Supporting Information for "Condensates in RNA Repeat Sequences are Heterogeneously Organized and Exhibit Reptation-like Dynamics"

Hung T. Nguyen[1], **Naoto Hori**[2], **and D. Thirumalai**[1,*]

[1]Department of Chemistry, The University of Texas at Austin, Austin, TX 78712
[2]School of Pharmacy, University of Nottingham, Nottingham, UK
[*]dave.thirumalai@gmail.com

ABSTRACT

RNA energy function

In order to describe the organization and dynamics of RNA condensates, we develop a new coarse-grained model that is sufficiently simple to simulate multiple chains while retaining sequence information. Such an approach is needed to investigate the mechanism of RNA condensation in arbitrary low complexity sequences. Because condensate formation requires multiple simulations involving many chains for sufficiently long times, we resorted to a low resolution coarse-grained description for RNA. We have shown previously that simulations based on related models are efficacious in predicting the outcomes of single molecule pulling experiments.[1,2]

Each nucleotide is represented by a single bead (Fig. S1c). The energy function for an isolated RNA chain has the following form:

$$U = U_{BA} + U_{EV} + U_{HB}, \tag{1}$$

where U_{BA} accounts for chain connectivity, consisting of bond and angle restraints between the connected beads. We used harmonic potentials to keep the bond lengths and the angles close to the A-form helix: $U_{BA} = \frac{1}{2}\sum_i k_{bond}\left(r_i - r_o\right)^2 + \frac{1}{2}\sum_j k_{angle}\left(\alpha_j - \alpha_o\right)^2$ where $k_{bond} = 15.0$ kcal/mol.Å^2, $k_{angle} = 10.0$ kcal/mol.radian^2, $r_o = 5.9$ Å, $\alpha_o = 2.618$ rad. In the simulations, the angles fluctuate due to soft restraints, and deviate from the values in the A-form helix.

The excluded volume interaction U_{EV} is given by the Weeks–Chandler–Andersen potential:

$$U_{EV} = \sum_{i,j} \Theta\left(\sigma - r_{ij}\right) \varepsilon \left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma}{r_{ij}}\right)^{6} + 1\right], \tag{2}$$

where Θ is the Heavyside step function, $\sigma = 10.0$ Å, and $\varepsilon = 2.0$ kcal/mol. The excluded term is only computed between two beads that are not involved in the bond or angle restraints, and are separated at least by two other beads (nucleotides) along the chain.

Hydrogen bond term U_{HB} is a many-body and short-ranged potential that mimics the canonical Watson–Crick base-pairing between A-U and G-C. The hydrogen bonding potential between the two beads i, j is

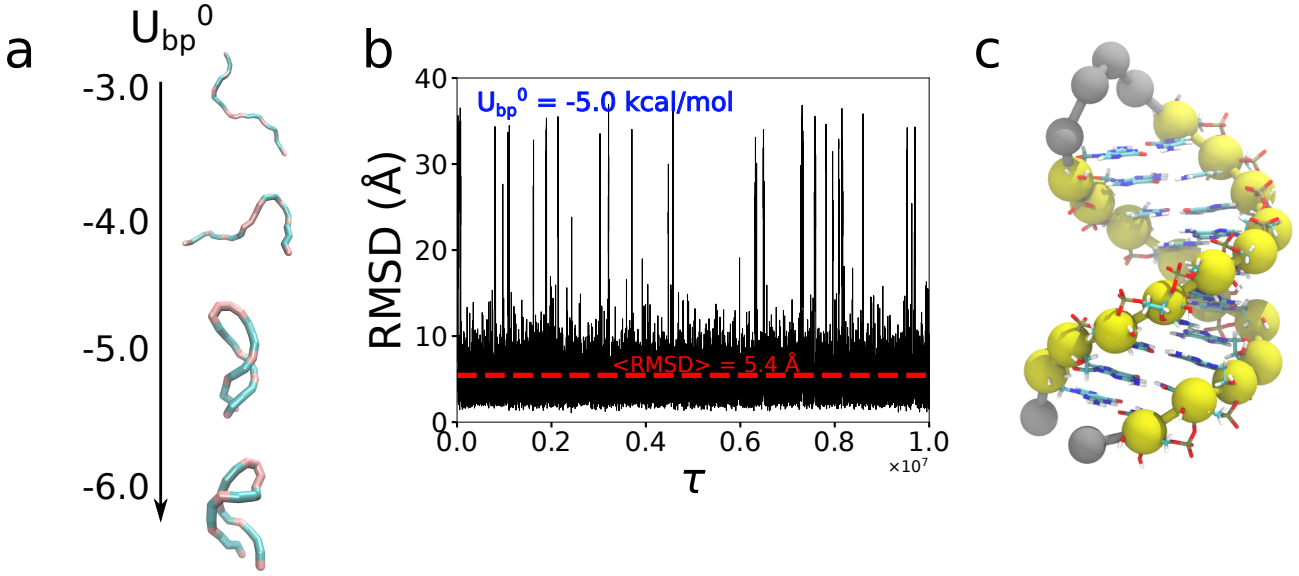

Figure S1. Calibration of the base pair interaction strength U_{bp}^o. **a**, Structural dependence of a small CAG repeat sequence (AGGCAGCAGCCAAAAGGCAGCAGCCA) on U_{bp}^o. The sequence we chose to calibrate U_{bp}^o is almost identical to the X-ray structure (PDB 3NJ6),[3] except with the addition of a AAAA tetraloop and two terminal A nucleotides. The sequence adopts extended conformations for small U_{bp}^o ($U_{bp}^o < 4.5$ kcal/mol), and folds into hairpin conformations in base pair interaction range, $4.5 < U_{bp}^o < 6.0$ kcal/mol. We set $U_{bp}^o = -5.0$ kcal/mol. **b**, Root mean squared deviation (RMSD) between the simulations and the X-ray crystal structure, shown for $U_{bp}^o = -5.0$ kcal/mol. The averaged value for RMSD is around 5 Å, which is reasonable given the coarse-grained nature of the model. **c**, Superposition of the simulated structure (yellow and grey beads) onto the X-ray structure for the lowest RMSD (around 1 Å).

taken to be,

$$U_{HB} = U_{bp}^o \exp\left[U_{hb,bond} + U_{hb,angle} + U_{hb,dihedral}\right], \tag{3a}$$

$$U_{hb,bond} = -k_r \left(r_{ij} - r_{hb,o}\right)^2, \tag{3b}$$

$$U_{hb,angle} = -k_\theta \left(\theta_{i,j,j-1} - \theta_1\right)^2 - k_\theta \left(\theta_{i-1,i,j} - \theta_1\right)^2 - k_\theta \left(\theta_{i,j,j+1} - \theta_2\right)^2 - k_\theta \left(\theta_{i+1,i,j} - \theta_2\right)^2, \tag{3c}$$

$$U_{hb,dihedral} = -k_\phi \left[1 + \cos\left(\phi_{j-1,j,i,i-1} + \phi_1\right)\right] - k_\phi \left[1 + \cos\left(\phi_{j+1,j,i,i+1} + \phi_2\right)\right], \tag{3d}$$

where $\theta_{a,b,c}$ is the angle formed between beads a, b and c; $\phi_{a,b,c,d}$ is the dihedral angle formed between beads a, b, c and d; $r_{hb,o} = 13.8$ Å, $k_r = 3.0$ Å^{-2}, $k_\theta = 1.5$ rad^{-2}, $k_\phi = 0.5$, $\theta_1 = 1.8326$ rad, $\theta_2 = 0.9425$ rad, $\phi_1 = 1.8326$ rad, $\phi_2 = 1.1345$ rad.

The functional form of U_{HB} is designed to capture both the Watson–Crick base-pairing, and the helical nature of the A-form RNA at the single-nucleotide resolution. The only unknown parameter, the strength of the hydrogen bond interactions, is adjusted to reproduce the structure of a small CAG repeat sequence[3] ($U_{bp}^o = -5.0$ kcal/mol for GC base pair (equivalently ≈ -1.67 kcal/mol per hydrogen bond), Fig. S1).

Inter-chain interactions: Interactions between different chains also involve excluded volume and hydrogen bonding interactions, which are treated identically as in intramolecular interactions. Our model, therefore, does not distinguish between intra- and inter-molecular hydrogen bond interactions. A nucleotide could form WC base-pair interactions with others either within the same chain or in a different chain.

Therefore, this is a general model for studying condensate formation in a generic RNA, provided that a single bead resolution suffices. As always, the validity of the model can only be assessed by comparisons to experiments.

Isolated repeat RNAs form hairpin-like structures

We first characterized the structural ensembles of an isolated $(CAG)_n$, which serves as a reference when comparisons to RNAs within the condensates are made. Interestingly, the mean end-to-end distance, R_{ee}, is independent of n (Fig. S2**a**), which accords well with experiments and theoretical predictions for mRNA and long non-coding RNA.[4–7] The peak values in the distributions of R_{ee} is around 35 Å. The R_{ee} distributions for the three RNA chains with different lengths deviate from the Gaussian distribution for ideal chains, exhibiting long tails. This is somewhat surprising because the radius of gyration R_g distribution (shown in Fig. S2**b**) shows little deviation from the ideal chain behavior.

Unstructured RNAs (poly-rA or poly-rC) that do not favor base pair formation follow the expected R_{ee} distribution for a random coil.[7] In contrast, CAG repeats form intramolecular base pair interactions, bringing the 5' and 3' ends to proximity. Experiments have reported that $(CAG)_n$ repeat sequences with small n could form stable hairpins containing GC base pairs (bps) with A:A mismatches.[3,8] Formation of GC bps in longer repeats also requires the chain to fold upon itself. Folding would be favored if the overall free energy gain by forming GC bps compensates for the chain bending penalty. For the CAG repeat sequence, extensive GC bp formation could mitigate the unfavorable interactions, leading to the formation of hairpin-like structures.[3,9] The probabilistic contact map (Fig. S2**d**) shows that the majority of interactions are along the diagonal, consistent with the formation of a hairpin structure. The two ends have the highest probability of being in proximity. Representative snapshots from the simulations, shown in Fig. S2**a**, confirm that the RNA monomer mostly samples a set of hairpin-like structures.

It is worth emphasizing that we did not adjust any parameter in the RNA model to constrain $(CAG)_n$ to form hairpins. The structures found in the simulations are the result of the generic tendency of C and G to form WC base pairs. Due to the repeat nature of the sequence, the RNA maintains an ensemble of hairpin-like structures by sliding one end over another without paying significant energetic penalty.

The formation of helical structures is also reflected in the bond-bond orientational correlation function,

$$\langle \cos\theta(s) \rangle = \left\langle \vec{b}_i . \vec{b_{i+s}} \right\rangle / l_b^2, \tag{4}$$

where $\vec{b}_i$ is the unit vector of bond i, with the bond length l_b. $\cos\theta(s)$ shows periodicity at small $s = |i - j|$, where i and j are the indices of the nucleotides in the $(CAG)_n$ sequence (Fig. S2**c**). At a short length scale (up to 5 to 6 nucleotides), the structure of CAG is roughly rigid, with $R(s)$ scaling almost linearly with s (inset of Fig. S2**c**). At larger separations, $(s \geq 6)$, $R(s) \propto s^{0.56}$, suggesting that the RNA is flexible, and hence could fold upon itself to generate hairpin-like structures. At a separation corresponding to the end-to-end distance, there is an abrupt decrease in $R(s)$ as a function of s because the 5' and 3' ends are close.

Similar results are also observed for $(CUG)_n$ (Fig. S3). Our simulations show that the hairpin-like structures in $(CXG)_n$ (X = A or U) are a consequence of multiple canonical Watson–Crick base pair formation.

Simulation details and analyses

Simulation details. We placed 64 $(CAG)_n$ (n = 20, 31 or 47) molecules in a cubic box, whose size varied from 50-200 nm depending on the RNA concentration. Simulations were performed on GPU

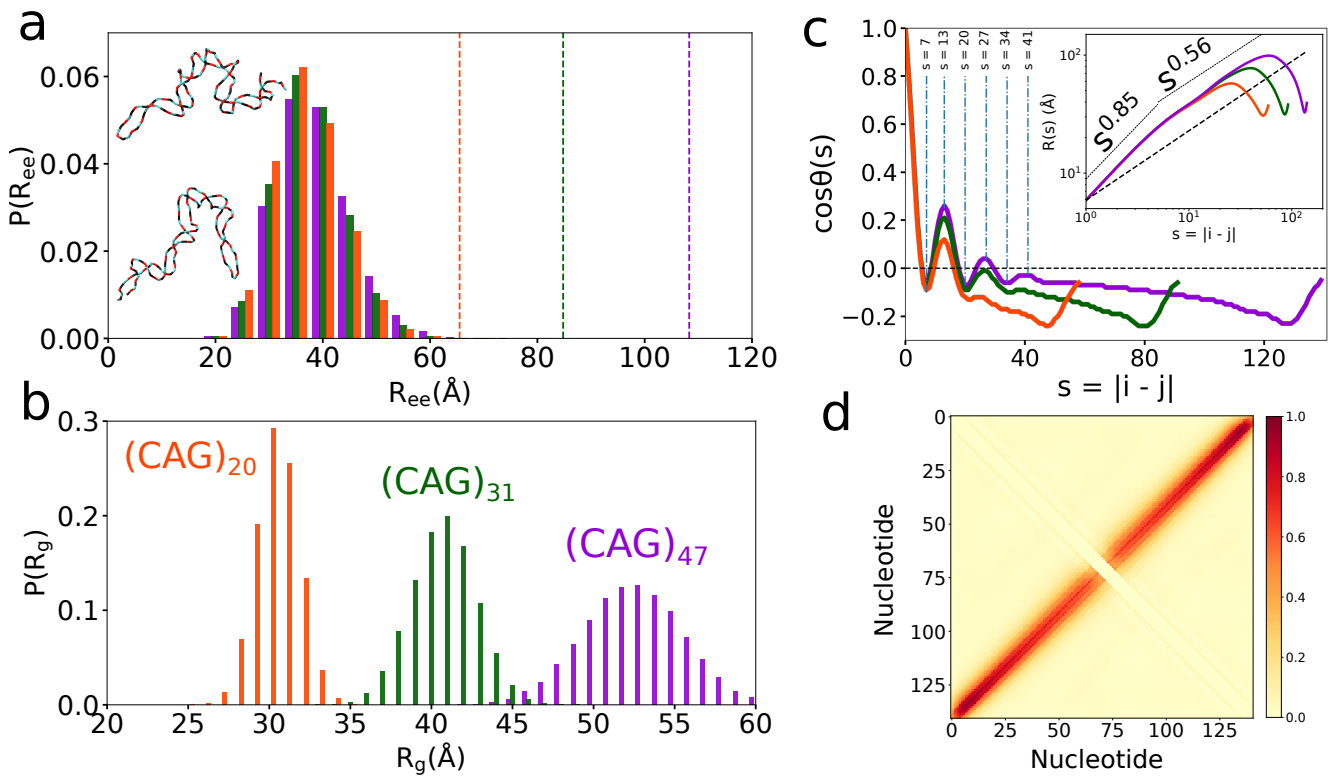

Figure S2. Structures of isolated $(CAG)_n$ monomers. **a**, Distribution of the end-to-end distance R_{ee} and **b**, radius of gyration R_g of $(CAG)_n$ with n=47, 31 and 20. The vertical dash lines in **a** indicate mean values for self-avoiding random walk chains with the same n. Snapshots are for $(CAG)_{47}$. Cytosine is in cyan, adenine is in red and guanine is in black. **c**, Bond-bond orientational correlation function $\cos\theta(s)$ as a function of the sequence distance s. The periodicity, as indicated by the vertical lines, is unmistakable. The inset shows average inter-nucleotide distances $R(s)$ vs. s. The dashed line shows $R(s)$ for a self-avoiding polymer ($R(s) \propto s^{0.588}$). At large s, there is an an abrupt drop in $R(s)$ because the two ends strongly interact with each other, thus bringing them to proximity. **d**, Contact map for $(CAG)_{47}$ shows that the majority of interactions occur along the diagonal, indicating the formation of hairpin structures.

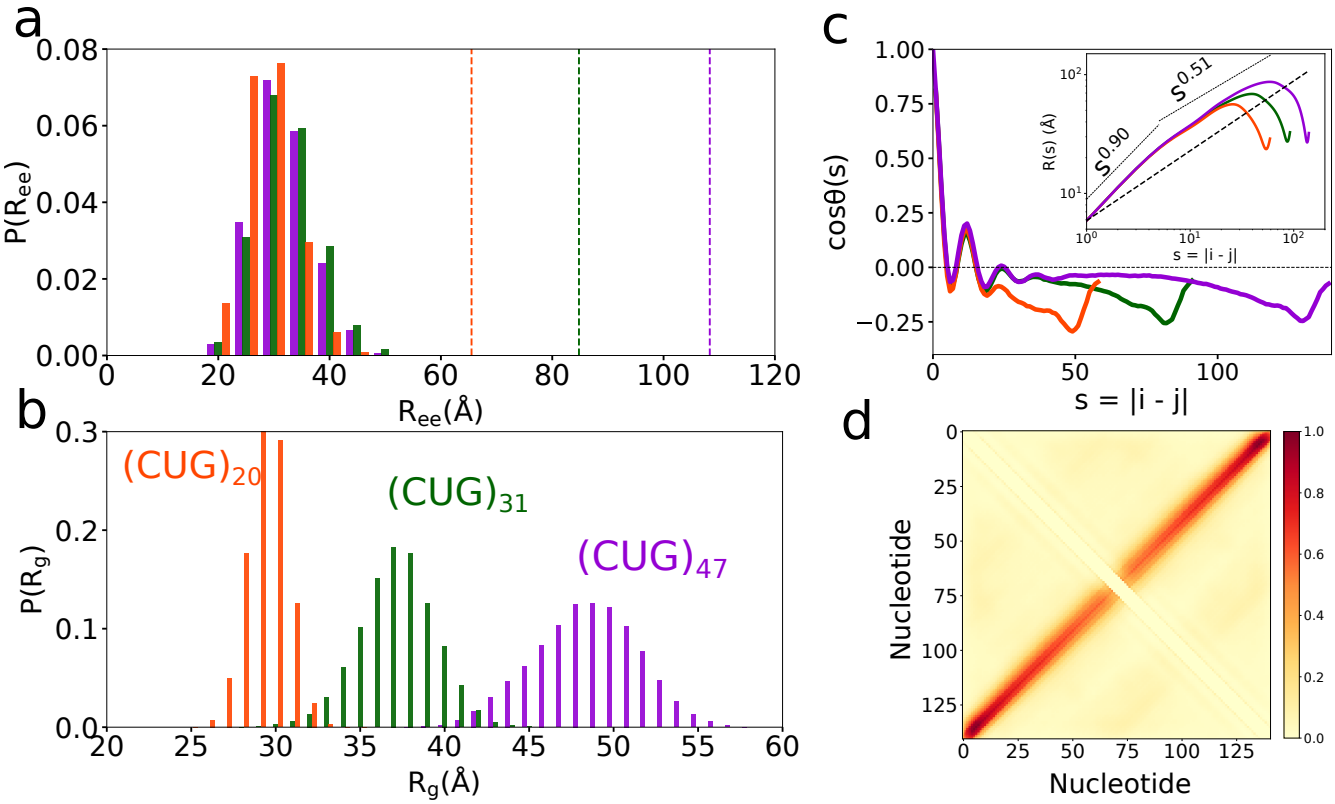

Figure S3. Same as Fig. S2, but for $(CUG)_n$ monomers.

using the OpenMM code to speed up sampling of the conformational space.[10] We used low-friction Langevin dynamics in which the viscosity of water was reduced by a factor of 100 in order to further speed up the sampling efficiency.[11] Even using a primitive RNA model, the simulations are computationally intensive because there are many chains in the system. Simulations were conducted for ~ 100 days for each trajectory using NVIDIA Quadro RTX 5000 graphics card on the Frontera supercomputer (Texas Advanced Computing Center). Snapshots were recorded every 10,000 steps, which were subsequently used to calculate several quantities of interest.

Clustering. For each snapshot, we grouped the monomers that are within 20 Å from each other to the same cluster. A monomer belongs to a cluster if the distance between any of its nucleotide to any nucleotide in the cluster is less than 20 Å. For comparison, the equilibrium base pairing distance in the CG model is $r_{hb,o} = 13.8$ Å. Thus, the cutoff distance for clustering is about $\approx 40\%$ larger than $r_{hb,o}$.

Form factor. The form factor of RNAs was calculated as:

$$S_c = \left\langle \left| \sum_k \exp\left(i\mathbf{q}.\mathbf{r_k}\right) \right|^2 \right\rangle_N , \tag{5}$$

where the brackets denote averaging over the chains at different orientation.

Radius of gyration and shape parameters. R_g and shape parameters for an RNA molecule was calculated using the gyration tensor, with each element is defined as:

$$S_{xy} = \frac{1}{2N^2} \sum_i \sum_j \left(x_i - x_j\right)\left(y_i - y_j\right). \tag{6}$$

The radius of gyration, R_g^2, is then calculated by summing over the three eigenvalues of the gyration tensor $R_g^2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$. To characterize the shape of RNAs, we calculated the relative shape anisotropy κ^2 and the shape parameter S using,

$$\kappa^2 = \frac{3}{2}\frac{\lambda_1^4 + \lambda_2^4 + \lambda_3^4}{\left(\lambda_1^2 + \lambda_2^2 + \lambda_3^2\right)^2} - \frac{1}{2}, \tag{7}$$

$$S = \prod_{i=1,2,3} \frac{\lambda_1^2 - \bar{\lambda^2}}{\bar{\lambda^2}}, \tag{8}$$

where $\bar{\lambda^2} = \frac{1}{3}\left(\lambda_1^2 + \lambda_2^2 + \lambda_3^2\right)$. κ^2 is bounded between 0 and 1; $\kappa^2 = 0$ implies that the molecule is perfectly spherical, and $\kappa^2 = 1$ if every point lies on a straight line. S satisfies $-1/4 \leq S \leq 2$. Negative values of S correspond to oblate ellipsoids, while prolate ellipsoids have positive S.

Concentrations of the two phases. To determine the concentration inside a droplet, we first calculate the volume of a droplet. We assume the droplet is an ellipsoid with the three effective radii, characterized by the three eigenvalues of the gyration tensor. Thus, the volume of the i$^{\text{th}}$ droplet is,

$$V_i = \frac{4}{3}\pi abc = 4\pi\sqrt{3}\lambda_1\lambda_2\lambda_3. \tag{9}$$

Note that the gyration tensor is calculated for the whole droplet, and not individual RNA molecule as the previous section. The concentration inside the droplets is computed by averaging all the concentrations inside medium and large droplets:

$$C_{droplet} = \bar{C}_i, \tag{10}$$

with $C_i = \frac{N_i}{V_i}$ is the concentration of the i$^{\text{th}}$ droplet with N_i RNA molecules and volume V_i. We calculate C_i only for $N_i \geq 5$.

The concentration of the aqueous solution, consisting of monomers and oligomers outside the droplets, is then calculated using:

$$C_{solution} = \frac{N - \sum_i N_i}{V - \sum_i V_i}. \tag{11}$$

Mean squared displacement. The mean squared displacement (MSD) of the center-of-mass of a chain is given by,

$$\Delta(\tau - \tau_o) = \left\langle \left(\vec{r}_i(\tau) - \vec{r}_i(\tau_o)\right)^2 \right\rangle_N, \tag{12}$$

where τ_o is the initial time.

The time-averaged MSD is calculated using:

$$\Delta(\Delta\tau) = \left\langle \left(\vec{r}_i(\tau + \Delta\tau) - \vec{r}_i(\tau)\right)^2 \right\rangle_{\tau,N}. \tag{13}$$

We also computed the MSD of individual nucleotides to probe the dynamics of the RNA chains in large droplets (Fig. 6c in the main text). For this purpose, we chose the largest droplet formed in the simulations. Before computing the nucleotide MSD, the position of the droplet in each snapshot was sequentially superimposed to the one in the previous time frame, starting from the instance the droplet is formed (τ_o). This procedure ensures that the MSD accounts solely for the dynamics of chains within the droplet, but not for collective translational and rotational motion of the entire droplet in the simulation box. The MSD was then calculated in the same way using Eq. 12 by averaging over all the nucleotides except for 10 nucleotides at both the $5'$ and $3'$ ends. Note that τ_o depends on when the droplet forms.

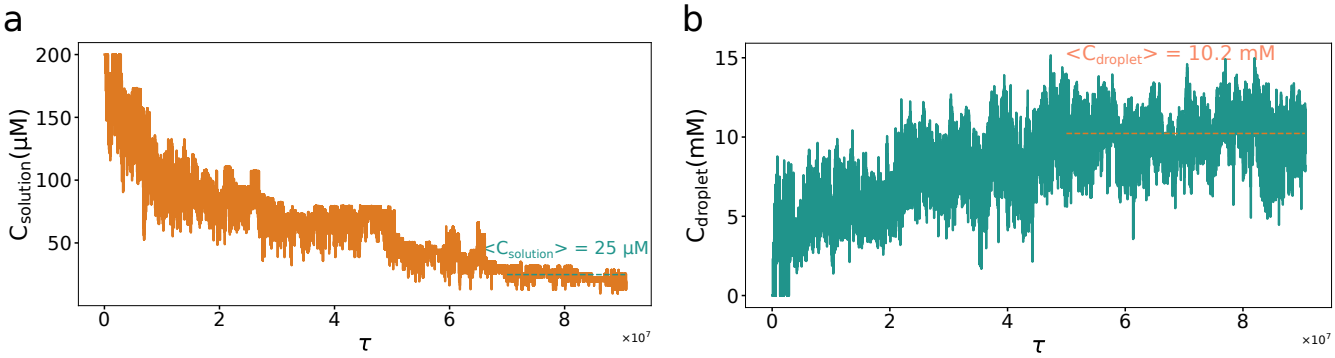

Figure S4. Determination of the concentrations of the two co-existing phases. Results are shown for $(CAG)_{47}$. Time dependent changes in the concentrations in the aqueous phase is on the left and for the droplet is on the right. The plateau values near the end are used to calculate the concentrations at which the two phases coexist.

Supplementary data

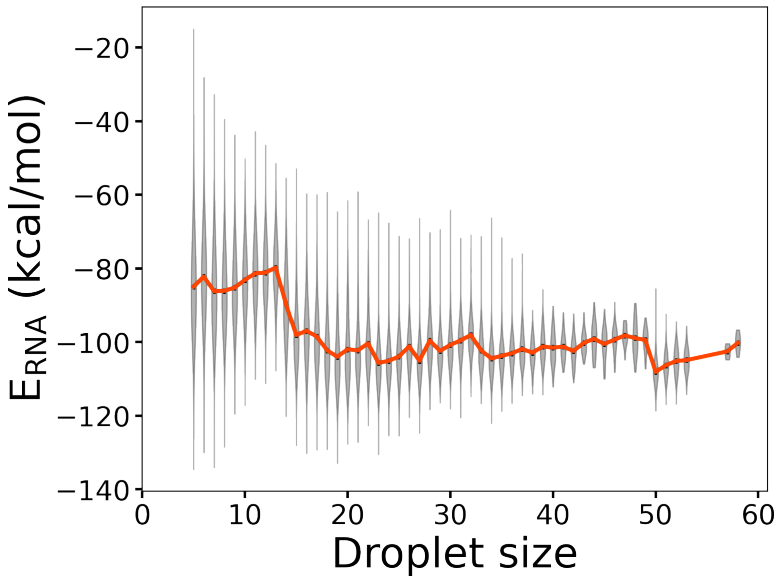

Figure S5. Energy per chain vs droplet size. As the size of the condensate increases, the energy per chain decreases, which indeed is the reason for phase separation from a thermodynamic perspective.

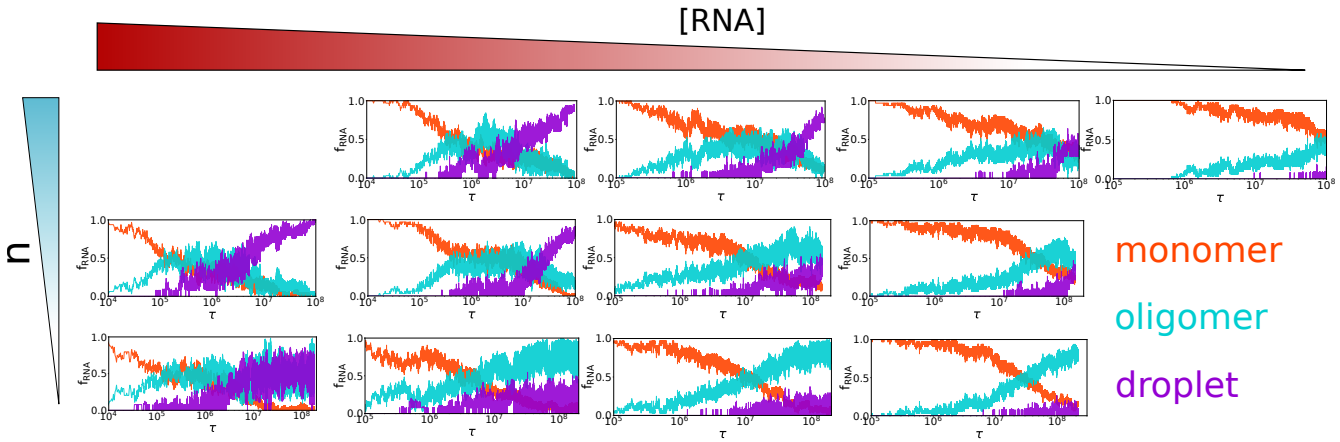

Figure S6. Fraction of monomers, oligomers and droplets as a function of (CAG) repeat length and concentration.

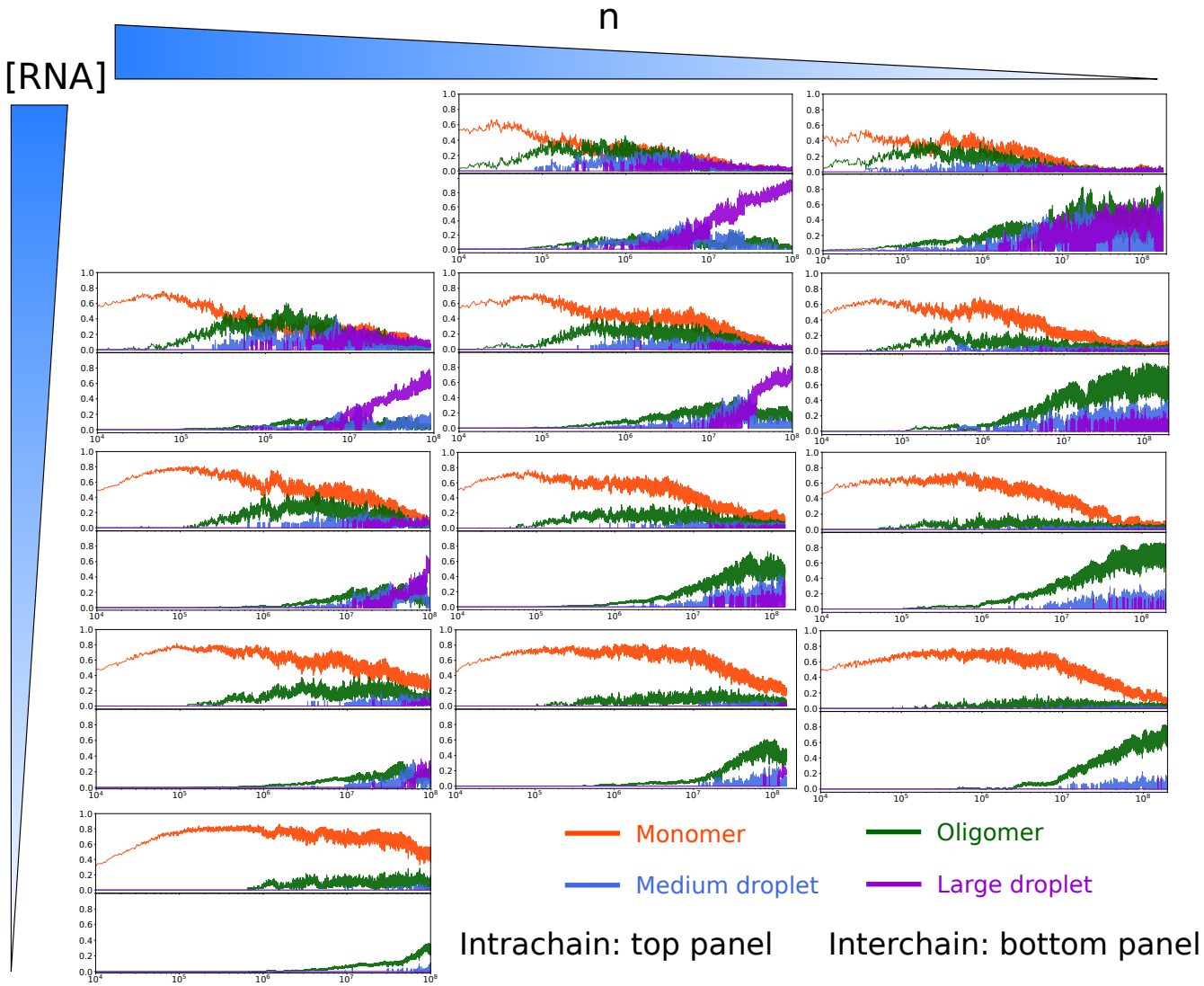

Figure S7. Fraction of (both intra- and intermolecular) hydrogen bonds as function of time. The conversion of intra- to intermolecular interactions is the dominant factor driving the phase separation.

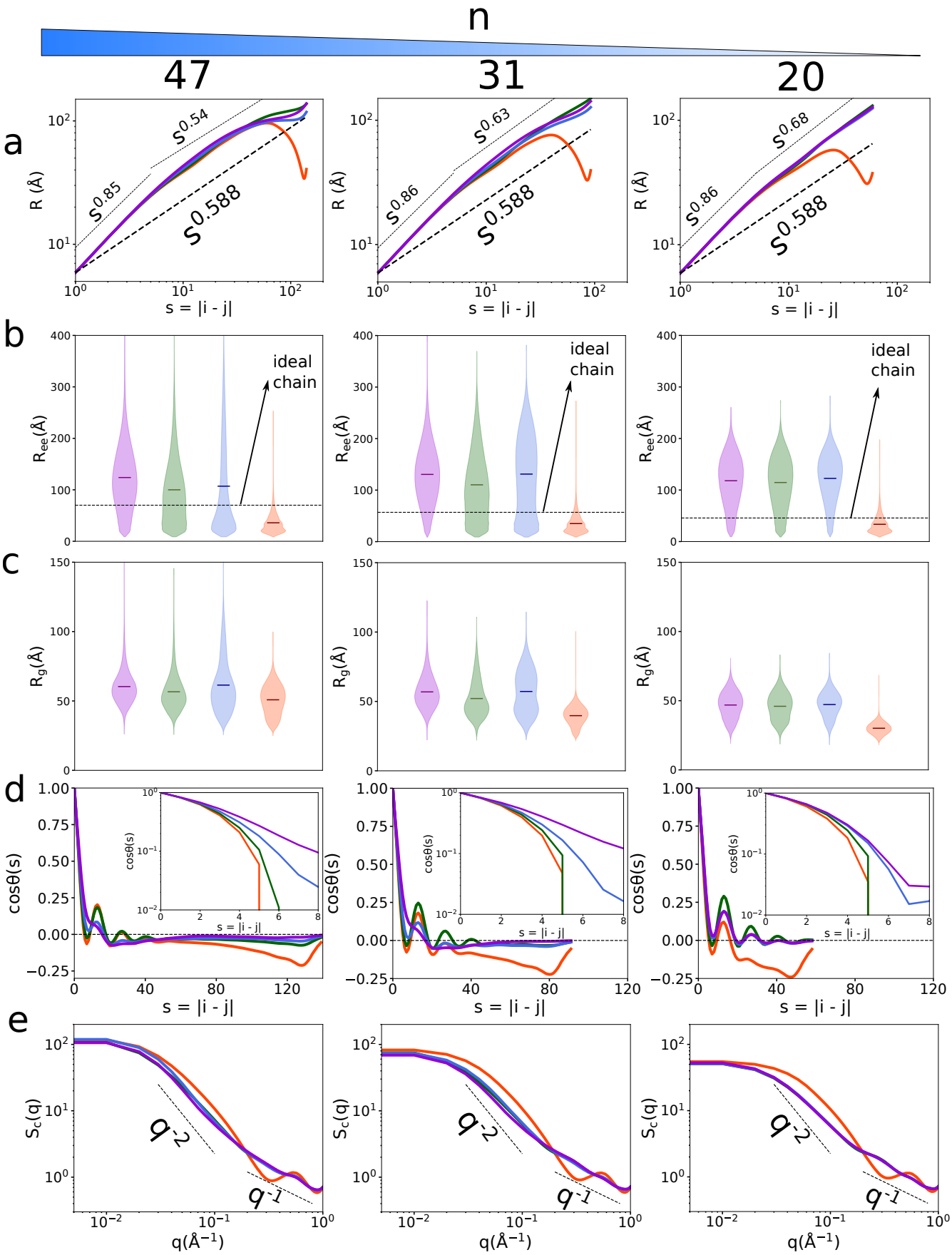

Figure S8. RNA undergoes significant conformational changes upon entering droplets.

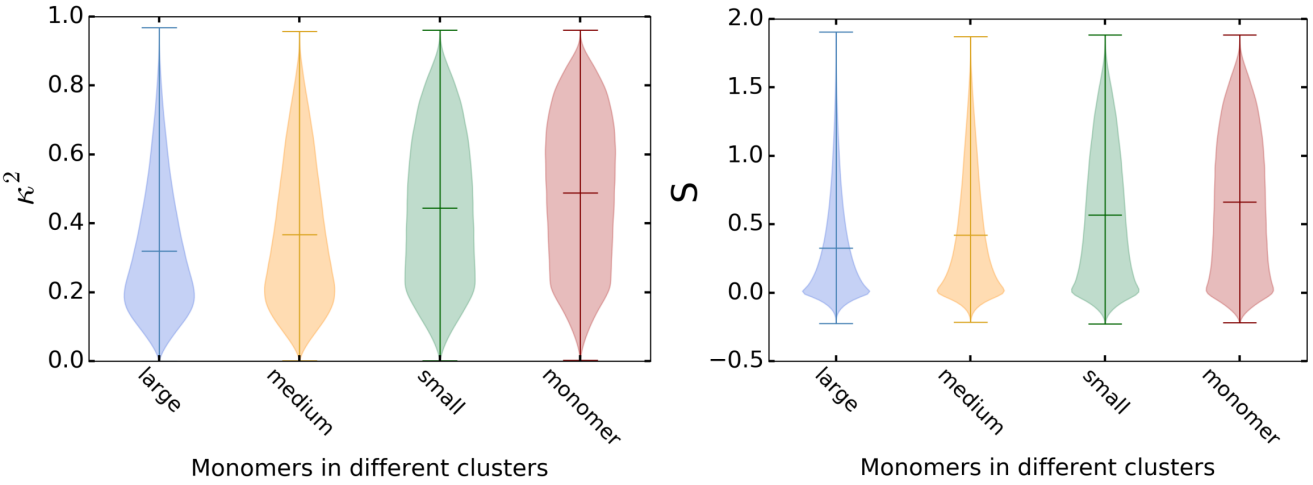

Figure S9. Comparison of shape parameters of monomeric RNAs in solution and in droplets.

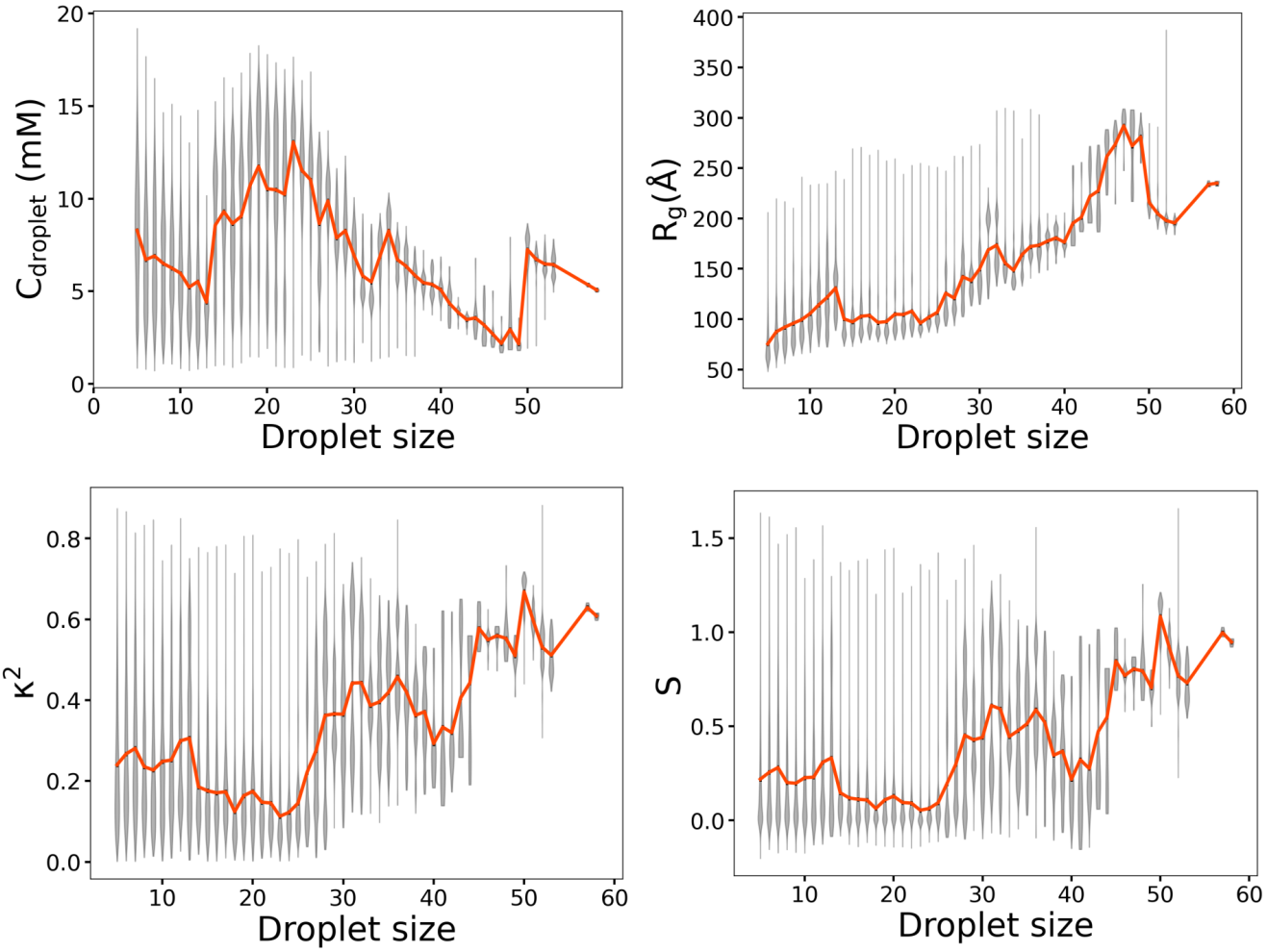

Figure S10. Characterization of droplet properties as a function of droplet size.

Supplemental Movies

Movie S1. Dynamics of phase separation in $(CAG)_{47}$ from monomers to condensates. The monomer RNA concentration is 200 μM. At the initial time ($\tau = 0$), the chains exist as monomers. At intermediate times, oligomers form, which subsequently fuse together resulting in large droplets as time progresses. See Figures 1a, 3a, and 5 for the corresponding trajectories. Note that some of the individual RNA chains may look less helical shape due to trajectory-smoothing needed for easier visualization. The movie shows both fusion as well as fission of the droplets.

Movie S2. Growth and fusion of condensates. The movie shows two small clusters of RNA chains (green and magenta) fuse to form a droplet. In the process, the cluster in magenta is once partially dissolved but eventually coalesce with the other cluster (green).

Movie S3. Jamming dynamics of two labelled RNA chains inside a droplet. The movie shows the late-stage dynamics in a simulation trajectory at 200 μM $(CAG)_{47}$ after the formation of a large droplet. For clarity, focus is on two RNA chains shown in yellow and red colors, whereas all other chains in the same droplet are shown as transparent blue chains. The motions of these RNA chains are highly restricted, resulting in reptation-like dynamics (Figure 6).

References

1. Hyeon, C. & Thirumalai, D. Mechanical Unfolding of RNA: From Hairpins to Structures with Internal Multiloops. *Biophys. J.* **92**, 731–743 (2007). Publisher: Elsevier.
2. Lin, J.-C. & Thirumalai, D. Relative Stability of Helices Determines the Folding Landscape of Adenine Riboswitch Aptamers. *J. Am. Chem. Soc.* **130**, 14080–14081 (2008). Publisher: American Chemical Society.
3. Kiliszek, A., Kierzek, R., Krzyzosiak, W. J. & Rypniewski, W. Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nucl. Acids Res.* **38**, 8370–8376 (2010).
4. de Gennes, P. G. Statistics of branching and hairpin helices for the dAT copolymer. *Biopolymers* **6**, 715–729 (1968).
5. Yoffe, A. M., Prinsen, P., Gelbart, W. M. & Ben-Shaul, A. The ends of a large RNA molecule are necessarily close. *Nucl. Acids Res.* **39**, 292–299 (2011).
6. Clote, P., Ponty, Y. & Steyaert, J.-M. Expected distance between terminal nucleotides of RNA secondary structures. *J. Math. Biol.* **65**, 581–599 (2012).
7. Lai, W.-J. C. *et al.* mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances. *Nat. Comm.* **9**, 4328 (2018).
8. Marquis Gacy, A., Goellner, G., Juranić, N., Macura, S. & McMurray, C. T. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* **81**, 533–540 (1995).
9. de Mezer, M., Wojciechowska, M., Napierala, M., Sobczak, K. & Krzyzosiak, W. J. Mutant CAG repeats of Huntingtin transcript fold into hairpins, form nuclear foci and are targets for RNA interference. *Nucl. Acids Res.* **39**, 3852–3863 (2011).
10. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017).
11. Honeycutt, J. D. & Thirumalai, D. The nature of folded states of globular proteins. *Biopolymers* **32**, 695–709 (1992).