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

Journal
Protein Science, 29(9)

ISSN
0961-8368

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

Publication Date
2020-09-01

DOI
10.1002/pro.3909

Peer reviewed

DOI10.1002/pro.3909

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

Received: Jun 06, 2020;Revised: Jul 02, 2020;Accepted: Jul 03, 2020

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 protein, which packages viral RNA into new virions. We determined a 1.4 Å resolution crystal structure of this protein's N2b domain, revealing a compact, intertwined dimer similar to that of related coronaviruses including SARS-CoV. While the N2b domain forms a dimer in solution, addition of the C-terminal spacer B/N3 domain mediates formation of a homotetramer. 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 self-associates and cooperates with the N2b domain to mediate tetramer formation. Finally, we map the locations of amino acid substitutions in the N protein from over 38,000 SARS-CoV-2 genome sequences. We find that these substitutions are strongly clustered in the protein's N2a linker domain, and that substitutions 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, with implications for both drug design and antibody-based testing.

Keywords: SARS-CoV-2, Coronavirus, COVID-19, nucleocapsid, crystal structure

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 10.5 million people and caused more than 500,000 deaths as of July 1, 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 interacts with the membrane (M) protein and is packaged into virions. Electron microscopy analysis of related betacoronaviruses has suggested that the RNP complex adopts a helical filament structure,[23–28] but recent cryo-electron tomography analysis of intact SARS-CoV-2 virions has revealed a beads-on-a-string like arrangement of globular RNP complexes that sometimes assemble into stacks resembling helical filaments.[29] 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 to the surface spike protein.[30,31] 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 the N protein into higher-order RNP complexes is not well understood, but likely involves cooperative interactions between the dimerization domain and other regions of the protein, plus the bound RNA.[32–40] 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 or N-terminal domain/NTD) and dimerization (N2b or C-terminal domain/CTD) domains separated by short regions with high predicted disorder [N1a, N2a, and spacer B/N3; Fig. 1(A)]. 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,[32] and the purified full-length protein was shown to self-associate into predominantly dimers in solution.[33] The structures of the N2b domain of SARS-CoV and several related coronaviruses confirmed the obligate homodimeric structure of this domain,[34–40] and other work showed that the region C-terminal to this domain mediates further self-association into tetramer, hexamer, and potentially higher oligomeric forms.[41–43] Other studies have suggested that the protein's N-terminal region, including the RNA-binding N1b domain, can also self-associate,[44,45] 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).[46,47] 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 Materials and 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 [Fig. 1(B)]. 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 contain two N_{2b} dimers, giving four crystallographically independent views of the N_{2b} dimer. These four dimers differ only slightly, showing overall $C\alpha$ r.m.s.d values of 0.15-0.19 Å and with most variation arising from loop regions [Fig. S1(A)]. Our structures also overlay closely with four other recently-deposited structures of the SARS-CoV-2 N2b domain (PDB IDs 6WJI, 6YUN, 6ZCO, and 7C22; all unpublished). One of these structures (PDB ID 7C22) adopts the same space group and unit cell parameters as our structure of untagged N_{2b}. Including all of these structures, there are now nine independent crystallographic views of the SARS-CoV-2 N2b domain dimer (17 total protomers; the 6ZCO dimer is assembled from crystal symmetry) in five crystal forms at pH 4.5, 7.5, 7.8, and 8.5 (crystallization pH for 6ZCO is not reported). All of these structures overlay closely, with an overall $C\alpha$ r.m.s.d of 0.15-0.31 Å [Fig. S1(A)].

The structure of N_{2b} closely resembles that of related coronaviruses, including SARS-CoV, Infectious Bronchitis Virus (IBV), MERS-CoV, and HCoV-NL63.[34–39] 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 $C\alpha$ r.m.s.d of 0.314 Å across the N_{2b} dimer [Fig. 1(C)].

A crystal structure of the SARS-CoV N protein revealed a helical assembly of N2b domain dimers that was proposed as a possible structure for the observed helical nucleocapsid filaments in virions.[34] We therefore examined the packing of N2b domain dimers in the six crystal structures of this domain, five of which show distinct space groups and unit cell parameters. We identified two dimer-dimer packing modes that appear in multiple crystal forms, with packing mode 1 appearing in five structures, and packing mode 2 appearing in four [Fig. S1(B)]. Neither of these packing modes would result in the assembly of a helical filament if repeated, nor do the dimer-dimer interfaces strongly correlate with conserved surfaces on the N2b domain. This evidence, combined with our finding that N_{2b} forms solely dimers in solution (see below), suggests that packing of N2b domain dimers does not underlie higher-order assembly of SARS-CoV-2 N protein filaments.

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] over 38,000 full genomic sequences have been deposited in public databases (as of June 4, 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 38,318 genome sequences from the China National Center for Bioinformation, 2019 Novel Coronavirus Resource. Among these sequences, there are 10,979 instances of amino acid substitutions spread across 250 of the 419 amino acids of the N protein [Fig. 2(A), Table S2]. While many of these substitutions arise only once in our dataset and may therefore reflect errors in sequencing or sequence assembly, most likely reflect true variation among circulating strains of SARS-CoV-2. As a whole, the reported substitutions are enriched in the three intrinsically disordered domains (N1a, N2a, and spacer B/N3), with a particularly high density of substitutions in the serine/arginine-rich subdomain of N2a [SR in Fig. 2(A)]. 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 [~4,100 of the 38,318 sequences in our dataset; Fig. S2(A), S2(B)]. While positions 203 and 204 accounted for over two-thirds of the total individual amino acid substitutions in this dataset, the N2a domain shows a strong enrichment of mutations even when these positions are not considered [Fig. 2(A)]. In contrast to the enrichment of missense mutations in the N2a domain, synonymous mutations were distributed relatively equally throughout the protein [Fig. S2(C), Table S2]. 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 345 instances of amino acid variants in the RNA-binding N1b domain, and 315 instances in the N2b domain. We mapped these onto high-resolution structures of both domains [Fig. 2(B), 2(C)]. Two high-resolution crystal structures of the SARS-CoV-2 N1b domain have been determined (PDB ID 6M3M and 6VYO),[48] and a recent NMR study determined a solution structure of the domain and defined its likely RNA binding surface [Fig. 2(B)].[49] In keeping with its functional importance, the identified RNA binding surface shows only a single mutation in this dataset [Fig. 2(B); 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 [Fig. 2(C)].

Finally, the 38,318 SARS-CoV-2 genome sequences contain nine sequences with reported nonsense/premature stop codons in the N protein. Two of these are located at position 256 within the N2b domain, while the remaining seven are located in the spacer B/N3 region between positions 372-418 [Fig. 2(A)].

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 systematically investigated the molecular basis for higher-order self-assembly of the SARS-CoV-2 nucleocapsid. We first purified the full-length N protein (N_{FL}) for analysis of its oligomeric state. While our initial attempts at purification yielded large aggregates significantly contaminated with nucleic acid [Fig. S3(A)], purification of the protein in high-salt buffer (1M NaCl) and in the presence of both DNase and RNase yielded pure N_{FL} [Fig. S3(B)]. Size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) of purified N_{FL} revealed a heterogeneous population that is predominantly homotetrameric [Fig. 3(A)].

To determine the molecular basis for homotetramer assembly, we purified a series of truncation constructs encompassing the ordered N1b and N2b domains and their associated linker domains [N1a, N2a, and spacer B/N3; Fig. 1(A)]. We characterized four truncations encompassing the protein's N-terminal regions, including N_{1ab} (residues 2-175), N_{1b} (residues 49-175), N_{1ab2a} (residues 2-246), and N_{1b2a} (residues 49-246). All four of these truncations are monomeric in solution as determined by SEC-MALS [Fig. 3(B), 3(C), S3(C), S3(D)]. We next analyzed N2b, which forms a homodimer in our crystal structures. As expected, N2b is dimeric in solution [Fig. 3(D)].

Finally, we analyzed the contribution of the C-terminal spacer B/N3 region to N protein self-assembly. Prior work with the Murine Hepatitis Virus (MHV) N protein showed that this region can, on its own, incorporate into nucleocapsid structures that lack the associated Membrane (M) protein, suggesting that the region mediates a homotypic interaction between N proteins.[47] Other work with SARS-CoV and HCoV-229E N proteins also found that the C-terminal spacer B/N3 region is required for higher-order assembly of tetramers and larger oligomers.[41–43] We purified a construct encoding N2b and the spacer B/N3 region (N_{2b3}, residues 247-419) and found that it forms a homotetramer [Fig. 3(E)]. We also analyzed self-assembly of the spacer B/N3 region on its own by performing SEC-MALS analysis on this isolated region (N_3, residues 365-419) fused to a His_6-maltose binding protein (MBP) tag. Initial purification of His_6-MBP-N_3 yielded two peaks on the final size exclusion column, which we separated pooled and analyzed by SEC-MALS. We found that these two peaks correspond to a monomer and a dimer, respectively [Fig. 3(F)]. The pooled dimer population also showed a small population of potentially higher-order oligomers [Fig. 3(F)]. Together, these data suggest that assembly of betacoronavirus N protein filaments 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 [Fig. 3(G), 3(H)]; and (3) oligomer/filament assembly mediated by cooperative RNA binding and potential higher-order self-association of N homotetramers.

To gain structural insight into how the spacer B/N3 region mediates N protein self-association, we performed hydrogen-deuterium exchange mass spectrometry (HDX-MS) on N_{2b} and N_{2b3} (Fig. 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 [Fig. 4(A), 4(C), 4(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 [Fig. 4(B)]. 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; Fig. 4(E)]. This finding suggests that at least part of the spacer B/N3 domain possesses secondary structure and may mediate N_{2b3} tetramer formation. Indeed, a recent molecular dynamics simulation of the N3 domain suggests the existence of an α-helix spanning residues 400-411 in this domain,[50] and our own analysis using the PSI-PRED server[51] suggests an α-helix spanning residues 400-416 [Fig. 1(A)].

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. Contrary to this expectation, we found that the H-D exchange rates within the N2b domain were nearly identical between the two constructs, varying at most 2% in fractional deuterium uptake in individual peptides ([Fig. 4(B), 4(D)]. While these data do not rule out the possibility that the spacer B/N3 region docks against N2b, they nonetheless support our SEC-MALS data showing that spacer B/N3 independently self-associates to mediate N protein tetramer formation. We attempted to test this idea by measuring the association of His_6-MBP-N_3 with N_{2b}, N_{2b3}, and N_{FL} in a pulldown assay [Fig. S4(B)]. We were unable to detect binding of His_6-MBP-N_3 to any of these three constructs. As both N_{2b3} and N_{FL} likely exist as pre-formed tetramers, their failure to interact with additional His_6-MBP-N_3 is not surprising. The inability of N_{2b} to interact with His_6-MBP-N_3, however, is consistent with the idea that the spacer B/N3 domain self-interacts rather than binding the N2b domain.

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. We envision two possible modes of N protein tetramer assembly based on either parallel or antiparallel arrangement of the putative α-helices in the N3 domain [Fig. 3(H)]. How these tetramers interact with viral RNA and self-assemble into either helical filaments or the more recently-observed globular viral RNP complexes[29] will require higher-level reconstitution and/or high-resolution analysis of the internal structure of SARS-CoV-2 virions.

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 betacoronaviruses, 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 over 38,000 individual genome sequences deposited in public databases as of early June, 2020. We used this set of genome sequences to identify over 10,000 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 ~650 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.

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 N1b and N2b domains 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.

Materials and 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_{1b} (49-175), N_{1b2a} (49-246), 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*. N_3 (residues 365-419) was similarly inserted into UC Berkeley Macrolab vector 2C-T (AmpR, N-terminal His_6-MBP fusion; Addgene #29706). 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.[52] 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.[53] 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[54] for indexing and data reduction, and the CCP4 programs AIMLESS[55] and TRUNCATE[56] for scaling and conversion to structure factors. The structure of untagged N_{2b} was determined by molecular replacement in PHASER[57] using a dimer of the SARS-CoV N2b domain (PDB ID 2GIB)[35] as a template. The resulting model was manually rebuilt in COOT[58] and refined in phenix.refine[59] 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.

Nickel pulldown

Nickel pulldown assays were performed in buffer A plus 300 mM NaCl and 10 mM imidazole pH 8.0. Ten μg bait (His_6-MBP-N_3) was mixed with equal weights of each prey protein in 50 μl total reaction volume and incubated on ice for 30 minutes (5 μl “load” sample removed). 30 μl of a

50% slurry of Ni-NTA Superflow beads (Qiagen) was added and the mixture was incubated with occasional mixing on ice for 30 minutes. Beads were washed three times with 1 mL buffer, then bound proteins were eluted with the addition of 30 μl buffer A plus 300 mM NaCl and 250 mM imidazole pH 8.0. Ten μl of each eluate was analyzed by SDS-PAGE alongside the load samples.

Bioinformatics

To examine variation in SARS-CoV-2 sequences, we downloaded a list of variants in the N gene in 38,318 genome sequences from China National Center for Bioinformation, 2019 Novel Coronavirus Resource (https://bigd.big.ac.cn/ncov?lang=en; downloaded June 3, 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 S2(A). We used the NextStrain database[60] 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 July 2, 2020).

Supplementary Material

Supplementary material for this article includes four Supplemental Figures (Figures S1-S4) and two Supplemental Tables (Tables S1 and S2). Figures S1-S4 and Table S1 can be found in the combined Supplementary Material PDF file, and Table S2 can be found in a separate Microsoft Excel file.

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, R. Lumpkin for assistance with the DECA software, 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.

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Chang Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395:497–506.

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W (2020) A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733.

3. Drosten C, Günther S, Preiser W, Van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RAM, Bergere A, Burguière A-M, Cinatl J, Eickmann M, Escriou N, Grywna K, Kramme S, Manuguerra J-C, Müller S, Rickerts V, Stürmer M, Vieth S, Klenk H-D, Osterhaus ADME, Schmitz H, Doerr HW (2003) Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 348:1967–1976.

4. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim W, Rollin PE, Dowell SF, Ling A-E, Humphrey CD, Shieh W-J, Guarner J, Paddock CD, Rota P, Fields B, DeRisi J, Yang J-Y, Cox N, Hughes JM, LeDuc JW, Bellini WJ, Anderson LJ, SARS Working Group (2003) A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med 348:1953–1966.

5. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM (2012) Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 367:1814–1820.

6. Zhou P, Yang X Lou, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273.

7. Goyal P, Choi JJ, Pinheiro LC, Schenck EJ, Chen R, Jabri A, Satlin MJ, Campion TR, Nahid M, Ringel JB, Hoffman KL, Alshak MN, Li HA, Wehmeyer GT, Rajan M, Reshetnyak E, Hupert N (2020) Clinical characteristics of Covid-19 in New York City. N Engl J Med 382:2372-2374.

8. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, Liu L, Shan H, Lei C, Hui DSC, Du B, Li L, Zeng G, Yuen K-Y, Chen R, Tang C, Wang T, Chen P, Xiang J, Li S, Wang J-l, Liang Z, Peng Y, Wei L, Liu Y, Hu Y-h, Peng P, Wang J-m, Liu J, Chen Z, Li G, Zheng Z, Qiu S, Luo J, Ye C, Zhu S, Zhong N (2020) Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 382:1708–1720.

9. Bialek S, Boundy E, Bowen V, Chow N, Cohn A, Dowling N, Ellington S, Gierke R, Hall A, MacNeil J, et al. (2020) Severe outcomes among patients with coronavirus disease 2019

(COVID-19) — United States, February 12–March 16, 2020. Morb Mortal Wkly Rep 69:343–346.

10. Warren TK, Jordan R, Lo MK, Ray AS, Mackman RL, Soloveva V, Siegel D, Perron M, Bannister R, Hui HC, Larson N, Strickley R, Wells J, Stuthman KS, Van Tongeren SA, Garza NL, Donnelly G, Shurtleff AC, Retterer CJ, Gharaibeh D, Zamani R, Kenny T, Eaton BP, Grimes E, Welch LS, Gomba L, Wilhelmsen CL, Nichols DK, Nuss JE, Nagle ER, Kugelman JR, Palacios G, Doerffler E, Neville S, Carra E, Clarke MO, Zhang L, Lew W, Ross B, Wang Q, Chun K, Wolfe L, Babusis D, Park Y, Stray KM, Trancheva I, Feng JY, Barauskas O, Xu Y, Wong P, Braun MR, Flint M, McMullan LK, Chen S-S, Fearns R, Swaminathan S, Mayers DL, Spiropoulou CF, Lee WA, Nichol ST, Cihlar T, Bavari S (2016) Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. Nature 531:381–385.

11. Siegel D, Hui HC, Doerffler E, Clarke MO, Chun K, Zhang L, Neville S, Carra E, Lew W, Ross B, Wang Q, Wolfe L, Jordan R, Soloveva V, Knox J, Perry J, Perron M, Stray KM, Barauskas O, Feng JY, Xu Y, Lee G, Rheingold AL, Ray AS, Bannister R, Strickley R, Swaminathan S, Lee WA, Bavari S, Cihlar T, Lo MK, Warren TK, Mackman RL (2017) 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.

12. Yin W, Mao C, Luan X, Shen D-D, Shen Q, Su H, Wang X, Zhou F, Zhao W, Gao M, Chang S, Xie Y-C, Tian G, Jiang H-W, Tao S-C, Shen J, Jiang Y, Jiang H, Xu Y, Zhang S, Zhang Y, Xu HE (2020) Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. Science 368:1499-1504.

13. Agostini ML, Andres EL, Sims AC, Graham RL, Sheahan TP, Lu X, Smith EC, Case JB, Feng JY, Jordan R, Ray AS, Cihlar T, Siegel D, Mackman RL, Clarke MO, Baric RS, Denison MR (2018) Coronavirus susceptibility to the antiviral remdesivir (GS-5734) is mediated by the viral polymerase and the proofreading exoribonuclease. MBio 9:e00221-18.

14. Gordon CJ, Tchesnokov EP, Feng JY, Porter DP, Götte M (2020) The antiviral compound remdesivir potently inhibits RNA-dependent RNA polymerase from Middle East respiratory syndrome coronavirus. J Biol Chem 295:4773–4779.

15. Grein J, Ohmagari N, Shin D, Diaz G, Asperges E, Castagna A, Feldt T, Green G, Green ML, Lescure F-X, Nicastri E, R. Oda, Yo K, Quiros-Roldan E, Studemeister A, Redinski J, Ahmed S, Bernett J, Chelliah D, Chen D, Chihara S, Cohen SH, Cunningham J, D'Arminio Monforte A, Ismail S, Kato H, Lapadula G, L'Her E, Maeno T, Majumder S, Massari M, Mora-Rillo M, Mutoh Y, Nguyen D, Verweij E, Zoufaly A, Osinusi AO, DeZure A, Zhao Y, Zhong L, Chokkalingam A, Elboudwarej E, Telep L, Timbs L, Henne I, Sellers S, Cao H, Tan SK, Winterbourne L, Desai P, Mera R, Gaggar A, Myers RP, Brainard DM, Childs R, Flanigan T (2020) Compassionate use of Remdesivir for patients with severe Covid-19. N Engl J Med 382:2327-2336.

16. Williamson BN, Feldmann F, Schwarz B, Meade-White K, Porter DP, Schulz J, Doremalen N van, Leighton I, Yinda CK, Pérez-Pérez L, Okumura A, Lovaglio J, Hanley PW, Saturday G, Bosio CM, Anzick S, Barbian K, Cihlar T, Martens C, Scott DP, Munster VJ, de Wit E (2020) Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. bioRxiv:2020.04.15.043166.

17. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, Becker S, Rox K, Hilgenfeld R (2020) Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science 368:409-412.

18. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181:281-292.

19. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367:1260–1263.

20. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367:1444–1448.

21. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu NH, Nitsche A, Müller MA, Drosten C, Pöhlmann S (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181:271-280.

22. Duan K, Liu B, Li C, Zhang H, Yu T, Qu J, Zhou M, Chen L, Meng S, Hu Y, Peng C, Yuan M, Huang J, Wang Z, Yu J, Gao X, Wang D, Yu X, Li L, Zhang J, Wu X, Li B, Xu Y, Chen W, Peng Y, Hu Y, Lin L, Liu X, Huang S, Zhou Z, Zhang L, Wang Y, Zhang Z, Deng K, Xia Z, Gong Q, Zhang W, Zheng X, Liu Y, Yang H, Zhou D, Yu D, Hou J, Shi Z, Chen S, Chen Z, Zhang X, Yang X (2020) Effectiveness of convalescent plasma therapy in severe COVID-19 patients. Proc Natl Acad Sci USA 117:9490–9496.

23. Davies HA, Dourmashkin RR, Macnaughton MR (1981) Ribonucleoprotein of avian infectious bronchitis virus. J Gen Virol 53:67–74.

24. Macnaughton MR, Davies HA, Nermut M V. (1978) Ribonucleoprotein-like structures from coronavirus particles. J Gen Virol 39:545–549.

25. Masters PS (2006) The molecular biology of coronaviruses. Adv Virus Res 65:193–292.

26. de Haan CAM, Rottier PJM (2005) Molecular interactions in the assembly of coronaviruses. Adv Virus Res 64:165–230.

27. Fehr AR, Perlman S Coronaviruses: An overview of their replication and pathogenesis. In: Coronaviruses: Methods and Protocols. Vol. 1282. Springer New York; 2015. pp. 1–23.

28. Bárcena M, Oostergetel GT, Bartelink W, Faas FGA, Verkleij A, Rottier PJM, Koster AJ, Bosch BJ (2009) Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. Proc Natl Acad Sci USA 106:582–587.

29. Klein S, Cortese M, Winter SL, Wachsmuth-Melm M, Neufeldt CJ, Cerikan B, Stanifer ML, Boulant S, Bartenschlager R, Chlanda P (2020) SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. bioRxiv:2020.06.23.167064.

30. Hachim A, Kavian N, Cohen CA, Chin AW, Chu DK, Mok CKP, Tsang OT, Yeung YC, Perera RA, Poon LL, Peiris MJS, Valkenburg SA. (2020) Beyond the Spike: identification of viral targets of the antibody response to SARS-CoV-2 in COVID-19 patients. medRxiv:2020.04.30.20085670.

31. Burbelo PD, Riedo FX, Morishima C, Rawlings S, Smith D, Das S, Strich JR, Chertow DS, Davey RT, Cohen JI (2020) 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.

32. Surjit M, Liu B, Kumar P, Chow VTK, Lal SK (2004) 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.

33. Luo H, Ye F, Sun T, Yue L, Peng S, Chen J, Li G, Du Y, Xie Y, Yang Y, Shen J, Wang Y, Shen X, Jiang H (2004) In vitro biochemical and thermodynamic characterization of nucleocapsid protein of SARS. Biophys Chem 112:15–25.

34. Chen CY, Chang C ke, Chang YW, Sue SC, Bai HI, Riang L, Hsiao CD, Huang T huang (2007) 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.

35. Yu IM, Oldham ML, Zhang J, Chen J (2006) 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.

36. Takeda M, Chang C ke, Ikeya T, Güntert P, Chang Y hsiang, Hsu Y lan, Huang T huang, Kainosho M (2008) 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.

37. Jayaram H, Fan H, Bowman BR, Ooi A, Jayaram J, Collisson EW, Lescar J, Prasad BV (2006) X-Ray structures of the N- and C-terminal domains of a coronavirus nucleocapsid protein: Implications for nucleocapsid formation. J Virol 80:6612–6620.

38. Nguyen TH Van, Lichière J, Canard B, Papageorgiou N, Attoumani S, Ferron F, Coutard B (2019) Structure and oligomerization state of the C-terminal region of the Middle East respiratory syndrome coronavirus nucleoprotein. Acta Cryst D75:8–15.

39. Szelazek B, Kabala W, Kus K, Zdzalik M, Twarda-Clapa A, Golik P, Burmistrz M, Florek D,

Wladyka B, Pyrc K, Dubin G (2017) Structural characterization of human coronavirus NL63 N protein. J Virol 91:e02503-16.

40. Ma Y, Tong X, Xu X, Li X, Lou Z, Rao Z (2010) 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.

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

42. Chang C, Chen C-MM, Chiang M, Hsu Y, Huang T (2013) Transient oligomerization of the SARS-CoV N protein – implication for virus ribonucleoprotein packaging. PLoS One 8:e65045.

43. Lo Y-S, Lin S-Y, Wang S-M, Wang C-T, Chiu Y-L, Huang T-H, Hou M-H (2013) Oligomerization of the carboxyl terminal domain of the human coronavirus 229E nucleocapsid protein. FEBS Lett 587:120–127.

44. Fan H, Ooi A, Tan YW, Wang S, Fang S, Liu DX, Lescar J (2005) The nucleocapsid protein of coronavirus infectious bronchitis virus: Crystal structure of its N-terminal domain and multimerization properties. Structure 13:1859–1868.

45. Cong Y, Kriegenburg F, De Haan CAM, Reggiori F (2017) Coronavirus nucleocapsid proteins assemble constitutively in high molecular oligomers. Sci Rep 7:1–10.

46. Hurst KR, Ye R, Goebel SJ, Jayaraman P, Masters PS (2010) 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.

47. Hurst KR, Koetzner CA, Masters PS (2009) Identification of in vivo-interacting domains of the murine coronavirus nucleocapsid protein. J Virol 83:7221–7234.

48. 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 Pharm Sin B https://doi.org/10.1016/j.apsb.2020.04.009.

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

50. Cubuk J, Alston JJ, Jeremías Incicco J, Singh S, Stuchell-Brereton MD, Ward MD, Zimmerman MI, Vithani N, Griffith D, Wagoner JA, Bowman GR, Hall KB, Soranno A, Holehouse AS (2020) The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. bioRxiv:2020.06.17.158121.

51. Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 41:W349-357.

52. Tropea JE, Cherry S, Waugh DS (2009) Expression and purification of soluble His(6)-tagged TEV protease. Methods Mol Biol 498:297–307.

53. Lumpkin RJ, Komives EA (2019) DECA, a comprehensive, automatic post-processing program for HDX-MS data. Mol Cell Proteomics 18:2516–2523.

54. Kabsch W (2010) XDS. Acta Cryst D66:125–132.

55. Evans PR, Murshudov GN (2013) How good are my data and what is the resolution? Acta Cryst D69:1204–1214.

56. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS (2011) Overview of the CCP4 suite and current developments. Acta Cryst D67:235-242.

57. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Cryst 40:658–674.

58. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Cryst D66:486–501.

59. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with *phenix.refine*. Acta Cryst D68:352-367.

60. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callendar C, Sagulenko P, Bedford T, Neher RA (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123.

61. Livingstone CD, Barton GJ (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 9:745–56.

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

Figure Legends

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

(A) Domain structure of the SARS-CoV-2 Nucleocapsid protein, as defined previously,[46,47] with plot showing the Jalview alignment conservation score (three-point smoothed; gray)[61] and DISOPRED3 disorder propensity (red)[62] 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.[47] All purified truncations are noted at bottom.

(B) Top-down view 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. See Figure S1(A) for comparison with other structures of this domain.

(C) Structural overlay of the SARS-CoV-2 N_{2b} dimer (blue) and the equivalent domain of SARS-CoV-N (PDB ID 2CJR).[34]

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 4,000 times due to a prevalent trinucleotide substitution [Fig. S2(A), S2(B)]. Red tick marks indicate the locations of seven premature stop mutations detected (two sequences contained stop codons at residue 256; not graphed). *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).[49] At right, two views colored by the number of variants at each position observed in a set of 38,318 SARS-CoV-2 genomes. The two most frequently-mutated residues are shown in stick view and labeled. Only one mutation (A50E, observed in one sequence) overlaps the putative RNA binding surface.

(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 38,318 SARS-CoV-2 genomes. The four 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 (Superose 6 Increase 10/300 GL; void volume=8.4 mL, total volume=20.5 mL) coupled to multi-angle light scattering (SEC-MALS) analysis of full-length SARS-CoV-2 N. The measured MW of 190.0 kDa closely matches that of a tetramer (182.5 kDa). See Figure S3(B) for SDS-PAGE analysis of all purified proteins.

(B) Size exclusion chromatography (Superdex 200 Increase 10/300 GL; void volume=7.3 mL, total volume=20.6 mL; used for panels B-F) 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 [Fig. S3(B)].

(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) SEC-MALS analysis of MBP-SARS-CoV-2 N_3 ("peak 1" black/dark blue; "peak 2" gray/light blue) The measured MW of peak 1 (101.9 kDa) and peak 2 (48.9 kDa) closely match those of a homodimer (101.7 kDa) and a monomer (50.9 kDa). The small peak at 10.5 mL suggests higher-order self-assembly.

(G) Schematic summary of size exclusion and SEC-MALS results on N protein constructs. See Figures S3(C) and S3(D) for SEC-MALS analysis of N_{1b} (residues 49-174) and N_{1b2a} (residues 49-246).

(H) Possible configurations of a SARS-CoV-2 N protein tetramer. Dimerization is mediated by the N2b domain, and these dimers self-associate through the N3 region to form homotetramers. *Left:* Parallel arrangement of the putative N3 domain α-helices; *Right:* antiparallel arrangement.

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

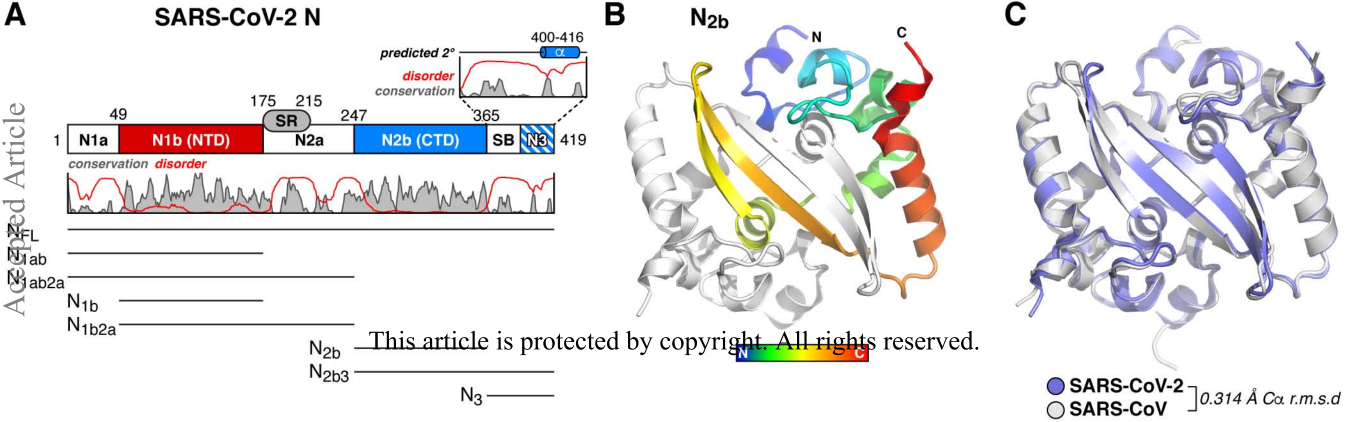
(A) Schematic showing the N_{2b} sequence and structure, plus protein regions detected by HDX-MS. Each peptide is colored by its fractional deuterium uptake during the course of the experiment (blue-white-magenta = 0-100% fractional uptake).

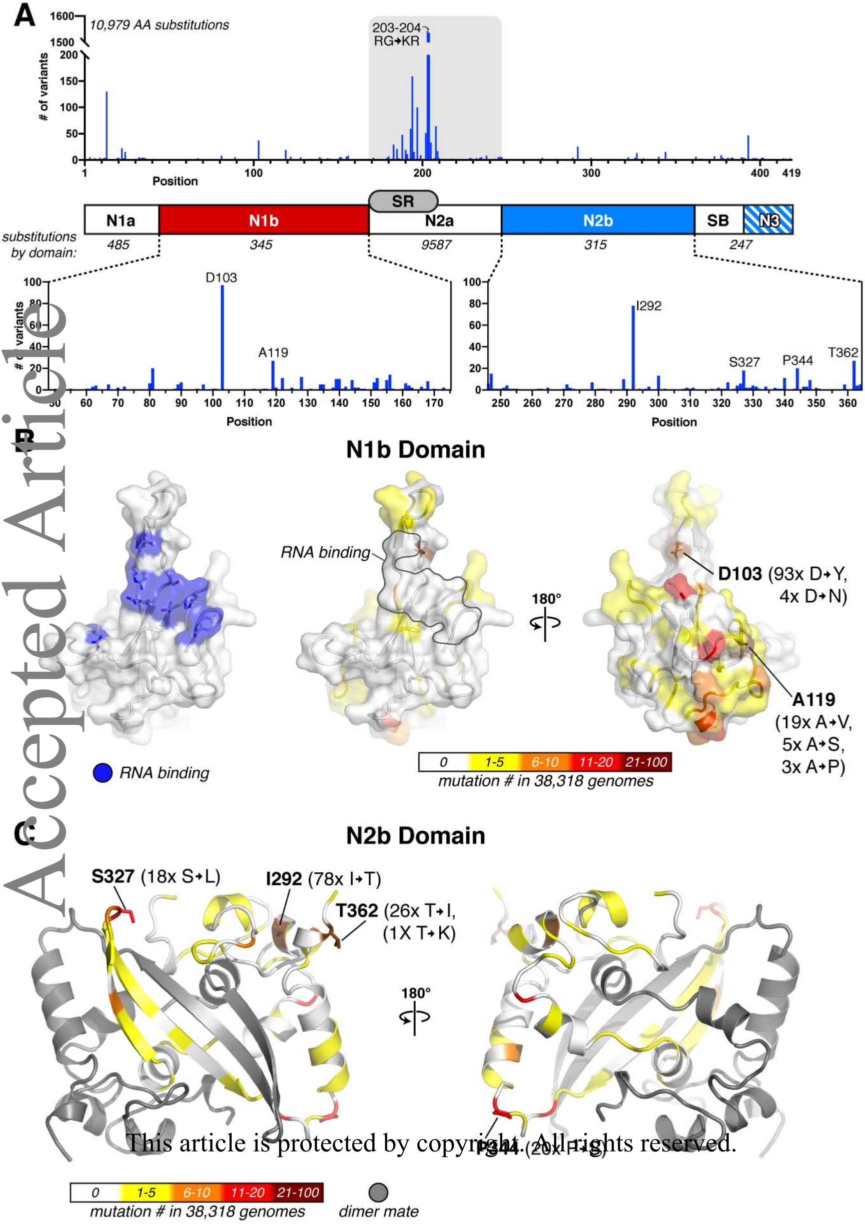
(B) Schematic showing the N_{2b3} sequence and inferred structure (the α-helix spanning residues 400-416 is predicted by PSI-PRED), plus protein regions detected by HDX-MS. Two sets of exchange rates are shown: fractional deuterium uptake in N_{2b3} (upper box) colored as in panel A, and relative uptake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

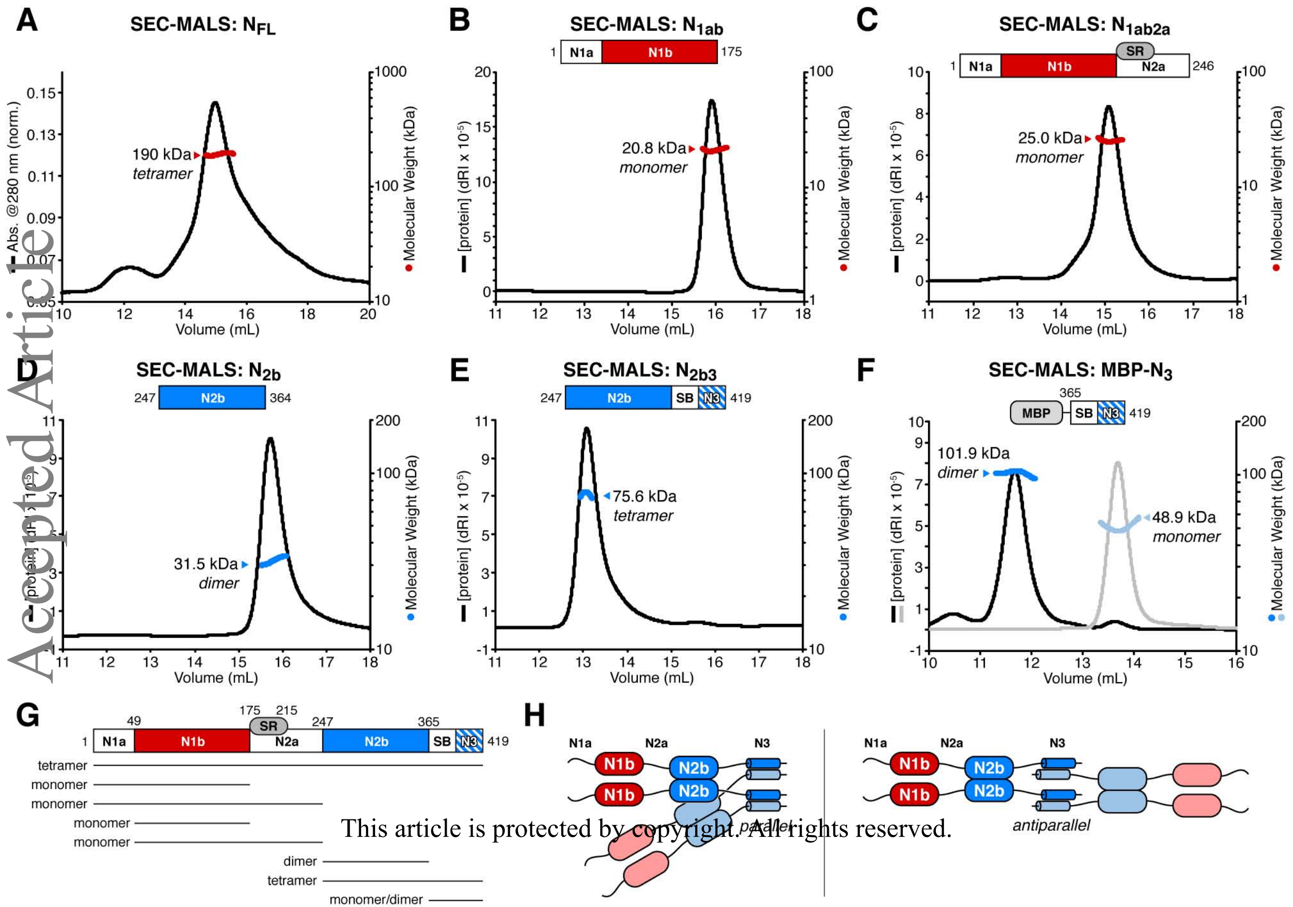
(C) Structure of the N_{2b} dimer, with one monomer colored by fractional deuterium uptake (blue-white-magenta = 0-75% fractional uptake).

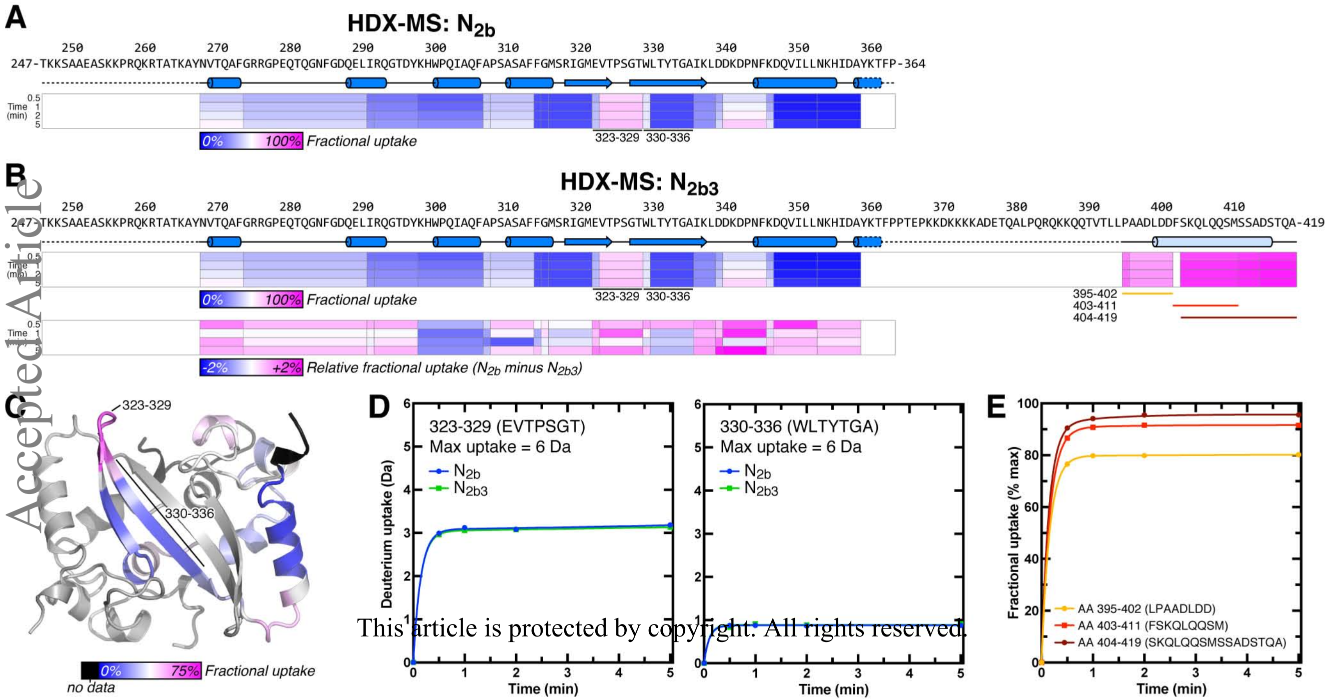
(D) Uptake plots for two peptides within the ordered N2b domain, with uptake in N_{2b} indicated in blue and uptake in N_{2b3} indicated in green. The peptide covering residues 323-329 (located within a loop) is relatively exposed, while the peptide covering residues 330-336 (within a β-strand) is strongly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

(E) Uptake plots for three peptides in the C-terminal region of N_{2b3}, plotted by fractional deuterium uptake. Peptides covering residues 395-402 (yellow) and 403-411 (red) show more protection than residues 404-419, suggesting that this region is partially structured. See Figure S4(A) for each peptide plotted by absolute deuterium uptake.


A
SARS-CoV-2 N
400-416
predicted 2°
α
disorder
conservation
175
215
49
SR
247
365
1
N1a
N1b (NTD)
N2a
N2b (CTD)
SB
N3
419
conservation
disorder
N_{FL}
N_{1ab}
N_{1ab2a}
N_{1b}
N_{1b2a}
N_{2b}
N_{2b3}
N_3
B
N_{2b}
N
C
N
C
C
SARS-CoV-2
SARS-CoV
0.314 Å Cα r.m.s.d


A
10,979 AA substitutions
203-204
RG→KR
# of variants
1600
1500
200
150
100
50
0
1
100
200
300
400
419
Position
SR
N1a
N1b
N2a
N2b
SB
N3
substitutions by domain:
485
345
9587
315
247
D103
A119
I292
S327
P344
T362
# of variants
100
80
60
40
20
0
50
60
70
80
90
100
110
120
130
140
150
160
170
Position
250
260
270
280
290
300
310
320
330
340
350
360
Position
B
N1b Domain
RNA binding
180°
D103 (93x D→Y, 4x D→N)
A119 (19x A→V, 5x A→S, 3x A→P)
RNA binding
0
1-5
6-10
11-20
21-100
mutation # in 38,318 genomes
C
N2b Domain
S327 (18x S→L)
I292 (78x I→T)
T362 (26x T→I, (1X T→K)
180°
P344 (20x P→S)
0
1-5
6-10
11-20
21-100
mutation # in 38,318 genomes
dimer mate


A
SEC-MALS: N$_{FL}$
190 kDa
tetramer
Abs. @280 nm (norm.)
0.05
0.07
0.09
0.11
0.13
0.15
Molecular Weight (kDa)
10
100
1000
Volume (mL)
10
12
14
16
18
20
B
SEC-MALS: N$_{1ab}$
1
N1a
N1b
175
20.8 kDa
monomer
[protein] (dRI x 10^{-5})
0
5
10
15
20
Molecular Weight (kDa)
1
10
100
Volume (mL)
11
12
13
14
15
16
17
18
C
SEC-MALS: N$_{1ab2a}$
1
N1a
N1b
SR
N2a
246
25.0 kDa
monomer
[protein] (dRI x 10^{-5})
0
2
4
6
8
10
Molecular Weight (kDa)
1
10
100
Volume (mL)
11
12
13
14
15
16
17
18
D
SEC-MALS: N$_{2b}$
247
N2b
364
31.5 kDa
dimer
[protein] (dRI x 10^{-5})
-1
1
3
5
7
9
11
Molecular Weight (kDa)
10
100
200
Volume (mL)
11
12
13
14
15
16
17
18
E
SEC-MALS: N$_{2b3}$
247
N2b
SB
N3
419
75.6 kDa
tetramer
[protein] (dRI x 10^{-5})
-1
1
3
5
7
9
11
Molecular Weight (kDa)
10
100
200
Volume (mL)
11
12
13
14
15
16
17
18
F
SEC-MALS: MBP-N$_3$
365
MBP
SB
N3
419
101.9 kDa
dimer
48.9 kDa
monomer
[protein] (dRI x 10^{-5})
-1
1
3
5
7
9
10
Molecular Weight (kDa)
10
100
200
Volume (mL)
10
11
12
13
14
15
16
G
49
175
215
247
365
1
N1a
N1b
SR
N2a
N2b
SB
N3
419
tetramer
monomer
monomer
monomer
monomer
dimer
tetramer
monomer/dimer
H
N1a
N2a
N3
N1b
N1b
N2b
N2b
parallel
N1a
N2a
N3
N1b
N1b
N2b
N2b
antiparallel


A
HDX-MS: N_{2b}
250 260 270 280 290 300 310 320 330 340 350 360
247-TKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFP-364
Time (min)
0.5
1
2
5
0% 100% Fractional uptake
323-329
330-336
B
HDX-MS: N_{2b3}
250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410
247-TKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA-419
Time (min)
0.5
1
2
5
0% 100% Fractional uptake
323-329
330-336
395-402
403-411
404-419
Time
0.5
1
5
-2% +2% Relative fractional uptake (N_{2b} minus N_{2b3})
C
323-329
330-336
0% 75% Fractional uptake
no data
D
323-329 (EVTPSGT)
Max uptake = 6 Da
N_{2b}
N_{2b3}
Deuterium uptake (Da)
0 1 2 3 4 5 6
Time (min)
0 1 2 3 4 5
330-336 (WLTYTGA)
Max uptake = 6 Da
N_{2b}
N_{2b3}
0 1 2 3 4 5 6
Time (min)
0 1 2 3 4 5
E
Fractional uptake (% max)
0 20 40 60 80 100
AA 395-402 (LPAADLDD)
AA 403-411 (FSKQLQQSM)
AA 404-419 (SKQLQQSMSSADSTQA)
Time (min)
0 1 2 3 4 5