Role of spatial patterning of N-protein interactions in SARS-CoV-2 genome packaging

Ian Seim[1,2,3], Christine A. Roden[1], and Amy S. Gladfelter[1,2,*]

[1]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, USA
[2]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, USA
[3]Department of Applied Physical Sciences, University of North Carolina at Chapel Hill, Chapel Hill, USA
[*]Contact: amyglad@unc.edu

Abstract
Viruses must efficiently and specifically package their genomes while excluding cellular nucleic acids and viral sub-genomic fragments. Some viruses use specific packaging signals, which are conserved sequence/structure motifs present only in the full-length genome. Recent work has shown that viral proteins important for packaging can undergo liquid-liquid phase separation (LLPS), where one or two viral nucleic acid binding proteins condense with the genome. The compositional simplicity of viral components lends itself well to theoretical modeling compared to more complex cellular organelles. Viral LLPS can be limited to one or two viral proteins and a single genome that is enriched in LLPS-promoting features. In our previous study, we observed that LLPS-promoting sequences of SARS-CoV-2 are located at the 5′ and 3′ ends of the genome, whereas the middle of the genome is predicted to consist mostly of solubilizing elements. Is this arrangement sufficient to drive single genome packaging, genome compaction, and genome cyclization? We addressed these questions using a coarse-grained polymer model, LASSI, to study the LLPS of nucleocapsid protein with RNA sequences that either promote LLPS or solubilization. With respect to genome compaction and cyclization, we find the most optimal arrangement restricts LLPS-promoting elements to the 5′ and 3′ ends of the genome, consistent with the native spatial patterning. Single genome packaging is possible for diverse arrangements of LLPS-promoting sequences in the genome, but only in limited conditions at the edge of the phase boundary. These results suggest that many and variably positioned LLPS promoting signals can support packaging in the absence of a singular packaging signal which argues against necessity of such a feature. We hypothesize that this model should be generalizable to multiple viruses as well as cellular organelles like paraspeckles, which enrich specific, long RNA sequences in a defined order.

Statement of significance
The COVID-19 pandemic has motivated research of the basic mechanisms of coronavirus replication. A major challenge faced by viruses such as SARS-CoV-2 is the selective packaging of a large genome in a relatively small capsid while excluding host and subgenomic nucleic acids. Genomic RNA of SARS-CoV-2 can condense with the Nucleocapsid (N)-protein, a protein component critical for packaging of many viruses. Notably, certain regions of the genomic RNA drive condensation of N-protein while other regions solubilize it. Here, we explore how the spatial patterning of these opposing elements promotes single genome compaction, packaging, and cyclization. This model informs future *in silico* experiments addressing spatial patterning of

genomic features that are experimentally intractable because of the extensive size of the genome.

Introduction

Biomolecular condensation is a simple and versatile way for cells to spatially and temporally control biochemistry. It is now clear that a wide variety of compartments likely form using the process of liquid-liquid phase separation (LLPS) which leads to a condensation of specific components out of bulk cytosol or nucleoplasm (7). The protein components of condensates tend to contain intrinsically disordered or low complexity sequences and RNA-binding domains (8). Many condensates also contain nucleic acids, and indeed RNA can promote phase separation in many instances (9). However, the contributions of specific RNA sequences and structures in condensate assembly, contents, and material properties is poorly understood (9).

Viruses present a powerful system to examine sequence specificity for both proteins and nucleic acids in phase separation because of their highly compact genomes and limited protein coding genes. Indeed, reports have emerged for VSV (10), respiratory syncytial virus (11), rabies (12), measles (13), and HIV (14) components showing the capacity to undergo LLPS.

SARS-CoV-2 is a positive strand RNA virus that has an exceptionally large genome of 30kb which is packaged into a relatively small capsid estimated to be ~100 nm in diameter (20). How the genome is selected out of the sub-genomic RNAs generated by the virus and the host transcriptome and sufficiently compressed to fit into a virion is not yet understood. A packaging signal for SARS-CoV-1 is still not clear, although one sequence has been found to be sufficient but not necessary to package RNA (21).

The nucleocapsid protein of SARS-CoV-2 undergoes LLPS (1,16,26,27), and our work found this occurs in an RNA sequence-specific manner with different regions of the genome (1). Remarkably, RNAs of the same length can either promote or limit phase separation depending on their sequences. The sequences with differing behavior also show distinct patterns of binding of N-protein with LLPS-promoting sequences having discrete patterned N-protein interactions, while RNA sequences that limit phase separation are uniformly coated in N-protein. The regions that promote phase separation are in the 5′ and 3′ ends of the genome, prompting us to speculate that phase separation could be relevant to packaging, as these LLPS-promoting sequences are present specifically on the whole genome and would not be together on sub-genomic or host RNAs.

Here, we developed a coarse-grained model to test the hypothesis that phase separation could be a relevant process for selecting and compacting a single genome. Our goal was to examine how the linear location of different RNA sequences in the genome generates spatially segregated and condensed RNA molecules. We first explore fragments of the SARS-CoV-2 genome that have opposing phase behavior when mixed with N-protein, as shown in (1). Specifically, the 5′ and 3′ ends of the genome promote phase separation, while the frameshifting element (FE) and central regions of the genome solubilize N-protein (Fig 1a). We next examine the spatial patterning of these opposing elements within a full genome model

and quantify its effects on phase separation, packaging of single genomes, genome compaction, and genome cyclization. We find that in this model localization of LLPS-promoting features to the 5′ and 3′ ends of the genome is sufficient to drive LLPS-based single genome packaging and genome compaction, and is necessary for genome cyclization.

Methods

LASSI model parameterization

Simulations were performed using LASSI (2) and run on the Longleaf computer cluster at UNC-Chapel Hill and on the Comet XSEDE cluster at the San Diego Supercomputer Center (22). Each simulation was run independently on a single compute node with 4GB RAM. The following parameter sets were used for all simulations:

Isotropic Binding Energies	RBD1	RBD2/dimer	5′ end RNA	FE RNA	5′ gRNA terminus	3′ gRNA terminus
RBD1	-0.5	0.0	-1.2	0.0	0.0	0.0
RBD2/dimer	0.0	0.0	0.0	0.0	0.0	0.0
5′ end RNA	-1.2	0.0	1.0	1.0	1.0	1.0
FE RNA	0.0	0.0	1.0	1.0	1.0	1.0
5′ gRNA terminus	0.0	0.0	1.0	1.0	1.0	0.0
3′ gRNA terminus	0.0	0.0	1.0	1.0	0.0	1.0

Table 1: Isotropic binding energies.

Anisotropic Binding Energies	RBD1	RBD2/dimer	5′ end RNA	FE RNA	5′ gRNA terminus	3′ gRNA terminus
RBD1	0.0	0.0	0.0	0.0	0.0	0.0
RBD2/dimer	0.0	-3.0	0.0	-5.0	0.0	0.0
5′ end RNA	0.0	0.0	0.0	0.0	0.0	0.0
FE RNA	0.0	-5.0	0.0	0.0	0.0	0.0
5′ gRNA terminus	0.0	0.0	0.0	0.0	0.0	-3.0
3′ gRNA terminus	0.0	0.0	0.0	0.0	-3.0	0.0

Table 2: Anisotropic binding energies.

The appropriate subsets of these interactions were used for the simulations involving N-protein alone, N-protein and 5′ end RNA, N-protein and FE RNA, and all spatial rearrangements of beads in genomic and 6-bead RNA chain simulations. Simulations of N-protein alone were run at 15 temperatures linearly spaced between 1 a.u. and 2 a.u., and 15 concentrations logarithmically spaced between 1e-4 and 1e-0.3, each for 1e9 total time steps with the first 5e6 timesteps discarded as equilibration. For all ternary systems, simulations were run at a single temperature, 1 a.u., for 3e9 timesteps and 5e6 steps discarded as equilibration. Concentrations

of each component were chosen to be logarithmically spaced across the ranges of interest, with hundreds of ([protein], [RNA]) coordinates sampled for each system (Fig S1). For N-protein with 6-bead RNA chains, the maximum number of N-protein and RNA chains were used such that the total beads in the system never exceeded 40000 and the target stoichiometries were satisfied. For simulations with gRNA and N-protein, 10 gRNA molecules were always used, and the number of N-protein chains was altered to match target stoichiometries. At the lowest gRNA and highest N-protein volume fractions, the number of gRNA molecules was gradually scaled down to 1 due to computational limitations on the sizes of the systems. At each coordinate, 2 independent simulations were run with and without interactions, for a total of 4 simulations at each concentration coordinate.

Simulation analysis

Analysis was performed using scripts within LASSI and custom scripts that relied upon the ovito python module (23). Specifically, LASSI outputs a global density inhomogeneity value, $\bar{\rho}$, for each simulation, which is used to determine whether phase separation has occurred. We used the ovito module to calculate cluster sizes using a maximum cutoff of $3^{1/2}$ since this is the farthest apart two interacting particles can be in a cubic lattice. The radius of gyration for each gRNA molecule was calculated according to the equation in the text. Ovito was also used to calculate end-to-end distances of genomic RNAs. We counted genomes as cyclized if their terminal beads were within $3^{1/2}$ units of each other, since this is the farthest apart two adjacent particles can be in a cubic lattice. For all of these metrics, a snapshot from the final frame of each simulation was used for calculation. A single frame supplied sufficient statistics, since we did not see significant variation among independent runs for each system (Fig S2). $\bar{\rho}$, cluster metrics, $\langle R_g \rangle$, cyclized genome percentages were interpolated to a grid logarithmically spaced between 1e-4 and 1 with a discretization of 80 points along the protein and RNA axes. Interpolation was performed using interpolate.griddata from scipy with a two-dimensional piecewise cubic, continuously differentiable, and approximately curvature-minimizing polynomial surface. Contours at $\bar{\rho} = 0.025$ were used to determine phase boundaries. Plots were made using matplotlib.

Results

Simulations recapitulate known phase behavior and binding patterns of N-protein with FE and 5′ end RNA

We first sought to characterize the phase behavior of different regions of the viral genome by focusing on simulations involving the N-protein with either the first 1000 nucleotides of the genome (5′ end RNA) or 1000 nucleotides around the frameshifting region located between ORF1A and ORF1B (FE RNA). Simulations were carried out using LASSI, which employs coarse-grained representations and generates full phase diagrams and thermodynamic readouts (2). The first consideration was how to represent the N-protein and RNA in a coarse-grained manner based on existing data. The N-protein has two RNA binding domains (RBDs), a dimerization domain which overlaps with RBD2, and three intrinsically disordered regions (Fig 1a). RBD1 is conserved across multiple betacoronavirus genomes and has previously been

shown to interact with the conserved sequences and structures in the 5′ UTR (15). We previously demonstrated that a single point mutant within RBD1, Y109A, greatly reduces N-protein phase separation and changes N-protein interactions with 5′ end RNA, while only minimally affecting protein binding and phase behavior with FE RNA (1). Since FE RNA primarily solubilized N-protein, and deletion of the RBD2/dimerization domain blocked N protein LLPS (16), we postulated that FE may block N-protein phase separation by specifically interacting with RBD2 and preventing N-protein dimerization. Thus, we hypothesize that N-protein RBD1 primarily binds to 5′ end RNA, while RBD2 primarily binds to FE RNA.

Based on these data, we represented the N-protein as simply as possible using two spheres. The first sphere participates in weak isotropic interactions with other N-proteins (representing the association of the IDRs) and with the 5′ end RNA (representing RBD1). The second sphere participates in strong anisotropic interactions with other N-proteins to capture dimerization and the interaction with the FE RNA via RBD2 (Fig 1a). These latter anisotropic interactions operate under the assumption that N-protein dimerization competes with binding to FE RNA, since anisotropic binding in this model is one-to-one. To model charge effects, the RNA molecules experience an isotropic repulsive force amongst themselves. Using these specifications, we sought to qualitatively reproduce the phase behavior among these molecules established experimentally in (1). We found that, relative to N-protein phase separation on its own (Fig 1b), addition of 5′ end RNA promotes enhanced phase separation across a wide range of protein and RNA concentrations (Fig 1c), while addition of FE RNA does not promote phase separation and solubilizes N-protein at sufficiently high concentrations of RNA (Fig 1d).

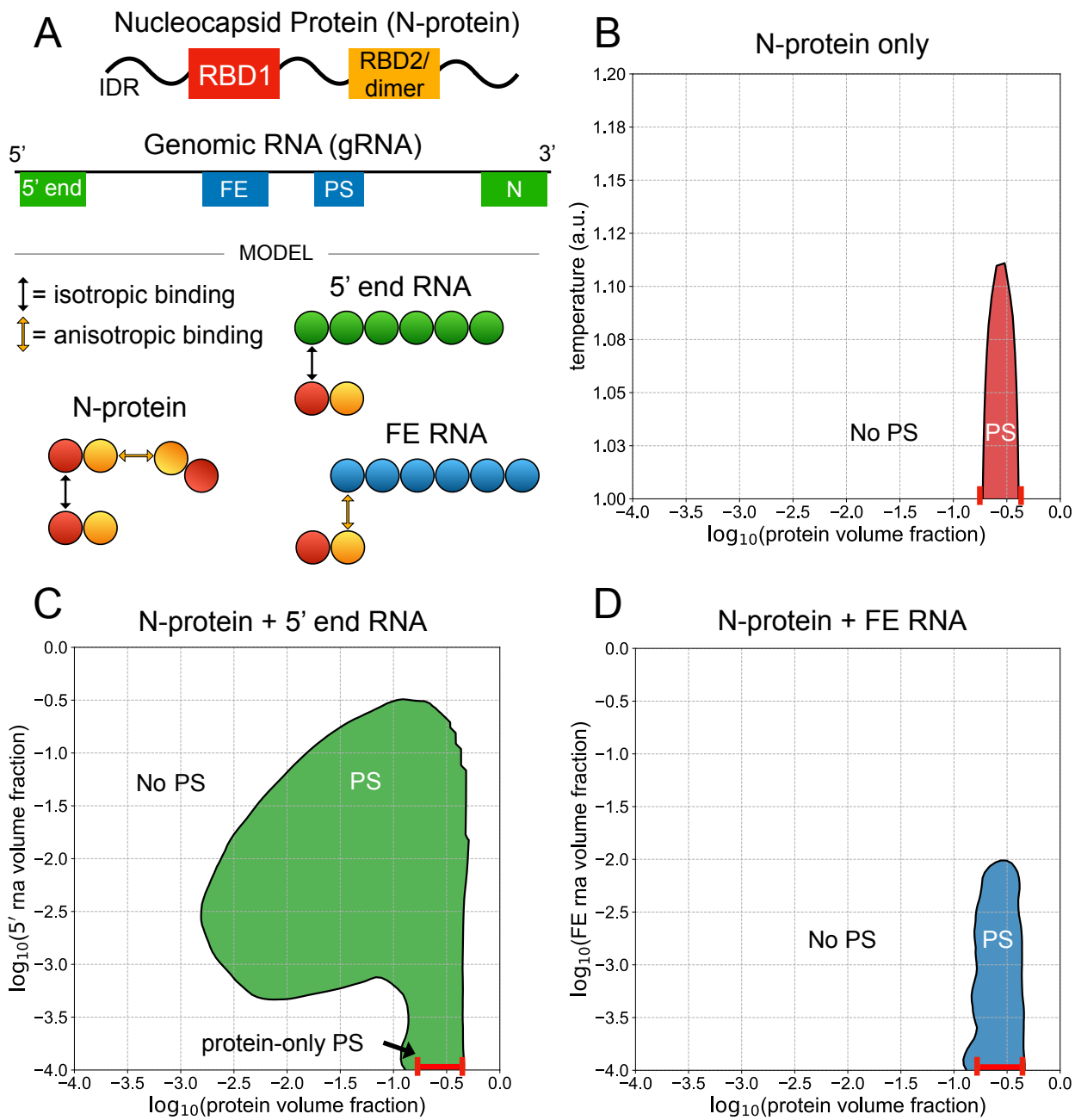

Figure 1: 5′ end and FE RNA with N-protein have opposing phase behavior. A) N-protein is represented as a two-bead chain, with the first bead participating in isotropic homotypic interactions, and the second bead participating in anisotropic homotypic interactions. Both 5′ end and FE RNA segments are roughly three times larger than N-protein and are represented by six beads each. N-protein interacts with all 5′ end RNA beads via isotropic binding with its first bead, and it interacts with all FE RNA beads via anisotropic binding with its second bead. This interaction with FE competes with N-protein dimerization. B) N-protein phase separates (PS) in a narrow concentration and temperature range on its own. C) N-protein with 5′ end RNA at temperature 1 a.u. phase separates across a wider concentration range than on its own. D) N-

protein with FE RNA at temperature 1 a.u. is solubilized at sufficiently high FE RNA concentrations.

In addition to opposing phase behavior, the binding of N-protein to 5′ end and FE RNA as a function of N-protein concentration was shown to be distinctly patterned based on protein crosslinking in (1). While FE RNA is uniformly coated with protein across a wide range of protein concentrations, 5′ end RNA has a few discrete binding sites and is only gradually coated more with protein as the protein concentration is increased. The simulations report a similar distinct pattern of protein interactions. For a fixed RNA concentration and at low bulk protein concentrations, 5′ end RNA is initially coated with very little protein (Fig 2a). As the bulk protein concentration is increased, 5′ end RNA sharply transitions to binding large amounts of protein (Fig 2a). In contrast, FE RNA binds more protein at low protein concentrations, relative to 5′ end RNA, and experiences a more gradual transition to high amounts of bound protein (Fig 2b). In all, the distinct protein binding behavior of these two RNA elements is consistent with that found in (1) (Fig 2c). Thus, the simulation results are consistent with the experimental system and support that the coarse-grained representations and binding energies are reasonable approximations of the actual system.

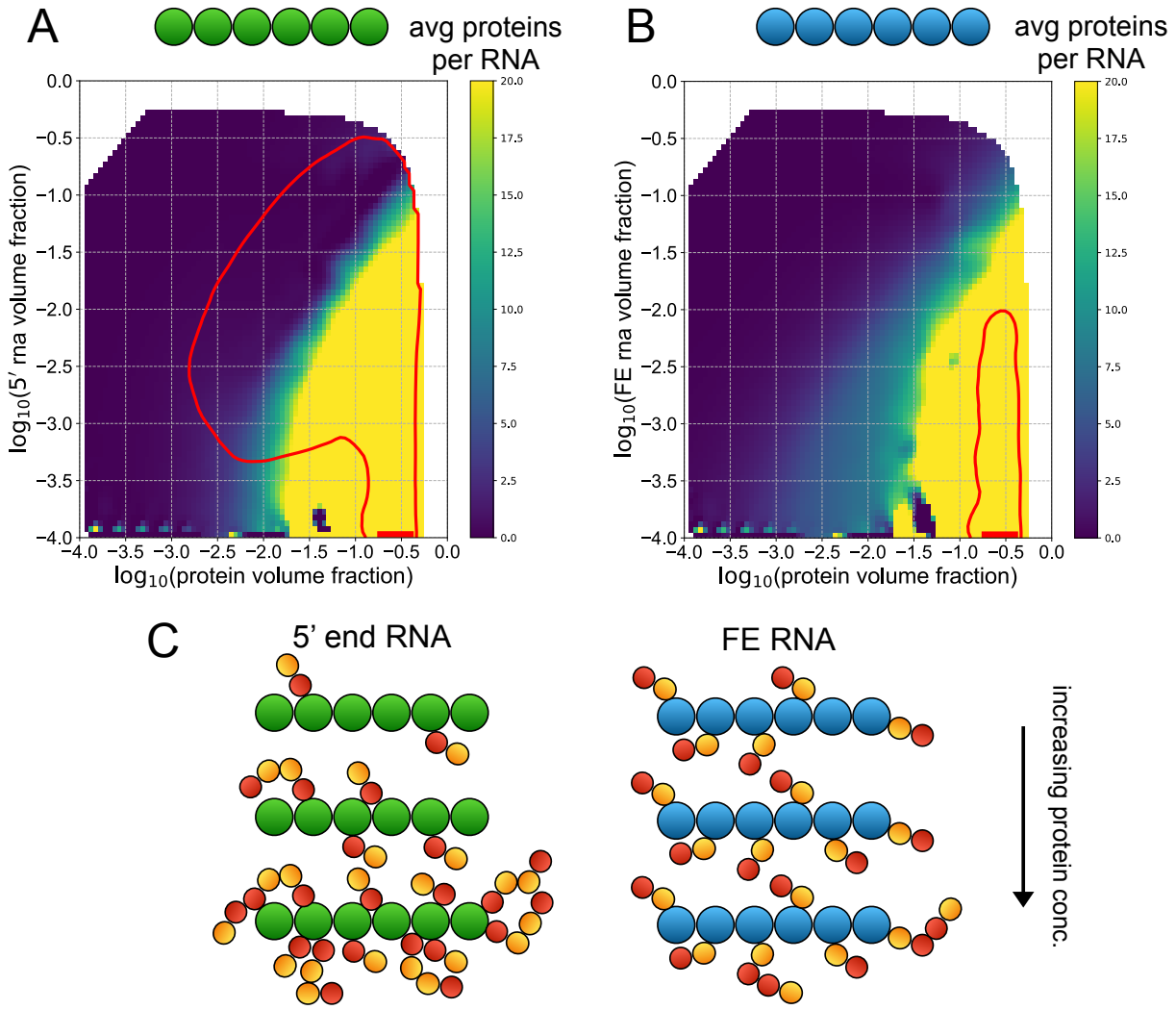

Figure 2: 5′ end and FE RNA have distinct N-protein binding behavior. A) and B) Phase boundaries for 5′ end RNA with N-protein (A) and FE RNA (B) with N-protein are shown in red. The heatmaps indicate the average number of protein chains per RNA chain for each cluster identified in each simulation. If more than one RNA chain is in a cluster, the ratio of protein chains to RNA chains within that cluster is reported. C) For any given fixed RNA concentration, as the protein concentration is increased, 5′ end RNA shows a sharp transition to highly bound protein, while FE shows a more gradual and transition, with more proteins bound at low protein concentrations relative to the 5′ end system.

Whole genome simulations reveal effects of phase separation on single genome packaging, genome compaction, and genome cyclization

The large size (30kb) of the genome makes it challenging to synthesize *in vitro* for experiments. We therefore were eager to use this simulation space to ask questions about how the different RNA-sequence elements will behave when present *in cis* on the same polymer, as they are found in the native virus. To address if specific arrangements of RNA encoded features could be sufficient for packaging a single genome, we utilized the same representations of the 5′ end and FE RNA and the N-protein described above and assembled a system that represents the N-protein and the entire viral genome (gRNA) (Fig 3a). In addition to the 5′ end of the gRNA, the 3′ end was also found to promote phase separation with N-protein and shared similar protein binding behavior and RNA sequence features (Fig 1a) (1). Central regions of the gRNA that were studied behaved similarly to the FE RNA, and further, the internal portions of the genome were predicted to be more similar to the FE than 5′ end RNA (1). Taking into account the relative sizes of nucleotides and amino acids, the gRNA is roughly 90x larger than N-protein, so our representation involves 180-bead chains for each genome. An additional strong, anisotropic interaction between the terminal beads of each gRNA chain is added to represent known nucleotide complementarity between the 5′ and 3′ ends of the gRNA and propensity to crosslink *in vivo* (3).

For a given protein concentration, phase separation of N-protein in the gRNA system occurs over a smaller range of RNA concentrations relative to the 5′ end system (Fig 1c, Fig 3b). This is consistent with experiments in which 5′ end and FE RNAs were combined *in trans* which led to more limited phase separation compared to 5′ end RNA alone (1). However, phase separation can occur at lower concentrations of protein and gRNA due to the length of the gRNA chains, which is also consistent with experiments that were performed with RNA purified from infected cells that contained gRNA (1). We quantified the average amount of protein bound to a single genome and found that the contour delineating an equal volume fraction of protein and gRNA per cluster aligns well with the high-RNA concentration edge of the phase envelope (white dotted line, Fig 3b). The area to the right of this contour indicates bulk concentrations of N-protein and gRNA that lead to a majority of volume fraction per cluster occupied by N-protein. This region includes almost all of the phase separating regime and likely captures the most relevant stoichiometries of gRNA and N-protein during infection in host cells and virion assembly (4).

During virion assembly, single genomes must be packaged within a capsid built of structural proteins and N-protein (5), so we also quantified how many gRNA chains were in each phase separated cluster. We defined a simple metric to quantify single gRNA packaging; the number of gRNA-containing clusters in the system is divided by the total number of gRNA chains. Therefore, the metric is 1 when single-genome packaging is perfect, and approaches 0 as multiple genomes are clustered together. Our analysis indicates that, for the most part, phase separation hinders packaging of single gRNAs in clusters (Fig 3c). However, at the lowest N-protein and gRNA concentrations within the phase boundary and along the edge of the boundary, single packaging can occur. This finding supports the possibility that phase separation can promote singularity in packaging, but only with specific bulk concentrations of protein and RNA.

If arranged linearly, the 30kb genome has an end-to-end length of roughly 10000nm. However, during virion assembly, this genome must be packaged into a viral particle with a diameter of about 100nm, representing an immense compaction challenge. We reasoned that phase separation with N-protein could provide a simple mechanism for gRNA compaction. To quantify compaction, we measured the average radius of gyration of gRNA in each simulation (17),

$$R_g^2 = \frac{1}{N}\sum_{k=1}^{N}(\vec{r}_k - \vec{r}_{\mathrm{mean}})^2$$

where $N = 180$ is the number of monomers in a chain, $\vec{r}_k$ is the position of monomer *k*, and $\vec{r}_{\mathrm{mean}}$ is the average position of monomers in the chain. R_g is computed for each gRNA in a simulation, and the average across gRNAs, $\langle R_g \rangle$, is reported for each simulation. We found that phase separation indeed leads to more compact genomes (Fig 3d). In particular, for intermediate bulk protein concentrations (between protein volume fractions of 0.01 and 0.1), genomes are especially compact, with $\langle R_g \rangle$ between 5 and 10 particle diameters. Notably, at the concentrations within the phase boundary where genomes are singularly packaged, there is a low degree of gRNA compaction, and vice versa, suggesting that these processes compete (Fig. 3d).

Genome cyclization is important for replication of many RNA viruses (6), and there is recent *in vivo* evidence of cyclization of the SARS-CoV-2 genome (3). We thus sought to characterize the potential role of phase behavior in genome cyclization in our model. We hypothesized that the similar RNA features at the 5′ and 3′ ends of the genome would promote cyclization when undergoing phase separation with N-protein. We defined a cyclized genome as one whose terminal beads occupy adjacent lattice sites. For the final frame of each simulation, we quantified the percentage of genomes that met this criterion. Strikingly, we see that up to 50% of genomes are cyclized for concentrations within the phase boundary, with additional but lesser cyclization beyond the equal N-protein gRNA volume fraction contour (Fig 3e). Thus, it appears that sufficient binding of N-protein is essential for cyclization of gRNA molecules, with phase separation providing additional efficiency.

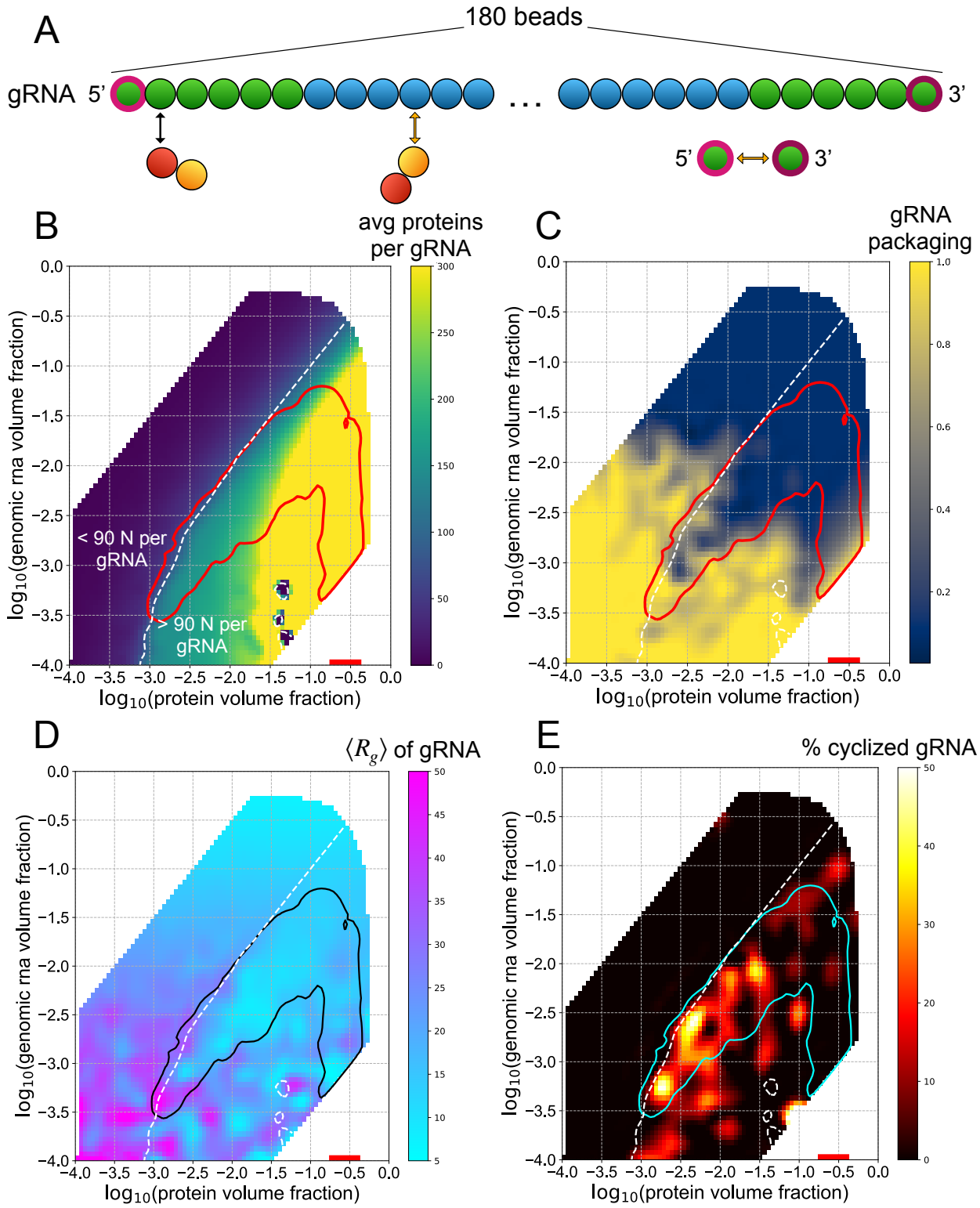

Figure 3: gRNA phase separates with N-protein which limits single genome packaging but promotes genome compaction and cyclization. A) Genomic RNA (gRNA) is represented as a chain with 180 beads. The terminal 6 beads on the 5′ and 3′ end are 5′ end-like beads, and the rest are FE-like beads. An additional anisotropic interaction among the terminal beads is added

to represent known nucleotide complementarity. B) The phase boundary is shown in red. As in Fig 2, the heatmap indicates the average number of protein chains per RNA chain for each cluster identified in each simulation. The white dotted line indicates the contour along which an equal volume fraction of N-protein and gRNA is found within a given cluster, i.e. 90 N-proteins for each gRNA. This contour is shown in all subsequent panels. C) The phase boundary is shown in red. The heatmap shows the single gRNA packaging metric, which is the number of clusters containing RNA divided by the total number of genomes in each simulation. A value of 1 represents perfect single-genome packaging. D) The phase boundary is shown in black. The heatmap shows the average radius of gyration of gRNA chains. E) The phase boundary is shown in cyan. The heatmap shows the percentage of cyclized genomes. Each gRNA chain is categorized as cyclized if its terminal beads are in adjacent lattice positions.

Spatial patterning gRNA mutants can enhance single genome packaging but limit compaction and prevent genome cyclization

Since the 5′ and 3′ ends of the gRNA were both found to promote phase separation with N-protein, we hypothesized that this spatial arrangement of phase separation-promoting elements at the ends of the genome may be relevant to packaging. We investigated the importance of the arrangement of phase separation-promoting sequences on the ends of the gRNA by designing mutants where these regions are rearranged. We created three mutants, all of which retain 12 5′ end-like beads and the anisotropic interaction among their terminal beads. The 5′ end-like beads are repositioned either in the middle of the genome (middle), uniformly throughout (uniform), or on one end (end) (Fig 4a). The phase boundaries of each of the mutants remain relatively unchanged with respect to the WT system, suggesting that the relative abundance and not the spatial patterning of 5′ end and FE RNA beads determines the bulk concentrations at which phase separation occurs (Fig 4). However, we observed differences in the genome packaging metrics relative to WT and among the mutants.

The uniform system is highly efficient at packaging single genomes into clusters, with middle gRNA slightly less effective, and end gRNA behaving similarly to WT (Fig 4b). Thus, it appears that dispersed or centrally located phase separating elements within the gRNA are preferred for single genome packaging by N-protein alone. Interestingly, however, effects on genome compaction appear to behave in an opposite manner. Uniform gRNA does not significantly compact upon phase separation, with middle gRNA compacting intermediately, and end gRNA compacting similarly to WT (Fig 4c). As in the WT gRNA simulations these results reiterate the possibility that single packaging and compaction of genomes by phase separation with N-protein arise from competing processes and will be further explored in the next section.

Given the positioning of the wildtype LLPS promoting sequences at the ends of the genome, we postulated that rearrangement of the location of these sequences would have the strongest impact on genome cyclization. To this end, we also quantified genome cyclization for these mutants, and found that none of them were able to cyclize genomes (Fig 4d). Importantly, each of these systems maintains an intrinsic bonding capability between the terminal beads of its gRNA chains. However, since the chains are so large, they cannot efficiently locate each other during the course of the simulation. Thus, the localization of phase separation to the 5′ and 3′

ends with N-protein in the WT system is necessary for the positioning of the genome ends for binding and cyclization (Fig 3e).

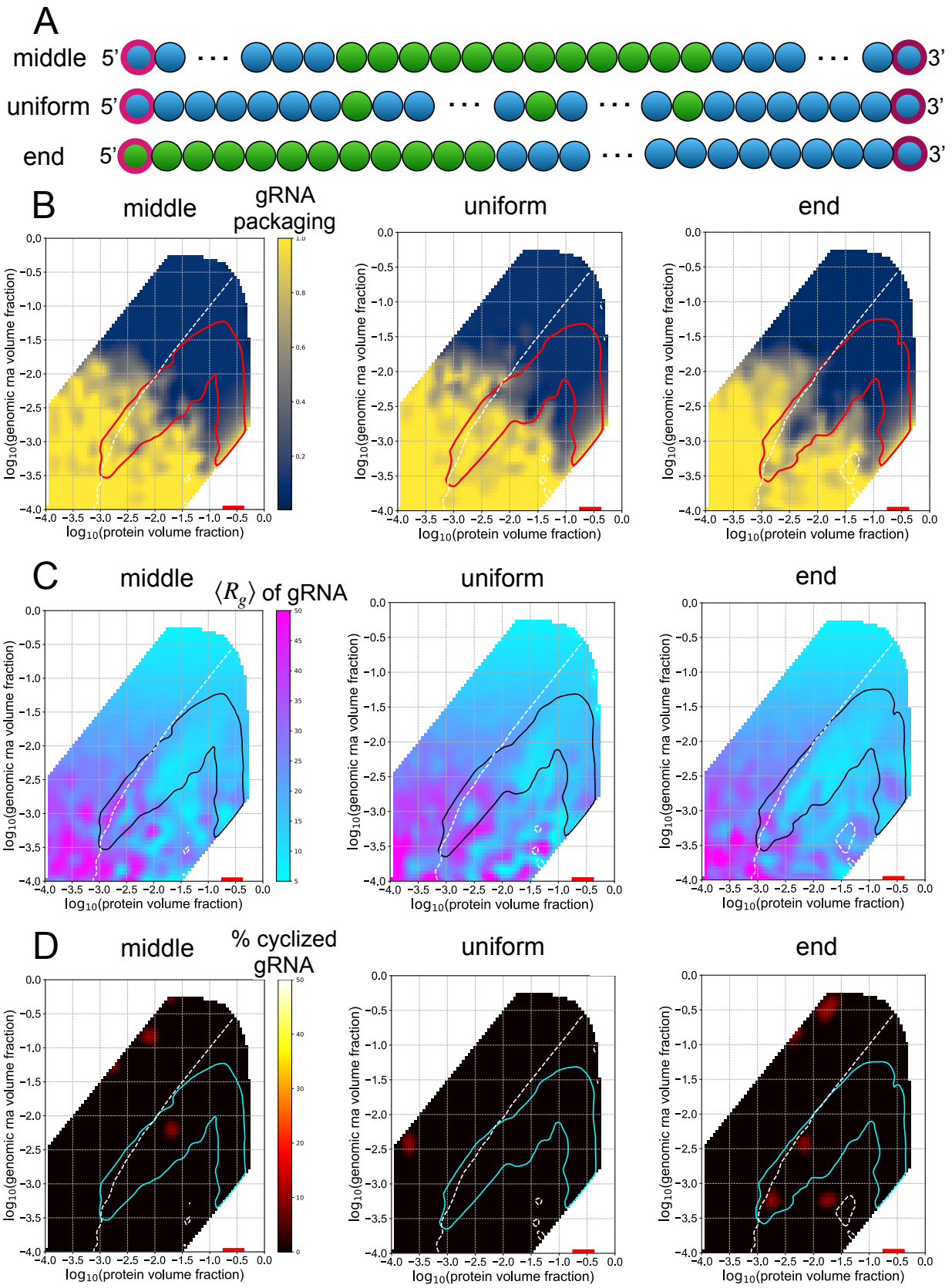
A
middle 5' 3'
uniform 5' 3'
end 5' 3'
B
middle
gRNA packaging
uniform
end
C
middle
$\langle R_g \rangle$ of gRNA
uniform
end
D
middle
% cyclized gRNA
uniform
end
$\log_{10}$(genomic rna volume fraction)
$\log_{10}$(protein volume fraction)

Figure 4: Spatial patterning gRNA mutants show genome packaging metrics. A) Three spatial patterning genome mutants were constructed. Each mutant retains the same number of 5′ end-like and FE-like beads as in WT. The terminal beads maintain their anisotropic interaction, regardless of their identity as 5′ end-like or FE-like. B) Phase boundaries are shown in red for each spatial patterning mutant system. The white dotted line indicates the contour along which an equal volume fraction of N-protein and gRNA is found within a given cluster and is included in all subsequent panels. The heatmaps show the single gRNA packaging metric. C) Phase boundaries are shown in black. The heatmaps show the average radius of gyration of gRNA chains. D) Phase boundaries are shown in cyan. The heatmaps show the percentage of cyclized genomes.

Phase behavior and gRNA packaging can be continuously tuned by titrating 5′ end-like and FE-like RNA elements

Our studies so far have assumed that 5′ end and FE RNA elements are perfectly distinct in their N-protein binding modes and thus phase behavior. The initial empirical studies on which these assumptions are based were limited in scope, but given the size of the genome, it is likely that further LLPS-promoting or inhibiting elements are also heterogeneously distributed throughout the gRNA. We investigated the phase behavior of six-bead RNA chains that gradually transition from a pure 5′ end identity to a FE one (Fig 5). As the identity of the RNA chains changes, the phase boundary gradually shifts from one extreme to the other, suggesting that phase behavior and resultant properties like single packaging, compaction, and cyclization can be precisely tuned by proper positioning of these elements in the gRNA.

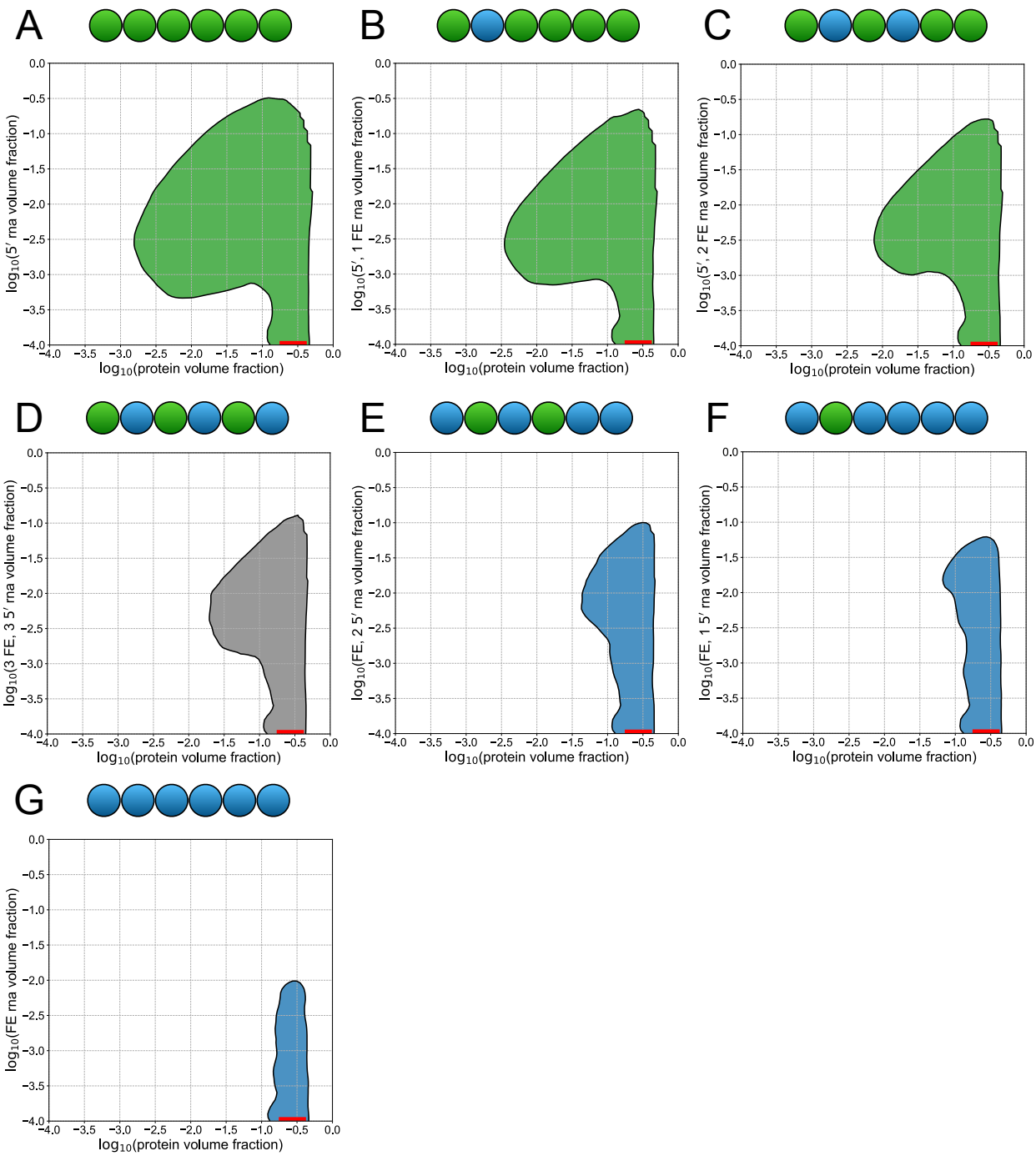

Figure 5: N-protein with FE-5′ end RNA blends show a continuous change in their phase diagrams. A)-G) Each system involves an RNA chain that includes 5′ end-like (green) and FE-like (blue) beads in the indicated positions, and N-protein. Phase diagrams for each system are shown.

With this understanding, in addition to the optimal packaging in the uniform system and compaction and cyclization in the WT system, we hypothesized that a genome could evolve that would optimally function according to all of these metrics. We designed a gRNA which has

the WT arrangement at the ends, with 12 additional 5′ end-like beads uniformly spaced throughout its length (Fig 6a). We found that phase separation occurs over a slightly broader range of concentrations, and that all studied genome packaging metrics are enhanced, relative to WT. Single genome packaging is slightly more preferable at low concentrations of N-protein and gRNA relative to WT, though efficient packaging is still limited to this corner of the phase envelope (Fig 6b). Single packaging is less efficient relative to the uniform mutant, which provides further evidence that the WT arrangement of phase separation promoting beads at the 5′ and 3′ ends counteracts the effects of uniformly distributed phase separation promoting beads. In our optimal gRNA, significant genome compaction occurs across much of the phase envelope, with an optimal range of N-protein extending almost an order of magnitude lower than in WT (Fig 6c). Unlike the results for single genome packaging, the genome compaction data suggest that the combination WT and uniform arrangements of 5′ end-like beads cooperatively promote compaction. Genome cyclization remains high, similar to WT, with most efficient cyclization occurring in the lower-left corner of the phase envelope (Fig 6d).

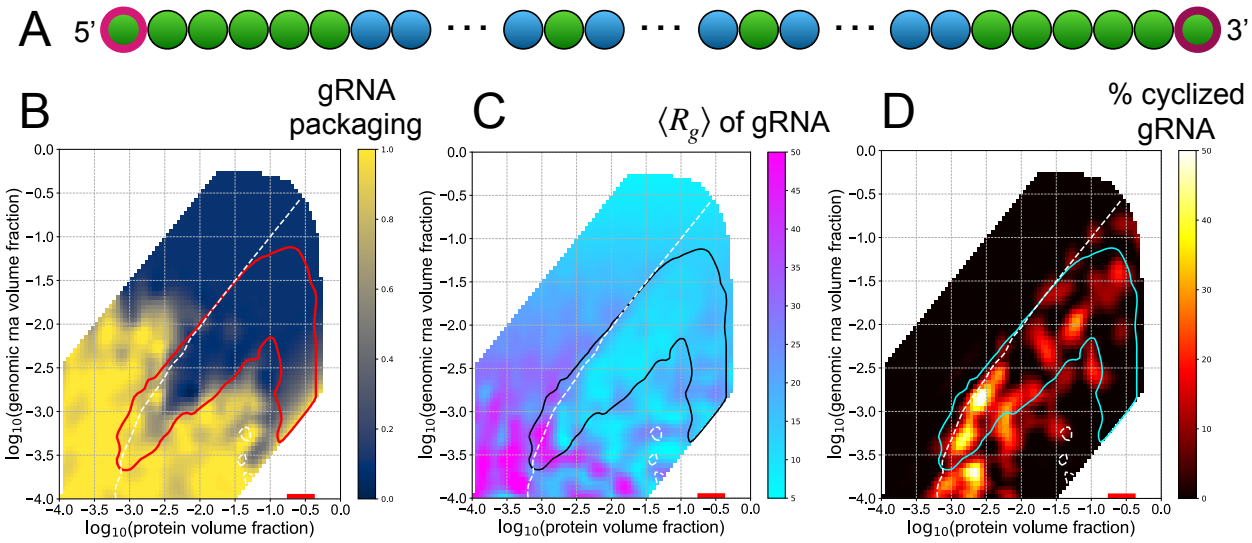

Figure 6: WT and uniform elements optimize genome packaging when combined. A) The optimal gRNA pattern for singular and compact packaging has 6 5′ end-like beads at each end, with 12 additional 5′ end-like beads distributed uniformly throughout the rest of the chain. B) The phase boundary is shown in red. The white dotted line indicates the contour along which an equal volume fraction of N-protein and gRNA is found within a given cluster and is included in all subsequent panels. The heatmap shows the single gRNA packaging metric. C) The phase boundary is shown in black. The heatmap shows the average radius of gyration of gRNA chains. D) The phase boundary is shown in cyan. The heatmap shows the percentage of cyclized genomes.

For this optimized gRNA design, there appears to be a concentration regime corresponding to the low gRNA and N-protein concentration corner of the phase envelope where single genome packaging, compaction, and cyclization can all occur efficiently. These results raise two predictions concerning the virion assembly process. First, there exists an optimal concentration range of gRNA and N-protein, just within the phase boundary, that promotes virion assembly.

Second, phase-separation promoting gRNA sequences may be located not only at the 5′ and 3′ ends, but also sparsely distributed throughout the genome to enhance genome packaging.

Discussion

It is clear that protein and RNA elements of SARS-CoV-2 can engage in phase separation with N-protein, but the functional consequences of this physical chemistry capacity for viral replication remain elusive. In this study we sought to explore how the spatial patterning of phase separation promoting or inhibiting RNA elements in the genome could facilitate the specificity and singularity of packaging the genome. Using coarse-grained simulations rooted in empirical observations, we find that phase separation can promote conditions that favor singular packaging, but that many spatial patterns of opposing phase separating elements can function similarly. However, the arrangement of phase separation promoting sequences at the ends of the genome is critical for genome compaction, and the positioning of these elements at both ends is necessary for cyclization.

Biophysical interpretation of the different protein binding modes leading to opposing phase behavior

An essential feature of our model is the different binding mode of N-protein with 5′ end and FE RNA. We found that weak, isotropic interactions with 5′ end RNA promote phase separation, while strong, anisotropic interactions with FE RNA lead to N-protein solubilization. Intrinsically, the valence of the isotropic interactions is 26 in a cubic lattice, while the valence is 1 for the anisotropic interactions. Additionally, since the anisotropic interactions have a higher binding energy than the isotropic ones, they last for longer, which effectively compounds the difference in valence between the 5′ end and FE beads. For the FE RNA system, this design allows for more N-protein binding at lower protein concentrations, but a less significant increase in protein binding as the protein concentration is increased. On the other hand, 5′ end RNA experiences a cooperative binding effect with N-protein, leading to greatly increased protein binding as protein concentrations are increased. The valence difference between 5′ end and FE RNA also includes competition between N-protein-FE RNA binding and N-protein dimerization, since a single bead in the model is responsible for both interactions. Therefore, while N-protein is bound to FE RNA, it can no longer dimerize, but while it is bound to 5′ end RNA, it is free to dimerize. We suspect that this competition also compounds the cooperative effect of the N-protein binding to each of these different RNAs and contributes to the distinct behaviors of these different polymers. The agreement between experimental results and our model suggests that there may be some underlying differences in the N-protein-RNA interactions in the 5′ end-like and FE-like regions in the gRNA that lead to distinct protein binding behavior. We hypothesize that FE-like regions have a lower effective valence with respect to N-protein than 5′ end-like regions, which will need to be explored experimentally in future work.

Boomerang shape of the phase boundaries

It has been previously reported that phase boundaries in ternary systems that are purely driven by heterotypic interactions are roughly elliptical in log-log space (2). However, the phase diagrams of the ternary systems studied in this work have distinct shapes. The 5′ end RNA with

N-protein phase diagram resembles an elliptical shape combined with a high N-protein concentration arm that corresponds well to the shape of the FE RNA with N-protein phase diagram. Each of these high N-protein concentration arms roughly correspond to the N-protein concentration range where phase separation occurs for N-protein alone at temperature 1 a.u., the temperature at which all ternary systems were studied (Fig 1b-d). The same superimposed shape is even more apparent in the gRNA with N-protein phase diagrams, with boundaries that resemble boomerangs (Fig 3b-e). We hypothesize that these phase diagrams arise from the union of two regimes of phase behavior that are driven by distinct forces. The elliptical portion that extends to low N-protein and RNA concentrations demonstrates phase separation that is driven by heterotypic interactions between RNA and N-protein, which aligns with reports in (2). On the other hand, the high concentration N-protein portion of the phase diagrams indicates phase separation that is driven by N-protein homotypic interactions. Here, RNA can partition into the dense phase, but it is neither necessary for phase separation, nor is it the driver. Therefore, the FE RNA with N-protein phase diagram consists only of this high N-protein concentration regime, which is eventually capped at high enough RNA concentrations (Fig 1d). As more FE RNA is present to sequester N-protein out of solution, the concentration of the available pool of N-protein is effectively decreased and phase separation can no longer occur. For systems with 5′ end RNA present, the phase diagrams have a complex reentrant character, passing in and out of the phase separating regime for certain fixed RNA concentrations as N-protein concentration is changed (Fig 1c, 3b-e). We speculate that this reentrant behavior is due to the interplay of phase separation driven by heterotypic or homotypic interactions, with intermediate regimes where neither are strong enough to drive phase separation. It is interesting to speculate whether such rich phase behavior may exist in other multicomponent systems, specifically those that involve long nucleotide polymers and RNA- or DNA-binding proteins.

Compatibility with the single packaging signal model

A recent model of single genome packaging has emerged as part of a study of the SARS-CoV-2 N-protein (18). The authors showed that for a 2-bead representation of N-protein that experienced isotropic attraction to itself and a long, 61-bead RNA molecule, large phase-separated clusters would form. However, if a much stronger binding site was added to the center of the RNA chain, mimicking a hypothetical packaging signal, N-protein and RNA chains would instead form kinetically trapped clusters that only very slowly coalesced into a single phase-separated droplet. This model is simpler than the one presented here, but it is reminiscent of the distinct effects seen here between 5′ end RNA and FE RNA. In our model, the most important effective difference between the 5′ end and FE RNA beads is their valence, with 5′ end RNA having a much higher valence than FE RNA. The model in (18) includes only a difference in binding energy, since all beads interact isotropically. However, the higher binding energy beads have a lower effective valence than the low energy beads at a given timescale, since they participate in bonds with fewer partners. In line with our results, the presence of lower-valence binding sites sequesters N-protein into clusters, opposing large-scale phase separation. Indeed, other groups have shown that a high enough valence is required for phase separation to occur *in silico* (28) and *in vitro* (24). In our model, however, most of the binding sites on the gRNA are of a low-valence character, which is distinct from the single, low-valence,

packaging signal site in (18). Despite this difference, both models provide evidence that such low-valence sites are essential for packaging tasks required during virion assembly, and that runaway phase separation must be tempered via alternative self-assembly pathways.

Generalization to other viruses and systems with long RNAs or DNA

Our models developed here are sufficiently coarse-grained to speculate that they may be applicable to other viruses and systems that involve long nucleotide chains and proteins. Components from several viruses have been shown to undergo phase separation, raising the possibility that spatial patterning of specific LLPS-promoting RNA or DNA sequences may have evolved to promote optimal genome packaging in other viruses in addition to SARS-CoV-2.

Many cellular phase separated bodies involve long RNAs or DNA and proteins that bind them. A particularly relevant example for our modeling is paraspeckles. Paraspeckles are highly ordered, phase separated nuclear bodies that require the presence of the long noncoding RNA, NEAT1. NEAT1 was recently shown to contain distinct functional domains, one of which is repetitive in its sequence and is necessary for paraspeckle formation (19). Paraspeckles also require several proteins, most of which contain RNA-binding domains and disordered regions (25). Recent work has shown that the central region of NEAT1 is necessary and sufficient for paraspeckle formation, and that it initiates assembly by binding several proteins (19). Specifically, the proteins NONO and SPFQ must first bind NEAT1, dimerize, and promote further polymerization via coiled-coil domains with other proteins for paraspeckle assembly to continue (25). There are thus many parallels with the SARS-CoV-2 system studied here; specific spatial patterning of protein binding elements along the RNAs is essential, and the protein partners must be able to dimerize/oligomerize for further assembly.

Conclusion

Identification of specific RNA sequences that promote ordered phase separated bodies via protein binding will not only advance our understanding of viruses, but also the many diverse cellular bodies and regions that contain long RNAs or DNA.

Author Contributions

I. S., A. S. G., and C. A. R. designed the conceptual models; I. S. implemented and performed simulations, data analysis, and created figures; I. S., A. S. G., and C. A. R. wrote the manuscript.

Acknowledgements

We thank Mikayla Feldbauer for helpful discussions and code drafting and the Pappu lab for initial introductions to working with LASSI. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. The Comet cluster at the San Diego Supercomputer Center was used through allocation nca106. We thank UNC Chapel Hill campus champion Mark Reed for helping I.S. access XSEDE resources on short notice. This work was supported by a FastGrant and C.A.R. is supported by NIH T32 CA 9156-43, F32GM136164, and L'OREAL USA for Women in Science Fellowship, and I. S. and A. S. G. were supported by NIH R01 GM081506 and the Air Force Office

of Scientific Research (grant FA9550-20-1-0241). A. S. G. serves on the scientific advisory board to Dewpoint Therapeutics. The authors declare no conflicts of interest.

References

1. Iserman C., Roden C. A., M. A. Boerneke, R. S. G. Sealfon, G. A. McLaughlin, I. Jungreis, E. J. Fritch, Y. J. Hou, J. Ekena, C. A. Weidmann, C. L. Theesfeld, M. Kellis, O. G. Troyanskaya, R. S. Baric, T. P. Sheahan, K. M. Weeks, and A. S. Gladfelter. 2020. Genomic RNA elements drive phase separation of the SARS-CoV-2 nucleocapsid. Mol. Cell. 80:1078-1091.
2. Choi J. M., F. Dar, and R. V. Pappu. 2019. LASSI: A lattice model for simulating phase transitions of multivalent proteins. PLoS Comput. Biol. 15(10): e1007028.
3. Ziv O., J. Price, L. Shalamova, T. Kamenova, I. Goodfellow, F. Weber, and E. A. Miska. 2020. The short- and long-range RNA-RNA interactome of SARS-CoV-2. Mol. Cell. 80:1067-1077.
4. Bar-On Y. M., A. Flamholz, R. Phillips, and R. Milo. 2020. Science Forum: SARS-CoV-2 (COVID-19) by the numbers. eLife. 9:e57309.
5. Masters P. S. 2019. Coronavirus genomic RNA packaging. Virology. 537:198-207.
6. Ziv O., M. M. Gabryelska, A. T. L. Lun, L. F. R. Gebert, J. Sheu-Gruttadauria, L. W. Meredith, Z.-Y. Liu, C. K. Kwok, C.-F. Qin, I. J. MacRae, I. Goodfellow, J. C. Marioni, G. Kudla, and E. A. Miska. 2018. COMRADES determines in vivo RNA structures and interactions. Nat. Methods 15:785-788.
7. Boeynaems S., S. Alberti, N. L. Fawzi, T. Mittag, M. Polymenidou, F. Rousseau, J. Schymkowitz, J. Shorter, B. Wolozin, L. V. D. Bosch, P. Tompa, and M. Fuxreiter. 2018. Protein phase separation: A new phase in cell biology. Trends Cell Biol. 28(6):420-435.
8. Brocca S., R. Grandori, S. Longhi, and V. Uversky. 2020. Liquid-liquid phase separation by intrinsically disordered protein regions of viruses: Roles in viral life cycle and control of virus-host interactions. Int. J. Mol. Sci. 21(23):9045.
9. Roden C. and A. S. Gladfelter. 2020. RNA contributions to the form and function of biomolecular condensates. Nat. Rev. Mol. 1-13.
10. Heinrich B. S., Z. Maliga, D. A. Stein, A. A. Hyman, and S. P. J. Whelan. 2018. Phase transitions drive the formation of vesicular stomatitis virus replication compartments. American Society for Microbiology 9(5):e02290-17.
11. Rincheval V., M. Lelek, E. Gault, C. Bouillier, D. Sitterlin, S. Blouquit-Laye, M. Galloux, C. Zimmer, J.-F. Eleouet, and M.-A. Rameix-Welti. 2017. Functional organization of cytoplasmic inclusion bodies in cells infected by respiratory syncytial virus. Nat. Commun. 8:563.
12. Nikolic J., R. L. Bars, Z. Lama, N. Scrima, C. Lagaudriere-Gesbert, Y. Gaudin, and D. Blondel. 2017. Negri bodies are viral factories with properties of liquid organelles. Nat. Commun. 8:58.
13. Guseva S., S. Milles, M. R. Jensen, N. Salvi, J.-P. Kleman, D. Maurin, R. W. H. Ruigrok, and M. Blackledge. 2020. Measles virus nucleo- and phosphoproteins form liquid-like phase-separated compartments that promote nucleocapsid assembly. Sci. Adv. 6:eaaz7095.

14. Monette A., M. Niu, L. Chen, S. Rao, R. J. Gorelick, and A. J. Mouland. 2020. Pan-retroviral nucleocapsid-mediated phase separation regulates genomic RNA positions and trafficking. Cell Rep. 31:107520.
15. Yang D. and J. L. Leibowitz. 2015. The structure and functions of coronavirus genomic 3' and 5' ends. Virus Res. 206:120-133.
16. Perdikari T. M., A. C. Murthy, V. H. Ryan, S. Watters, M. T. Naik, and N. L. Fawzi. 2020. SARS-CoV-2 nucleocapsid protein phase-separates with RNA and with human hnRNPs. EMBO J. 39:e106478.
17. Fixman M. 1962. Radius of gyration of polymer chains. J. Chem. Phys. 36:306-310.
18. Cubuk J., J. J. Alston, J. J. Incicco, S. Singh, M. D. Stuchell-Brereton, M. D. Ward, M. I. Zimmerman, N. Vithani, D. Griffith, J. A. Wagoner, G. R. Bowman, K. B. Hall, A. Soranno, and A. S. Holehouse. 2020. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. bioRxiv, doi: https://doi.org/10.1101/2020.06.17.158121 (preprint posted December 21, 2020).
19. Yamazaki T., S. Souquere, T. Chujo, S. Kobelke, Y. S. Chong, A. H. Fox, C. S. Bond, S. Nakagawa, G. Pierron, and T. Hirose. 2018. Functional domains of NEAT1 architectural lncRNA induce paraspeckle assembly through phase separation. Mol. Cell. 70:1038-1053.
20. Klein S., M. Cortese, S. L. Winter, M. Wachsmuth-Melm, C. J. Neufeldt, B. Cerikan, M. L. Stanifer, S. Boulant, R. Bartenschlager, and P. Chlanda. 2020. SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. Nat. Commun. 11:5885.
21. Hsieh P. K., S. C. Chang, C.-C. Huang, T.-T. Lee, C.-W. Hsiao, Y.-H. Kou, I.-Y. Chen, C.-K. Chang, T.-H. Huang, and M.-F. Chang. 2005. Assembly of severe acute respiratory syndrome coronavirus RNA packaging signal into virus-like particles is nucleocapsid dependent. J. Virol. 79(22):13848-55.
22. Towns J., T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. XSEDE: Accelerating Scientific Discovery. Comput. Sci. Eng. 16(5):62-74.
23. Stukowski A. 2010. Visualization and analysis of atomistic simulation data with OVITO – the open visualization tool. Modelling Simul. Materi. Sci. Eng. 18:015012.
24. Banani S. F., A. M. Rice, W. B. Peeples, Y. Lin, S. Jain, R. Parker, and M. K. Rosen. 2016. Compositional control of phase-separated cellular bodies. Cell. 166:651-663.
25. Hirose T., T. Yamazaki, and S. Nakagawa. 2019. Molecular anatomy of the architectural NEAT1 noncoding RNA: The domains, interactors, and biogenesis pathway required to build phase-separated nuclear paraspeckles. WIREs RNA. 10:e1545.
26. Savastano A., A. I. de Opakua, M. Rankovic, and M. Zweckstetter. 2020. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. Nat. Commun. 11:6041.
27. Chen H., Y. Cui, X. Han, W. Hu, M. Sun, Y. Zhang, P.-H. Wang, G. Song, W. Chen, and J. Lou. 2020. Liquid-liquid phase separation by SARS-CoV-2 nucleocapsid protein and RNA. Cell Res. 30:1143-1145.
28. Espinosa J. R., J. A. Joseph, I. Sanchez-Burgos, A. Garaizar, D. Frenkel, and R. Collepardo-Guevara. 2020. Liquid network connectivity regulates the stability and composition of biomolecular condensates with many components. PNAS. 117:13238-13247.