorporate into nucleocapsid structures that lack the associated Membrane (M) protein, suggesting that this domain mediates a homotypic interaction between N proteins (44). Other work with SARS-CoV and HCoV-229E N proteins also found that the region C-terminal to the N2b domain (containing the spacer B and N3 regions) is responsible for higher-order assembly of tetramers and larger oligomers (38–40). To test the role of different domains in SARS-CoV-2 N protein assembly, we purified the full-length N protein (N_{FL}) and a series of truncation constructs encompassing the ordered N1b and N2b domains and their associated linker domains (N1a, N2a, and spacer B/N3; **Figure 1A**). We found that N_{FL} formed a large oligomer that eluted in the void volume of a gel filtration column (**Figure 3A**). Despite having been purified in high-salt buffer (500 mM NaCl) and incubated with both DNase and RNase to remove associated nucleic acids, this protein retained significant bound nucleic acid as judged by an A_{260}/A_{280} ratio of ~1.6 (not shown). Thus, as in related betacoronavirus N proteins, this protein assembles into large oligomers in vitro. Because any associated nucleic acids are likely to be short fragments due to incubation with DNase and RNase, we propose that self-assembly is mediated mainly by protein-protein interactions with at most a minor contribution from bound nucleic acids.

We next purified two truncations encompassing the N-terminal N1a and N1b regions (N_{1ab}, residues 2-175) or these regions plus the disordered N2a linker domain (N_{1ab2a}, residues 2-246). As in related N proteins, N_{1ab} formed a monomer in solution as measured by size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS; **Figure 3B**). We observed significant proteolytic cleavage of N_{1ab2a} during purification with a prominent cleavage site within the N2a region (**Figure S3**), highlighting the disordered nature of this linker domain. Nonetheless, we were able to show by SEC-MALS that this construct, too, is monomeric in solution (**Figure 3C**).

Next, we analyzed N_{2b} and an extended construct containing both the N2b and spacer B/N3 regions (N_{2b3}, residues 247-419). While N_{2b} forms a dimer in agreement with our crystal structure (**Figure 3D**), we found that N_{2b3} strikingly forms a homotetramer (**Figure 3E**). While the spacer B/N3 domain has been found to mediate higher-order assembly in related coronavirus N proteins, the robustness of tetramer formation in SARS-CoV-2 N mediated by this domain is distinct from related proteins. This finding suggests that coronavirus N protein self-assembly likely proceeds through at least three steps, each mediated by different oligomerization interfaces: (1) dimerization mediated by the N2b domain; (2) tetramerization mediated by the spacer B/N3 region; and (3) helical filament assembly mediated by cooperative interaction of multiple N protein domains and bound RNA (**Figure 3F**).

To gain structural insight into how the spacer B/N3 region mediates N protein tetramer formation, we performed hydrogen-deuterium exchange mass spectrometry (HDX-MS) on N_{2b} and N_{2b3} (**Figure 4**). By probing the rate of exchange of amide hydrogen atoms with deuterium atoms in a D_2O solvent, HDX-MS provides information on the level of secondary structure and solvent accessibility across an entire protein. We found that H-D exchange rates within N_{2b}

largely agreed with our crystal structure: regions in β-strands or α-helices showed low exchange rates consistent with high order, while loop regions showed increased exchange rates consistent with their likely flexibility (**Figure 4A, C, D**).

Compared to N_{2b}, N_{2b3} contains an additional 56 amino acids (residues 365-419). While residues 360-394 were not detected in our HDX-MS analysis, we detected spectra for seven overlapping peptides spanning residues 395-419 at the protein's extreme C-terminus (**Figure 4B**). While all of these peptides exhibited higher levels of exchange than the ordered N2b domain, peptides spanning the N-terminal part of this region (particularly residues 395-402) showed a degree of protection compared to those at the extreme C-terminus (residues 404-419; **Figure 4E**). This finding suggests that at least part of the spacer B/N3 domain possesses secondary structure and may mediate N_{2b3} tetramer formation. Indeed, analysis by the PSI-PRED server (47) suggests that this region may adopt an α-helical conformation (**Figure 1A**).

We next compared HDX-MS exchange rates of N_{2b} versus N_{2b3} for peptides within the N2b domain. We reasoned that if the C-terminus of N_{2b3} mediates tetramer formation, it may do so by docking against a surface in the N2b domain, which may be detectable by reduced deuterium uptake in the involved region. Unexpectedly, we found that the H-D exchange rates within the N2b domain were nearly identical between the two constructs, varying at most ~1.5% in fractional deuterium uptake in individual peptides (**Figure 4B, D**). While these data do not rule out the possibility that the spacer B/N3 region docks against N2b, they nonetheless suggest that spacer B/N3 may instead self-associate to mediate N protein tetramer formation.

Discussion

Given the severity of the ongoing COVID-19 pandemic, a deep understanding of the SARS-CoV-2 life cycle is urgently needed. Here, we examine the architecture and self-assembly properties of the SARS-CoV-2 nucleocapsid protein, a key player in viral replication responsible for packaging viral RNA into new virions. Through two high-resolution crystal structures, we show that this protein's N2b domain forms a compact, strand-swapped dimer similar to that of related betacoronaviruses. While the N2b domain mediates dimer formation, we find that addition of the C-terminal spacer B/N3 domain mediates formation of a robust homotetramer. Finally, the full-length N protein assembles into large oligomers, likely through cooperative protein-protein interactions involving several regions of the protein, plus potentially bound nucleic acid.

Given the importance of nucleocapsid-mediated RNA packaging to the viral life cycle, small molecules that inhibit nucleocapsid self-assembly or mediate aberrant assembly may be effective at reducing the severity of infections and the infectivity of patients. The high resolution of our crystal structures will enable their use in virtual screening efforts to identify such nucleocapsid assembly modulators. Given the high conservation of the N2b domain in beta coronaviruses, these assembly modulators may also be effective at countering related viruses including SARS-CoV. As SARS-CoV-2 is unlikely to be the last betacoronavirus to jump from an animal reservoir to humans, the availability of such treatments could drastically alter the course of future epidemics.

The SARS-CoV-2 genome has been subject to unprecedented scrutiny, with nearly 28,000 individual genome sequences deposited in public databases as of May 17, 2020. We used a set of 16,975 genome sequences to identify over 4,400 instances of amino acid substitutions in the N protein, and showed that these variants are strongly clustered in the protein's N2a linker domain. The ~300 substitutions we identified in the N1b and N2b domains were clustered away from these domains' RNA binding and dimerization interfaces, reflecting the functional importance of these surfaces. Finally, the identification of nonsense mutations in the protein's spacer B/N3 region suggests that this region may not be absolutely required for viral replication; this idea remains to be experimentally validated.

Given the early and strong antibody responses to the nucleocapsid displayed by SARS-CoV-2 infected patients, the distribution of mutations within this protein should be carefully considered as antibody-based tests are developed. The high variability of the N2a domain means that individual patient antibodies targeting this domain may not be reliably detected with tests using the reference N protein; especially if these antibodies recognize residues 203 and 204, which are mutated in a large fraction of infections. At the same time, patient antibodies targeting the conserved N2b domain may in fact cross-react with nucleocapsids of related coronaviruses like SARS-CoV. The availability of a panel of purified N protein constructs now makes it possible to systematically examine the epitopes of both patient-derived and commercial anti-nucleocapsid antibodies.

Methods

Cloning and Protein Purification

SARS-CoV-2 N protein constructs (N_{FL} (residues 2-419), N_{1ab} (2-175), N_{1ab2a} (2-246), N_{2b} (247-364), N_{2b3} (247-419)) were amplified by PCR from the IDT 2019-nCoV N positive control plasmid (IDT cat. # 10006625; NCBI RefSeq YP_009724397) and inserted by ligation-independent cloning into UC Berkeley Macrolab vector 2B-T (AmpR, N-terminal His_6-fusion; Addgene #29666) for expression in *E. coli*. Plasmids were transformed into *E. coli* strain Rosetta 2(DE3) pLysS (Novagen), and grown in the presence of ampicillin and chloramphenical to an OD_{600} of 0.8 at 37°C, induced with 0.25 mM IPTG, then grown for a further 16 hours at 18°C prior to harvesting by centrifugation. Harvested cells were resuspended in buffer A (25 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$ 10% glycerol, 5 mM β-mercaptoethanol, 1 mM NaN_3) plus 500 mM NaCl and 5 mM imidazole pH 8.0. For purification, cells were lysed by sonication, then clarified lysates were loaded onto a Ni^{2+} affinity column (Ni-NTA Superflow; Qiagen), washed in buffer A plus 300 mM NaCl and 20 mM imidazole pH 8.0, and eluted in buffer A plus 300 mM NaCl and 400 mM imidazole. For cleavage of His_6-tags, proteins were buffer exchanged in centrifugal concentrators (Amicon Ultra, EMD Millipore) to buffer A plus 300 mM NaCl and 20 mM imidazole, then incubated 16 hours at 4°C with TEV protease (48). Cleavage reactions were passed through a Ni^{2+} affinity column again to remove uncleaved protein, cleaved His_6-tags, and His_6-tagged TEV protease. Proteins were concentrated in centrifugal concentrators and purified by size-exclusion chromatography (Superdex 200; GE Life Sciences) in gel filtration buffer (25 mM Tris-HCl pH 7.5, 300 mM NaCl, 5 mM $MgCl_2$, 10% glycerol, 1 mM DTT). Purified proteins were concentrated and stored at 4°C for analysis.

SEC-MALS

For size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS), 100 μL purified proteins at 2-5 mg/mL were injected onto a Superdex 200 Increase 10/300 GL column (GE Life Sciences) in gel filtration buffer. Light scattering and refractive index profiles were collected by miniDAWN TREOS and Optilab T-rEX detectors (Wyatt Technology), respectively, and molecular weight was calculated using ASTRA v. 6 software (Wyatt Technology).

HDX-MS

H-D exchange experiments were conducted with a Waters Synapt G2S system. 5 μL samples containing 10 μM protein in gel filtration buffer were mixed with 55 μL of the same buffer made with D_2O for several deuteration times (0 sec, 1 min, 2 min, 5 min, 10 min) at 15°C. The exchange was quenched for 2 min at 1°C with an equal volume of quench buffer (3M guanidine HCl, 0.1% formic acid). Proteins were cleaved with pepsin and separated by reverse-phase chromatography, then directed into a Waters SYNAPT G2s quadrupole time-of-flight (qTOF) mass spectrometer. Peptides were identified using PLGS version 2.5 (Waters, Inc.), deuterium uptake was calculated using DynamX version 2.0 (Waters Corp.), and uptake was corrected for back-exchange using DECA (49). Uptake plots were generated in Prism version 8.

Crystallization and Structure Determination

For crystallization of untagged N_{2b}, protein (40 mg/mL) in crystallization buffer (25 mM Tris-HCl pH 7.5, 200 mM NaCl, 5 mM $MgCl_2$, and 1 mM Tris(2-carboxyethyl)phosphine) was mixed 1:1 with well solution containing 100 mM sodium acetate pH 4.5, 50 mM sodium/potassium tartrate, and 34% polyethylene glycol (PEG) 3350 at 20°C in hanging drop format. For crystallization of His_6-tagged N_{2b}, protein (40 mg/mL) in crystallization buffer was mixed 1:1 with well solution containing 100 mM Tris-HCl pH 8.5, 50 mM Ammonium Sulfate, and 38% PEG 3350 at 20°C in hanging drop format. Both untagged and His_6-tagged N_{2b} formed large shard-like crystals, and were frozen in liquid nitrogen directly from the crystallization drop without further cryoprotection.

Diffraction data were collected at beamline 24ID-E at the Advanced Photon Source. Diffraction datasets were processed with the RAPD pipeline (https://github.com/RAPD/RAPD/), which uses XDS (50) for indexing and data reduction, and the CCP4 programs AIMLESS (51) and TRUNCATE (52) for scaling and conversion to structure factors. The structure of untagged N_{2b} was determined by molecular replacement in PHASER (53) using a dimer of the SARS-CoV N2b domain (PDB ID 2GIB) (32) as a template. The resulting model was manually rebuilt in COOT (54) and refined in phenix.refine (55) using positional, isotropic B-factor, and TLS (one group per chain) refinement. The structure of His_6-tagged N2b was determined by molecular replacement from the structure of untagged N_{2b}, then manually rebuilt and refined as above. Data collection statistics, refinement statistics, and database accession numbers for diffraction data and final structures can be found in **Table S1**. All structural figures were generated with PyMOL version 2.3.

Bioinformatics

To examine variation in SARS-CoV-2 sequences, we downloaded a list of variants in the N gene in 16,975 genome sequences from China National Center for Bioinformation, 2019 Novel Coronavirus Resource (https://bigd.big.ac.cn/ncov?lang=en; downloaded May 6, 2020). We tabulated all mis-sense and nonsense mutations, and graphed the number of amino acid variants at each codon in Prism version 8 (all variant frequencies are listed in **Table S2**). To examine the prevalence of the trinucleotide substitution at genome positions 28881-28883, we downloaded a set of 200 SARS-CoV-2 genomes from the NCBI Virus Resource: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049). We selected genomes with and without the substitution to show in **Figure S2A**. We used the NextStrain database (56) to visualize the prevalence of the N protein G204R mutation, which is diagnostic of the GGG→AAC trinucleotide substitution in positions 28881-28883. To visualize SARS-CoV-2 clade identity, we used the URL "https://nextstrain.org/ncov/global?c=clade_membership&l=unrooted". To color by N protein residue 204 identity, we used the URL "https://nextstrain.org/ncov/global?c=gt-N_204&l=unrooted" (screenshots taken May 8, 2020).

Acknowledgements

The authors thank the staff of Advanced Photon Source NE-CAT beamline 24ID-E for assistance with diffraction data collection, E. Komives for advice on HDX-MS interpretation, and J. Pogliano, M. Daugherty, A. Schmidt, and members of the Corbett lab for helpful discussions. K.D.C. acknowledges generous institutional support from UC San Diego. The Waters Synapt HDX-MS at the UCSD BPMS Facility is supported by NIH S10 OD016234. The Northeastern Collaborative Access Team beamlines at the Advanced Photon Source are funded by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165). The Eiger 16M detector on the 24-ID-E beam line is funded by a NIH-ORIP HEI grant (S10OD021527). The Advanced Photon Source is a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357.

Figures

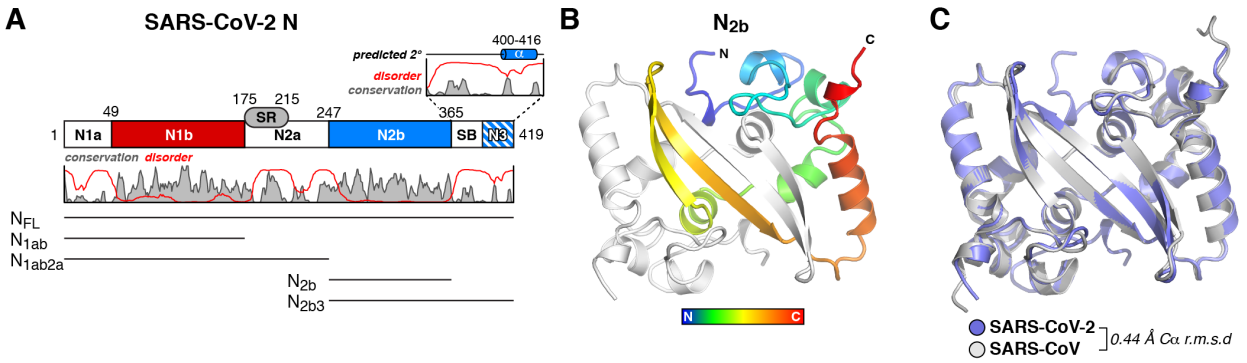

Figure 1. Structure of the SARS-CoV-2 Nucleocapsid dimerization domain

(A) Domain structure of the SARS-CoV-2 Nucleocapsid protein, as defined previously (43, 44), with plot showing the Jalview alignment conservation score (three-point smoothed; gray) (57) and DISOPRED3 disorder propensity (red) (58) for nine related coronavirus N proteins (SARS-CoV, SARS-CoV-2, MERS-CoV, HCoV-OC43, HCoV-HKU1, HCoV-NL63, and HCoV-229E, IBV (Infectious Bronchitis virus), and MHV (Murine Hepatitis virus)). SR: serine/arginine rich domain; SB; spacer B. The boundary between SB and N3 is not well-defined due to low identity between SARS-CoV/SARS-CoV-2 and MHV N proteins (44). All purified truncations are noted at bottom.

(B) Two views of the SARS-CoV-2 N_{2b} dimer, with one monomer colored as a rainbow (N-terminus blue, C-terminus red) and the other colored white.

(C) Structural overlay of the SARS-CoV-2 N_{2b} dimer (blue) and the equivalent domain of SARS-CoV-N (PDB ID 2GIB) (32).

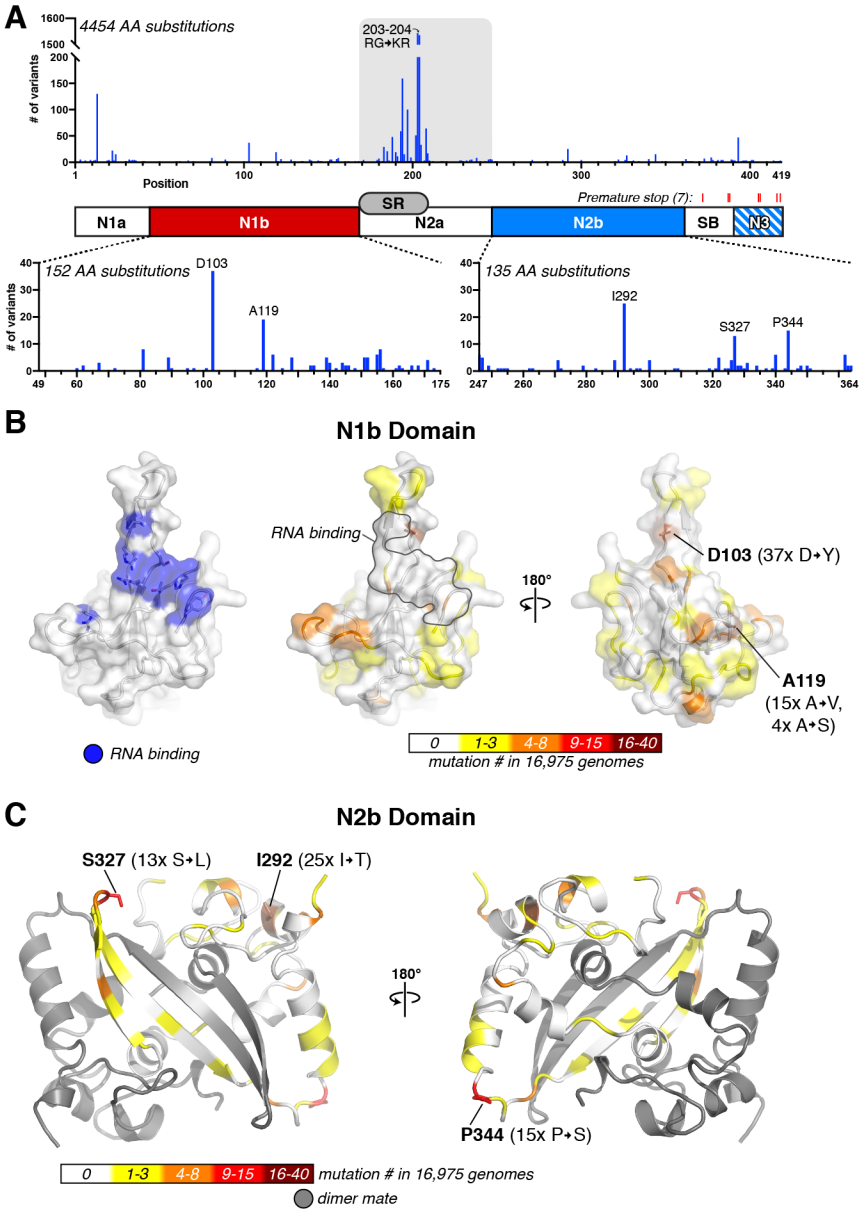

Figure 2. N protein variability in SARS-CoV-2 patient sequences

(A) *Top:* Plot showing the number of observed amino acid variants at each position in the N gene in 16,975 SARS-CoV-2 genomes (details in **Table S2**). The most highly-mutated positions are R203 and G204, which are each mutated more than 1500 times due to a prevalent trinucleotide substitution (**Figure S2**). *Bottom:* Plots showing amino acid variants in the N1b and N2b domains.

(B) Surface views of the N protein N1b domain (PDB ID 6VYO; Center for Structural Genomics of Infectious Diseases (CSGID), unpublished). At left, blue indicates RNA-binding residues identified by NMR peak shifts (A50, T57, H59, R92, I94, S105, R107, R149, and Y172) (46). At right, two views colored by the number of variants at each position observed in a set of 16,975 SARS-CoV-2 genomes. The two most frequently-mutated residues are shown in stick view and labeled.

(C) Cartoon view of the N protein N2b domain, with one monomer colored gray and the other colored by the number of variants at each position observed in a set of 16,975 SARS-CoV-2 genomes. The three most frequently-mutated residues are shown in stick view and labeled.

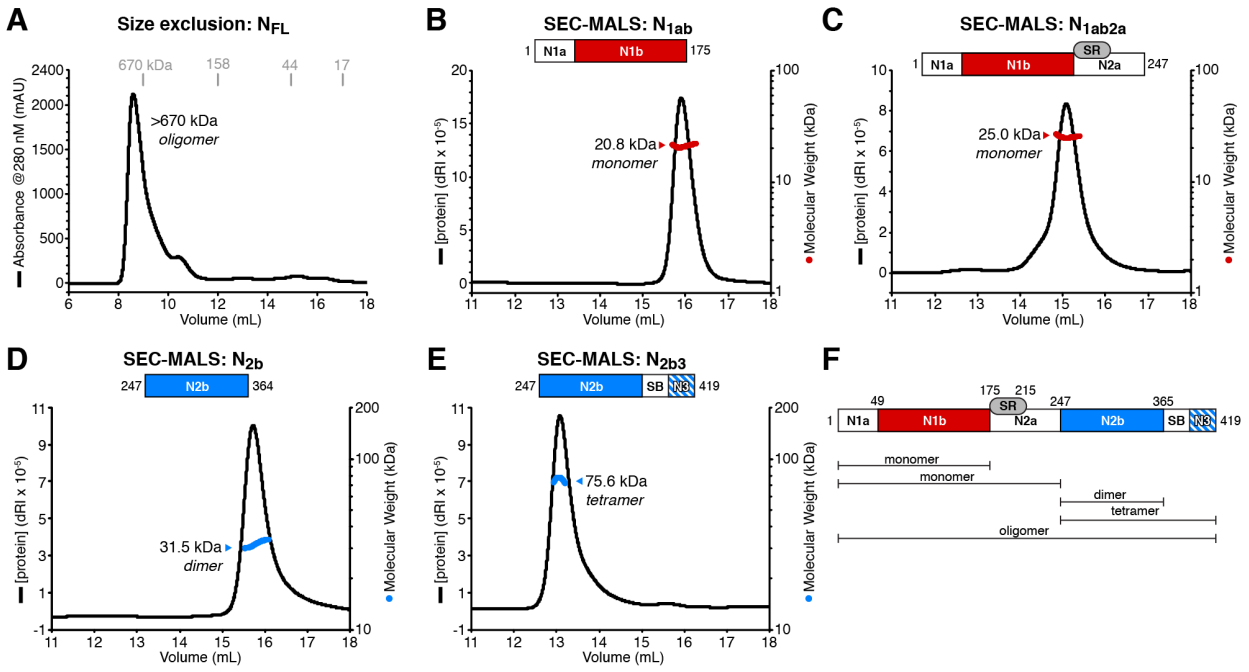

Figure 3. The C-terminus of N mediates tetramer formation

(A) Size exclusion chromatography (Superdex 200 Increase) of full-length SARS-CoV-2 N (MW 45.6 kDa), with elution volumes of molecular weight standards shown in gray. Migration in the void volume indicates formation of an oligomer with MW >670 kDa.

(B) Size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) analysis of SARS-CoV-2 N_{1ab} (residues 2-175). The measured MW of 20.8 kDa closely matches that of a monomer (18.9 kDa). dRI: differential refractive index.

(C) SEC-MALS analysis of SARS-CoV-2 N_{1ab2a} (residues 2-246). The measured MW of 25.0 kDa is slightly less than that of a monomer (26.2 kDa), reflecting partial proteolysis within the N2a domain (**Figure S3**).

(D) SEC-MALS analysis of SARS-CoV-2 N_{2b}. The measured MW (31.5 kDa) closely matches that of a homodimer (26.5 kDa).

(E) SEC-MALS analysis of SARS-CoV-2 N_{2b3}. The measured MW (75.6 kDa) closely matches that of a homotetramer (77.4 kDa).

(F) Schematic summary of size exclusion and SEC-MALS results on N protein constructs.

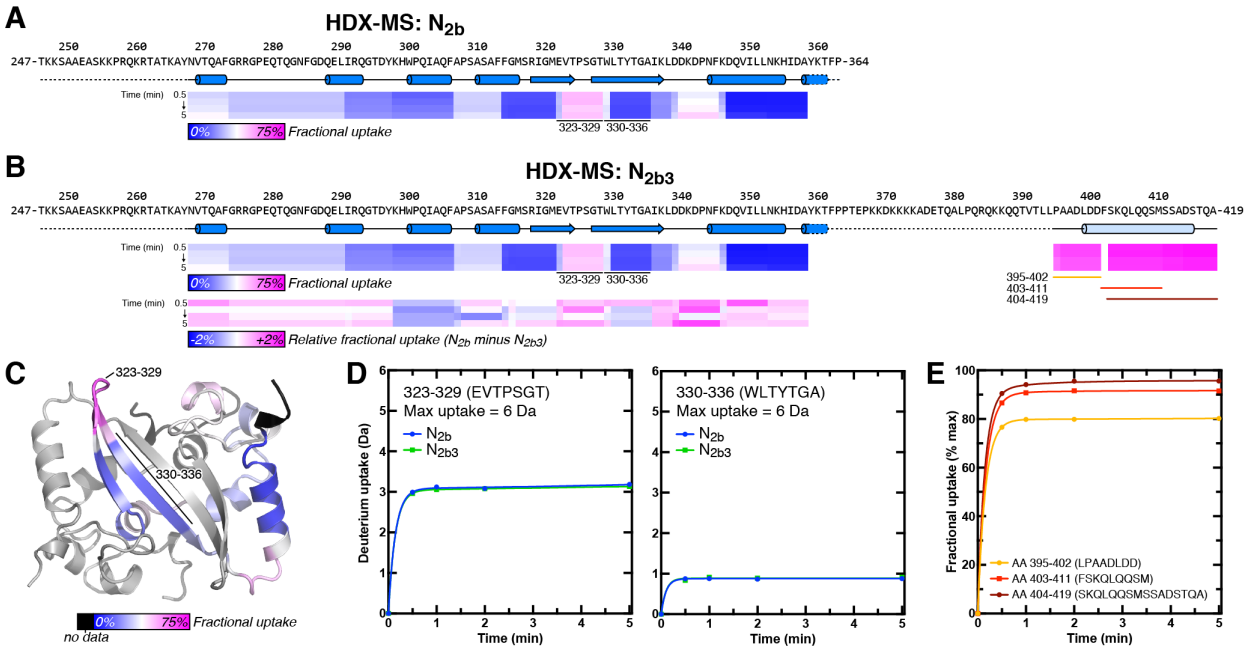

Figure 4. HDX-MS analysis of N_{2b} and N_{2b3}

(A) Schematic showing the N_{2b} sequence and structure, plus protein regions detected by HDX-MS. Each peptide is colored by its fractional deuterium uptake (not corrected for back-exchange) during the course of the experiment as indicated (time-points shown are 0.5, 1, 2, and 5 minutes).

(B) Schematic showing the N_{2b3} sequence and structure (the α-helix spanning residues 400-416 is predicted by PSI-PRED), plus protein regions detected by HDX-MS. Two sets of exchange rates are shown: fractional deuterium uptake in N_{2b3} (upper box) colored as in panel A, and relative exchange comparing N_{2b} and N_{2b3} (lower box).

(C) Structure of the N_{2b} dimer, with one monomer colored by fractional deuterium uptake as in panel A.

(D) Uptake plots (corrected for back-exchange) for two peptides within the ordered N2b domain, with uptake in N_{2b} indicated in blue and uptake in N_{2b3} indicated in green. The peptide covering residues 323-329 (located within a loop) is relatively exposed, while the peptide covering residues 330-336 (within a β-strand) is strongly protected from H-D exchange.

(E) Uptake plots for three peptides in the C-terminal region of N_{2b3}, plotted by fractional deuterium uptake (corrected for back-exchange). Peptides covering residues 395-402 (yellow) and 403-411 (red) show partial protection, suggesting that this region is structured, while the peptide covering residues 404-419 shows no protection. See **Figure S4** for each peptide plotted by absolute deuterium uptake.

Supplemental Figures

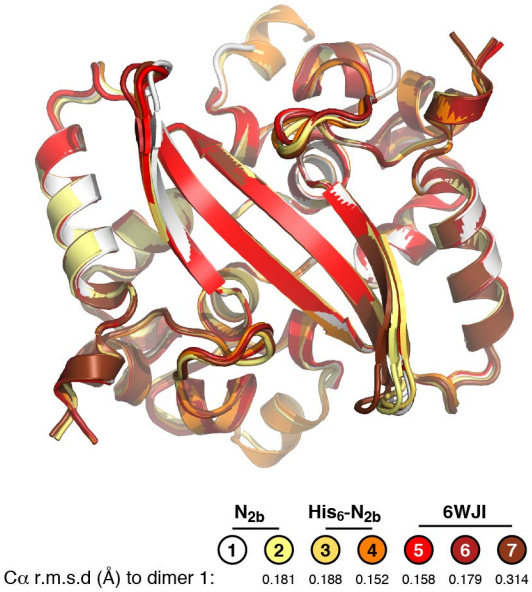

Figure S1. Overlay of SARS-CoV-2 N2b domain structures

Overlay of seven independent views of the SARS-CoV-2 N2b domain dimer, including the two structures determined in this work (dimer 1=untagged N_{2b} chains A+B; dimer 2=untagged N_{2b} chains C+D,dimer 3=His_6-tagged N_{2b} chains A+B, dimer 4=His_6-tagged N_{2b} chains C+D), and a recently-deposited structure of the same domain (PDB ID 6WJI; dimer 5=chains A+B, dimer 6=chains C+D, dimer 7=chains E+F). The overall Cα r.m.s.d values of all dimers overlaid on dimer 1 are shown at bottom.

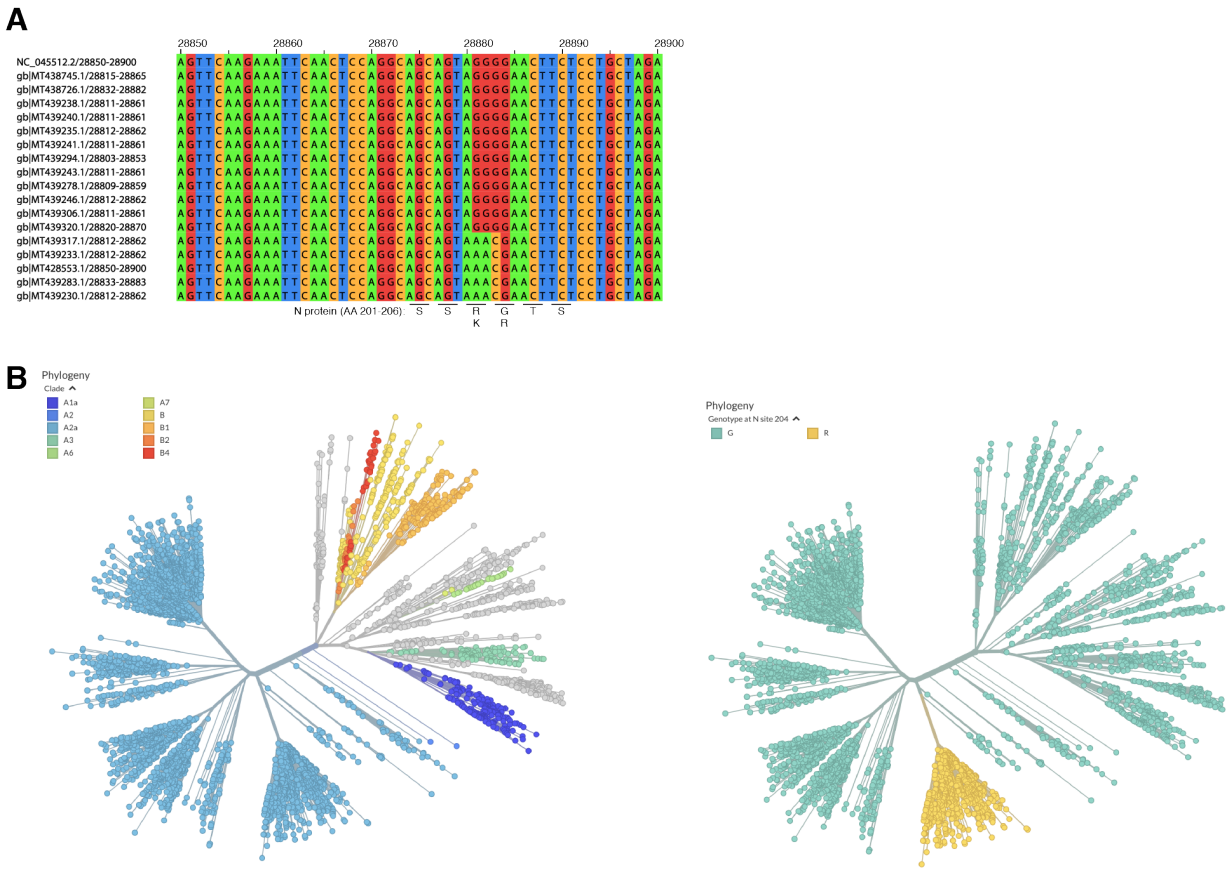

Figure S2. Distribution of a trinucleotide substitution in sequences SARS-CoV-2 isolates

(A) Sequence alignment of the SARS-CoV-2 reference genome (NCBI RefSeq NC_045512) and 17 selected sequences (from the NCBI Virus Resource; see **Methods**). Five genomes show the GGG→AAC trinucleotide substitution in position 28881-28883. Overall, roughly 10-15% of sequenced SARS-CoV-2 genomes show this trinucleotide substitution.

(B) Screenshots from the Nextstrain resource (56) showing an unrooted tree of SARS-CoV-2 genomes colored by clade assignment (left) or by colored by N protein residue 204 identity (right). The G204R mutation (yellow) present in a large fraction of SARS-CoV-2 clade A2a samples is diagnostic of the GGG→AAC trinucleotide substitution in position 28881-28883.

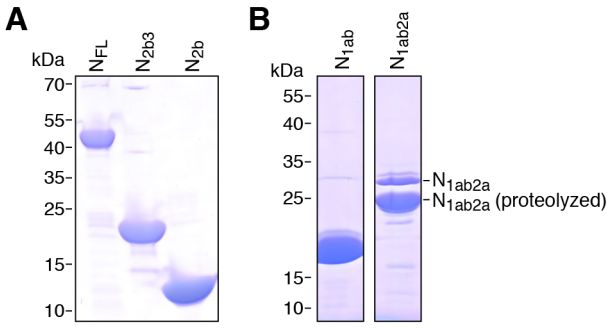

Figure S3. Purification of N protein constructs

SDS-PAGE analysis of purified SARS-CoV-2 N protein constructs: N_{FL} (residues 2-419), N_{1ab} (2-175), N_{1ab2a} (2-246), N_{2b} (247-364), and N_{2b3} (247-419). N_{1ab2a} shows evidence of proteolytic cleavage within the N2a region during purification.

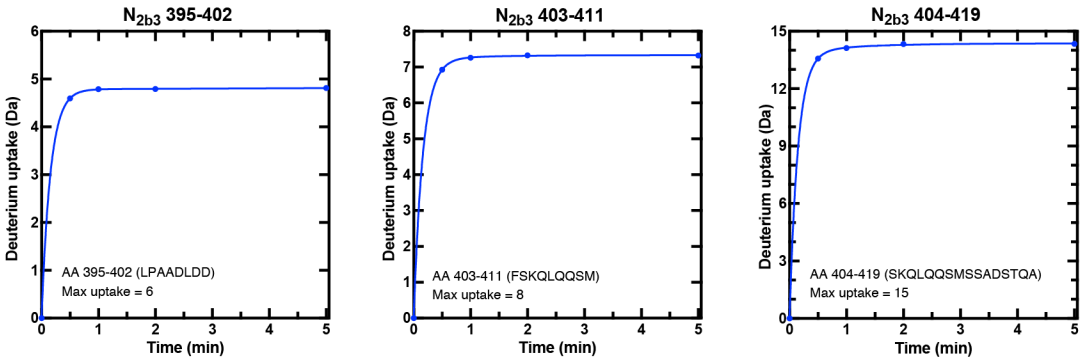

Figure S4. Uptake plots for N_{2b3} C-terminal peptides

Absolute deuterium uptake plots for three peptides in the extreme C-terminus of N_{2b3}, as shown in **Figure 4E**.

Supplemental Tables

Table S1. Crystallographic data collection and refinement

	N_{2b}	His_6-N_{2b}
Data collection		
Synchrotron/Beamline	APS 24ID-E	APS 24ID-E
Date collected	May 12, 2020	May 12, 2020
Resolution (Å)	66 – 1.42	74 – 1.45
Wavelength (Å)	0.97918	0.97918
Space Group	P1	$P2_1$
Unit Cell Dimensions (a, b, c) Å	43.72, 50.06, 69.34	71.15, 43.55, 74.23
Unit cell Angles (α,β,γ) °	106.50, 90.09, 97.14	90, 94.87, 90
I/σ (last shell)	16.8 (3.0)	16.0 (2.0)
[1]R_{sym} (last shell)	0.039 (0.310)	0.058 (0.802)
[2]R_{meas} (last shell)	0.046 (0.353)	0.063 (0.877)
[3]$CC_{1/2}$ (last shell)	0.999 (0.957)	0.998 (0.796)
Completeness (last shell) %	91.1 (75.0)	98.8 (91.9)
Number of reflections	352232	528562
unique	96446	80163
Multiplicity (last shell)	3.7 (3.6)	6.6 (6.0)
Refinement		
Resolution (Å)	66 – 1.42	74 – 1.45
No. of reflections	96393	80120
working	91660	76201
free	4733	3919
[4]R_{work} (last shell) (%)	15.73 (25.33)	16.73 (28.88)
[4]R_{free} (last shell) (%)	17.32 (24.32)	18.02 (30.96)
Structure/Stereochemistry		
No. of atoms	7526	7768
hydrogen	3410	3580
solvent	618	493
ligand	0	30
r.m.s.d. bond lengths (Å)	0.007	0.009
r.m.s.d. bond angles (°)	0.888	0.922
[5]SBGrid Data Bank ID	785	786
[6]Protein Data Bank ID	6WZO	6WZQ

[1]$R_{sym} = \sum\sum_j |I_j - \langle I \rangle| / \sum I_j$, where I_j is the intensity measurement for reflection j and $\langle I \rangle$ is the mean intensity for multiply recorded reflections.

[2]$R_{meas} = \sum_h [\sqrt{(n/(n\text{-}1))} \sum_j [I_{hj} - \langle I_h \rangle]] / \sum_{hj} \langle I_h \rangle$

where I_{hj} is a single intensity measurement for reflection h, $\langle I_h \rangle$ is the average intensity measurement for multiply recorded reflections, and n is the number of observations of reflection h.

[3]$CC_{1/2}$ is the Pearson correlation coefficient between the average measured intensities of two randomly-assigned half-sets of the measurements of each unique reflection. $CC_{1/2}$ is considered significant above a value of ~0.15.

[4]$R_{work,\ free} = \sum||F_{obs}| - |F_{calc}||/|F_{obs}|$, where the working and free R-factors are calculated using the working and free reflection sets, respectively.

[5]Diffraction data for each structure have been deposited with the SBGrid Data Bank (https://data.sbgrid.org) with the noted accession codes.

[6]Coordinates and structure factors for each structure have been deposited with the Protein Data Bank (http://www.pdb.org) with the noted accession codes.

Table S2. N protein variants in SARS-CoV-2 patient sequences

(see attached Excel spreadsheet)

References

1. C. Huang, *et al.*, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
2. N. Zhu, *et al.*, A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
3. C. Drosten, *et al.*, Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1967–1976 (2003).
4. T. G. Ksiazek, *et al.*, A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1953–1966 (2003).
5. A. M. Zaki, S. van Boheemen, T. M. Bestebroer, A. D. M. E. Osterhaus, R. A. M. Fouchier, Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *N. Engl. J. Med.* **367**, 1814–1820 (2012).
6. P. Zhou, *et al.*, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
7. P. Goyal, *et al.*, Clinical Characteristics of Covid-19 in New York City. *N. Engl. J. Med.*, https:/doi.org/10.1056/NEJMc2010419 (2020).
8. W. Guan, *et al.*, Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* (2020) https:/doi.org/10.1056/nejmoa2002032.
9. S. Bialek, *et al.*, Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020. *MMWR. Morb. Mortal. Wkly. Rep.* **69**, 343–346 (2020).
10. T. K. Warren, *et al.*, Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature* **531**, 381–385 (2016).
11. D. Siegel, *et al.*, Discovery and Synthesis of a Phosphoramidate Prodrug of a Pyrrolo[2,1-f][triazin-4-amino] Adenine C-Nucleoside (GS-5734) for the Treatment of Ebola and Emerging Viruses. *J. Med. Chem.* **60**, 1648–1661 (2017).
12. W. Yin, *et al.*, Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science (80-.).*, eabc1560 (2020).
13. M. L. Agostini, *et al.*, Coronavirus susceptibility to the antiviral remdesivir (GS-5734) is mediated by the viral polymerase and the proofreading exoribonuclease. *MBio* **9** (2018).
14. C. J. Gordon, E. P. Tchesnokov, J. Y. Feng, D. P. Porter, M. Gotte, The antiviral compound remdesivir potently inhibits RNA-dependent RNA polymerase from Middle East respiratory syndrome coronavirus. *J. Biol. Chem.* **295**, jbc.AC120.013056 (2020).
15. J. Grein, *et al.*, Compassionate Use of Remdesivir for Patients with Severe Covid-19. *N. Engl. J. Med.*, https:/doi.org/10.1056/NEJMoa2007016 (2020).
16. B. N. Williamson, *et al.*, Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. *bioRxiv*, 2020.04.15.043166 (2020).

17. L. Zhang, *et al.*, Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science (80-.).*, eabb3405 (2020).

18. A. C. Walls, *et al.*, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).

19. D. Wrapp, *et al.*, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (80-.).* **367**, 1260–1263 (2020).

20. R. Yan, *et al.*, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science (80-.).* **367**, 1444–1448 (2020).

21. M. Hoffmann, *et al.*, SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).

22. K. Duan, *et al.*, Effectiveness of convalescent plasma therapy in severe COVID-19 patients. *Proc. Natl. Acad. Sci. U. S. A.* (2020) https:/doi.org/10.1073/pnas.2004168117 (April 20, 2020).

23. P. S. Masters, The Molecular Biology of Coronaviruses. *Adv. Virus Res.* **65**, 193–292 (2006).

24. C. A. M. de Haan, P. J. M. Rottier, Molecular Interactions in the Assembly of Coronaviruses. *Adv. Virus Res.* **64**, 165–230 (2005).

25. A. R. Fehr, S. Perlman, "Coronaviruses: An overview of their replication and pathogenesis" in *Coronaviruses: Methods and Protocols*, (Springer New York, 2015), pp. 1–23.

26. M. Bárcena, *et al.*, Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 582–587 (2009).

27. A. Hachim, *et al.*, Beyond the Spike: identification of viral targets of the antibody response to SARS-CoV-2 in COVID-19 patients. *medRxiv*, 2020.04.30.20085670 (2020).

28. P. D. Burbelo, *et al.*, Detection of Nucleocapsid Antibody to SARS-CoV-2 is More Sensitive than Antibody to Spike Protein in COVID-19 Patients. *medRxiv*, 2020.04.20.20071423 (2020).

29. M. Surjit, B. Liu, P. Kumar, V. T. K. Chow, S. K. Lal, The nucleocapsid protein of the SARS coronavirus is capable of self-association through a C-terminal 209 amino acid interaction domain. *Biochem. Biophys. Res. Commun.* **317**, 1030–1036 (2004).

30. H. Luo, *et al.*, In vitro biochemical and thermodynamic characterization of nucleocapsid protein of SARS. *Biophys. Chem.* **112**, 15–25 (2004).

31. C. Y. Chen, *et al.*, Structure of the SARS Coronavirus Nucleocapsid Protein RNA-binding Dimerization Domain Suggests a Mechanism for Helical Packaging of Viral RNA. *J. Mol. Biol.* **368**, 1075–1086 (2007).

32. I. M. Yu, M. L. Oldham, J. Zhang, J. Chen, Crystal structure of the severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein dimerization domain reveals evolutionary linkage between Corona- and Arteriviridae. *J. Biol. Chem.* **281**, 17134–17139

(2006).

33. M. Takeda, *et al.*, Solution Structure of the C-terminal Dimerization Domain of SARS Coronavirus Nucleocapsid Protein Solved by the SAIL-NMR Method. *J. Mol. Biol.* **380**, 608–622 (2008).

34. H. Jayaram, *et al.*, X-Ray Structures of the N- and C-Terminal Domains of a Coronavirus Nucleocapsid Protein: Implications for Nucleocapsid Formation. *J. Virol.* **80**, 6612–6620 (2006).

35. T. H. Van Nguyen, *et al.*, Structure and oligomerization state of the C-terminal region of the Middle East respiratory syndrome coronavirus nucleoprotein. *Acta Crystallogr. Sect. D Struct. Biol.* **75**, 8–15 (2019).

36. B. Szelazek, *et al.*, Structural Characterization of Human Coronavirus NL63 N Protein. *J. Virol.* **91** (2017).

37. Y. Ma, *et al.*, Structures of the N- and C-terminal domains of MHV-A59 nucleocapsid protein corroborate a conserved RNA-protein binding mechanism in coronavirus. *Protein Cell* **1**, 688–697 (2010).

38. H. Luo, J. Chen, K. Chen, X. Shen, H. Jiang, Carboxyl terminus of severe acute respiratory syndrome coronavirus nucleocapsid protein: Self-association analysis and nucleic acid binding characterization. *Biochemistry* **45**, 11827–11835 (2006).

39. C. Chang, C.-M. M. Chen, M. Chiang, Y. Hsu, T. Huang, Transient Oligomerization of the SARS-CoV N Protein – Implication for Virus Ribonucleoprotein Packaging. *PLoS One* **8**, e65045 (2013).

40. Y.-S. Lo, *et al.*, Oligomerization of the carboxyl terminal domain of the human coronavirus 229E nucleocapsid protein. *FEBS Lett.* **587**, 120–127 (2013).

41. H. Fan, *et al.*, The nucleocapsid protein of coronavirus infectious bronchitis virus: Crystal structure of its N-terminal domain and multimerization properties. *Structure* **13**, 1859–1868 (2005).

42. Y. Cong, F. Kriegenburg, C. A. M. De Haan, F. Reggiori, Coronavirus nucleocapsid proteins assemble constitutively in high molecular oligomers. *Sci. Rep.* **7**, 1–10 (2017).

43. K. R. Hurst, R. Ye, S. J. Goebel, P. Jayaraman, P. S. Masters, An Interaction between the Nucleocapsid Protein and a Component of the Replicase-Transcriptase Complex Is Crucial for the Infectivity of Coronavirus Genomic RNA. *J. Virol.* **84**, 10276–10288 (2010).

44. K. R. Hurst, C. A. Koetzner, P. S. Masters, Identification of In Vivo-Interacting Domains of the Murine Coronavirus Nucleocapsid Protein. *J. Virol.* **83**, 7221–7234 (2009).

45. S. Kang, *et al.*, Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *bioRxiv*, 2020.03.06.977876 (2020).

46. D. C. Dinesh, D. Chalupska, J. Silhan, V. Veverka, E. Boura, Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *bioRxiv*, 2020.04.02.022194 (2020).

47. D. W. A. Buchan, F. Minneci, T. C. O. Nugent, K. Bryson, D. T. Jones, Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349--57 (2013).

48. J. E. Tropea, S. Cherry, D. S. Waugh, Expression and purification of soluble His(6)-tagged TEV protease. *Methods Mol. Biol. (Clifton, NJ)* **498**, 297–307 (2009).

49. R. J. Lumpkin, E. A. Komives, DECA, a comprehensive, automatic post-processing program for HDX-MS data. *Mol. Cell. Proteomics* **18**, 2516–2523 (2019).

50. W. Kabsch, XDS. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **66**, 125–132 (2010).

51. P. R. Evans, G. N. Murshudov, How good are my data and what is the resolution? *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **69**, 1204–1214 (2013).

52. M. D. Winn, *et al.*, Overview of the CCP4 suite and current developments. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **67**, 235–242 (2011).

53. A. J. McCoy, *et al.*, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

54. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **66**, 486–501 (2010).

55. P. V Afonine, *et al.*, Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. Sect. D, Biol. Crystallogr.* **68**, 352–367 (2012).

56. J. Hadfield, *et al.*, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

57. C. D. Livingstone, G. J. Barton, Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–56 (1993).

58. D. T. Jones, D. Cozzetto, DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).