nature chemistry

ARTICLES

https://doi.org/10.1038/s41557-022-00934-z

Check for updates

Condensates in RNA repeat sequences are heterogeneously organized and exhibit reptation dynamics

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

Although it is known that RNA undergoes liquid–liquid phase separation, the interplay between the molecular driving forces and the emergent features of the condensates, such as their morphologies and dynamic properties, is not well understood. We introduce a coarse-grained model to simulate phase separation of trinucleotide repeat RNAs, which are implicated in neurological disorders. After establishing that the simulations reproduce key experimental findings, we show that once recruited inside the liquid droplets, the monomers transition from hairpin-like structures to extended states. Interactions between the monomers in the condensates result in the formation of an intricate and dense intermolecular network, which severely restrains the fluctuations and mobilities of the RNAs inside large droplets. In the largest densely packed high-viscosity droplets, the mobility of RNA chains is best characterized by reptation, reminiscent of the dynamics in polymer melts. Our work provides a microscopic framework for understanding liquid–liquid phase separation in RNA, which is not easily discernible in current experiments.

The discovery that germline P granules—assembled from RNA and RNA-binding proteins in a single-cell embryo and implicated in embryo development—form liquid droplets resembling compartments without membranes[1] has resulted in a revolution, providing an impetus for a deeper understanding of how the cytoplasm and the nucleus are organized. In turn, this finding has inspired a large number of studies in a variety of unrelated systems that undergo liquid–liquid phase separation (LLPS), postulated to be the mechanism by which such compartments form[2–14]. Almost all these studies have focused on liquid organelles composed of intrinsically disordered proteins/regions (IDPs/IDRs) or IDPs interacting with RNA molecules that could engage in multivalent interactions[5]. These findings are typically explained using analogies to sol–gel phase transitions known in synthetic polymers, although there are differences due to the heteropolymer nature of biological sequences[4,15–18]. However, the molecular details of how proteins and nucleic acids, with diverse sequences, coalesce and undergo LLPS to form condensates, and potentially mature into gel-like or even ordered states, are still elusive. Furthermore, the conformations and dynamics of the molecules inside the condensates are essentially unknown either in experiments or in simulations.

That interactions between RNA and IDPs, such as FUS containing an RNA-binding C-terminal domain and a low-complexity N-terminal IDP domain, promote phase separation is well documented[19–21]. It is also known that RNA plays an important role in driving the formation, stability and morphology of the biomolecular condensates[6,12,13,21–31]. Recent experiments have also shown that RNA alone is sufficient to self-organize into condensates and gel-like states in vitro[24,27,32,33]. For example, poly(rU) with a small addition of short cationic polypeptides can drive reversible LLPS.[24,27,32,34] Even in germ granules, containing proteins, there is evidence for well-organized RNA clusters stabilized by homotypic interactions[35,36]. Despite the importance of RNA–RNA interactions in driving LLPS, not much is known quantitatively about the organization and dynamics of RNA condensates.

Nucleotide repeat expansion is known to cause several neurological and neuromuscular disorders such as Huntington disease, muscular dystrophy and amyotrophic lateral sclerosis[37–40]. The disease-associated repeat sequences have been found in both the coding and non-coding regions of the transcripts[38]. However, determining exactly how these repeat expansion sequences cause diseases is still elusive. Recently, Jain and Vale[33] showed that the trinucleotide repeat sequences $(CAG)_n$, $(CUG)_n$ and repeats of the hexanucleotide (G_4C_2) form biomolecular condensates. The main findings of the Jain and Vale study are: (1) the high GC content sequences $(CAG)_n$, $(CUG)_n$ and G_4C_2 repeats form droplets, which over time transform to a gel-like state at high RNA concentrations; (2) phase separation occurs only if the number of repeats exceeds a critical value, which is similar to the value where diseases appear[33,41]; (3) in vivo, the gel-like state is abolished (except in the repeats of G_4C_2), and only a liquid-like state is found. Presumably, this is due to active forces generated by an energy source (ATP binding or hydrolysis, for example). Although RNA transcripts are usually exported to the cytoplasm, the $(CAG)_n$ foci, once formed, are preferentially retained in the nucleus and co-localize with nuclear speckles that sequester splicing factors. These findings provide the required impetus to investigate the relationship between the tendency to undergo LLPS and the neurotoxicity of the repeat RNA sequences.

To provide insights into the mechanism of condensate formation of the RNA repeat sequences, the organization of the RNA chains in the condensates and the associated dynamics, we created a minimal coarse-grained model, representing each nucleotide by a single interaction site (SIS), to investigate the molecular mechanism of LLPS in RNA repeat sequences. We incorporated only features that are essential for intra- and intermolecular interactions of RNA. In creating the model, we adopted a top-down approach with the goal that it is sufficiently simple to make multichain simulations feasible. The resulting SIS model has only one parameter that sets the energy scale for base-pair interactions, which was chosen to reproduce the known structures of a short $(CAG)_2$ duplex[42].

[1]Department of Chemistry, The University of Texas at Austin, Austin, TX, USA. [2]School of Pharmacy, University of Nottingham, Nottingham, UK.
✉e-mail: dave.thirumalai@gmail.com

We then perform multichain simulations to decipher the LLPS mechanism in the sequences studied by Jain and Vale without adjusting any parameter to match experiments. We quantify the concentrations of the two phases and show unambiguously that phase separation indeed occurs in the simulations, which accords well with experiments. Our simulations recapitulate the length and concentration dependence of the phase separation, in agreement with the in vitro experiments. We show that intermolecular base-pair interactions drive phase separation. Unexpectedly, we find that once RNA molecules are recruited in the droplets, they undergo large conformational change, from a hairpin-like conformation in isolation to a stretched state, to form an extensive network of intermolecular interactions. This soft network, in turn, constrains the RNA conformational fluctuation and mobility. The RNA chains in the high-density viscous droplets are conformationally and dynamically heterogeneous. In the largest densely packed high-viscosity condensates, the RNA chains move predominantly by 'slithering' along their contour lengths. Such motions are reminiscent of reptation in polymer melts envisioned 50 years ago by de Gennes[43]. Our work provides important microscopic details of the mechanism of LLPS in the RNA repeat sequences, which currently are not easily accessible in experiments.

Results

LLPS depends on RNA concentration and size. Figure 1a illustrates the transition from monomers to condensates as a function of time, τ, for the $(CAG)_{47}$ polymer. At $\tau = 0$, the CAG chains exist as monomers. At intermediate times, oligomers form, which subsequently fuse together, resulting in large droplets as time progresses (see Supplementary Movie 1 for a vivid illustration of the phase-separation process). The fraction of RNA chains in the oligomers increases, reaching a peak value at intermediate times, and diminishes at long times. The increase in the droplet size coincides with a decrease in the number of RNA chains in the oligomers, suggesting that the small clusters must fuse to form large droplets.

We then examined the effect of chain length n on the phase separation in $(CAG)_n$. Figure 1b shows that reduction of n from 47 to 30 has little effect on the phase-transition behaviour. However, if n is smaller than a critical number $n^\star$, the propensity to form liquid droplets ceases entirely. In the $(CAG)_n$ system, $n^\star$ is somewhere between 20 and 30.

Next, we investigated the effect of RNA concentration, $C_{RNA} = N/V$ (where N is the number of RNA molecules and V is the volume of the simulation box), on the condensate formation. We reduced (increased) C_{RNA} by increasing (decreasing) V, while keeping the number of chains fixed, $N = 64$. Figure 1c shows that as C_{RNA} for $(CAG)_{47}$ decreases, there is a decrease in the droplet size. At 200 μM, almost all the molecules phase separate and enter the high-density phase, leaving very few RNA chains as monomers in the solution. The two largest droplets (blue and green curves) each with ~20–30 monomers frequently interact with each other. When C_{RNA} decreases to 50 and 100 μM, only small droplets (<20 monomers) form. The fluctuations in the size of these droplets are also much less pronounced, indicating that they are less likely to interact with each other and exchange monomers/oligomers.

We calculated the RNA concentrations of the coexisting low- and high-density phases using a procedure outlined in the Methods (see also Extended Data Fig. 1). The RNA concentration inside the droplets is enhanced by around 50- to 200-fold compared to the initial concentrations (shown in Fig. 1d), in agreement with the experimental value (163-fold).[33] Regardless of the initial C_{RNA}, the concentrations of the two phases are relatively unchanged. Once the initial concentration of $(CAG)_{47}$ reaches 20 μM, which is lower than the concentration of the aqueous (or dispersed) phase, stable droplets do not form. At low concentrations, the RNA molecules mostly exist either as monomers or oligomers, and the system as a whole is a single liquid phase. Thus, for a fixed n that is larger than $n^\star$, there is a threshold value of C_{RNA}, which has to be exceeded to form stable condensates. From the simulations at different n and C_{RNA}, we determined the putative phase diagram for $(CAG)_n$ (Fig. 1e). It shows that for $n \leq 20$, RNA molecules exist in a single phase. For $n > 30$, we predict coexistence of the high- and low-density phases, resulting from LLPS of the CAG repeat system.

RNA structure changes dramatically in the condensates. We then characterized the RNA conformations inside the condensates containing multiple chains utilizing numerous quantities widely used in polymer physics. The conformational changes of $(CAG)_n$ relative to the monomers is dramatic. (1) First, we calculate the mean distance between two nucleotides separated by $s = |i - j|$ (where i and j are the indices of these nucleotides along the chain). The $R(s)$ curve (shown in Fig. 2a) increases at large s, implying that the two ends are not in proximity, as in the isolated chain (Extended Data Fig. 2). (2) Interestingly, $R(s)$ versus s is independent of the droplet sizes. (Oligomers, 2–4 monomers; medium droplets, 5–10 monomers; large droplets, >10 RNA molecules.) (3) The bond orientational correlation, $\cos\theta(s) = \langle \mathbf{b}_i \cdot \mathbf{b}_{i+s} \rangle / l_b^2$ (where $\mathbf{b}_i$ is the i^{th} bond vector and l_b is the bond length), of the individual chains lacks periodicity inside the droplets, which is prominent in the monomers and oligomers (Fig. 2b). (4) The end-to-end distance (R_{ee}) distribution for chains inside the droplets shifts to higher values (Fig. 2c), signalling a disruption of the hairpin structures adopted by the isolated chains. Snapshots from the simulations (Fig. 2e) show that the RNA polymers populate extended conformations.

The form factors (equation (4) in Methods), which characterize the overall shape of the polymer in the reciprocal space, show that the chains inside the droplets are markedly different from the monomeric RNAs (Fig. 2d). For the chains in the condensates, at small scattering vector $q\left(q \ll \frac{2\pi}{R_g}\right)$, $S_c(q) \approx N \exp\left(-\frac{1}{3}q^2 \langle R_g^2 \rangle\right) \approx N\left(1 - \frac{1}{3}q^2 \langle R_g^2 \rangle\right)$ in the Guinier regime. At higher q, there is a crossover to a power law, $S(q) \approx q^{-1/\nu}$, with $\nu \approx 0.5$, suggesting that the RNAs may be characterized as ideal chains just as in polymer melts. This is because from the perspective of a single chain inside dense condensates, it is irrelevant whether the interactions arise intramolecularly or from other chains in the droplet. A similar reasoning led to the 'Flory ideality hypothesis' proposed for polymer melts (see Fig. 3 for the analyses of intra- and intermolecular interactions). At small spatial length scale, or high q, $S(q) \approx q^{-1}$ is recovered.

Intermolecular base-pairing drives RNA condensate formation. We calculated the fraction of formed base pairs, $f_{bp} = N_{bp}/N_{CAG}$ where N_{bp} is the number of base pairs and N_{CAG} is the number of CAG units in the simulations, to decipher the molecular details of the transition from low-order oligomers to condensates (see Methods for details). The results were decomposed into intra- and intermolecular base pairs (shown in Fig. 3a). Initially, the chains mostly adopt hairpin-like structures by forming exclusively intramolecular base pairs. Once small and medium-sized droplets form, intermolecular interactions increase at the expense of intramolecular interactions. For RNA chains to form inter-chain base pairs, the C and G nucleotides from two chains have to satisfy both the distance and orientation criteria, thus requiring population of extended conformations (Fig. 2). In this process, the self-interactions are replaced by the intermolecular interactions without having to compensate for bending on a short length scale, s.

We find that intermolecular base-pair interaction is the driving force in the coalescence of the RNA repeats[18,33,44,45]. For short chains ($n = 20$) or at low C_{RNA}, the number of intermolecular interactions is small (Supplementary Fig. 4). Thus, most of the molecules remain as monomers or form relatively small droplets with fewer than

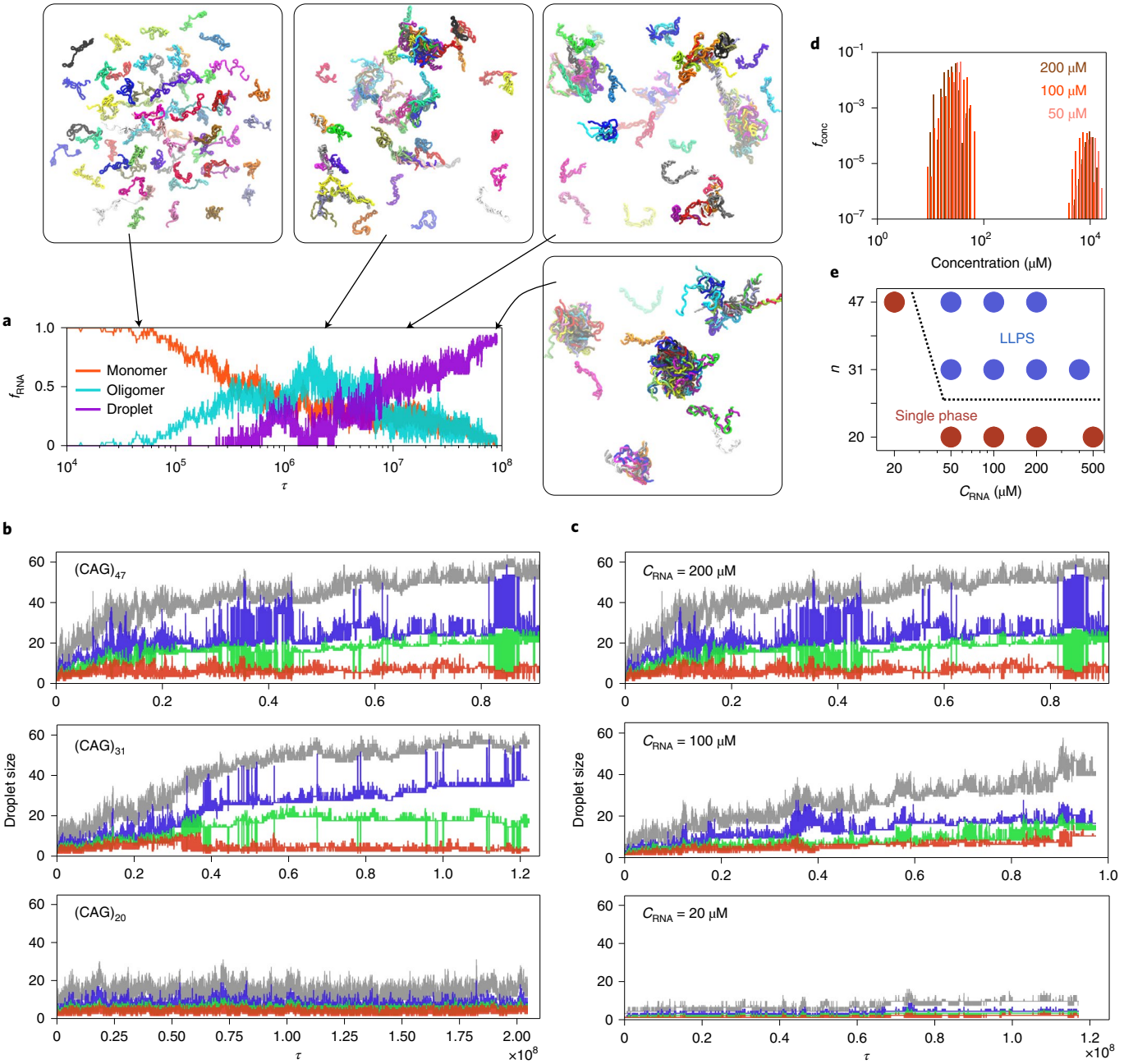


Fig. 1 | Liquid–liquid phase separation of CAG repeat RNAs. a, Fraction of $(CAG)_{47}$ in monomers, oligomers (dimers, trimers and tetramers) and droplets as a function of time. (Oligomers have 2–4 chains, droplets have ≥5 chains; see Fig. 3c for more details.) At $\tau = 0$, the RNA chains exist as monomers, whose structural characteristics are described in Extended Data Fig. 2. Subsequently, oligomers form and fuse together to form large droplets. Snapshots of $(CAG)_{47}$ pictorially illustrate phase separation. **b**, Evolution of the three largest droplet sizes (blue, green and orange, respectively) for different n. The droplet size is defined as the number of RNA chains inside the droplet. The grey curve is the sum of the three. $(CAG)_{47}$ and $(CAG)_{31}$ form several large droplets while $(CAG)_{20}$ shows little sign of phase separation. **c**, RNA concentration dependence of $(CAG)_{47}$ phase separation. At $C_{RNA} = 200\,\mu M$, very few monomers exist in the solution because they all interact with each other to form stable droplets. At $C_{RNA} = 20\,\mu M$, stable condensate is not observed. **d**, Concentrations of the two phases showing the coexistence of low- and high-density phases of $(CAG)_{47}$. The concentrations of the two (coexisting) phases are almost invariant, regardless of the initial concentrations, and are separated by almost three orders of magnitude. **e**, Putative phase diagram of $(CAG)_n$. For $n = 20$, there is only one phase even at elevated RNA concentrations.

eight chains. At high concentrations or for longer chains ($n = 47$ and $n = 31$), the increase in the number of intermolecular base pairs leads to phase separation, resulting in the formation of two discrete phases with vastly different RNA densities. The number of base pairs in the high-density phase is much higher than that in the low-density phase (Fig. 3b).

Intra- and intermolecular interactions depend on the droplet size (Fig. 3c). Unlike the self-interactions in the isolated RNA chain, which feature extensive Watson–Crick (WC) base-pairing along the anti-diagonal to form hairpin structures (Extended Data Fig. 2), those inside the droplets are structurally more diverse. The propensity to form interactions along the chains increases, beyond

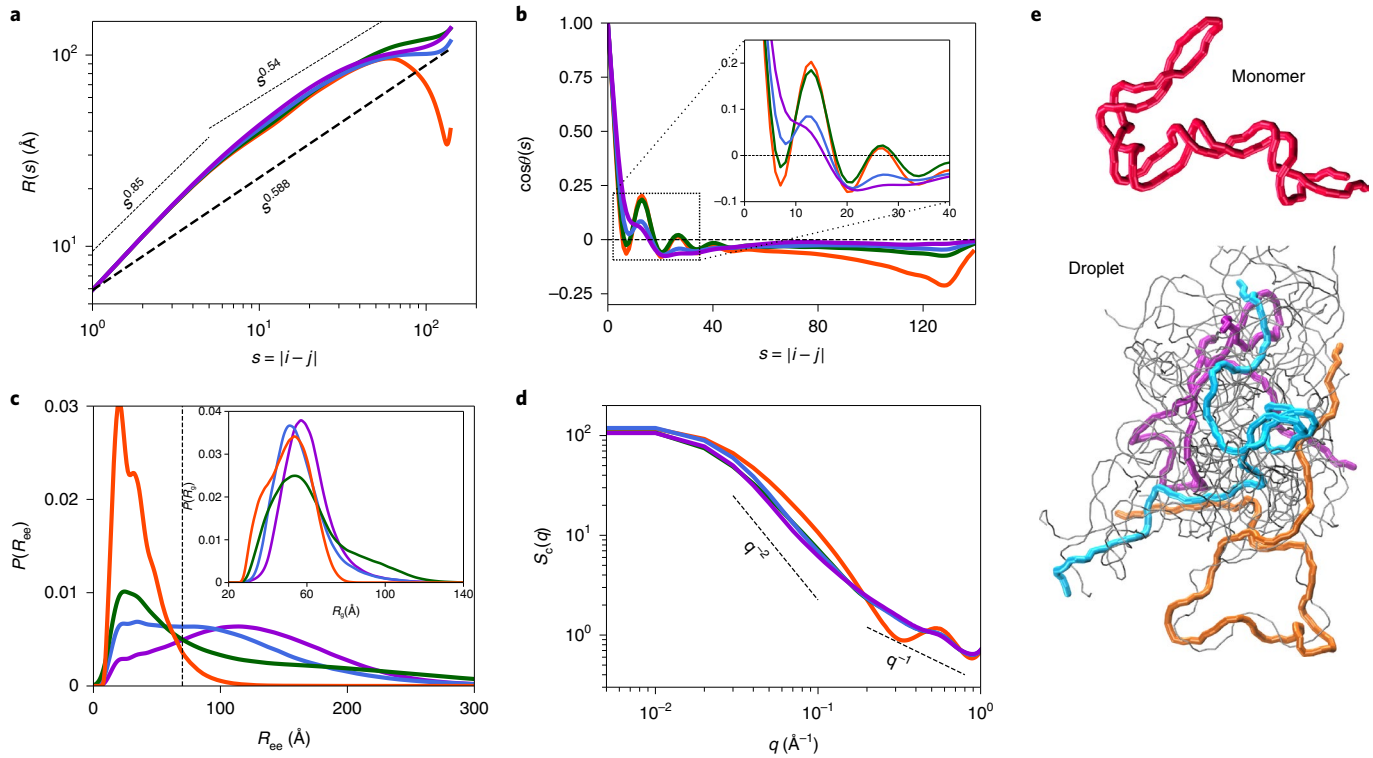


Fig. 2 | Comparison of the structures of $(CAG)_{47}$ inside the condensates and in isolation. The colours indicate if the molecule is a monomer, or in different sized droplets: orange, monomer; green, oligomer; blue, medium-sized droplet; purple, large droplet. **a**, $R(s)$ versus $s = |i-j|$. The dashed line shows $R(s) \propto s^{0.588}$ for a self-avoiding polymer. **b**, Bond orientational correlation function, $\cos\theta(s)$, as a function of s. Note the absence of periodicity (see inset for an enlarged view) in the large droplet (purple), which is prominent in the monomers and oligomers. **c**, Distributions of the end-to-end distance, R_{ee}. The vertical dashed line is the average value for an ideal chain. The R_{ee} values inside the condensates are substantially larger compared with the values for monomers, and the distribution could be fit using a broad Gaussian (see the inset in Fig. 4a for the fit). Inset: distributions of the radius of gyration (R_g). **d**, Form factors of RNA chains show that the RNA molecules inside the droplets are similar to an ideal polymer. **e**, Snapshots highlight the conformational differences between isolated chains (top) and chains inside the condensates (bottom).

those that form along the anti-diagonal. The total number of self-interactions of RNAs inside the droplets is smaller than in the monomers (Fig. 3a) due to the formation of extended conformations. Because the RNA conformations are considerably stretched inside the condensates, the interactions between different chains ought to increase. Intermolecular interactions between RNA molecules in droplets are also position-independent, reflecting the repeat nature of the sequence. However, for oligomers, specific 5′ to 3′ interactions still dominate, similar to what is found in a hybridized duplex.

RNA conformations in the droplets are highly heterogeneous. Due to extensive intermolecular base-pair formation inside the droplets, we expect the RNA mobility to be seriously hindered. Each RNA could be kinetically trapped, and only sample conformations around a local minimum in the free energy landscape, much like in a glass or a jammed system[46]. Figure 4 reveals that this is indeed the case. The distributions of R_{ee} and R_g of the individual chains have large dispersions, exceeding their usual values, although the histogram averaged over all the chains (shown in the inset) shows behaviour resembling that expected for ideal chains (red curves). Therefore, averaging over the ensemble of structures and over the number of RNA monomers in the droplet conceals the conformational heterogeneity. Some of the chains sample conformations with R_{ee} exceeding 200 Å, which is much greater than the dimensions expected for a random coil. As a consequence, there is a great degree of heterogeneity in the conformations that are sampled in the droplets. The unusual conformations that are accessed could arise because of the droplet is arrested in a non-ergodic phase with long overall relaxation times.

Condensates are dynamic. A hallmark of LLPS is that the droplets (or foci) interact with one another, which could result in two (or more) droplets fusing to form a larger droplet by the Ostwald ripening mechanism. The reverse process in which a large droplet disintegrates into smaller ones could also occur. Our simulations capture both these processes in the phase separation of $(CAG)_n$. Figure 5b shows the time evolution of each droplet size in the system, illustrating that the condensates formed in the simulations are dynamic, undergoing continuous growth, fusion and fission. The small droplets frequently interact and fuse with one another to form a larger droplet (snapshots in Fig. 5a and Supplementary Movie 2). However, not all fusion events are successful. There are instances of failed events in which the two droplets come together, interact for a while but subsequently dissociate (Fig. 5a). Such fusion–fission processes occur frequently, suggesting that the droplets are dynamic.

The dynamic nature of the droplets is also highlighted by the lack of persistent internal order, measured by the nematic order parameter (Fig. 5c). For small droplets, the internal RNA molecules need to position themselves in a preferred orientation, which would result in a loss of entropy. An extreme example is a dimer, in which the two strands are parallel to maximize the interaction energy (illustrated in Fig. 3c). However, in large droplets the RNA could form base-pair interactions with other chains without sacrificing orientational entropy, which results in a decrease in the nematic order parameter.

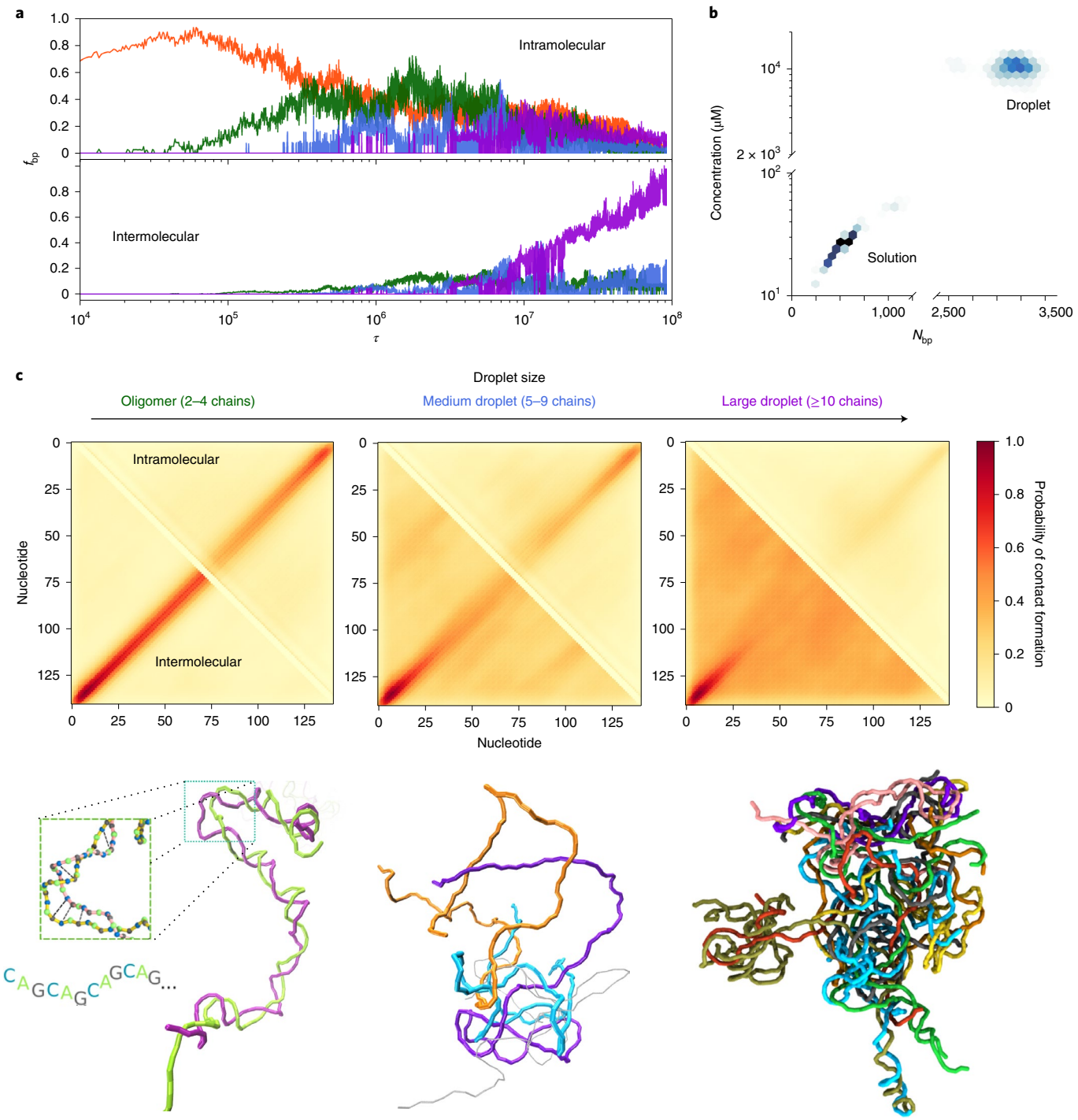


Fig. 3 | Intermolecular hydrogen bonding drives LLPS in RNA. a, Fraction of base pairs, f_{bp}, between G and C nucleotides in $(CAG)_{47}$ at 200 μM (results for other concentrations are shown in Supplementary Fig. 4). Different colours indicate if the molecule is a monomer (orange) or in different sized droplets: oligomer (green), medium (blue) or large (purple) droplets. Decomposition into intra- and intermolecular contributions (top and bottom panels, respectively). **b**, Concentration versus the number of base pairs formed in the two coexisting phases. **c**, Probability of contact formation due to intramolecular (top) and intermolecular (bottom) interactions. The contact between the two nucleotides is formed whenever the distance between them is less than 20 Å. In oligomers, intermolecular interactions between the RNAs generate structures similar to RNA duplexes, with signals along the anti-diagonal (from the bottom left to the top right) in the contact map. In larger droplets, the specific interactions decrease, and non-specific interactions predominate. Representative snapshots for different droplets with different sizes are at the bottom.

Figure 5a shows exquisitely the dynamic nature of the condensate formation. Each row corresponds to a single chain, with the colour denoting the droplet to which it is recruited. Each colour represents a unique droplet in Fig. 5b. At the early stage ($\tau < 2 \times 10^7$), most RNA molecules, at one instance or another, interact with almost all the droplets. Their residence times within a single droplet are relatively short. The monomers exchange with the bulk and are subsequently recruited by another (or the same) droplet. In contrast,

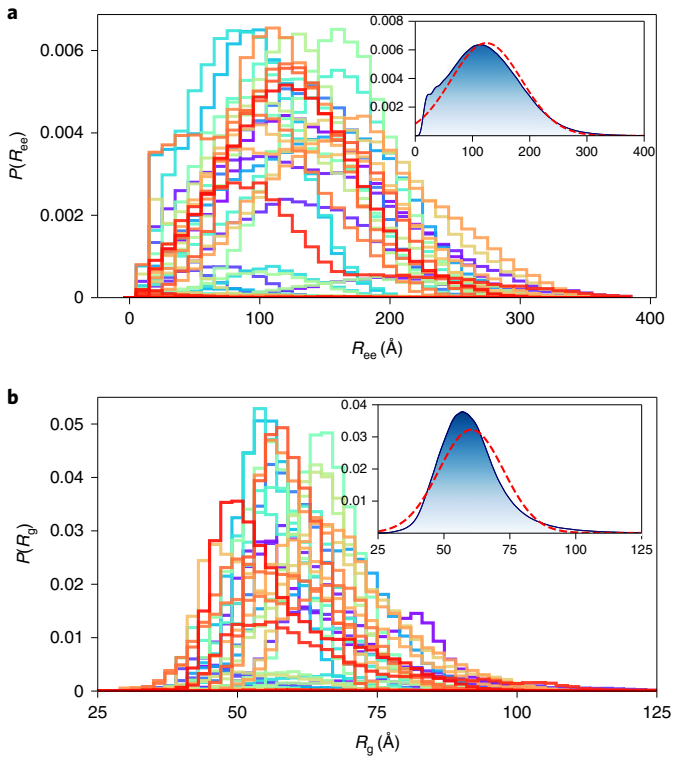


Fig. 4 | Structural heterogeneity of the RNA chains in large and dense $(CAG)_{47}$ droplets. a,**b**, Distributions of R_{ee} (**a**) and R_g (**b**) for each individual RNA chain. The insets show the average distributions, and the red dashed curves are the Gaussian fit to the data, indicating that the behaviour of chains inside large droplets is similar to that of ideal chains. Averaging over the monomers completely masks the remarkable heterogeneity of the monomer structures in the droplets.

at later stages, the chains are mostly restricted inside large droplets (orange and blue droplets in Fig. 5a). Despite the restricted movement, the interactions between the RNA chains inside large droplets are nevertheless transient, as illustrated in Fig. 5d, which shows the contact lifetime of one molecule with all other chains inside a single droplet. Although some of the contacts are stable, most of them are intermittent and are disrupted over time. The lifetime of contacts decays following a power law (Fig. 5e). This is because the chains shift their internal positions frequently, and therefore interact with different chains at various times (Supplementary Movie 2). Due to the absence of droplet fusion and lack of monomer exchange as time progresses, we surmise that the large droplets in our simulations could be near the onset of coarsening into a rigid gel-like state, which is in accord with in vitro experiments.[33]

A major advantage of our simulations is that it is possible to describe in vivid detail the behaviour of the chains during the phase-transition process where the monomers become part of the droplets. We found that there is no average or typical way a chain is integrated into a droplet, which underscores the importance of conformational and dynamic heterogeneity. Figure 6 illustrates one possibility (out of a large number) of how a droplet could grow, and what happens to the RNA chains during the process. Figure 6a shows the number of base pairs for a particular chain in the $(CAG)_{47}$ simulation. As the chain enters the droplet, the number of base pairs it forms with other chains increases, while the number of intramolecular base pairs diminishes. The recruitment of the monomer to a droplet started from a dangling end or an overhang of the hairpin (Fig. 6d,e). The end of the chain then hybridized with other chains on the surface of the droplet, and gradually penetrated into the droplet interior. Interestingly, the penetration of the monomer into the droplet was associated with slippage events of the stem region of the monomer hairpin. It is possible that the nature of the repeat sequences makes this kind of integration energetically feasible, as the monomer could retain intra-base-pairs that stabilize the stem on one side, while the other end hybridizes with other chains in the droplet. We wish to emphasize that the mechanism of incorporation is different for other chains. In addition, the droplet could grow by recruiting not only monomers but also oligomers and even smaller droplets (Extended Data Fig. 3 and Supplementary Movies 1 and 2). The growth mechanism is, therefore, probably highly heterogeneous.

Monomer dynamics inside the condensates is sluggish: evidence of reptation. We quantify the dynamics of RNA movement using the mean squared displacement (MSD) of the RNA centre-of-mass (Fig. 7a). The average MSDs for all the RNA chains increase linearly with time regardless of C_{RNA}, implying that the overall motion is diffusive. However, the dynamics are highly heterogeneous upon undergoing LLPS. The RNA chains in the aqueous phase undergo normal diffusion, but the movement of the polymers in the droplets is strongly affected, especially in large droplets. The MSD exponents are less than unity, and decrease as the droplet size increases (Fig. 7b), which is suggestive of subdiffusive behaviour. Indeed, the slowing down of the RNA dynamics inside large droplets could signal the onset of formation of a gel-like state. Given sufficient time, the large droplets could potentially undergo subsequent maturation through a sol–gel transition.[20,46–55]

To probe the jamming dynamics of RNA chains inside large droplets, we calculate the MSDs for each nucleotide in the RNA chain (Fig. 7c). The average values (the solid line in Fig. 7c) show that the MSD scales as $\Delta(\tau) \propto \tau^{1/4}$ at long times. In monomers, we find that $\Delta(\tau) \propto \tau^{0.78}$ (Supplementary Fig. 1). The $\tau^{1/4}$ behaviour, which was first predicted theoretically[43] and observed in simulations of polymer melts[43,56,57], is suggestive of reptation-like movement. The motion of these nucleotides is physically constrained along the contour length of the RNA, i.e. the chain movement is thought to occur in a low (one)-dimensional 'tube' generated by topological constraints imposed by other chains (see Supplementary Movie 3 for illustration). To exhibit reptation-like dynamics, the RNAs should form an intricate and entangled network of interactions, which arises due to intermolecular base-pair interactions. Interestingly, the formation of such a network has recently been demonstrated in AU-rich mRNAs[27,58]. The network of intermolecular interactions between the RNAs, therefore, plays a crucial role in dictating the slow heterogeneous dynamics of RNAs inside the droplets.

Discussion and conclusion

We developed a minimal coarse-grained SIS model that captures many of the observed features in the LLPS of repeat RNA sequences. In accord with experiments, we find that long $(CAG)_n$ repeats ($n = 47, 31$) phase separate at relatively low concentrations, whereas short chains ($n = 20$) remain in a single liquid phase even at elevated RNA concentrations. It is worth noting that there is only a single parameter in the SIS model that was determined using the structures of the $(CAG)_2$ duplex. In probing condensate formation in CAG repeats, the SIS model is essentially parameter-free, and hence is transferable to probe LLPS in other RNA sequences as well. As an example, we show in Extended Data Fig. 4 that the tendency to undergo LLPS in one scrambled sequence is diminished compared to the CAG repeat sequence, which is in qualitative agreement with experiments where there are no droplets observed.[33] There could be two reasons—one thermodynamic and another kinetic—that explain the quantitative difference between our simulations and experiments. One possibility is that a much higher concentration is needed in the experiment to induce phase separation in the

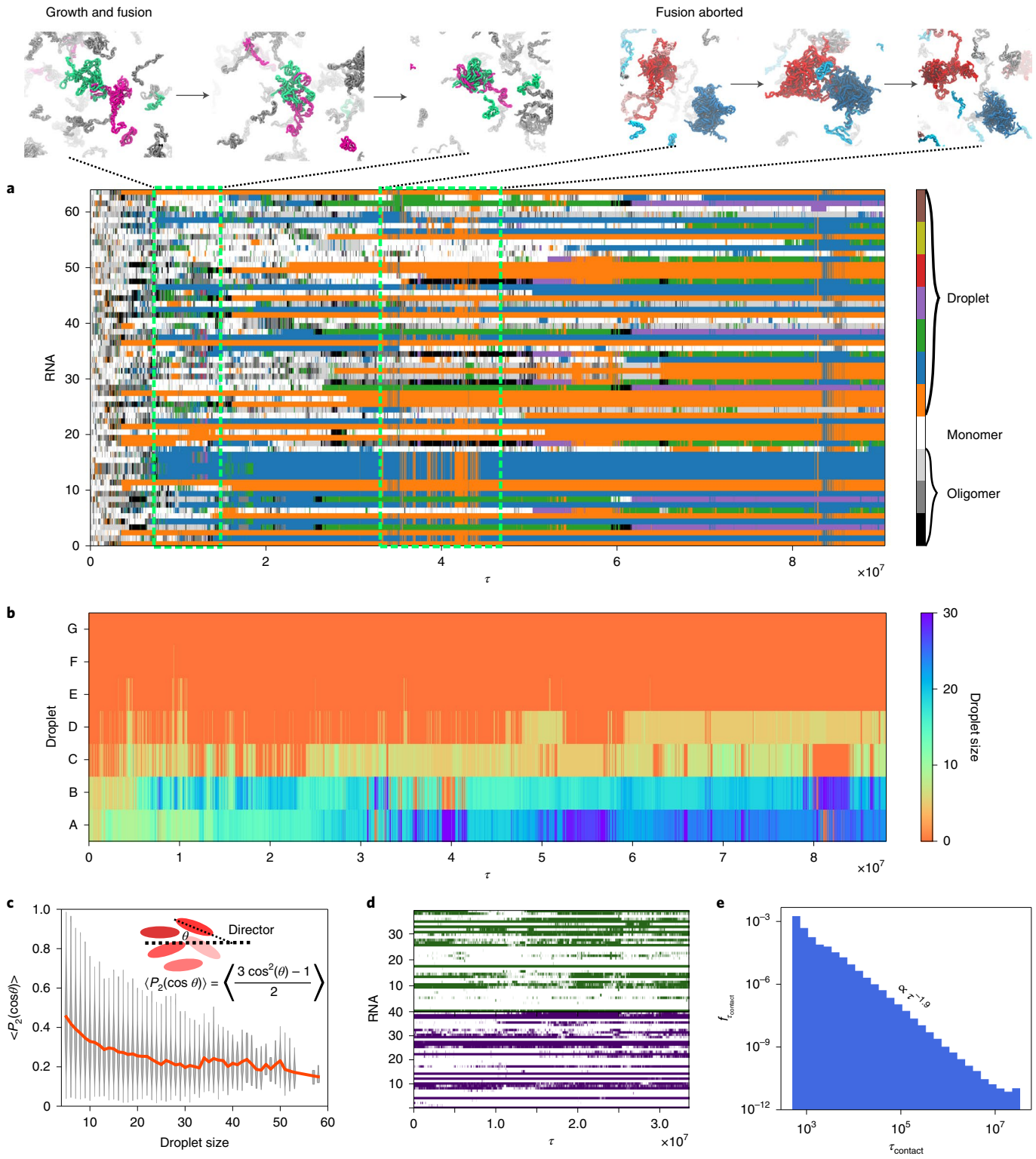


Fig. 5 | Condensate dynamics. a, Each line represents time evolution of a single chain. Colours encode the states of RNA. White is a monomer, oligomers are from grey to black. Seven droplets with different colours contain several RNA chains. Droplet growth and fusion shown on top left. Only a few chains are coloured, the rest are shown as grey molecules. An aborted fusion event is shown on the top right when the two droplets are in proximity and subsequently dissociate. See also Supplementary Movie 2. **b**, Evolution of droplets with the colour indicating the droplet size, with each row representing a unique droplet, corresponding to one colour in **a**. **c**, RNAs in small droplets have high nematic order, $< P_2 \cos(\theta) >$, which decreases as the droplet size increases (the orange curve shows average values). The grey lines show that there are substantial order parameter fluctuations. **d**, Contacts between one particular RNA with other chains inside the same droplet as a function of time. Colour code: white, no contact; colour, contact. Shown are two plots for two different chains in the same droplet. **e**, Frequency of contact lifetime between chains inside large droplets. Most contacts are transient, and the distribution of $\tau_{contact}$ follows a power-law decay.

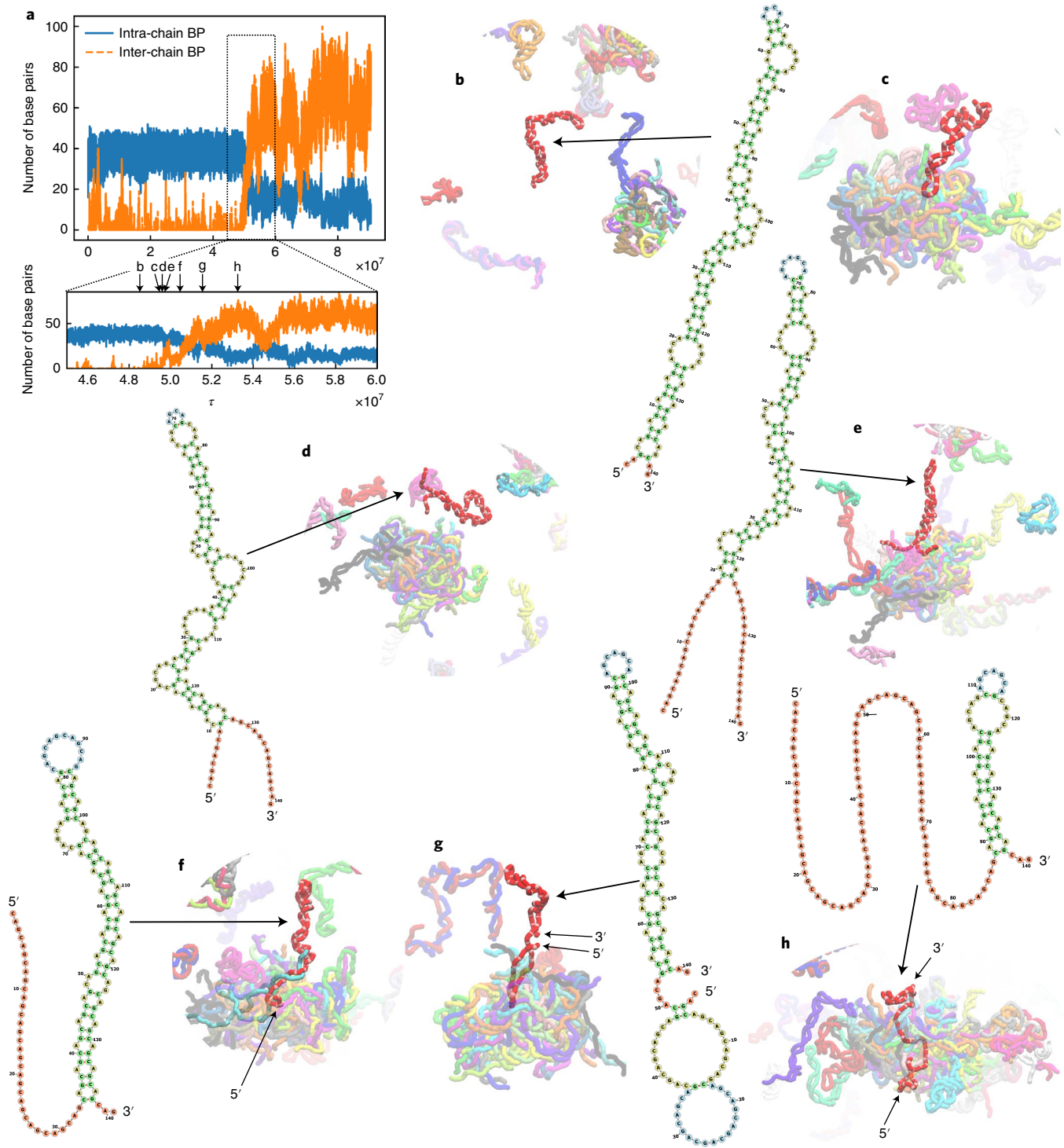


Fig. 6 | Visualization of the dynamics of monomer recruited into a droplet for $(CAG)_{47}$. This is merely one scenario that illustrates the changes in the dynamics of the monomer as it is recruited into the droplet. See Supplementary Movies 1 and 2 for additional scenarios. **a**, Number of base pairs of one specific chain as the recruitment process occurs. The base pairs are classified as either intra-chain (blue) or inter-chain (orange). In the bottom panel, the time is enlarged with arrows indicating time points of snapshots (**b**–**h**). **b**–**h**, Snapshots taken from the monomer recruiting event, accompanied by the RNA secondary structure of the recruited chain. Initially, the chain (in red) is in the hairpin conformation (**b**). The chain approaches a droplet from the tip of the loop. The attempt to coalesce is unsuccessful, and the chain eventually dissociates from the droplet (**c**). The 5′ and 3′ ends open, creating overhangs (**d**). The opened ends start interacting with other chains on the surface of the droplet (**e**). The 5′ end starts expanding and hybridizes with another chain (cyan) in the droplet. The base pairs of the hairpin stem slip due to the nature of the repeat sequence. As a result, the 3′ end shortens (**f**). At this stage, the 5′ end has been entirely integrated into the droplet. The 3′ end appears at the periphery, and transiently forms intramolecular base pairs (**g**). The RNA chain completely penetrates into the droplet from one side to the other. The 3′ end still has a short stem that is stabilized by intra-chain base pairs. This residual structure remains for the rest of the simulation (**h**). The secondary-structure diagrams were generated with Forna[79].

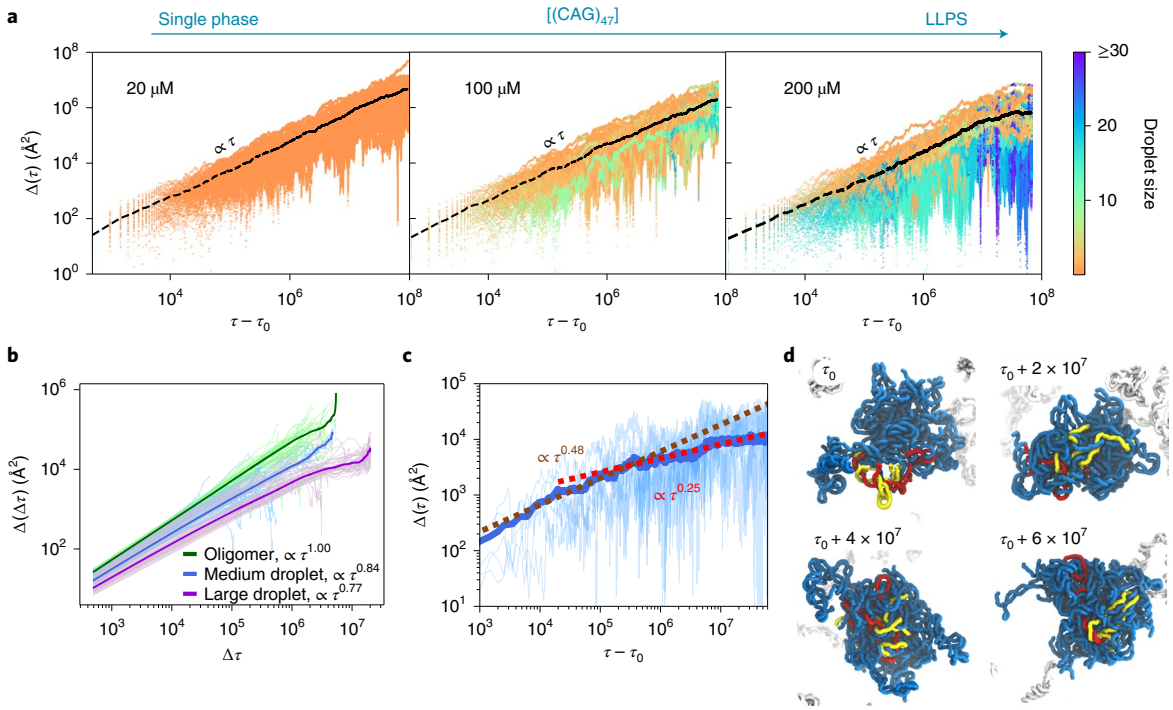


Fig. 7 | Dynamics of RNA chains in $(CAG)_{47}$ condensates is heterogeneous. a, MSD of the RNA centre-of-mass at different concentrations, increasing from left to right, shown on a log scale. Results are shown for individual RNA molecules with colour denoting the droplet size. Averaged values are plotted as black lines. Most of the fast-moving RNA chains are in oligomeric and monomeric states, while sluggish movement arises from RNA chains inside large droplets. **b**, Time-averaged MSD calculated for RNA chains inside oligomers, medium droplets and large droplets. The RNA movement in oligomers is diffusive, while for medium and large droplets, the molecules undergo subdiffusive motion. Results for individual RNA molecules are plotted in pale colour lines, and the average is shown as solid lines. **c**, Nucleotide MSD in a large viscous droplet. The solid line is the averaged value. In the intermediate regime ($\tau - \tau_0 < 10^5$, where τ_0 is the initial time when the droplet forms), the MSD exponent is ~0.5, indicating that the chains are jammed. There is a change in slope at $\tau - \tau_0 \approx 10^5$, where $\Delta \propto \tau^{1/4}$. This suggests that the movement of connected nucleotides is reminiscent of reptation, which implies that chain dynamics are mainly restricted along the contour length of the RNA due to entanglement. The reptation dynamics is most clearly illustrated in Supplementary Movie 3. **d**, Representative snapshots show movement of individual chains inside large droplets.

scrambled sequence. Another reason follows from the observation in the simulations that the scrambled sequence tends to form multiple smaller droplets instead of a few relatively large ones as in the CAG repeat sequence. It is conceivable that due to the smaller droplet sizes in the scrambled sequence, they are below the critical threshold detectable in the experiment (or one would need to wait longer for the droplets to grow to detectable sizes).

Recruitment of monomers to condensates involves unwinding of hairpin-like structure. Due to the high GC content of the sequence, the monomeric RNAs adopt an ensemble of hairpin-like structures with a small end-to-end distance (Extended Data Fig. 2), as is the case in a number of RNA molecules[42,59,60]. In sharp contrast, RNA chains inside the condensates are extended with large 5′–3′ distances. To form intermolecular interactions with other chains, the RNA monomer has to unwind, exposing the bases to form WC base pairs (or π–π stacking and non-canonical base pairs depending on the sequence[24,27,34]). We expect this finding to be a general feature of not only repeat RNAs, but also in any binary biomolecular system that undergoes phase separation (with solvent being another component), as was shown previously in oligomer formation in a fragment of Aβ peptides.[61]

Topological constraints, conformational and dynamical heterogeneity. As a result of a large number of intermolecular interactions between the RNA chains within the condensates, a given RNA molecule is topologically constrained by the neighbouring chains. The local environment of individual RNA chains differs greatly, which results in wide variability in the individual distributions of various shape parameters (Supplementary Fig. 6), even though the average distribution could be explained using polymer theory. Because of such topological constraints, the dynamics of RNA chains is also strongly affected. The mobility of the RNA chains is greatly diminished even though diffusion of the system as a whole is liquid-like[62,63]. In large droplets in which long RNA polymers are tightly packed with elevated density and viscosity, the movement of RNA molecules is reminiscent of reptation in polymer melts. RNA chains slither along their contour length due to translational inhibition in other dimensions. Strong entanglement in large droplets could further facilitate the maturation process, eventually driving the system to a gel-like state at long times under in vitro conditions. Under cellular conditions, it is likely that active processes (ATP binding and/or hydrolysis) regulate the formation of gel-like states by remodelling the RNA structures.[64–67] For instance, the recruitment of additional protein clients[29] or the presence of ATP-consuming enzymes (for example, helicases or RNA chaperones) in the nucleoplasm that reorganize the RNA base pairings[33,67] could play a role in maintaining the condensates in a liquid-like state.

Limitations of the study. The simulations, which use the SIS model without any parameter to fit the experiments, explain the findings on condensate formation in the repeat RNA sequence. However, the SIS model has limitations.

(1) Counterions, which play an important role in driving condensate formation[33,68], are not included explicitly. We assume that the cumulative effects of the Tris buffer, and the added monovalent

and divalent ions used in the experiments effectively screen the electrostatic repulsions between the nucleotides. It has been shown by Oosawa and Manning and confirmed by experiments that the polyelectrolyte charge is neutralized by nearly 88% whenever divalent ions are present (compared to 76% for monovalent ions alone)[69–71]. Under these conditions, the electrostatic repulsion between the phosphate groups is effectively 1% of its original magnitude, justifying our assumption of complete neutralization of RNA charge by counterions. Our simulations are valid at high ion concentrations at which the charge on the phosphate is negligible. The impact of ions, especially divalent cations, has to be examined, possibly using a theory proposed recently[72], for a more complete theoretical understanding of the RNA phase separation.

(2) Our simulations only consider the canonical WC base-pair formation, and completely neglect the possibility that non-canonical base pairs can form. This assumption is rooted in the observation that only 20–30% of nucleotides form non-canonical base pairs in the RNA structures[73,74]. Another reason is that the difference between these types of base pairs mostly arises from the alternative conformations of the base, sugar and phosphate groups within the same nucleotide. Since the SIS model represents a nucleotide by a single bead, any distinction at the single nucleotide resolution cannot be modelled explicitly. To take into account the possibility of non-canonical base-pair formation, one needs to utilize the more detailed three-interaction-site model in which a nucleotide is represented by three beads for the base, sugar and phosphate groups[75,76]. Incorporation of such non-canonical base pairs in the simulations could introduce alternative conformations and intermediate structures during the phase-transition process (Extended Data Fig. 5), altering the kinetics of hairpin unwinding and the timescale of conversion from intra- to intermolecular base pairs. However, we anticipate that such modifications would not remarkably change the outcomes of the simulations.[77]

(3) It is unclear whether the in-silico-generated condensates adopt gel-like states as has been suggested using flourescence recovery after photobleaching experiments, which we should point out are not always accurate. We did provide evidence that the condensates start the transition from liquid-like to gel-like at long times. For instance, as time progresses, we have shown that the droplets fail to fuse once they are close and the monomers inside droplets stop exchanging with the diluted phase. These criteria are qualitatively similar to how gel-like or liquid-like natures of the condensates are characterized in experiments. Our simulations, therefore, suggest that the condensates first form via LLPS, and then subsequently coarsen into 'gels', which could be the state observed in the Jain–Vale experiment. The time scale of the maturation process is likely to be long compared with the simulation time scale presented here. Because of the difficulty in distinguishing between the sol state with very slow dynamics and a gel-like state, material properties (by applying shear, for example) have to be determined before further characterization of the RNA condensates.[46]

Final remarks. We have used molecular simulations to probe condensate formation in repeat RNA sequences, and have uncovered general concepts that are difficult to obtain using experiments alone. Our study establishes that in the process of self-assembly of low-complexity RNA sequences into droplets, RNA chain makes a transition from hairpin-like structures to extended conformations, thereby maximizing the number of intermolecular base-pair interactions. The droplets are stabilized by a network of soft intermolecular interactions, which repeatedly break and reform as the chains move within the confines of the droplets. The dense network, which is most prominent in large, high-density droplets, restricts the mobility of the chain mostly along the contour. As a result, motion occurs by a process that is similar to the snake-like movement envisioned in the context of polymer melts[43]. Taken together, our results show that the structural heterogeneity and the pinning of RNA chains by their neighbours makes the dynamics sufficiently sluggish that the droplet may be in a non-ergodic phase.

The entanglement of RNA chains arises in part because the CXG sequence could engage in both intra- and intermolecular WC base-pair formation. Because such specific interactions are probably not possible in poly(rU) sequences[78], which also form droplets stabilized possibly by π–π stacking interactions, we believe that the only requirement for droplet formation in low-complexity RNA sequences is the presence of many weak inter-chain interactions. This would suggest that the length dependence for condensate formation is likely to be different in poly(rU) sequences to that in CAG or CUG polymers. Future studies are needed to provide definite answers about the effect of base stacking and non-canonical base-pair formation on RNA phase separation. Finally, it is worth emphasizing that our computational framework is not only well suited to examine the mechanism of self-assembly in other RNA sequences but also is general enough to account for ion effects and other cellular factors.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41557-022-00934-z.

Received: 5 March 2021; Accepted: 24 March 2022;
Published online: 2 May 2022

References

1. Brangwynne, C. P. et al. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009).
2. Hyman, A. A., Weber, C. A. & Julicher, F. Liquid–liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **30**, 39–58 (2014).
3. Brangwynne, C. P., Tompa, P. & Pappu, R. V. Polymer physics of intracellular phase transitions. *Nat. Phys.* **11**, 899–904 (2015).
4. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).
5. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
6. Langdon, E. M. & Gladfelter, A. S. A new lens for RNA localization: liquid–liquid phase separation. *Annu. Rev. Microbiology* **72**, 255–271 (2018).
7. Berry, J., Brangwynne, C. P. & Haataja, M. Physical principles of intracellular organization via active and passive phase transitions. *Rep. Prog. Phys.* **81**, 046601 (2018).
8. Boeynaems, S. et al. Protein phase separation: a new phase in cell biology. *Trends Cell Biol.* **28**, 420–435 (2018).
9. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and challenges in studying liquid–liquid phase separation and biomolecular condensates. *Cell* **176**, 419–434 (2019).
10. Choi, J.-M., Holehouse, A. S. & Pappu, R. V. Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.* **49**, 107–133 (2020).
11. Dignon, G. L., Best, R. B. & Mittal, J. Biomolecular phase separation: from molecular driving forces to macroscopic properties. *Annu. Rev. Phys. Chem.* **71**, 53–75 (2020).
12. Rhine, K., Vidaurre, V. & Myong, S. RNA droplets. *Annu. Rev. Biophys.* **49**, 247–265 (2020).
13. Roden, C. & Gladfelter, A. S. RNA contributions to the form and function of biomolecular condensates. *Nat. Rev. Mol. Cell Biol.* **22**, 183–195 (2021).
14. Sabari, B. R., Dall'Agnese, A. & Young, R. A. Biomolecular condensates in the nucleus. *Trends Biochem. Sci.* **45**, 961–977 (2020).
15. Stockmayer, W. H. Theory of molecular size distribution and gel formation in branched-chain polymers. *J. Chem. Phys.* **11**, 45–55 (1943).
16. Flory, P. J. *Statistical Mechanics of Chain Molecules* (Interscience, 1969).
17. Semenov, A. N. & Rubinstein, M. Thermoreversible gelation in solutions of associative polymers. 1. Statics. *Macromolecules* **31**, 1373–1385 (1998).
18. Li, P. et al. Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336–340 (2012).

19. Han, T. W. et al. Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell* **149**, 768–779 (2012).
20. Patel, A. et al. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077 (2015).
21. Maharana, S. et al. RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science* **360**, 918–921 (2018).
22. Schwartz, J. C., Wang, X., Podell, E. R. & Cech, T. R. RNA seeds higher-order assembly of FUS protein. *Cell Rep.* **5**, 918–925 (2013).
23. Banerjee, P. R., Milin, A. N., Moosa, M. M., Onuchic, P. L. & Deniz, A. A. Reentrant phase transition drives dynamic substructure formation in ribonucleoprotein droplets. *Angew. Chem. Int Ed.* **129**, 11512–11517 (2017).
24. van Treeck, B. et al. RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *Proc. Natl Acad. Sci. USA* **115**, 2734–2739 (2018).
25. van Treeck, B. & Parker, R. Emerging roles for intermolecular RNA–RNA interactions in RNP assemblies. *Cell* **174**, 791–802 (2018).
26. Langdon, E. M. et al. mRNA structure determines specificity of a polyQ-driven phase separation. *Science* **360**, 922–927 (2018).
27. Boeynaems, S. et al. Spontaneous driving forces give rise to protein–RNA condensates with coexisting phases and complex material properties. *Proc. Natl Acad. Sci. USA* **116**, 7889–7898 (2019).
28. Tauber, D., Tauber, G. & Parker, R. Mechanisms and regulation of RNA condensation in RNP granule formation. *Trends Biochem. Sci.* **45**, 764–778 (2020).
29. Guillén-Boixet, J. et al. RNA-induced conformational switching and clustering of G3BP drive stress granule assembly by condensation. *Cell* **181**, 346–361.e17 (2020).
30. Sanders, D. W. et al. Competing protein–RNA interaction networks control multiphase intracellular organization. *Cell* **181**, 306–324.e28 (2020).
31. Kaur, T. et al. Sequence-encoded and composition-dependent protein–RNA interactions control multiphasic condensate morphologies. *Nat. Commun.* **12**, 872 (2021).
32. Aumiller, W. M., Pir Cakmak, F., Davis, B. W. & Keating, C. D. RNA-based coacervates as a model for membraneless organelles: formation, properties, and interfacial liposome assembly. *Langmuir* **32**, 10042–10053 (2016).
33. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).
34. Aumiller, W. M. & Keating, C. D. Phosphorylation-mediated RNA/peptide complex coacervation as a model for intracellular liquid organelles. *Nat. Chem.* **8**, 129–137 (2016).
35. Trcek, T. et al. *Drosophila* germ granules are structured and contain homotypic mRNA clusters. *Nat. Comm.* **6**, 7962 (2015).
36. Trcek, T. et al. Sequence-independent self-assembly of germ granule mRNAs into homotypic clusters. *Mol. Cell* **78**, 941–950.e12 (2020).
37. Gatchel, J. R. & Zoghbi, H. Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.* **6**, 743–755 (2005).
38. La Spada, A. R. & Taylor, J. P. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.* **11**, 247–258 (2010).
39. McMurray, C. T. Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.* **11**, 786–799 (2010).
40. Krzyzosiak, W. J. et al. Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucl. Acids Res.* **40**, 11–26 (2012).
41. Lee, D.-Y. & McMurray, C. T. Trinucleotide expansion in disease: why is there a length threshold? *Curr. Opin. Genet. Dev.* **26**, 131–140 (2014).
42. 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).
43. de Gennes, P. G. Reptation of a polymer chain in the presence of fixed obstacles. *J. Chem. Phys.* **55**, 572–579 (1971).
44. 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).
45. Ciesiolka, A., Jazurek, M., Drazkowska, K. & Krzyzosiak, W. J. Structural characteristics of simple RNA repeats associated with disease and their deleterious protein interactions. *Front. Cell. Neurosci.* **11**, 97 (2017).
46. Jawerth, L. et al. Protein condensates as aging Maxwell fluids. *Science* **370**, 1317–1323 (2020).
47. Kato, M. et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012).
48. Molliex, A. et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015).
49. Lin, Y., Protter, D. S. W., Rosen, M. K. & Parker, R. Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. *Mol. Cell* **60**, 208–219 (2015).
50. Murray, D. T. et al. Structure of FUS protein fibrils and its relevance to self-assembly and phase separation of low-complexity domains. *Cell* **171**, 615–627.e16 (2017).
51. Wang, J. et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174**, 688–699.e16 (2018).
52. Franzmann, T. M. et al. Phase separation of a yeast prion protein promotes cellular fitness. *Science* **359**, eaao5654 (2018).
53. Wegmann, S. et al. Tau protein liquid-liquid phase separation can initiate tau aggregation. *EMBO J.* **37**, e98049 (2018).
54. Ray, S. et al. α-Synuclein aggregation nucleates through liquid–liquid phase separation. *Nat. Chem.* **12**, 705–716 (2020).
55. Pytowski, L., Lee, C. F., Foley, A. C., Vaux, D. J. & Jean, L. Liquid–liquid phase separation of type II diabetes-associated IAPP initiates hydrogelation and aggregation. *Proc. Natl Acad. Sci. USA* **117**, 12050–12061 (2020).
56. Kremer, K. & Grest, G. S. Dynamics of entangled linear polymer melts: a molecular-dynamics simulation. *J. Chem. Phys.* **92**, 5057–5086 (1990).
57. Hsu, H.-P. & Kremer, K. Static and dynamic properties of large polymer melts in equilibrium. *J. Chem. Phys.* **144**, 154907 (2016).
58. Ma, W., Zheng, G., Xie, W. & Mayr, C. In vivo reconstitution finds multivalent RNA–RNA interactions as drivers of mesh-like condensates. *eLife* **10**, e64252 (2021).
59. 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).
60. Lai, W.-J. C. et al. mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances. *Nat. Commun.* **9**, 4328 (2018).
61. Nguyen, P. H., Li, M. S., Stock, G., Straub, J. E. & Thirumalai, D. Monomer adds to preformed structured oligomers of Aβ-peptides by a two-stage dock–lock mechanism. *Proc. Natl Acad. Sci. USA* **104**, 111–116 (2007).
62. Elbaum-Garfinkle, S. et al. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl Acad. Sci. USA* **112**, 7189–7194 (2015).
63. Moon, S. L. et al. Multicolour single-molecule tracking of mRNA interactions with RNP granules. *Nat. Cell Biol.* **21**, 162–168 (2019).
64. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).
65. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **15**, 469–479 (2014).
66. Guo, J. U. & Bartel, D. P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* **353**, aaf5371 (2016).
67. Tauber, D. et al. Modulation of RNA condensation by the DEAD-Box protein eIF4A. *Cell* **180**, 411–426.e16 (2020).
68. Onuchic, P. L., Milin, A. N., Alshareedah, I., Deniz, A. A. & Banerjee, P. R. Divalent cations can control a switch-like behavior in heterotypic and homotypic RNA coacervates. *Sci. Rep.* **9**, 12161 (2019).
69. Manning, G. S. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quart. Rev. Biophys.* **11**, 179–246 (1978).
70. Bloomfield, V. A. DNA condensation by multivalent cations. *Biopolymers* **44**, 269–282 (1997).
71. Bai, Y. et al. Quantitative and comprehensive decomposition of the ion atmosphere around nucleic acids. *J. Am. Chem. Soc.* **129**, 14981–14988 (2007).
72. Nguyen, H. T., Hori, N. & Thirumalai, D. Theory and simulations for RNA folding in mixtures of monovalent and divalent cations. *Proc. Natl Acad. Sci. USA* **116**, 21022–21030 (2019).
73. Lemieux, S. & Major, F. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucl. Acid Res.* **30**, 4250–4263 (2002).
74. Yang, H. et al. Tools for the automatic identification and classification of RNA base pairs. *Nucl. Acid Res.* **31**, 3450–3460 (2003).
75. Denesyuk, N. A. & Thirumalai, D. Coarse-grained model for predicting RNA folding thermodynamics. *J. Phys. Chem. B* **117**, 4901–4911 (2013).
76. Denesyuk, N. A. & Thirumalai, D. How do metal ions direct ribozyme folding? *Nat. Chem.* **7**, 793–801 (2015).
77. Best, R. B., Hummer, G. & Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl Acad. Sci. USA* **110**, 17874–17879 (2013).
78. Chen, H. et al. Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc. Natl Acad. Sci. USA* **109**, 799–804 (2012).
79. Kerpedjiev, P., Hammer, S. & Hofacker, I. L. Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 3377–3379 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

Methods

RNA energy function. 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 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[80,81].

We introduce the SIS model in which each nucleotide is represented by a single bead (Extended Data Fig. 6c). The energy function for an isolated RNA chain has the following form:

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

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: $U_{\mathrm{BA}} = \frac{1}{2}\sum_i k_{\mathrm{bond}}(r_i - r_0)^2 + \frac{1}{2}\sum_j k_{\mathrm{angle}}(\alpha_j - \alpha_0)^2$ where $k_{\mathrm{bond}} = 15.0\,\mathrm{kcal\,mol^{-1}\,\AA^{-2}}$, $k_{\mathrm{angle}} = 10.0\,\mathrm{kcal\,mol^{-1}\,rad^2}$, $r_0 = 5.9$ Å, $\alpha_0 = 2.618$ 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[82]:

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

where Θ is the Heavyside step function, $\sigma = 10.0$ Å and $\varepsilon = 2.0\,\mathrm{kcal\,mol^{-1}}$. 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.

The base-pairing term U_{BP} is a many-body and short-range potential that mimics the canonical WC base-pairing between A-U and G-C. The base-pairing potential between the two beads i, j is taken to be:

$$U_{\mathrm{BP}} = U_{\mathrm{bp}}^{\circ} \exp\left[U_{\mathrm{bp,bond}} + U_{\mathrm{bp,angle}} + U_{\mathrm{bp,dihedral}}\right], \tag{3a}$$

$$U_{\mathrm{bp,bond}} = -k_r\left(r_{ij} - r_{\mathrm{bp},0}\right)^2, \tag{3b}$$

$$\begin{aligned} U_{\mathrm{bp,angle}} = {} & -k_\theta\left(\theta_{i,j,j-1} - \theta_1\right)^2 - k_\theta\left(\theta_{i-1,i,j} - \theta_1\right)^2 \\ & -k_\theta\left(\theta_{i,j,j+1} - \theta_2\right)^2 - k_\theta\left(\theta_{i+1,i,j} - \theta_2\right)^2, \end{aligned} \tag{3c}$$

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

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; $r_{\mathrm{bp},0} = 13.8$ Å, $k_r = 3.0\,\mathrm{\AA^{-2}}$, $k_\theta = 1.5\,\mathrm{rad^{-2}}$, $k_\phi = 0.5$, $\theta_1 = 1.8326$ rad, $\theta_2 = 0.9425$ rad, $\phi_1 = 1.8326$ rad, $\phi_2 = 1.1345$ rad.

The functional form of U_{BP} is designed to capture both the WC base-pairing and the helical nature of the A-form RNA at single-nucleotide resolution. The only unknown parameter, the strength of the base-pairing interactions, is adjusted to reproduce the structure of a small CAG repeat sequence[42] ($U_{\mathrm{bp}}^{\circ} = -5.0\,\mathrm{kcal\,mol^{-1}}$ for the GC base pair, equivalently $\sim -1.67\,\mathrm{kcal\,mol^{-1}}$ per hydrogen bond; Extended Data Fig. 6).

Inter-chain interactions. Interactions between different chains also involve excluded volume and base-pairing interactions, which are treated identically as in intramolecular interactions. Thus, the SIS model when used to simulate multiple RNA chains does not distinguish between intra- and intermolecular base-pair interactions. A nucleotide could form WC base-pair interactions with others either within the same chain or in a different chain. Therefore, the SIS model is general enough for studying condensate formation in a generic RNA, provided that single-bead resolution suffices. As always, the validity of the model, like all others, can only be assessed by comparison to experiments.

Counterion effects. Although counterions play an important role in driving RNA condensate formation (Extended Data Fig. 1h in ref. [33]), we do not include them explicitly in the current version of the model. One reason for neglecting them is that the charge on the phosphate ions has to be effectively neutralized by counterions for them to coalesce into droplets. Upon counterion condensation, electrostatic repulsion between phosphate groups is greatly diminished. A similar situation occurs in Ψ-condensation, where it has been shown that roughly 90% of the DNA charge needs to be neutralized for the DNA to condense[69,70]. The electrostatic repulsion between phosphate groups is around 1% of its original magnitude. Thus, it is not unreasonable to assume that the effective charge on the phosphate group is sufficiently small that it can be neglected. As a result, the simulations reported here are valid only at high monovalent (for example, Na^+) and divalent (for example, Mg^{2+}) concentrations. Future works, using a recently proposed theory[72], could incorporate ion effects more precisely to calculate the ion-dependent phase diagram.

Simulation details and analyses. *Simulation details.* We placed 64 $(CAG)_n$ (n = 20, 31 or 47) molecules evenly spaced in a cubic box, whose size varied from 50 to 200 nm depending on the RNA concentration. The initial conformations of the RNAs are deliberately expanded to minimize possible conformational biases. The initial conditions used in the simulations are a mimic of the experimental protocol in which the RNA molecules are denatured at an elevated temperature for a certain duration, and then the temperature is decreased to room temperature[33]. To ensure that our results are not affected by initial conditions, we also performed simulations where the temperature is decreased slowly starting from a high temperature. We observed little difference in the phase behaviour (Extended Data Fig. 7), which shows that the results do not depend on the initial conditions as long as the RNA molecules are unfolded. Simulations were performed on a graphics processing unit using the custom OpenMM code to speed up sampling of the conformational space[83]. We used low-friction Langevin dynamics in which the viscosity of water was reduced by a factor of 100 to further speed up the sampling efficiency[84]. Even using the simple SIS model for RNA, the simulations are computationally intensive because there are many chains in the system. Simulations were conducted for ~100 days for each trajectory using an 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 of 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 SIS model is $r_{\mathrm{bp},0} = 13.8$ Å. Thus, the cut-off distance for clustering is about ~40% larger than $r_{\mathrm{bp},0}$.

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

$$S_{\mathrm{c}} = \left\langle \left| \sum_k \exp\left(i\mathbf{q}\cdot\mathbf{r}_k\right) \right|^2 \right\rangle_N, \tag{4}$$

where $\mathbf{q}$ is the momentum transfer and $\mathbf{r}$ is the position of RNA beads. The brackets denote averaging over the chains at different orientations.

Radius of gyration and shape parameters. R_{g} and the shape parameters were calculated using the gyration tensor, with each element defined as:

$$S_{xy} = \frac{1}{2N^2}\sum_i\sum_j \left(x_i - x_j\right)\left(y_i - y_j\right). \tag{5}$$

R_{g}^2 is then calculated by summing over the three eigenvalues of the gyration tensor $R_{\mathrm{g}}^2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$. To characterize the shape of RNAs, we calculated the relative shape anisotropy κ^2 and the shape parameter S using:

$$\kappa^2 = \frac{3}{2}\frac{\lambda_1^4 + \lambda_2^4 + \lambda_3^4}{\left(\lambda_1^2 + \lambda_2^2 + \lambda_3^2\right)^2} - \frac{1}{2}, \tag{6}$$

$$S = \prod_{i=1,2,3} \frac{\lambda_1^2 - \bar{\lambda^2}}{\bar{\lambda^2}}, \tag{7}$$

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 $-1/4 \le S \le 2$. 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 ith droplet is:

$$V_i = \frac{4}{3}\pi abc = 4\pi\sqrt{3}\lambda_1\lambda_2\lambda_3. \tag{8}$$

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

$$C_{\mathrm{droplet}} = \bar{C_i}, \tag{9}$$

where $C_i = N_i/V_i$ is the concentration of the i^{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:

$$C_{\text{solution}} = \frac{N - \sum_i N_i}{V - \sum_i V_i}. \quad (10)$$

Mean squared displacement. The MSD of the centre-of-mass of a chain is given by:

$$\Delta(\tau - \tau_0) = \left\langle (\mathbf{r}_i(\tau) - \mathbf{r}_i(\tau_0))^2 \right\rangle_N, \quad (11)$$

where τ_0 is the initial time.

The time-averaged MSD is calculated using:

$$\Delta(\Delta\tau) = \left\langle (\mathbf{r}_i(\tau + \Delta\tau) - \mathbf{r}_i(\tau))^2 \right\rangle_{\tau,N}. \quad (12)$$

We also computed the MSD of individual nucleotides to probe the dynamics of the RNA chains in large droplets (Fig. 7c). 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_0). This procedure ensures that the MSD accounts solely for the dynamics of chains within the droplet, but not for the collective translational and rotational motion of the entire droplet in the simulation box. The MSD was then calculated in the same way using equation (11) by averaging over all the nucleotides except for 10 nucleotides at both the 5′ and 3′ ends. Note that τ_0 depends on when the droplet forms.

Criterion for base-pair formation. To assess if a base pair is formed between G and C, we rely on the base-pair energy U_{BP} defined in equation (3). If $U_{BP} < -3k_BT$, where k_BT is the thermal energy, between two nucleotides, then we consider that they form a base pair. Since the energy function consists of several terms that depend on the distance, angles and dihedral angles around the two nucleotides (equations (3b), (3c) and (3d)), the energy-based definition naturally captures the geometric criteria that base pairs satisfy in the ideal A-form RNA.

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 (Extended Data Fig. 2a), which accords well with experiments and theoretical predictions for mRNA and long non-coding RNA[60,85–87]. The peak value 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 distribution of the radius of gyration, R_g (shown in Extended Data Fig. 2b), shows little deviation from the ideal chain behaviour.

Unstructured RNAs (poly(rA) or poly(rC)) that do not favour base-pair formation follow the expected R_{ee} distribution for a random coil[60,78]. In contrast, CAG repeats form intramolecular base-pair interactions, bringing the 5′ and 3′ ends into proximity. Experiments have reported that $(CAG)_n$ repeat sequences with small n could form stable hairpins containing GC base pairs with A:A mismatches[42,59]. Formation of GC base pairs in longer repeats also requires the chain to fold upon itself. Folding would be favoured if the overall free energy gain by forming GC base pairs compensates for the chain-bending penalty. For the CAG repeat sequences, extensive GC base-pair formation could mitigate the unfavourable interactions, leading to the formation of hairpin-like structures[42,44]. The probabilistic contact map (Extended Data Fig. 2d) shows that the majority of interactions are along the anti-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 Extended Data Fig. 2a, 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 SIS 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 a large energetic penalty.

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

$$\langle \cos\theta(s) \rangle = \langle \mathbf{b}_i \cdot \mathbf{b}_{i+s} \rangle / l_b^2, \quad (13)$$

where $\mathbf{b}_i$ is the vector of bond i^{th}, with the bond length l_b. $\text{Cos}\theta(s)$ shows periodicity at small $s = |i - j|$, where i and j are the indices of the nucleotides in the $(CAG)_n$ sequence (Extended Data Fig. 2c). At a short length scale (up to five or six nucleotides), the structure of CAG is roughly rigid, with $R(s)$ scaling almost linearly with s (inset of Extended Data Fig. 2c). 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 $(CUG)_n$ (Extended Data Fig. 8). Our simulations show that the hairpin-like structures in $(CXG)_n$ (X = A or U) are a consequence of multiple canonical WC base-pair formation.

Non-canonical base-pairing has a minimal effect on the conformations of the repeat monomers. In our model, only canonical WC base pairs are allowed; that is, G only pairs with C, and A only pairs with U. It is known that RNA bases do form a wide variety of other pairings, which may be classified as non-canonical base pairs[88]. In principle, it is possible to include non-canonical base pairs within our framework. However, by analysing the RNA structures in the PDB, it is found that the frequency of non-canonical base-pairs is small compared to WC base pairs; 70–80% base pairs in the PDB are WC, and the remaining fraction forms non-canonical base pairs[73,74]. Furthermore, while there is only one potential WC base pair between two bases (*cis*-WC–WC, using the Leontis–Westhof notation), there are 11 additional ways to form non-canonical base pairs by orienting different edges of the base (WC, Hoogsteen or sugar). Therefore, one could infer that the non-canonical base pairs are much less stable compared to the WC base pairs, which is the basis for our assumption that WC base-pairing is the dominant force in driving phase separation of the RNA repeats. In addition, the difference between these types of base pairs mostly arises from the alternative conformations of the base, sugar and phosphate groups within the same nucleotide. Since the SIS model represents a nucleotide by a single bead, any distinction within the nucleotide resolution cannot be modelled explicitly. Therefore, we believe that in order to quantitatively account for the effects of non-canonical base-pair formation, one needs the three-interaction-site model in which a nucleotide is represented by three beads for the base, sugar and phosphate groups[72,75,76,89]. The drawback is that a higher resolution naturally raises the computational demand, and making it difficult to simulate multi-chains to probe LLPS using the currently available computing resources.

Having given the justification for neglecting non-canonical base pairs, here we explored a revised SIS model that includes non-canonical base pairs between any two bases that cannot form WC base pairs. The base-pairing energy function in the revised model is the same as in the SIS model (equation (3)), except we include all other possible combinations of base pairs with a smaller $U_{bp}^{\circ} = -1.67$ kcal mol^{-1} (which is a third of the value used for a GC base pair). When used the revised model to the CAG repeats, this means that A could form non-canonical base pairs with C, A or G. Even though the modification is a minor addition, it takes much longer time to perform simulations compared with the SIS model. Thus, in this preliminary calculation, we report results for the monomer $(CAG)_n$, shown in Extended Data Fig. 5. The chain is somewhat more compact (R_{ee} and R_g are smaller than in the SIS model) due to the ability of A nucleotides to form non-canonical base pairs. However, the conformations are globally hairpin-like, which is exactly the same as in the model that neglects non-canonical base-pair formation. Therefore, introduction of the non-canonical base-pair formation in the phase-separation simulations could alter the kinetics of hairpin unwinding and the time scale of conversion from intra- to intermolecular base pairs. However, we anticipate that such modifications would not remarkably change the major findings in our work.

Effect of cooling rate on droplet formation. An annealing protocol was employed in the in vitro phase-separation experiments of the CAG repeats[33]. RNA molecules were denatured at 95 °C for 3 min and cooled at the rate of 1–4 °C min^{-1} to 37 °C. In our simulations, the initial conformations of the RNA molecules are deliberately expanded to eliminate biases in the structures that the RNA molecules adopt. The expanded initial condition used in the simulations is a mimic of the experimental protocol in which the RNAs are denatured at an elevated temperature for a short duration. Subsequently, the temperature is lowered to the desired value. To ensure that our findings are not affected by the initial conditions, we also performed simulations in which we slowly decreased the temperature, as in the experiments. We first ran simulations at 100 °C, which is high enough such that all the RNA molecules are unfolded. Then, after every δt, we lowered the temperature by 10 °C until the final temperature reached 20 °C. We could explore the cooling rate by varying δt. We hasten to add that limitations in the simulation times prevent us from reaching the cooling rates achieved in experiments. Extended Data Fig. 7 shows that the overall picture of phase separation is not qualitatively altered, regardless of the procedures followed in initiating the simulations. This indicates that our simulations are robust with respect to the initial conditions and simulation protocols, and that the phase separation of the repeat RNAs occurs spontaneously, just as in experiments. It should be noted that experimental studies from other groups also showed that RNA molecules alone undergo phase separation without using the aforementioned annealing method,[27,34] or using a rapid cooling rate as we did here[24]. Thus, it appears that initial conditions do not affect the mechanism of LLPS in RNA repeat sequences.

Droplet stability at low salt concentrations. To probe the droplet formation in low salt conditions, we included electrostatic effects. In addition to the bond, angle, excluded volume and base-pairing terms in equation (1), each nucleotide now

carries a charge of $-q$ $(0<q<1)$ to account for counterion condensation[69]. The electrostatic repulsion between them is modelled using the Debye–Hückel theory, accounting for the screening effect of the buffer solution. This approach has been successfully used to calculate accurately the thermodynamics of RNA folding[72,75]. To test if the phase separation occurs at low salt conditions, we performed simulations using a model that includes ion effects. The initial conditions correspond to a preformed condensate. The simulations show that the RNA droplets start dissociating into smaller droplets to form oligomers and monomers (Extended Data Fig. 9). Thus, in the presence of monovalent ions alone, the droplet is unstable, which recapitulates qualitatively the experimental findings.

Of course, these simulations using the Debye–Hückel theory only account for the effect of monovalent ions (Na^+, K^+, and so on). This approach could be extended to include divalent cations (Mg^{2+}, Ca^{2+} and so on) explicitly, while keeping monovalent ions at the continuum level, as we have shown previously for the RNA folding problem[72]. Although simulations using such a model would be computationally demanding, they would provide the complete phase diagram. The SIS model is sufficiently general to accommodate ion explicitly, as shown elsewhere[72,75].

Scrambled sequences are less likely to undergo LLPS. We then tried a different sequence to test the model transferability. We shuffled the sequence of $(CAG)_{47}$ to maintain the GC content and chose the sequence (out of an astronomically large number of sequences) CCGGGAAGAGACCGCAACAGAAGCAGCCG CGAGCGCGACAGCGACGAGCACGCGGCACAGACAGCAAGAGAAGGG AAGACAAAGAGCCGACGGAAGCCACGCAGAGCAAAAGGACGCCGGC GGAACGACAAGGAAAGAGAG. To the best of our knowledge, Jain and Vale did not report the order of nucleotides in their scrambled sequence, which forced us to create one for computational purposes. We then repeated the simulations at different bulk concentrations 200, 100, 50 and 20 μM. We observed that this particular scrambled sequence undergoes phase separation, which seems to be in apparent contradiction to the findings of the Jain–Vale experiment (Extended Data Fig. 4). However, in our simulations, the fraction of chains that are in the droplet state for the scrambled sequence is lower than in $(CAG)_{47}$ (Extended Data Fig. 4a). As a result, the concentration of the diluted solution in the case of the scrambled sequence is larger (45 μM, compared to 25 μM for $(CAG)_{47}$) (Extended Data Fig. 4c). We emphasize that the concentration of the diluted phase is the saturated concentration C_{sat}, above which phase separation occurs. The observation of $C_{sat}^{scrambled} > C_{sat}^{(CAG)_{47}}$ indicates that the propensity to undergo phase separation of the scrambled sequence is lower than in $(CAG)_{47}$. On the other hand, the concentration inside the droplets of the scrambled sequence is lower than in droplets composed of $(CAG)_{47}$ (5 versus 10 mM), suggesting less compaction of the scrambled sequence droplets. This decreased compaction is probably due to the random nature of the sequence, leading to a decreased probability of finding similarly patterned neighbouring chains to form base pairs. We also observed that the droplet sizes in the case of the scrambled sequence are smaller than $(CAG)_{47}$, and the system tends to form many smaller droplets instead of a few relatively large droplets (Extended Data Fig. 4).

Taken together, the simulations suggest that the propensity to undergo LLPS of the scrambled sequence is less than that of the repeat sequence, which may be in qualitative agreement with the experiment[33]. It is possible that a much higher concentration of the scrambled sequence is needed to induce phase separation. Even if the concentration is higher than C_{sat}, it is conceivable that due to the smaller droplet sizes in the scrambled sequence compared to the repeat sequence, it is below the critical threshold detectable in the experiment (or one would need to wait longer for the droplets to grow to detectable sizes).

Local behaviour of RNA chains growing from monomer to oligomer to droplet. A major advantage of simulations is that the microscopic behaviour of the RNA chains can be visualized and investigated if the simulations are validated against experiments. In particular, it is interesting to probe the sequences of events, which are not currently accessible in experiments, that occur as the monomer becomes part of the condensate. One difficulty is that the fate of each chain could be different, which implies that there may not be typical or most probable changes in the dynamics as the monomer becomes part of the droplet. With this caveat, we performed several analyses focusing on the local behaviour of chains.

Extended Data Fig. 3 illustrates examples of local dynamics while multiple chains form a single droplet. In the series of snapshots, we show a process of coalescence by 11 RNA chains that eventually form an 11-mer in the middle of the simulation at 200 μM $(CAG)_{47}$. We observed that chains often form small oligomers first, then they coalesce to form larger oligomers, leading eventually to a droplet. Oligomers can also grow by merging with monomers and dimers. In the sample trajectory, we show that a trimer and a tetramer coalesce into a heptamer, followed by integrating other dimers to finally form an 11-mer. A more detailed description of the event is provided in the caption of Extended Data Fig. 3.

Reptation does not occur for monomers. Supplementary Fig. 1 shows MSD for nucleotides in an RNA monomer in the diluted phase. The MSD scales as $t^{0.78}$, which is substantially larger than $t^{1/4}$ expected for reptation. This proves that reptation dynamics only occurs inside the droplets.

Data availability

All data are included in the paper and the Supplementary Information. The raw data are available on Zenodo at https://zenodo.org/record/5794441.[90]

Code availability

The codes to perform simulations and analyses are available at GitHub (https://github.com/tienhungf91/RNA_llps).

References

80. Hyeon, C. & Thirumalai, D. Mechanical unfolding of RNA: from hairpins to structures with internal multiloops. *Biophys. J.* **92**, 731–743 (2007).
81. 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).
82. Weeks, J. D., Chandler, D. & Andersen, H. C. Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Chem. Phys.* **54**, 5237–5247 (1971).
83. Eastman, P. et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017).
84. Honeycutt, J. D. & Thirumalai, D. The nature of folded states of globular proteins. *Biopolymers* **32**, 695–709 (1992).
85. de Gennes, P. G. Statistics of branching and hairpin helices for the dAT copolymer. *Biopolymers* **6**, 715–729 (1968).
86. 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).
87. Clote, P., Ponty, Y. & Steyaert, J.-M. Expected distance between terminal nucleotides of RNA secondary structures. *J. Math. Biol.* **65**, 581–599 (2012).
88. Leontis, N. B. & Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **7**, 499–512 (2001).
89. Hori, N., Denesyuk, N. A. & Thirumalai, D. Shape changes and cooperativity in the folding of the central domain of the 16S ribosomal RNA. *Proc. Natl Acad. Sci. USA* **118**, e2020837118 (2021).
90. Nguyen, H., Hori, N. & Thirumalai, D. Raw data for 'Condensates in RNA repeat sequences are heterogeneously organized and exhibit reptation dynamics' (2021); https://doi.org/10.5281/zenodo.5794441

Acknowledgements

We are indebted to A. D. Bowen at the Visualization Laboratory (Vislab), Texas Advanced Computing Center, for generating the videos. We are grateful to H. Maity, S. Myong, M. Mugnai, S. Sinha and R. Takaki for stimulating discussions and critical reading of the manuscript. This work was supported by National Science Foundation Grant (CHE 19-00093) and the Welch Foundation Grant (F-0019) through the Collie–Welch chair. We thank the Texas Advanced Computing Center for providing computational resources.

Author contributions

H.T.N. and D.T. conceived and designed research, H.T.N. conducted research, H.T.N., N.H. and D.T. analysed the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

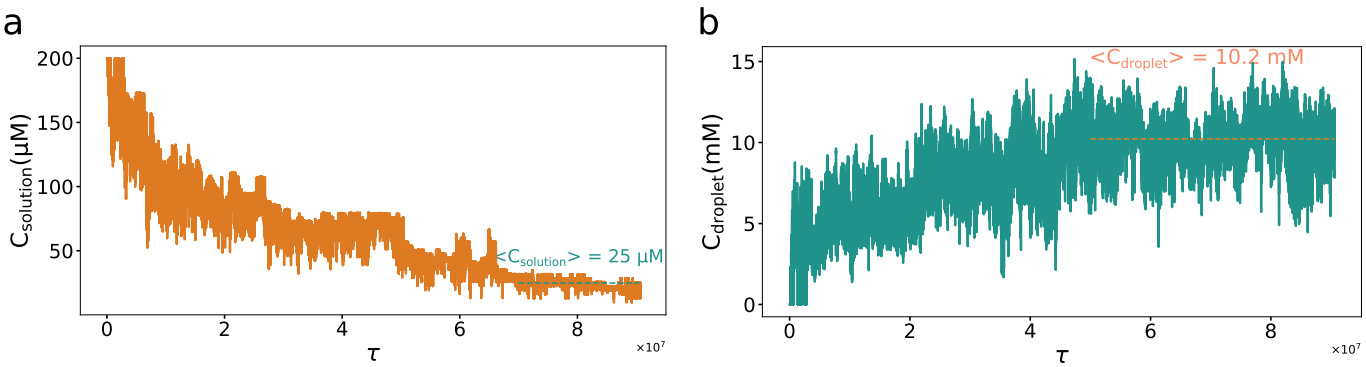
Extended data is available for this paper at https://doi.org/10.1038/s41557-022-00934-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-022-00934-z.

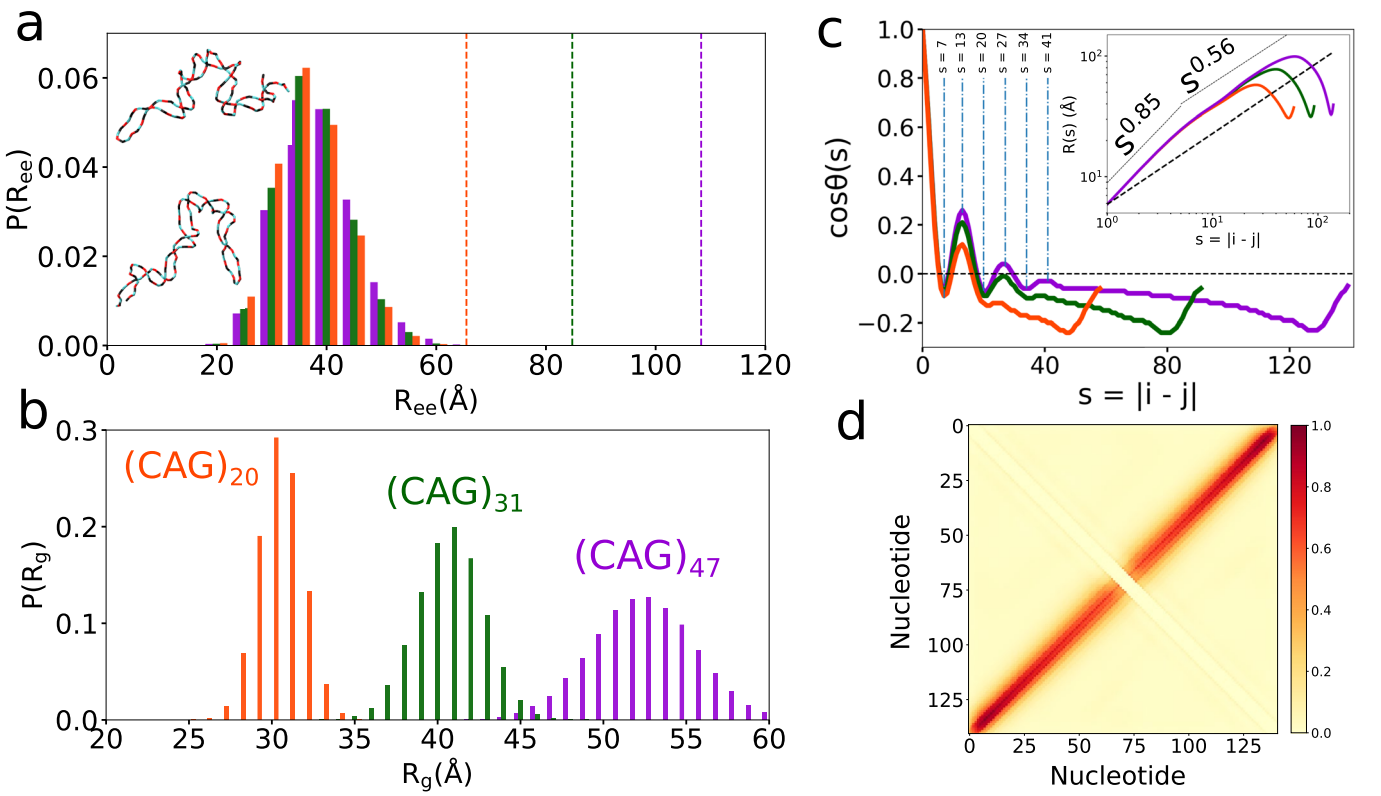
Correspondence and requests for materials should be addressed to D. Thirumalai.

Peer review information *Nature Chemistry* thanks the anonymous reviewers for their contribution to the peer review of this work.

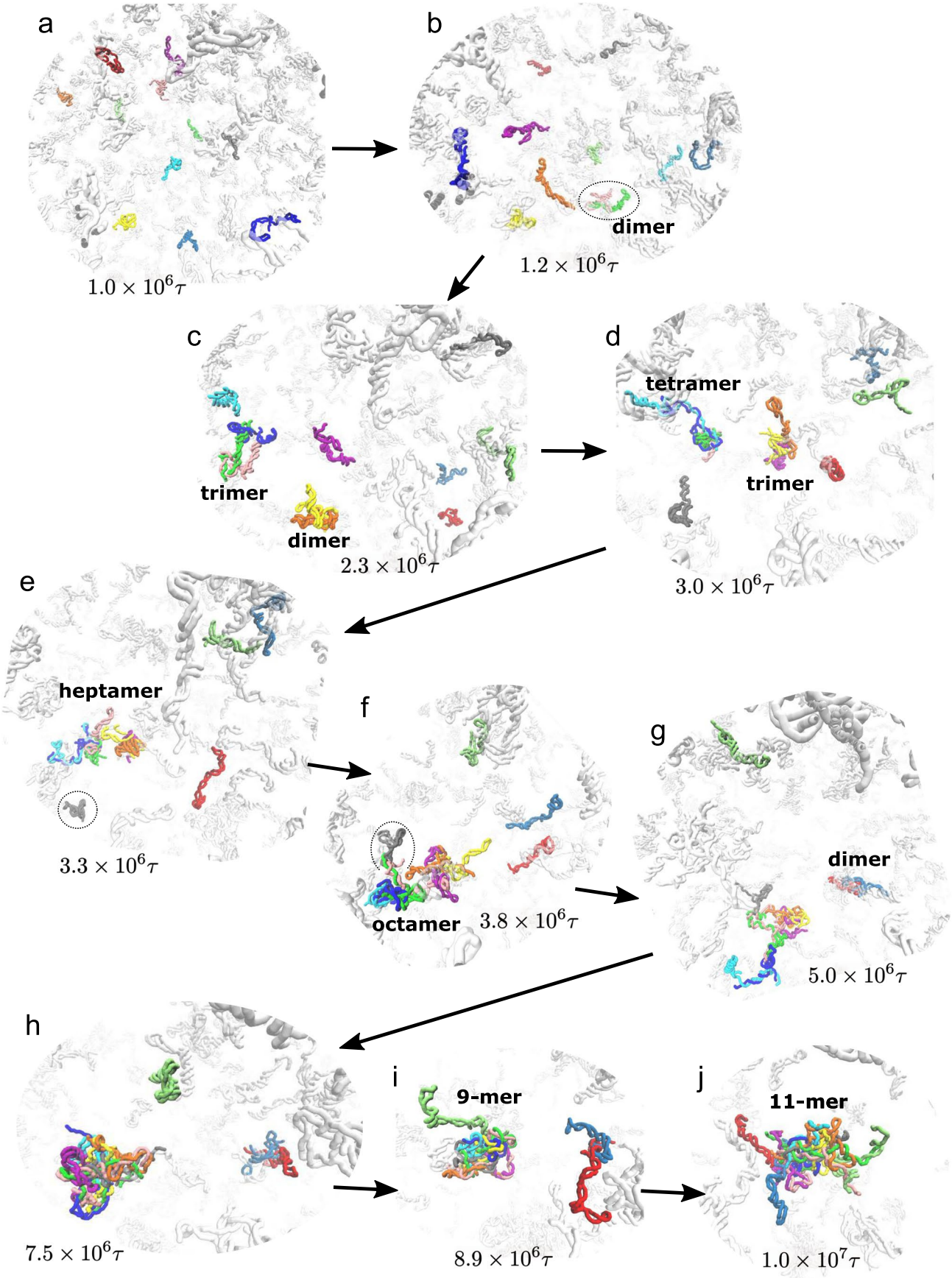
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Determination of the concentrations of the two coexisting 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.

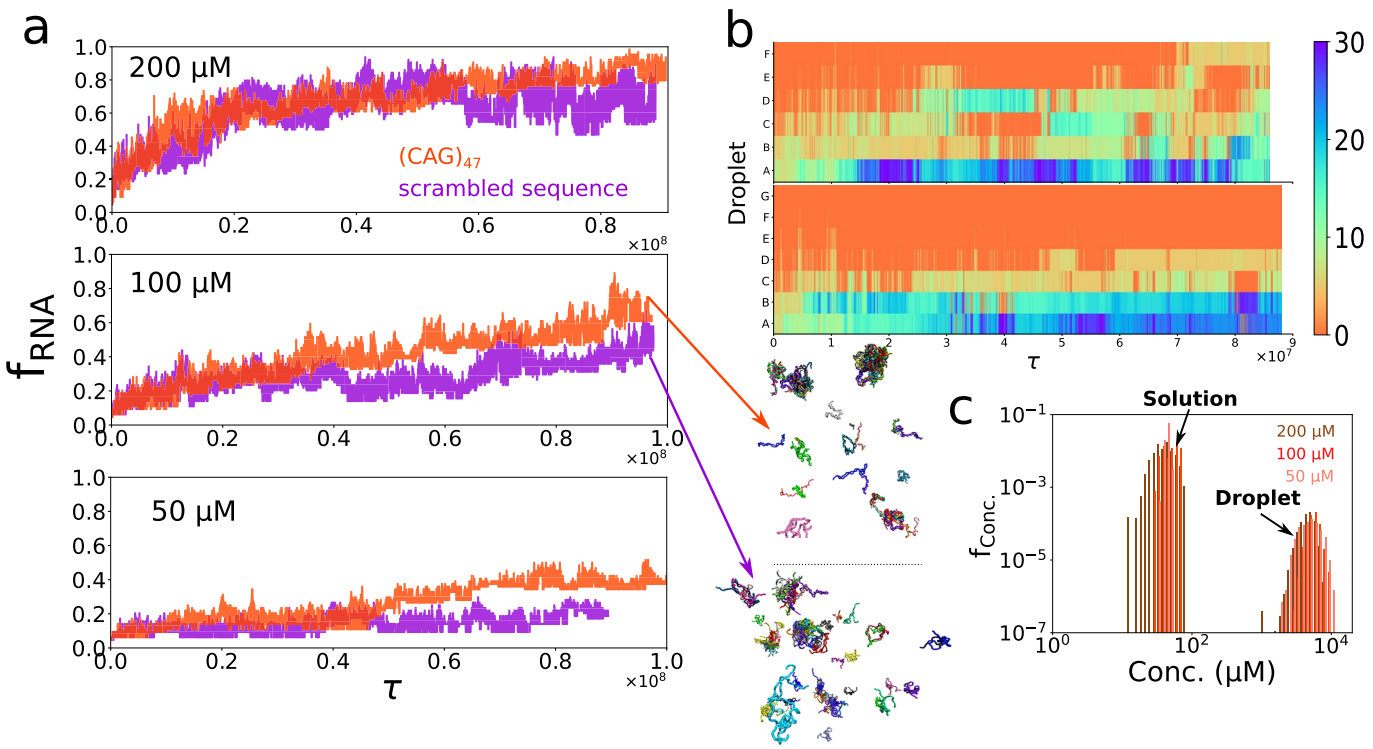


Extended Data Fig. 2 | Structures of isolated $(CAG)_n$ monomers. a, Distribution of the end-to-end distance R_{ee} and **b**, radius of gyration R_g of $(CAG)_n$ with n=47, 31 and 20. The vertical dash lines in **a** indicate mean values for self-avoiding random walk chains with the same *n*. Snapshots 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 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 anti-diagonal, indicating the formation of hairpin structures.

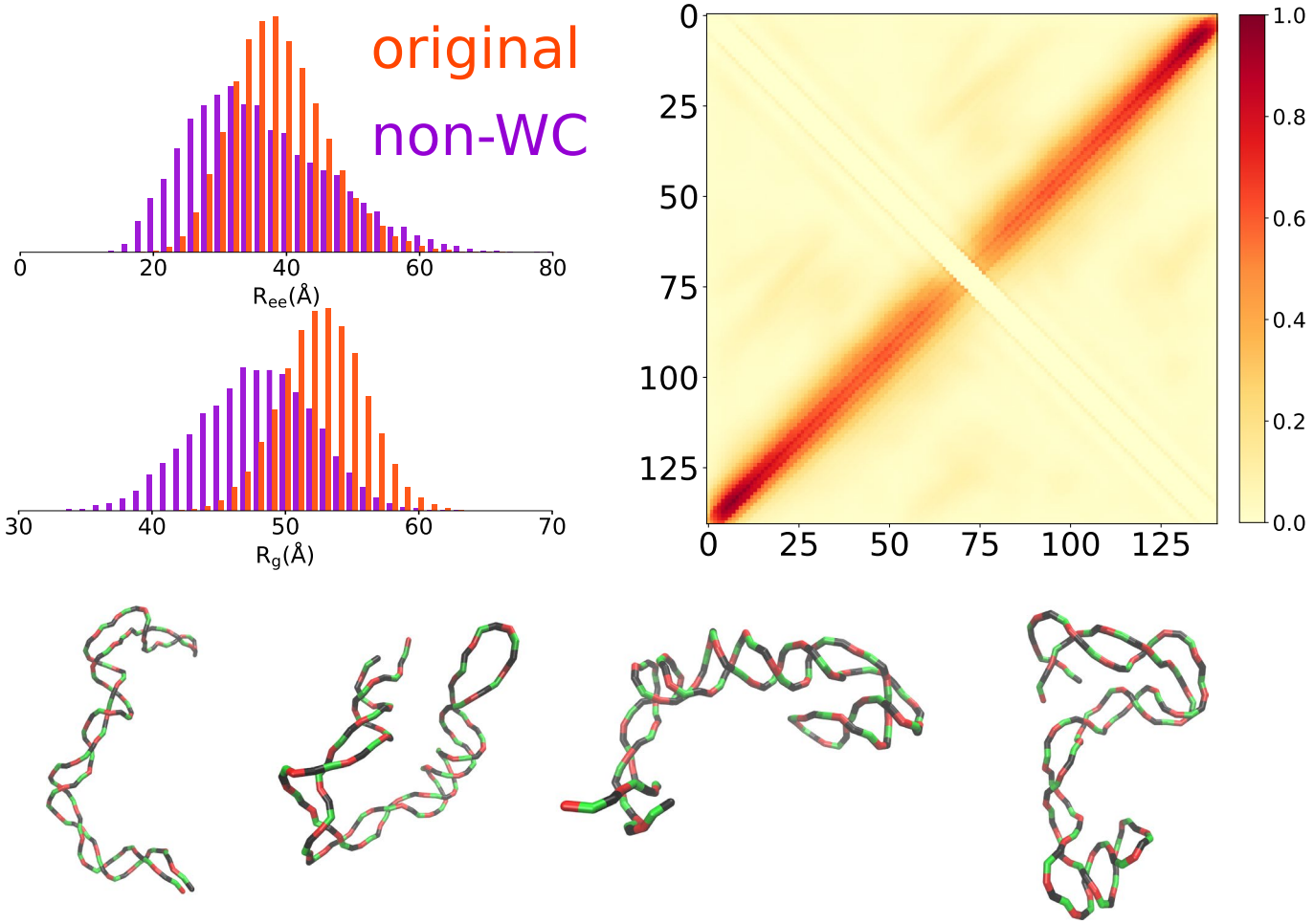


Extended Data Fig. 3 | See next page for caption.

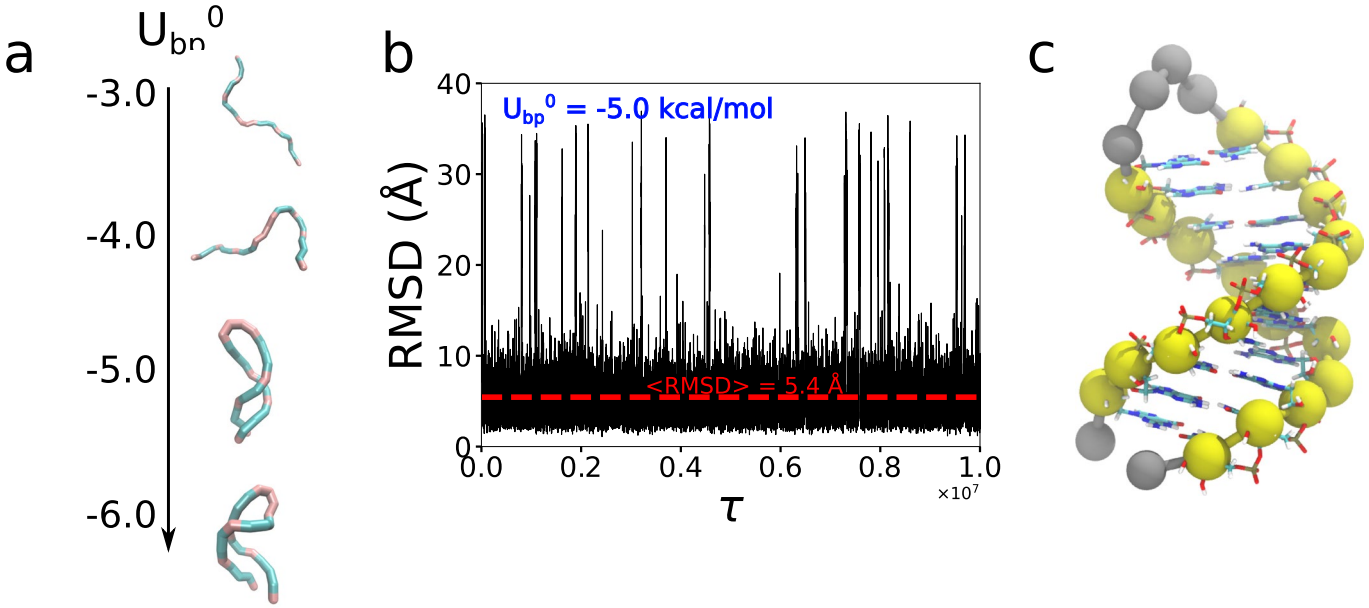
Extended Data Fig. 3 | Sequence of events in early droplet formation extracted from the simulation of $(CAG)_{47}$. Eleven RNA chains were chosen and coloured to see how individual chains form oligomers and grow to a single droplet. All other chains are in grey for clarity. Each panel has a label indicating the simulation time. (a) At the earliest times, all the eleven chains are monomers with no interactions between them. (b) Two chains merge to form a dimer. (c) The dimer captures another chain and becomes a trimer. There is another dimer that is formed around the same time. (d) The two oligomers further grow to a tetramer and trimer, respectively, by interacting with another chain. (e) The tetramer and trimer coalesce into a heptamer. There are still four other chains in the monomer form. (f) One of the remaining monomers joins the oligomer making it an octamer. (g) Two of the remaining monomers form a dimer. (h) It takes some time to the next event (~ $4 \times 10^6 \tau$ from (g) to (h)). (i) The octamer eventually captures the last monomer and becomes a nonamer. (j) The nonamer and the dimer finally coalesce into an 11-mer. The sequence of events is complicated, and is different for different chains.



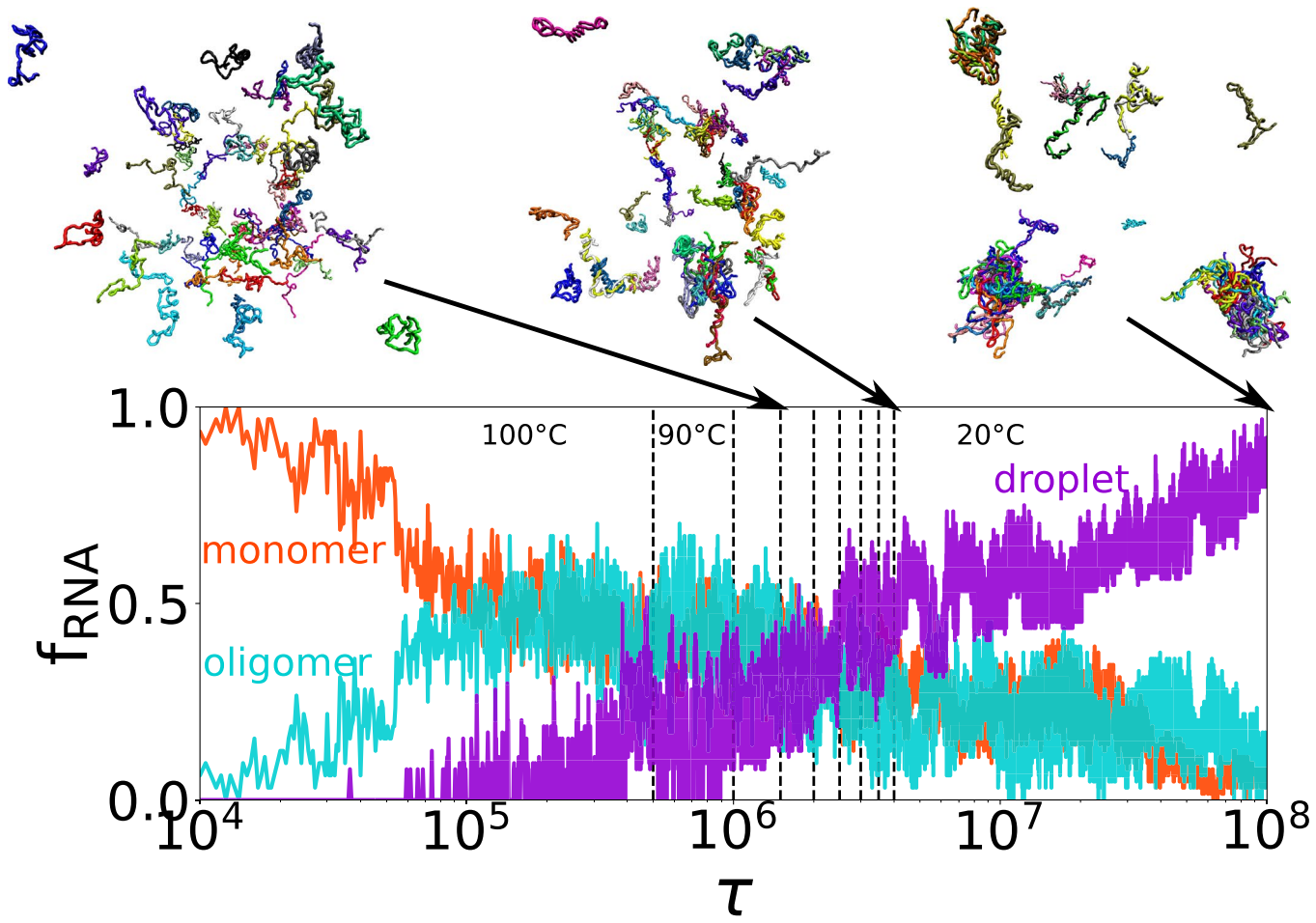
Extended Data Fig. 4 | Simulations for the scrambled sequence. a, Comparison of fraction of RNA chains inside the droplets for $(CAG)_{47}$ and the scrambled sequence at three different concentrations. Snapshots near the end of the simulations for the two sequences are shown. **b**, Droplet size evolution for the scrambled sequence (top) vs. $(CAG)_{47}$ (bottom). Each horizontal line corresponds to a specific droplet in the system. The size is denoted by the colour (colour scale is on the right). **c**, Concentrations of the two phases for the scrambled sequence.



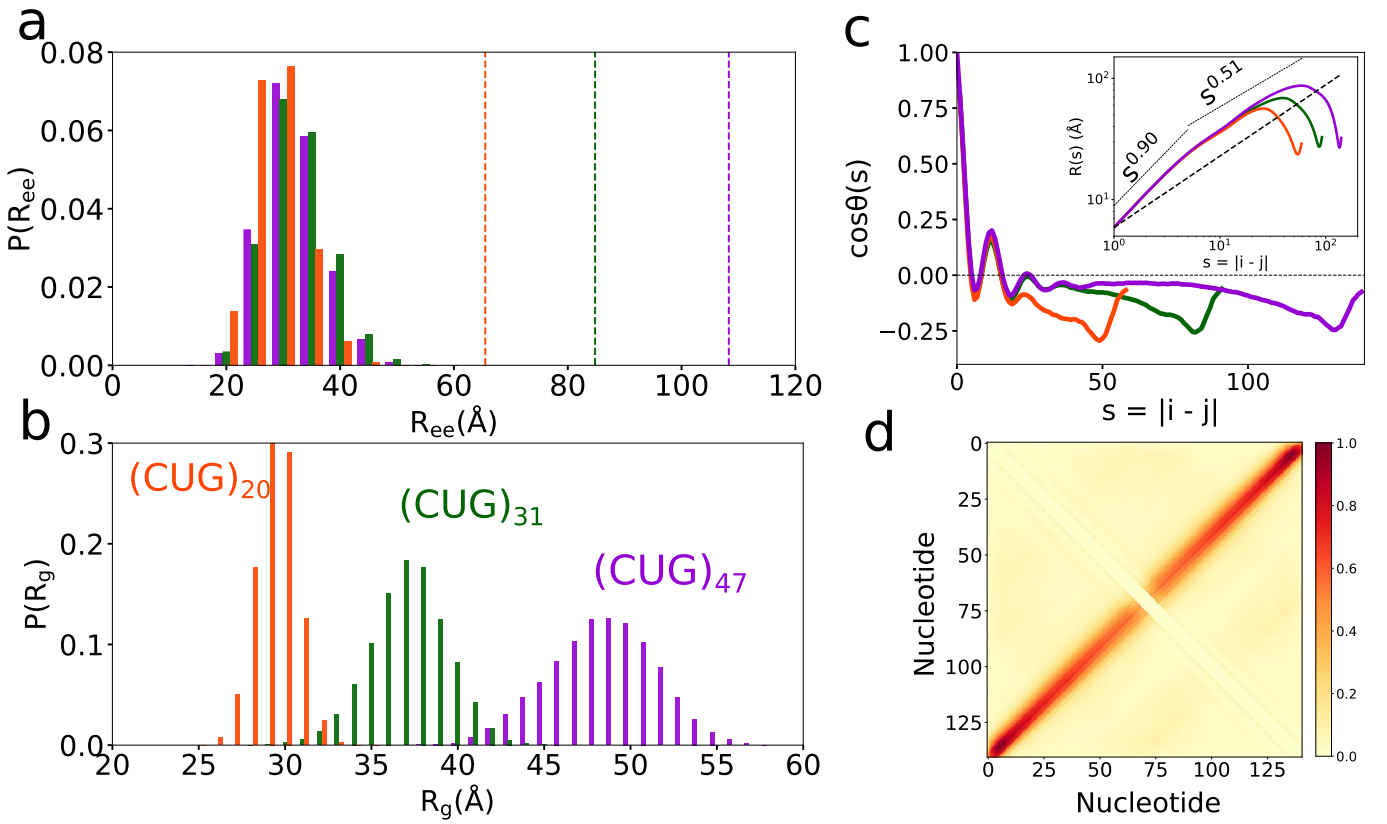
Extended Data Fig. 5 | Effect of non-canonical bps. Simulations for an isolated $(CAG)_{47}$ monomer where non-canonical bps are allowed (purple), compared with the original model where there are only WC bps (orange). Shown on the left are histograms of the end-to-end distance R_{ee} (top) and radius of gyration R_g (bottom). An intramolecular contact map is shown on the right. Some representative snapshots from the simulations are shown at the bottom.



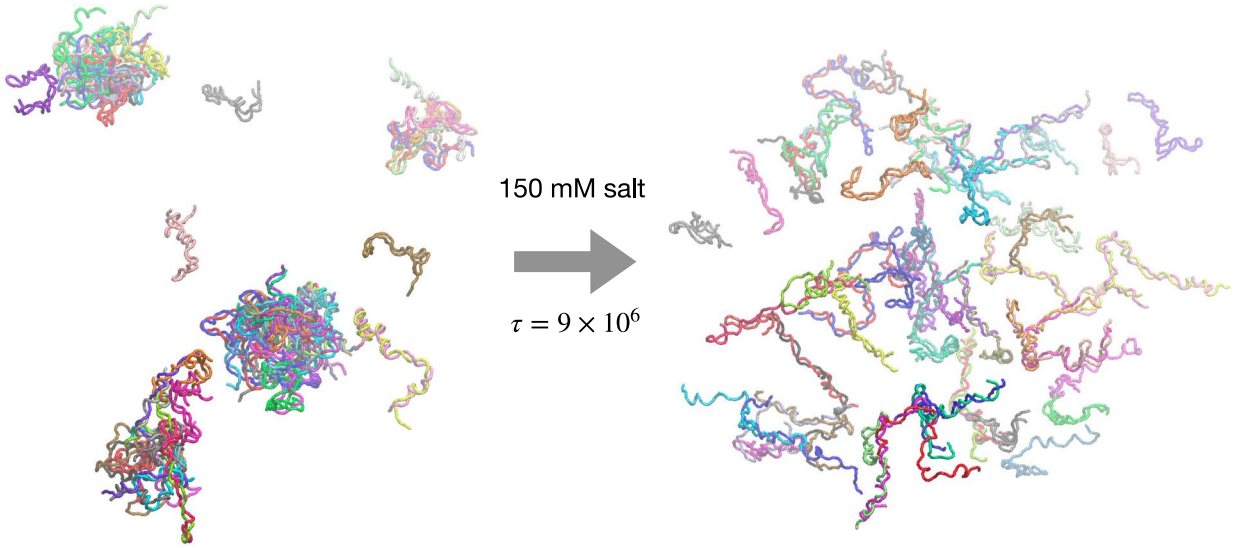
Extended Data Fig. 6 | Calibration of the bp interaction strength U_{bp}^o. **a**, Structural dependence of a small CAG repeat sequence (AGGCAGCAGCCAA AAGGCAGCAGCCA) on U_{bp}^o. The sequence we chose to calibrate U_{bp}^o is almost identical to the X-ray structure (PDB 3NJ6) (ref. [3]), except with the addition of an AAAA tetraloop and two terminal A nucleotides. The sequence adopts extended conformations for small U_{bp}^o (U_{bp}^o < 4.5 kcal mol^{-1}), and folds into hairpin conformations in the bp interaction range, 4.5 < U_{bp}^o < 6.0 kcal mol^{-1}. We set $U_{bp}^o = -5.0$ kcal mol^{-1}. **b**, Root mean squared deviation (RMSD) between the simulations and the X-ray crystal structure, shown for $U_{bp}^o = -5.0$ kcal mol^{-1}. 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).



Extended Data Fig. 7 | Condensate formation does not depend on the cooling rate. Fraction of chains in droplets or existing as oligomers/monomers. The vertical dashed lines indicate when the temperature is lowered (from 100°C to 20°C). Snapshots from left to right correspond to, respectively, the end of 80°C, the end of the cooling period and the final state of the simulation.



Extended Data Fig. 8 | Structures of isolated $(CUG)_n$ monomers. Same as Extended Data Fig. 2, but for $(CUG)_n$ monomers. In addition to WC G-C base pairs, Wobble base pairs between G-U could also form. **a**, Distribution of the end-to-end distance R_{ee} and **b**, radius of gyration R_g of $(CUG)_n$ 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*. **c**, Bond–bond orientational correlation function cos θ (s) as a function of the sequence distance *s*. 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 abrupt drop in *R*(*s*) because the two ends strongly interact with each other, thus bringing them to proximity. **d**, Contact map for $(CUG)_{47}$ shows that the majority of interactions occur along the anti-diagonal, indicating the formation of hairpin structures.



Extended Data Fig. 9 | Dissociation of RNA droplets at 150 mM NaCl. The simulations were started from the final configuration obtained in the droplet simulation of 200μM of $(CAG)_{47}$ (left). The repulsive electrostatic interactions between RNA nucleotides (due to the incomplete neutralization of phosphate charges) lead to the disassembly of the droplets, leaving only monomers and small oligomers (right).