Diversity of Short Linear Interaction Motifs in SARS-CoV-2 Nucleocapsid Protein

Peter Schuck* and Huaying Zhao

Laboratory of Dynamics of Macromolecular Assembly, National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health, Bethesda, MD 20892, USA

*Correspondence to schuckp@mail.nih.gov

SUMMARY

Molecular mimicry of short linear interaction motifs has emerged as a key mechanism for viral proteins binding host domains and hijacking host cell processes. Here, we examine the role of RNA-virus sequence diversity in the dynamics of the virus-host interface, by analyzing the uniquely vast sequence record of viable SARS-CoV-2 species with focus on the multi-functional nucleocapsid protein. We observe the abundant presentation of motifs encoding several essential host protein interactions, alongside a majority of possibly non-functional and randomly occurring motif sequences absent in subsets of viable virus species. A large number of motifs emerge *ex nihilo* through transient mutations relative to the ancestral consensus sequence. The observed mutational landscape implies an accessible motif space that spans at least 25% of known eukaryotic motifs. This reveals motif mimicry as a highly dynamic process with the capacity to broadly explore host motifs, allowing the virus to rapidly evolve the virus-host interface.

INTRODUCTION

Short linear interaction motifs (SLiMs) are stretches of several amino acids that serve as microdomains in intrinsically disordered regions (IDRs) mediating weak, transient protein-protein interactions with target domains (Sangster et al., 2022). SLiMs have emerged as a ubiquitous modules in the organization of protein-protein interaction networks. For example, SLiMs designate substrates for post-translational modification including kinases and phosphatases, target cellular localization, and control docking to adaptors, assembly of signaling complexes, and recruitment of enzymes to multi-protein complexes (Manna et al., 2018; Mohapatra and Dixit, 2022; Peti and Page, 2013; Ragusa et al., 2010; Shi et al., 2023; Sorgeloos et al., 2022; Srivastava et al., 2023; Wang et al., 2016). The number of eukaryotic SLiM classes is in the hundreds and rapidly expanding, and it has been estimated there may be more than a hundred thousand of instances of such motifs in the human proteome (Davey et al., 2023; Kumar et al., 2022; Tompa et al., 2014).

Molecular mimicry of SLiMs is a key mechanism for viral proteins to hijack and modulate host cell processes and is broadly exploited among many viruses (Davey et al., 2011; Hagai et al., 2014; Mihalič et al., 2023). Therefore, the distribution and evolution of viral motifs is of significant interest in the search for broad-spectrum anti-viral drug targets (Shuler and Hagai, 2022; Simonetti et al., 2023). It has been proposed that the evolution of SLiMs is facilitated by their compact size, in combination with the unusually high abundance of IDRs in viral proteins; the latter, by virtue of fewer constraints, generally exhibiting high mutation frequencies that would allow the efficient *de novo* formation of SLiMs through as little as a single amino acid change (Brown et al., 2002; Davey et al., 2015; Fuxreiter et al., 2007; Gitlin et al., 2014; Neduva and Russell, 2005; Tokuriki et al., 2009). The conservation of disorder has been hypothesized to facilitate change of interaction partners and thereby provide an evolutionary advantage (Mosca et al., 2012). In eukaryotes, the generation and loss of SLiMs has been confirmed in sequence analyses of evolutionarily related species, as well as in rare random mutations individual human patients (Cordeddu et al., 2009; Davey et al., 2015). In viruses, substantial heterogeneity of motif content has been observed across different viral families, with instances of convergent evolution, pointing to a highly dynamic repertoire of motif usage and evolutionary adaptation (Hagai et al., 2014).

On the other hand, a salient feature of RNA-viruses is their high intracellular and intrahost sequence diversity due to their evolved low transcription fidelity and resulting quasispecies nature (Domingo, 2019; Eigen, 1993; Holland and Domingo, 1997; Lauring and Andino, 2010). How this sequence diversity impacts the virus-host interface and the evolution of SLiMs has remained unexplored. An opportunity to study this question recently arose with the unprecedented, vast collection SARS-CoV-2 genomes in GISAID (Elbe and Buckland-Merrett, 2017). While it provides a basis for monitoring the evolution of mutations distinguishing emergent clades, the repository contains a majority of random transient mutations that exhaustively explore the mutational landscape of viral proteins (Bloom et al., 2023; Saldivar-Espinoza et al., 2023; Zhao et al., 2022). In the present work, we exploit the diversity of SARS-CoV-2 sequences and examine the distribution of viral motif content on the infected host population level, which we propose may serve as a model also for intrahost viral motif diversity.

Eukaryotic motif mimicry for cell entry has been found in the RBD of the spike protein of SARS-CoV-2 (Mészáros et al., 2021). In the present work we focus on the nucleocapsid (N-)protein, which has several favorable properties as platform for SLiMs, having the highest expression level of all SARS-CoV-2 proteins at ≈1 % of total protein in infected cells (Tugaeva et al., 2021), and containing three IDRs spanning nearly half of the protein (**Figure 1A**). Even though it provides a major antigen, it is not immuno-dominant as the spike protein. Besides its eponymous structural role in viral assembly (Carlson et al., 2022; Masters, 2019; Zhao et al., 2023), N-protein is highly multi-functional with a large host interactome (Gordon et al., 2020; Kruse et al., 2021; Wu et al., 2023; Zheng et al., 2021), including interactions with proteins of the type 1 interferon signaling pathway (Li et al., 2020; Mu et al., 2020; Yelemali et al., 2022), the inflammasome (Pan et al., 2021), complement activation (Gao et al., 2022), lipid metabolism (Yuan et al., 2021), and expression and binding to cytokines (Karwaciak et al., 2021; López-Muñoz et al., 2022). Among interactions described in greatest biophysical detail are the complex formation with G3BP1 leading to rewiring of stress granules (Biswal et al., 2022; Kruse et al., 2021; Yang et al., 2023), binding of 14-3-3 (Eisenreichova and Boura, 2022; Tugaeva et al., 2021, 2023), and the interaction with host kinases leading to extensive phosphorylation particularly in the linker IDR of intracellular N-protein (Carlson et al., 2020; Syed et al., 2023; Yaron et al., 2022). Other posttranslational modifications include ubiquination, proteolytic cleavage, sumoylation, and ADP-ribosylation (Fung and Liu, 2018; Madahar et al., 2023; Mao et al., 2023).

The basis for the present work is a dataset of ≈5 million SARS-CoV-2 consensus sequences, which we first characterize regarding the mutation frequencies and mutational landscape of N-protein over time. We have recently exploited the exhaustive map of viable amino acid mutations as a tool in the structural analysis of assembly roles of the linker IDR (Zhao et al., 2022, 2023). For the analysis of motif content we examine the ancestral consensus sequence and utilize a previously introduced statistical approach (Hagai et al., 2014) to identify possibly random motif instances. We then determine the distribution of motif content in the ensemble of viral sequences sampled across the infected host population. This analysis reveals a highly dynamic motif presentation where most of the ancestral motifs are abandoned in at least some sequence subsets, while in others large numbers of new motifs arise that potentially interact with various host pathways. Finally, based on the mutational landscape, we estimate a measure of the available sequence space and corresponding accessible motif space, which suggests that even for the smallest N-protein IDR a large fraction of known motifs may be presented. While many or most of these may not be functionally interacting with host proteins, it reveals a mechanism for extensive probing of viral proteins for potentially beneficial host protein interactions. We discuss possible implications for the virus-host interface considering intrahost and intracellular sequence diversity.

RESULTS

The mutational landscape of SARS-CoV-2 nucleocapsid protein

The present study is based on 5.06 million high-quality consensus SARS-CoV-2 sequences retrieved on January 20, 2023 from Nextstrain (Hadfield et al., 2018). Due to their origin from COVID19 patient samples we may assume these to be sequences of viable and infectious virus. The sequences exhibit significant diversity, with ≈56% of sequences being distinct, ≈30% unique, and ≈8% spatio-temporally distant repeats. The set contains ≈43 million instances of N-protein mutations relative to the ancestral Wuhan-Hu-1 isolate. Most of the mutations are different from the defining substitutions of the variants of concern (**Figure 1B**), and instead occur transiently and are distributed across ≈92% of all N-protein residues. This suggests these mutations may have a significant impact on the N-protein motif repertoire.

Figure 1C shows the frequency of these evolutionarily inconsequential mutations at different positions and as a function of time (where time is plotted in the ordinate transformed to a cumulative sequence number to compensate for uneven sampling frequency (**Figure 1D**)). As may be discerned from the barcode-like vertical patterns in the frequency plot **Figure 1C**, the local mutation frequencies are in first approximation constant outside the defining variant substitutions, with minor evidence of limited stochastic transmission events causing temporal variation. By contrast, there is significant structure across different positions: Higher frequencies are generally observed in the IDRs, and lower mutation frequencies correlate with biophysical functional constraints in the folded domains and IDRs (Zhao et al., 2022, 2023). Short of a detailed evolutionary analysis, we may roughly group sequences into three distinct sets of Omicron variants, Delta variants, and those preceding the Delta-variants, which constitute the vast majority of sequences in different periods of time (**Figure 1D**). Their residue mutation frequency pattern is nearly identical outside the defining substitutions, which shows that major constraining biophysical properties of N-protein are unaltered. This is highlighted further in the detailed comparison of average residue mutation frequencies as a function of position among the three groups (**Figure 1E**).

A detailed chart of the observed amino acid mutations at all positions is shown in **Figure 1F**. On average ≈5.5 different amino acids may occupy each position, ranging from zero mutations at 14 of the 37 conserved positions across related coronaviruses, to a maximum of 12 different amino acids that may occupy the most variable positions of the IDRs. Their pattern defines a characteristic mutational landscape, which is similar to that reported previously (Zhao et al., 2022), despite the fact that the current data set includes ≈2.6 million Omicron sequences (separately shown in **Figure S1**) that were not yet available previously. We conclude that the different waves of SARS-CoV-2 variants independently reproduce very similar mutational landscapes, as would be expected for exhaustively sampled N-protein in a steady-state.

The mutational landscape is a comprehensive set of all tolerable, non-lethal mutations (Bloom and Neher, 2023) and as such it reflects detailed biophysical constraints and provides complementary information to traditional structural tools (Zhao et al., 2022, 2023). For example, binding to G3BP1/2 was identified as an essential N-protein function and a crystal structure shows G3BP1 binding the φ-x-F motif in the N-arm (Biswal et al., 2022). Accordingly, F17 is an almost completely conserved residue in the mutational landscape of the otherwise highly variable IDR (**Figure 1F**). Similarly, in the leucine-rich region of the linker IDR the oligomerization of transient helices to form coiled-coils was recently identified as an essential assembly function, and structural requirements for oligomerization were found to be reflected in the nature of the limited set of amino acid mutations in positions 221-233, in the otherwise highly variable linker IDR (Zhao et al., 2023).

Different combinations of N-protein mutations define 40,988 distinct N-protein sequences. Since linear motifs are prevalent in IDRs, we focus on the subset of distinct IDR sequences. Using a threshold condition that each sequence is observed in at least 10 different genomes, there are 512 distinct sequences for N-arm carrying an average of 2.75 mutations, 979 for the linker IDR with an average of 2.90 mutations, and 556 for the C-arm IDR with an average of 1.77 mutations. Each sequence was examined with regard to their SLiM content using the Eukaryotic Linear Motif (ELM) Resource for Functional Sites in Proteins (Kumar et al., 2022), which can search for occurrences of regular expressions of 327 documented motif classes.

Motif content of the ancestral SARS-CoV-2 N-protein sequence

As a starting point to parse the results we consider first the motif content in the linker IDR of the ancestral Wuhan-Hu-1 sequence. **Figure 2** lists in bold the predicted ancestral motifs and the inset shows their location. Many motif classes occur in multiple instances, their number charted as white crosses. The motif set is dominated by sites for kinases, in particular in the SR-rich region. This is not surprising, considering the high degree of phosphorylation experimentally observed (Yaron et al., 2022). Kinase motifs significantly overlap, which may produce allovalency and allow for cooperativity and increase effective affinity of the sites for kinase binding (Klein et al., 2003). In addition, several other motifs overlap in both the SR-rich and the transiently helical L-rich region of the linker (**Figure 1B**), which may not preclude their function considering the large number of intracellular copies N-protein. A previously reported 14-3-3 motif in the linker (Eisenreichova and Boura, 2022; Tugaeva et al., 2023) is reproduced, and a variety of motifs for different posttranslational modifications and binding functions are found. A similar preponderance of phosphorylation motifs is found in the C-arm (**Figure 3**) and N-arm (**Figure 4**) IDRs (the latter missing the above mentioned G3BP1 binding motif not contained in the ELM database). While some of the other motifs for protein modification and host protein interactions seem plausible, such as those related to de-ubiquination, sumoylation, autophagy, and apoptosis, others appear unlikely to describe real interactions, for example, several glycosylation motifs (Shajahan et al., 2021) and those targeting proteins of different organisms.

As a measure for the likelihood that some of these motifs may appear just stochastically we employ a strategy previously developed by Hagai and co-workers (Hagai et al., 2014): For each of the IDRs we generated a set of 10,000 randomly scrambled sequences with the same amino acid content, and for each motif, we determined the frequency of it occurring in the randomized set. The average number of sites with standard deviation is depicted as blue bars in **Figures 2-4**. As shown in the first three rows of **Figure 2**, there is a relatively high probability of generating multiple GSK3, CK1, and PKA phosphorylation sites in the linker IDR by chance, which may be expected given the amino acid composition particularly of the SR-rich region (**Figure 2**). However, their actual number in the ancestral sequence (white crosses) exceeds the statistical expectation, consistent with the important role of phosphorylation for intracellular N-protein (Carlson et al., 2020; Yaron et al., 2022). Other motifs of the ancestral sequence that have a high probability to occur by chance given the linker amino acid composition are sites for binding of 14-3-3 protein, USP7, and glycosaminoglycan attachment. Interestingly, a motif for interaction with yeast KEX2 protease has a high statistical chance to occur, and indeed is created through the defining R203K/G204R mutation of Alpha, Gamma, and Omicron variants. In the C-arm and N-arm IDRs, the majority of motifs displayed in the ancestral sequence are also likely to occur by chance given the respective IRD amino acid compositions. Unfortunately, the statistical analysis breaks down for motifs utilizing amino acids not part of the ancestral sequence, such as the defining 215C mutation of Delta variant linker, which creates a likely non-functional N-glycosylation motif.

Distribution of motifs in the mutant spectrum of SARS-CoV-2 N-protein

The distribution of motifs across the observed sequences can be depicted as a histogram of the site multiplicity. Accordingly, for each motif in **Figures 2-4**, the color and size of the circles is scaled from large black to small red according to the frequency of sequences exhibiting different numbers of instances of that motif, as indicated numerically. A complete list of SLiMs and their frequencies can be found in the **Supplementary Information**. Strikingly, the major phosphorylation motifs in the linker (**Figure 2**) exhibit a great polydispersity in site numbers, for example, ranging from 6 to 11 with a mode of 10 for GSK3, from 5 to 10 with a mode of 8 for CK1, and from 2 to 5 with a mode of 3 for PKA sites. The sum of phosphorylation motifs ranges from 15 to 27. Even though a higher than statistically expected number of phosphorylation motifs is conserved across all sequences, it appears as if the detailed phosphorylation events are not critical for viable virus, as judged by the fact that individually none of the phosphorylation sites in the SR-rich region of the linker IDR are conserved in the mutational landscape (**Figure 1F**). Interestingly, most of the sequences have one less predicted PKA site than the ancestral sequence, which is caused by defining mutations R203K/G204R in Alpha, Gamma, and Omicron, and the R203M mutation in Delta variant; these mutations has have been shown experimentally to cause reduced phosphorylation and enhanced assembly functions and were hypothesized to reflect viral evolution (Syed et al., 2023).

A similar picture of prominent polydispersity emerges in the motif distributions of the C-arm and N-arm IDRs. Regarding the likelihood of motifs occurring by chance given the amino acid composition of the linker IDR, it is interesting to note that motifs with high statistical chance are indeed found in greater numbers in sequence subsets, which may be discerned for the glycosaminoglycan attachment, 14-3-3 binding, and USP7 binding motifs in the linker, as well as the GSK3 binding and glycosaminoglycan attachment sites in the N-arm.

A striking aspect of the motif content in the mutant spectrum is that nearly all motifs appear dispensable, which is indicated in the distinct sequence populations with zero sites in **Figures 2-4**. The only exceptions are the phosphorylation motifs of the linker and N-arm, the FHA binding motif in the C-arm IDR, and probably the G3BP1 binding motif in the N-arm not included in the ELM resource, as judged by the strong conservation of F17 in the mutational landscape. All other motifs are absent in sometimes sizeable fractions of mutant virus sequences, ostensibly suggesting that they may not describe real host protein interactions, or that these are not an essential part of the virus-host interface (see **Discussion**). For example, 1.1% of all linker sequences lack the 14-3-3 site, and 9.9% lack the CDC14 phosphatase dephosphorylation site, 25.4% of C-arm sequences lack both CK1 phosphorylation sites, and 32.2% of N-arm sequences lack the GSK3 site. Similarly, the defining Δ31-33 mutation in Omicron sequences destroys a likely non-functional glycosaminoglycan motif.

Conversely, many motifs that do not exist in the ancestral sequence are formed *ex nihilo* in subsets of sequences due to their particular constellation of mutations. In addition to several motifs arising from defining mutations of variants of concern (such as the yeast KEX2 protease site mentioned above), a large number of motifs occur in only a small fraction of sequences. However, since only sequences occuring in at least 10 different genomes were included in the analysis, a frequency of ≈1% of all linker sequences translates to on the order of 100 instances in the genome database.While many of those are plausible host protein interactions, others are less likely and may be random matches with regular expressions.

Estimated accessible sequence and motif space of N-protein IDRs

The total number of motif classes displayed in the database sequences is 72 in the linker, 62 in the C-arm, and 53 in the N-arm IDR, out of a total of 327 motif classes currently contained in the ELM database. This raises the question of how efficiently random mutations can create new motifs, and what fraction of the total currently known motif space is accessible to N-protein. Since the viable amino acid landscape (**Figure 1F**) has been exhaustively explored during the pandemic so far, it is possible to make a back-of-the-envelope estimate of the theoretical maximal size of the associated sequence space by permutation through viable amino acid mutations at each position. For the N-arm (1:44) – the smallest of the N-protein IDRs – allowing for 3 mutations per sequence, which is close to the average of ≈2.8 observed in the existing database, there are ≈1.8×10^6 different mutant sequence permutations. Although likely not all will be viable due to epistatic effects, this upper limit is more than three orders of magnitude larger than the 512 distinct N-arm sequences observed so far in the genomic database, and far exceeds our capacity for computational determination of the associated motif space.

Nonetheless, limited approximate sequence spaces can be searched when focusing on only the most frequently encountered amino acid mutations. For example, considering only the mutations observed in > 1,000 instances (i.e., in 0.02% of genomes; depicted in bold red in **Figure 1F**), the associated sequence space with three mutations consists of 10,500 sequences, which is searchable and describes 26 motif classes from the ELM database, nearly twice the number of different motifs in the ancestral N-arm sequence. Lowering the mutation frequency threshold leads to a rapidly growing sequence space and associated motif space (**Figure 5**). For example, at a mutation threshold of > 100 instances (mutation frequence of 0.002% in all genomes) 231,519 possible sequences with three mutations cover a range of 85 motif classes, or 26% of all in the ELM database. Consideration of more rare mutations occurring in the viable amino acid landscape further increases the theoretical sequence space and thereby the accessible motif space.

DISCUSSION

The virus-host interface is crucial for viral survival and a promising area for the development of antiviral therapeutics. Significant recent advances were driven by increased understanding of the evolutionary role of IDRs and the recognition of SLiMs as ubiquitous interaction modules that can be hijacked through viral mimicry. Unavoidably, both large-scale experimental and bioinformatics studies contributing to this picture were limited to consensus sequences of viral species, lacking opportunity to account for the quasispecies nature of RNA viruses, and thus leaving the salient feature of sequence diversity previously unexplored with regard to the role of SLiMs in the virus-host interface. However, an important recent development is the assembly of a vast SARS-CoV-2 genomic repository at GISAID (Elbe and Buckland-Merrett, 2017), which exceeds the number of available influenza sequences by more than one order of magnitude. This allows for the first time the exhaustive characterization of the amino acid mutation landscape (Bloom et al., 2023; Saldivar-Espinoza et al., 2023; Zhao et al., 2022), and is exploited here to power an analysis of the viral sequence space. This illuminates the highly dynamic motif space of viral IDRs, providing new insights in the unexpected efficiency of *ex nihilo* motif creation, the extent of viral motif mimicry, and the potential size of the virus-host interface.

We have started the present work with the assembly of the amino acid mutational landscapes from the host population-wide ensemble of consensus sequences. In a first approximation, we may consider the amino acid landscape as a reflection of all mutations consistent with vital biophysical, functional constraints of the viral protein, from which random sequence samples arise with certain mutation frequencies. Grouping all SARS-CoV-2 N-protein sequences in three major waves from different periods of the pandemic, groups essentially representing independent repeats of deep mutational scans, produces nearly identical mutational landscapes and local mutation frequencies. This suggests that the basic biophysical properties of N-protein overall have not significantly changed, as in an evolutionary stable steady-state. In the first approximation, this view justifies considering the derived motif space to be similarly in steady-state, and to represent an intrinsic property of N-protein.

This is notwithstanding fitness modulations from localized N-protein such as 203K/204R (Syed et al., 2023) and 215C (Zhao et al., 2022) that seem secondary to evolution of the immunodominant spike protein. Interestingly, in both Delta and Omicron waves the defining mutations lead to the destruction of one PKA motif that may impact the extent of linker phosphorylation and modulate the switch between intracellular and assembly functions (Syed et al., 2023). On the other hand, the observed variation in the content of kinase motifs is very large, and none of the potential phosphorylation sites other than S184 is strongly conserved in the mutational landscape, apparently without compromising virus viability. This points to a distributed phosphorylation threshold in the linker IDR rather than specific structural requirements (Zarin et al., 2021) for N-protein to be viable, which may be fine-tuned for fitness optimization.

The data for the available time-scale depict highly parallel random exploration of many motifs. On the level of single proteins, multiple overlapping repeat instances of motifs are displayed along the IDRs, a feature frequently encountered RNA viruses (Mihalič et al., 2023), which may lead to cooperativity and effective enhancement (Klein et al., 2003; Watson et al., 2022). Similarly, different overlapping motifs provide multi-functionality with little competition due to the large expression level with 10^8 copies of N-protein in the infected cells (Tugaeva et al., 2021). Across the mutant spectrum, we observe highly effective motif creation, spanning an astonishingly wide range theoretically covering 20% or more of the known eukaryotic motif space even in the shortest N-protein IDR. Conversely, most of the motifs displayed in the ancestral sequence are destroyed in at least a subset of the viable mutant spectrum.

The origin of the sequence diversity in the consensus sequences considered here is rooted in the error-prone transcription and the intracellular quasispecies. However, it is unclear to what extent the observed SARS-CoV-2 sequence space reflects intracellular quasispecies and intrahost 'meta-quasispecies' (Domingo, 2019), perhaps constituting a 'hyper-quasispecies' as the ensemble of consensus sequences across the host-population, with the obvious exclusion of non-viable species. Even though the amino acid mutation landscape (**Figure 1F**) is sufficiently sampled to reflect biophysical features (Zhao et al., 2023), only a very small fraction of the implied theoretical sequence space has been sampled. Even less is known about the quasispecies, with deep sequencing capabilities probing intrahost minority sequences currently limited to a level of 0.1% (Martínez-González et al., 2022). However, several studies report that most of the mutations of intrahost minority species are independently reflected in the GISAID repository of consensus sequences (Martínez-González et al., 2022; Siqueira et al., 2021; Tonkin-Hill et al., 2021), demonstrating at least significant overlap. If the motif space observed in the present work is any indication of intracellular diversity, this would strongly further leverage the viral motif range. For example, it raises the possibility of cooperation between species (Vignuzzi et al., 2006), and one could envision minor species to interact with host proteins through motifs that the master sequence has abandoned. Thus, interaction motifs that disappear in a subset of consensus sequences in our study may nonetheless still be essential for viable virus. Also, the simultaneous presence of viral protein species with different motif sets attacking host processes at multiple entry points may exert synergistic effects on redundant interaction networks (Levy et al., 2010).

Whether the observed motif diversity extends intracellularly or not, the present work demonstrates that single sequence-based studies of viral protein/host interactions under neglect of viral diversity likely vastly underestimates the abundance of SLiM-based protein-protein interactions in the virus-host interface. On the other hand, many of the displayed motifs may not be functional host protein interactions, due to incorrect sequence context, localization, or even species specificity. This does not diminish the role of the dynamics of random motif generation, which may serve as a fertile basis to search for beneficial interactions for further optimization and host adaptation. Non-functional interactions as a consequence of 'evolutionary noise' may have little fitness penalty and should be expected to occur, as pointed out by Levy and colleagues (Levy et al., 2009), and this should hold true in particular for viral quasispecies.

ACKNOWLEDGMENTS

We thank the GISAID, Nextstrain, and ELM Resource teams and its contributors for maintaining these tools and databases. This work was supported by the Intramural Research Program of the National

Institute of Biomedical Imaging and Bioengineering (ZIA EB000095-02), National Institutes of Health. This work utilized the computational resources of the National Institutes of Health HPC Biowulf cluster for sequence analyses.

REFERENCES

Biswal, M., Lu, J., and Song, J. (2022). SARS-CoV-2 nucleocapsid protein targets a conserved surface groove of the NTF2-like domain of G3BP1. J. Mol. Biol. 167516.

Bloom, J.D., and Neher, R.A. (2023). Fitness effects of mutations to SARS-CoV-2 proteins. BioRxiv 2023.01.30.526314.

Bloom, J.D., Beichman, A.C., Neher, R.A., and Harris, K. (2023). Evolution of the SARS-CoV-2 Mutational Spectrum. Mol. Biol. Evol. *40*, 2022.11.19.517207.

Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., and Keith Dunker, A. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. J. Mol. Evol. *55*, 104–110.

Carlson, C.R., Asfaha, J.B., Ghent, C.M., Howard, C.J., Hartooni, N., Safari, M., Frankel, A.D., and Morgan, D.O. (2020). Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for its Dual Functions. Mol. Cell *80*, 1092-1103.e4.

Carlson, C.R., Adly, A.N., Bi, M., Howard, C.J., Frost, A., Cheng, Y., and Morgan, D.O. (2022). Reconstitution of the SARS-CoV-2 ribonucleosome provides insights into genomic RNA packaging and regulation by phosphorylation. J. Biol. Chem. *298*, 102560.

Cordeddu, V., Di Schiavi, E., Pennacchio, L.A., Ma'ayan, A., Sarkozy, A., Fodale, V., Cecchetti, S., Cardinale, A., Martin, J., Schackwitz, W., et al. (2009). Mutation of SHOC2 promotes aberrant protein N-myristoylation and causes Noonan-like syndrome with loose anagen hair. Nat. Genet. *41*, 1022–1026.

Davey, N.E., Travé, G., and Gibson, T.J. (2011). How viruses hijack cell regulation. Trends Biochem. Sci. *36*, 159–169.

Davey, N.E., Cyert, M.S., and Moses, A.M. (2015). Short linear motifs - Ex nihilo evolution of protein regulation Short linear motifs - The unexplored frontier of the eukaryotic proteome. Cell Commun. Signal. *13*, 9–11.

Davey, N.E., Simonetti, L., and Ivarsson, Y. (2023). The next wave of interactomics: Mapping the SLiM-based interactions of the intrinsically disordered proteome. Curr. Opin. Struct. Biol. *80*, 102593.

Domingo, E. (2019). Virus as Populations (Academic Press).

Eigen, M. (1993). Viral Quasispecies. Sci. Am. *269*, 42–49.

Eisenreichova, A., and Boura, E. (2022). Structural basis for SARS-CoV-2 nucleocapsid (N) protein recognition by 14-3-3 proteins. J. Struct. Biol. *214*, 107879.

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob. Challenges *1*, 33–46.

Erdos, G., Pajkos, M., and Dosztányi, Z. (2021). IUPred3: Prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res.

49, W297–W303.

Fung, T.S., and Liu, D.X. (2018). Post-translational modifications of coronavirus proteins: Roles and function. Future Virol. *13*, 405–430.

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. Bioinformatics *23*, 950–956.

Gao, T., Zhu, L., Liu, H., Zhang, X., Wang, T., Fu, Y., Li, H., Dong, Q., Hu, Y., Zhang, Z., et al. (2022). Highly pathogenic coronavirus N protein aggravates inflammation by MASP-2-mediated lectin complement pathway overactivation. Signal Transduct. Target. Ther. *7*, 318.

Gitlin, L., Hagai, T., LaBarbera, A., Solovey, M., and Andino, R. (2014). Rapid Evolution of Virus Sequences in Intrinsically Disordered Protein Regions. PLoS Pathog. *10*.

Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J., Rezelj, V. V., Guo, J.Z., Swaney, D.L., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 1–13.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). NextStrain: Real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123.

Hagai, T., Azia, A., Babu, M.M., and Andino, R. (2014). Use of Host-like Peptide Motifs in Viral Proteins Is a Prevalent Strategy in Host-Virus Interactions. Cell Rep. *7*, 1729–1739.

Holland, J.J., and Domingo, E. (1997). RNA virus mutations and fitness for survival. Annu. Rev. Microbiol. *51*, 151–178.

Karwaciak, I., Sałkowska, A., Karaś, K., Dastych, J., and Ratajewski, M. (2021). Nucleocapsid and Spike Proteins of the Coronavirus SARS-CoV-2 Induce IL6 in Monocytes and Macrophages-Potential Implications for Cytokine Storm Syndrome. Vaccines *9*.

Klein, P., Pawson, T., and Tyers, M. (2003). Mathematical Modeling Suggests Cooperative Interactions between a Disordered Polyvalent Ligand and a Single Receptor Site. Curr. Biol. *13*, 1669–1678.

Kruse, T., Benz, C., Garvanska, D.H., Lindqvist, R., Mihalic, F., Coscia, F., Inturi, R., Sayadi, A., Simonetti, L., Nilsson, E., et al. (2021). Large scale discovery of coronavirus-host factor protein interaction motifs reveals SARS-CoV-2 specific mechanisms and vulnerabilities. Nat. Commun. *12*, 1–13.

Kumar, M., Michael, S., Alvarado-Valverde, J., Mészáros, B., Sámano-Sánchez, H., Zeke, A., Dobson, L., Lazar, T., Örd, M., Nagpal, A., et al. (2022). The Eukaryotic Linear Motif resource: 2022 release. Nucleic Acids Res. *50*, D497–D508.

Lauring, A.S., and Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. PLoS Pathog. *6*, 1–8.

Levy, E.D., Landry, C.R., and Michnick, S.W. (2009). How perfect can protein interactomes be? Sci. Signal. *2*, pe11.

Levy, E.D., Landry, C.R., and Michnick, S.W. (2010). Cell signaling. Signaling through cooperation. Science *328*, 983–984.

Li, J.Y., Liao, C.H., Wang, Q., Tan, Y.J., Luo, R., Qiu, Y., and Ge, X.Y. (2020). The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. Virus Res. *286*, 198074.

López-Muñoz, A.D., Kosik, I., Holly, J., and Yewdell, J.W. (2022). Cell surface SARS-CoV-2 nucleocapsid protein modulates innate and adaptive immunity. Sci. Adv. *8*.

Madahar, V., Dang, R., Zhang, Q., Liu, C., Rodgers, V.G.J., and Liao, J. (2023). Human Post-Translational SUMOylation Modification of SARS-CoV-2 Nucleocapsid Protein Enhances Its Interaction Affinity with Itself and Plays a Critical Role in Its Nuclear Translocation. Viruses *15*, 1600.

Manna, A., Zhao, H., Wada, J., Balagopalan, L., Tagad, H.D., Appella, E., Schuck, P., and Samelson, L.E. (2018). Cooperative assembly of a four-molecule signaling complex formed upon T cell antigen receptor activation. Proc. Natl. Acad. Sci. *115*, 201817142.

Mao, S., Cai, X., Niu, S., Wei, J., Jiang, N., Deng, H., Wang, W., Zhang, J., Shen, S., Ma, Y., et al. (2023). TRIM21 promotes ubiquitination of SARS-CoV-2 nucleocapsid protein to regulate innate immunity. J. Med. Virol. *95*.

Martínez-González, B., Soria, M.E., Vázquez-Sirvent, L., Ferrer-Orta, C., Lobo-Vega, R., Mínguez, P., de la Fuente, L., Llorens, C., Soriano, B., Ramos-Ruíz, R., et al. (2022). SARS-CoV-2 Mutant Spectra at Different Depth Levels Reveal an Overwhelming Abundance of Low Frequency Mutations. Pathogens *11*, 1–22.

Masters, P.S. (2019). Coronavirus genomic RNA packaging. Virology *537*, 198–207.

Mészáros, B., Sámano-Sánchez, H., Alvarado-Valverde, J., Čalyševa, J., Martínez-Pérez, E., Alves, R., Shields, D.C., Kumar, M., Rippmann, F., Chemes, L.B., et al. (2021). Short linear motif candidates in the cell entry system used by SARS-CoV-2 and their potential therapeutic implications. Sci. Signal. *14*, 1–26.

Mihalič, F., Simonetti, L., Giudice, G., Sander, M.R., Lindqvist, R., Peters, M.B.A., Benz, C., Kassa, E., Badgujar, D., Inturi, R., et al. (2023). Large-scale phage-based screening reveals extensive pan-viral mimicry of host short linear motifs. Nat. Commun. *14*, 2409.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682.

Mohapatra, T., and Dixit, M. (2022). IQ Motif Containing GTPase Activating Proteins (IQGAPs), A-Kinase Anchoring Proteins (AKAPs) and Kinase Suppressor of Ras Proteins (KSRs) in Scaffolding Oncogenic Pathways and Their Therapeutic Potential. ACS Omega *7*, 45837–45848.

Mosca, R., Pache, R.A., and Aloy, P. (2012). The role of structural disorder in the rewiring of protein interactions through evolution. Mol. Cell. Proteomics *11*, M111.014969-1-M111.014969-8.

Mu, J., Fang, Y., Yang, Q., Shu, T., Wang, A., Huang, M., Jin, L., Deng, F., Qiu, Y., and Zhou, X. (2020). SARS-CoV-2 N protein antagonizes type I interferon signaling by suppressing phosphorylation and nuclear translocation of STAT1 and STAT2. Cell Discov. *6*, 65.

Neduva, V., and Russell, R.B. (2005). Linear motifs: Evolutionary interaction switches. FEBS Lett. *579*, 3342–3345.

Pan, P., Shen, M., Yu, Z., Ge, W., Chen, K., Tian, M., Xiao, F., Wang, Z., Wang, J., Jia, Y., et al. (2021). SARS-CoV-2 N protein promotes NLRP3 inflammasome activation to induce hyperinflammation. Nat. Commun. *12*, 1–17.

Papadopoulos, J.S., and Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics *23*, 1073–1079.

Peti, W., and Page, R. (2013). Molecular basis of MAP kinase regulation. Protein Sci. *22*, 1698–1710.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. *30*, 70–82.

Ragusa, M.J., Dancheck, B., Critton, D.A., Nairn, A.C., Page, R., and Peti, W. (2010). Spinophilin directs protein phosphatase 1 specificity by blocking substrate binding sites. Nat. Struct. Mol. Biol. *17*, 459–464.

Saldivar-Espinoza, B., Garcia-Segura, P., Novau-Ferré, N., Macip, G., Martínez, R., Puigbò, P., Cereto-Massagué, A., Pujadas, G., and Garcia-Vallve, S. (2023). The Mutational Landscape of SARS-CoV-2. Int. J. Mol. Sci. *24*.

Sangster, A.G., Zarin, T., and Moses, A.M. (2022). Evolution of short linear motifs and disordered proteins Topic: yeast as model system to study evolution. Curr. Opin. Genet. Dev. *76*, 101964.

Shajahan, A., Pepi, L.E., Rouhani, D.S., Heiss, C., and Azadi, P. (2021). Glycosylation of SARS-CoV-2: structural and functional insights. Anal. Bioanal. Chem. *413*, 7179–7193.

Shi, G., Song, C., Torres Robles, J., Salichos, L., Lou, H.J., Lam, T.T., Gerstein, M., and Turk, B.E. (2023). Proteome-wide screening for mitogen-activated protein kinase docking motifs and interactors. Sci. Signal. *16*.

Shuler, G., and Hagai, T. (2022). Rapidly evolving viral motifs mostly target biophysically constrained binding pockets of host proteins. Cell Rep. *40*, 111212.

Simonetti, L., Nilsson, J., McInerney, G., Ivarsson, Y., and Davey, N.E. (2023). SLiM-binding pockets: an attractive target for broad-spectrum antivirals. Trends Biochem. Sci. *48*, 420–427.

Siqueira, J.D., Goes, L.R., Alves, B.M., Carvalho, P.S.D., Cicala, C., Arthos, J., Viola, J.P.B., De Melo, A.C., and Soares, M.A. (2021). SARS-CoV-2 genomic analyses in cancer patients reveal elevated intrahost genetic diversity. Virus Evol. *7*, 1–11.

Sorgeloos, F., Peeters, M., Hayashi, Y., Borghese, F., Capelli, N., Drappier, M., Cesaro, T., Colau, D., Stroobant, V., Vertommen, D., et al. (2022). A case of convergent evolution: Several viral and bacterial pathogens hijack RSK kinases through a common linear motif. Proc. Natl. Acad. Sci. U. S. A. *119*, 1–7.

Srivastava, G., Choy, M.S., Bolik-Coulon, N., Page, R., and Peti, W. (2023). Inhibitor-3 inhibits Protein Phosphatase 1 via a metal binding dynamic protein–protein interaction. Nat. Commun. *14*.

Syed, A.M., Ciling, A., Chen, I.P., Carlson, C.R., Adly, A., Martin, H., Taha, T.Y., Khalid, M.M., Bouhaddou, M., Ummadi, M., et al. (2023). SARS-CoV-2 evolution balances conflicting roles of N protein phosphorylation. Available SSRN Https//Ssrn.Com/Abstract=4472729.

Tokuriki, N., Oldfield, C.J., Uversky, V.N., Berezovsky, I.N., and Tawfik, D.S. (2009). Do viral proteins possess unique biophysical features? Trends Biochem. Sci. *34*, 53–59.

Tompa, P., Davey, N.E., Gibson, T.J., and Babu, M.M. (2014). A Million peptide motifs for the molecular biologist. Mol. Cell *55*, 161–169.

Tonkin-Hill, G., Martincorena, I., Amato, R., Lawson, A.R., Gerstung, M., Johnston, I., Jackson, D.K., Park, N., Lensing, S. V., Quail, M.A., et al. (2021). Patterns of within-host genetic diversity in SARS-COV-2. Elife *10*, 1–25.

Tugaeva, K. V., Hawkins, D.E.D.P., Smith, J.L.R., Bayfield, O.W., Ker, D.-S., Sysoev, A.A., Klychnikov, O.I., Antson, A.A., and Sluchanko, N.N. (2021). The Mechanism of SARS-CoV-2 Nucleocapsid Protein Recognition by the Human 14-3-3 Proteins. J. Mol. Biol. *433*, 166875.

Tugaeva, K. V., Sysoev, A.A., Kapitonova, A.A., Smith, J.L.R., Zhu, P., Cooley, R.B., Antson, A.A., and Sluchanko, N.N. (2023). Human 14-3-3 Proteins Site-selectively Bind the Mutational Hotspot Region of SARS-CoV-2 Nucleoprotein Modulating its Phosphoregulation. J. Mol. Biol. *435*, 167891.

Vignuzzi, M., Stone, J.K., Arnold, J.J., Cameron, C.E., and Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature *439*, 344–348.

Wang, X., Bajaj, R., Bollen, M., Peti, W., and Page, R. (2016). Expanding the PP2A Interactome by Defining a B56-Specific SLiM. Structure *24*, 2174–2181.

Watson, M., Almeida, T.B., Ray, A., Hanack, C., Elston, R., Btesh, J., McNaughton, P.A., and Stott, K. (2022). Hidden Multivalency in Phosphatase Recruitment by a Disordered AKAP Scaffold. J. Mol. Biol. *434*, 167682.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269.

Wu, W., Cheng, Y., Zhou, H., Sun, C., and Zhang, S. (2023). The SARS-CoV-2 nucleocapsid protein: its role in the viral life cycle, structure and functions, and use as a potential target in the development of vaccines and diagnostics. Virol. J. *20*, 6.

Yang, Z., Johnson, B.A., Meliopoulos, V.A., Ju, X., Zhang, P., Hughes, M.P., Wu, J., Koreski, K.P., Chang, T.-C., Wu, G., et al. (2023). Interaction between host G3BP and viral nucleocapsid protein regulates SARS-CoV-2 replication. BioRxiv Doi:10.1101/2023.06.29.546885.

Yaron, T.M., Heaton, B.E., Levy, T.M., Johnson, J.L., Jordan, T.X., Cohen, B.M., Kerelsky, A., Lin, T., Liberatore, K.M., Bulaon, D.K., et al. (2022). Host protein kinases required for SARS-CoV-2 nucleocapsid phosphorylation and viral replication. Sci. Signal. *15*, 1–17.

Yelemali, P., Hao, L., and Liu, Q. (2022). Mechanisms of host type I interferon response modulation by the nucleocapsid proteins of alpha- and betacoronaviruses. Arch. Virol. *167*, 1925–1930.

Yuan, S., Yan, B., Cao, J., Ye, Z., Liang, R., Tang, K., Luo, C., Cai, J., Chu, H., Chung, T.W., et al. (2021). SARS-CoV-2 exploits host DGAT and ADRP for efficient replication. Cell Discov. *7*, 100.

Zarin, T., Strome, B., Peng, G., Pritišanac, I., Forman-Kay, J.D., and Moses, A.M. (2021). Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. Elife *10*, 1–36.

Zhao, H., Nguyen, A., Wu, D., Li, Y., Hassan, S.A., Chen, J., Shroff, H., Piszczek, G., and Schuck, P. (2022). Plasticity in structure and assembly of SARS-CoV-2 nucleocapsid protein. PNAS Nexus *1*.

Zhao, H., Wu, D., Hassan, S.A., Nguyen, A., Chen, J., Piszczek, G., and Schuck, P. (2023). A conserved oligomerization domain in the disordered linker of coronavirus nucleocapsid proteins. Sci. Adv. *9*.

Zheng, X., Sun, Z., Yu, L., Shi, D., Zhu, M., Yao, H., and Li, L. (2021). Interactome Analysis of the Nucleocapsid Protein of SARS-CoV-2 Virus. Pathogens *10*, 1155.

METHODS

Mutational landscape

Mutation data were based on consensus sequences of SARS-CoV-2 isolates submitted to the GISAID , and downloaded on January 20, 2023 as database files metadata.tsv and nextclade.tsv preprocessed by Nextstrain (Hadfield et al., 2018). These contained ≈7.23 million genomes, of which only 5.06 million high quality sequences based on multiple criteria evaluated in the Nextstrain workflow were included here. As described previously (Zhao et al., 2022), for inspection of the mutational landscape a threshold of 10 observations of any mutation was used to filter adventitious sequencing errors. 746 sequences exhibiting insertions in the N protein were omitted. The Wuhan-Hu-1 isolate (GenBank QHD43423) (Wu et al., 2020) was used as the ancestral reference. Alignment of SARS and related betacoronavirus sequences was carried out with COBALT at NLM (Papadopoulos and Agarwala, 2007). SARS-CoV-2 sequences were grouped in sets of Omicron variants, Delta variants (Nextstrain 21J and later Delta clades), and sequences preceding 21J including Alpha, Beta, and other variant as well as ancestral sequences (termed pre-Delta). Processing, plotting, and analysis of the sequence data was performed with MATLAB (Mathworks, Natick, MA).

Prediction of N-protein disorder and visualization

The N-protein structure was predicted using ColabFold (Mirdita et al., 2022). Since no confidence is achieved for residue angles in disordered regions, these were artificially stretched closer to 180° for better visualization. The resulting structure was plotted using ChimeraX (Pettersen et al., 2021) and colored according to the predicted IUPred3 score (Erdos et al., 2021).

Analysis of SLiMs

SLiMs pattern recognition was carried out on distinct amino acid sequences with mutations in the disordered regions of interest, i.e., the N-arm (1-44), linker (180-247) and C-arm (364-419). To this end, all 5.06 million sequences were classified according to their N-protein IDR amino acid sequences. 512 distinct classes of N-arm sequences, 979 distinct linker sequences, and 556 C-arm sequences were identified that occurred in more than a threshold of 10 genomes.

In a preprocessing step, deletions were removed, sequence regions of interest were extended by 10 aa and, for efficiency to reduce server traffic, concatenated with AAAAA spacers to create composite sequences of length ≈1,000 aa prior to submission to the Eukaryotic Linear Motif Resource server (http://elm.eu.org) (Kumar et al., 2022). This concatenation process excludes from the statistics all N- and C-terminal-specific SLiMs, such as LIG_BIR_II_1 at the N-terminus of the N-arm and LIG_PDZ_Class_1 at the C-terminus of the C-arm, but avoids creating artificial termini at the limits of the IDR sequences of interest. System CURL commands were issued from a MATLAB script to send and receive data, carry out communication error control, and to parse the returned text to extract motif data. Motif content was mapped back and cropped onto the original sequence framework of interest, and motif name, multiplicity, and starting and ending positions tabulated for statistical analysis. Motifs with alternate regular expressions that completely overlapped were counted as a single instance. The frequency distribution of motif multiplicities across the different IDR sequence classes were derived from the table of motif multiplicity in each sequence.

To assess the probability of any SLiM occurring by chance, motif searches on 10,000 random sequences matching the amino acid content of the ancestral sequence of disordered linker and arms, respectively, were carried out. This was accomplished by performing random permutations reassigning amino acids to random positions within each IDR. The resulting sequences were analyzed for their motif content using the same computational pipeline as described for the sequence data above.

The analysis of hypothetical sequence space was carried out for the N-arm on the basis of the ancestral Wuhan-Hu-1 sequence, and the mutational landscape of observed amino acid mutations. In a first step, a set of possible mutations was created given a threshold of instances (or mutation frequency) for mutations to be considered. This creates an amino acid mutation table a_{pm} of possible replacements of the ancestral residue by mutation m at position $p,$ with $0 \leq m \leq M_p$, where M_p is the total number of mutations above the threshold frequency f at that position p. In a second step, the set of possible sequences with 3 mutations was created in MATLAB by permutation through all combinations (a_{ix}, a_{jy}, a_{kz}) with $i < j < k$ with $1 \leq x \leq M_i$, $1 \leq y \leq M_j$, and $1 \leq z \leq M_k$, avoiding redundant symmetric permutations. The resulting amino acids at positions i, j, and k replaced the amino acids in the ancestral sequence. To determine the motif space associated with the accessible sequence space, each of the resulting sequences was subjected to the same motif analysis pipeline described above. This analysis was repeated for different threshold frequencies, which creates an extended set of considered amino acid mutations. For efficiency in the analysis of next lower threshold frequencies, only the new sequences not already contained in the previous set of higher threshold mutations were subjected to motif analysis; and results were merged with those already obtained at the higher threshold.

FIGURES

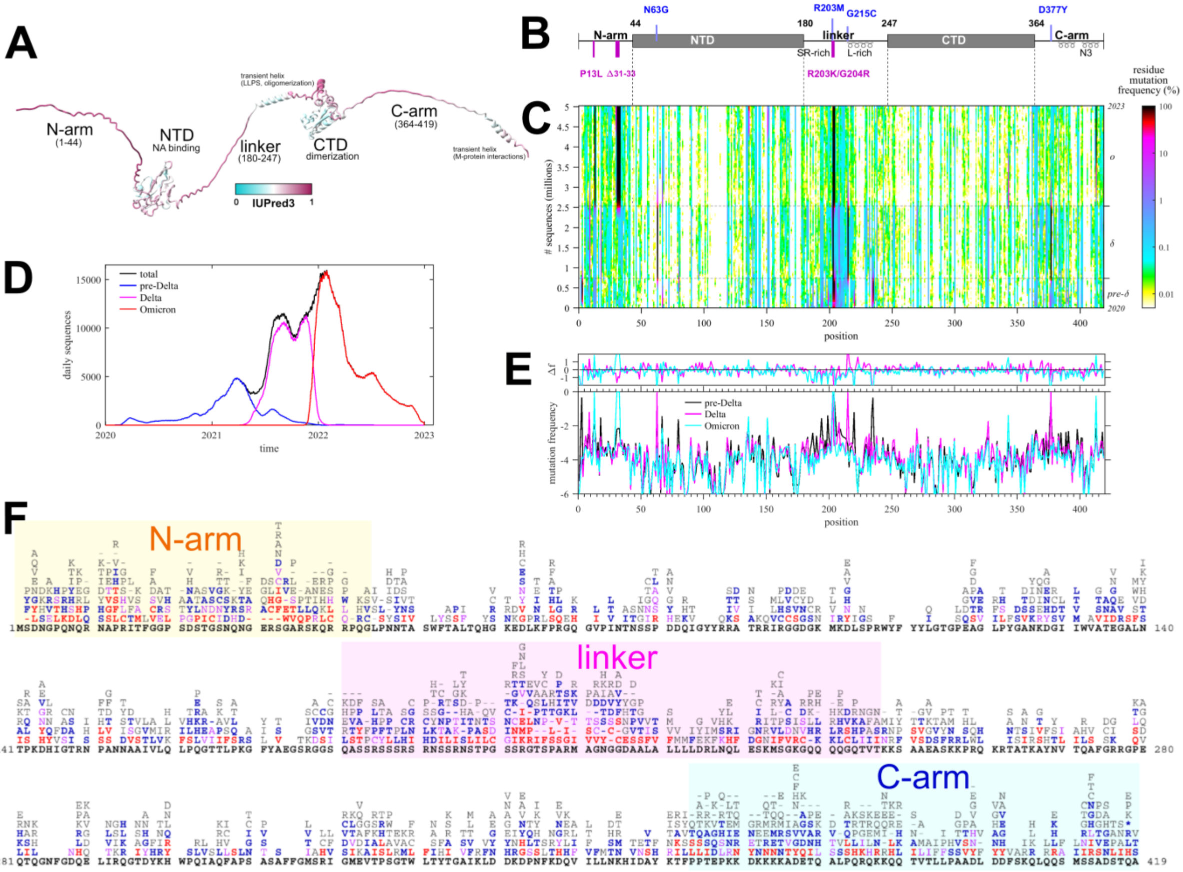

Figure 1. Mutational landscape of N-protein

(A) Predicted AlphaFold structure of N-protein, with extended IDRs for clarity, colored according to the disorder score. Labeled are the folded domains (NTD, CTD) and the three IDRs (N-arm, linker, and C-arm) with known assembly functions. (B) Schematic organization of N-protein with defining mutations of Delta (clade 21J, blue) and Omicron variant (magenta). (C) Time-dependent mutation frequencies (colors) at different positions (abscissa) plotted as a function of cumulative sequences (ordinate). The dotted horizontal lines are the time-points where Delta- and Omicron-variants rose to majority, as indicated in the history of daily deposited sequence numbers grouped by pre-Delta (comprising ancestral, Alpha, Beta, and other variants predating Delta 21J), Delta, and Omicron variants (D). (E) Average mutation frequencies of residues along N-protein positions, grouped by major variants as in (D). The upper panel shows the difference in average mutation frequencies of Delta and Omicron relative to the pre-Delta group. (F) Detailed list of observed amino acid mutations in each position, colored by the number of observed instances (≥ 1,000 red, ≥ 500 magenta, ≥ 100 blue, ≥ 10 grey, ancestral black).

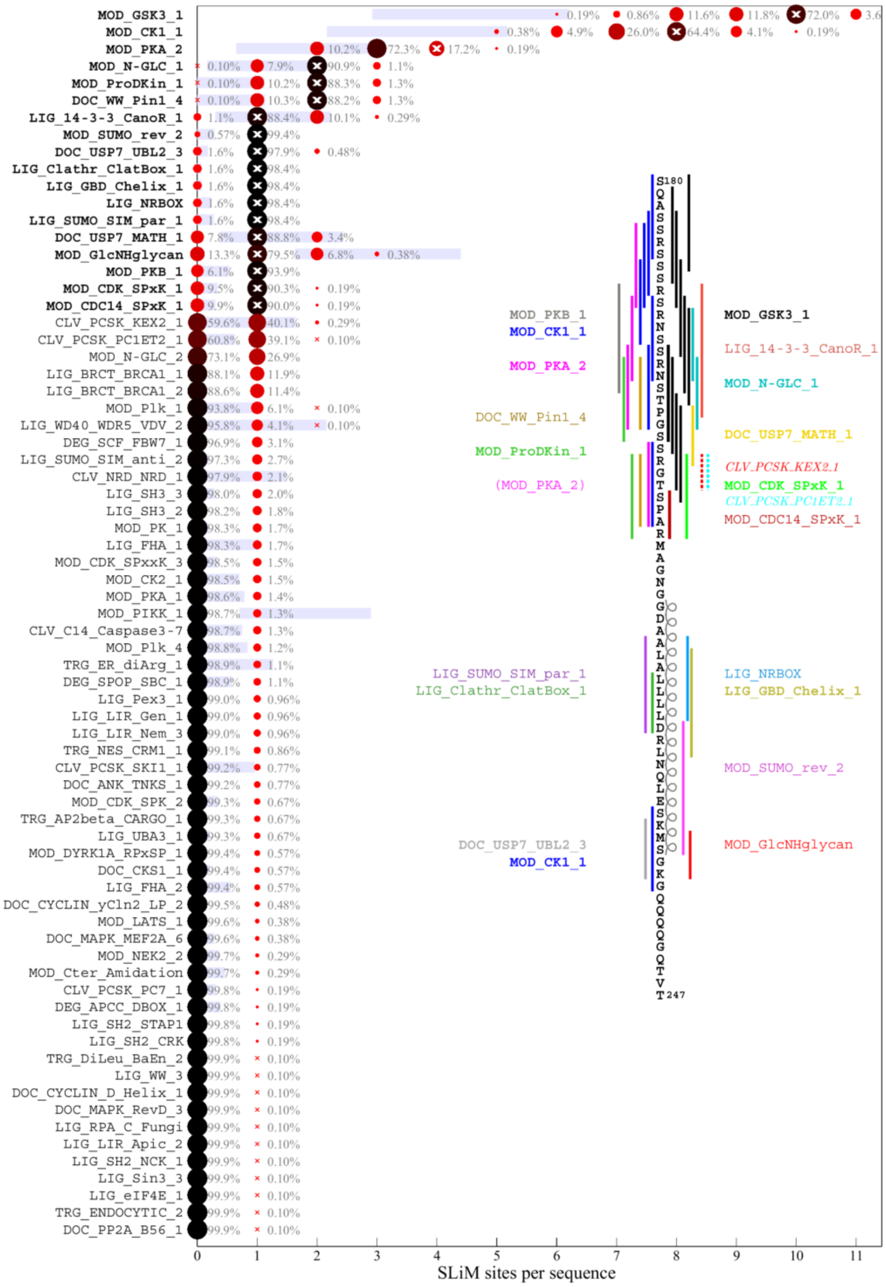

Figure 2. Predicted SLiM diversity in the linker IDR

Each row presents a histogram for the number of sites (abscissa) of a motif class in the ancestral reference sequence (bold) or in any of the 979 distinct mutant linker sequences. The frequency of the site numbers across the ensemble of sequences is indicated by symbol size and color, as well as the listed percentage value. The motif site number in the ancestral sequence is indicated by a white cross. The blue bars represent the mean ± standard deviation of the abundance of each motif in 10,000 randomly permutated reference sequences. The inset shows the location of the ancestral motifs as vertical bars and correspondingly colored motif name. New motifs emerging in the Omicron variant due to the defining R203K/G204R mutation are indicated as dotted lines and italicized motif name, and a disappearing motif is indicated in parenthesis.

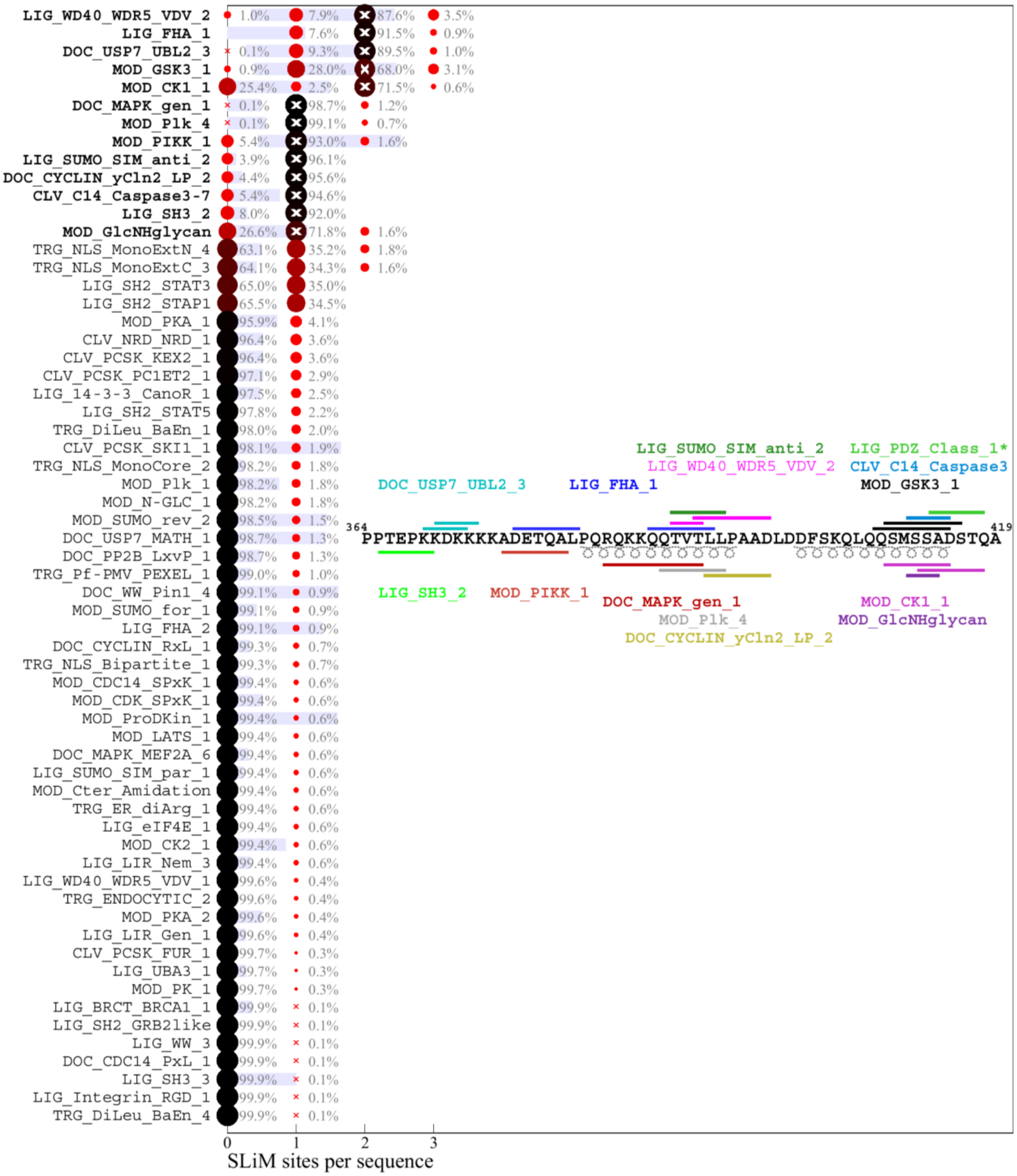

Figure 3. Predicted SLiM diversity in the C-arm IDR

Histograms for the number of sites of motif classes in the ancestral reference sequence (bold) or in any of the 556 distinct mutant C-arm sequences. Symbols and labels are as in **Figure 2**. The inset shows the location of the ancestral motifs as horizontal bars and correspondingly colored motif name. *LIG_PDZ_Class_1 is excluded from the distribution analysis (see **Methods**).

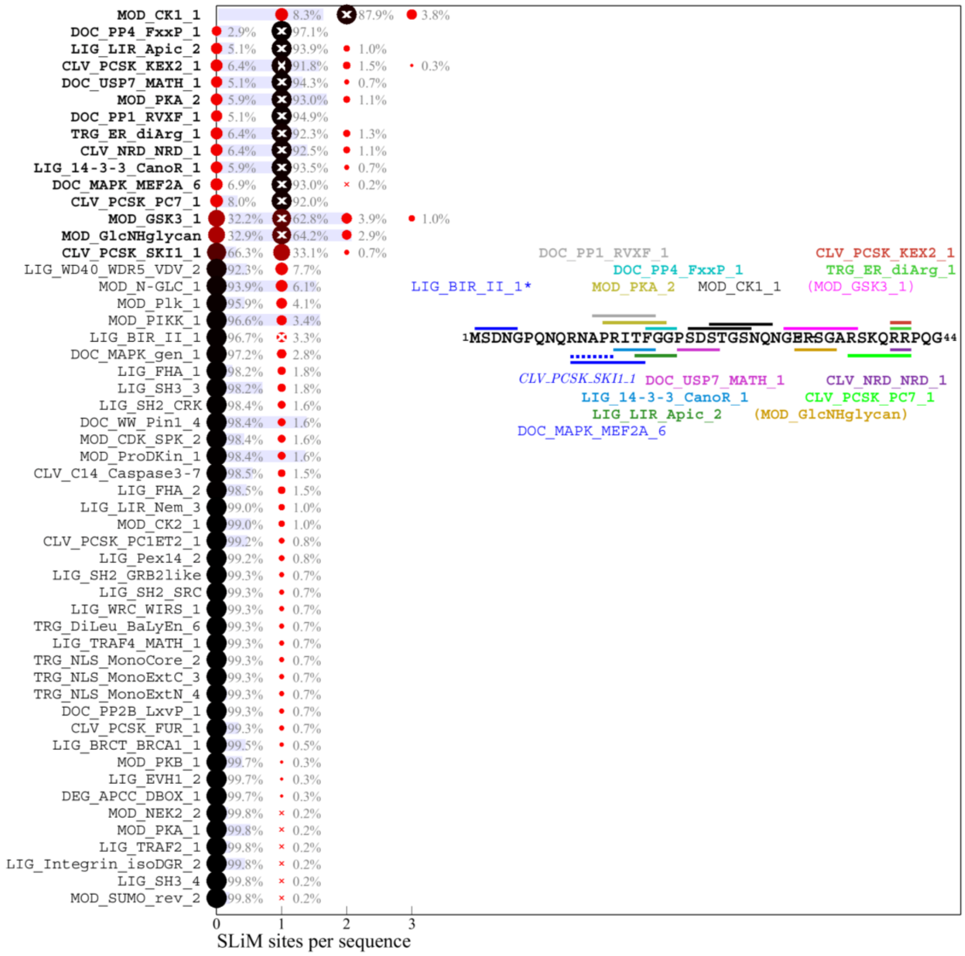

Figure 4. Predicted SLiM diversity in the N-arm IDR

Histograms for the number of sites of motif classes in the ancestral reference sequence (bold) or in any of the 512 distinct mutant N-arm sequences. Symbols and labels are as in **Figure 2**. The inset shows the location of the ancestral motifs as horizontal bars and correspondingly colored motif name. Omicron sequences have a defining substitution P13L and deletion in position 31-33. As a consequence, motifs in parenthesis do not occur in Omicron sequences, and motifs written in italics emerged. *LIG_BIR_II_1 is excluded from the distribution analysis (see **Methods**).

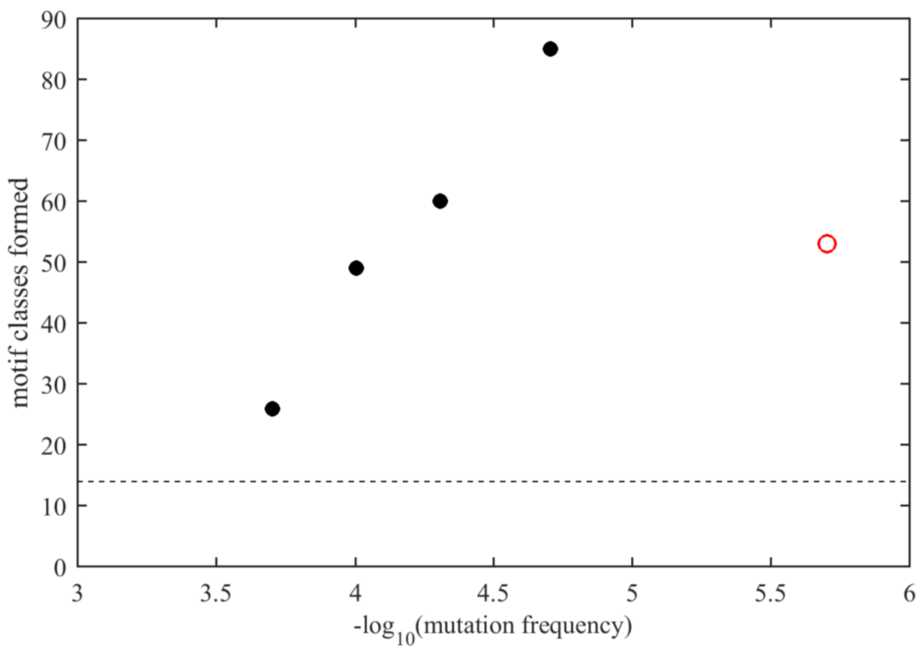

Figure 5. Estimated accessible motif space of N-protein N-arm IDR

Number of motif classes identified from the ELM database presented in the N-arm IDR examining a theoretical sequence space formed by three mutations per sequence permutated from the amino acid landscape (**Figure 1F**), considering only amino acid replacements with minimum frequencies indicated in the abscissa (number of observed mutation instances relative to a total of 5.06 million sequences). Black circles are completely evaluated theoretical sequence spaces, and the red circle is based on 512 distinct N-arm sequences contained in the GISAID database, comprising ≈0.03% of the theoretical sequence space. The dashed horizontal line is the number of motif classes displayed in the ancestral N-arm sequence.