Check for updates

OPEN

Sequence-dependent surface condensation of a pioneer transcription factor on DNA

Jose A. Morin [1,2,6], Sina Wittmann [1,6], Sandeep Choubey[1,3,6], Adam Klosin[1], Stefan Golfier[1,3], Anthony A. Hyman [1,4 ✉], Frank Jülicher [3,4,5 ✉] and Stephan W. Grill [1,2,4 ✉]

Biomolecular condensates are dense assemblies of proteins that form distinct biochemical compartments without being surrounded by a membrane. Some, such as P granules and stress granules, behave as droplets and contain many millions of molecules. Others, such as transcriptional condensates that form on the surface of DNA, are small and contain thousands of molecules. The physics behind the formation of small condensates on DNA surfaces is still under discussion. Here we investigate the nature of transcription factor condensates using the pioneer transcription factor Krüppel-like factor 4 (Klf4). We show that Klf4 can phase separate on its own at high concentrations, but at low concentrations, Klf4 only forms condensates on DNA. Using optical tweezers, we demonstrate that these Klf4 condensates form on DNA as a type of surface condensation. This surface condensation involves a switch-like transition from a thin adsorbed layer to a thick condensed layer, which shows hallmarks of a prewetting transition. The localization of condensates on DNA correlates with sequence, suggesting that the condensate formation of Klf4 on DNA is a sequence-dependent form of surface condensation. Prewetting together with sequence specificity can explain the size and position control of surface condensates. We speculate that a prewetting transition of pioneer transcription factors on DNA underlies the formation and positioning of transcriptional condensates and provides robustness to transcriptional regulation.

Recent works suggest that the regulation of gene expression involves the formation of biomolecular condensates on DNA[1–12]. Condensation from solution is an attractive concept to explain the spatial and temporal organization of transcription and the high local density of proteins at the transcription site. This concept invokes the collective behaviours of many molecules that emerge from their interactions, such as phase separation. However, the biophysics of the collective properties of transcription factors binding to DNA remain unresolved. What is the physical nature of condensates on DNA? What are the collective properties of their molecular components and how are they guided by DNA sequence[11,13–16]? Well-developed concepts from soft matter physics, such as wetting and prewetting[17–19], provide a powerful framework to understand the relationship between droplet formation in bulk solution and condensation on surfaces.

Here we use optical tweezers to directly observe the condensation of the pioneer transcription factor[20] Krüppel-like factor 4 (Klf4) on DNA in vitro. We demonstrate that Klf4 forms sequence-dependent liquid-like condensates that are enabled by interaction with the DNA surface. This sets their typical size and allows them to form below the saturation concentration for phase separation. By combining experiments with theory, we show that these condensates form via a switch-like transition similar to prewetting, a precursor to wetting that occurs below the saturation concentration for bulk-phase separation[17–19]. This transition amplifies the sequence specificity of Klf4 binding to DNA. Polymer-surface-mediated condensation reconciles several observations that were previously thought to be at odds with the idea of phase separation as an organizing principle in the nucleus.

The human pioneer transcription factor Klf4, one of the Yamanaka factors, is a driver of differentiation, cell growth and proliferation[21,22]. Klf4 has a domain organization typical for transcription factors: an activation domain predicted to be disordered and a structured DNA-binding region[23]. Human Klf4 with a C-terminal green fluorescent protein (GFP) tag purified from insect cells binds to DNA oligonucleotides in a sequence-specific manner[24–27] (Fig. 1a,b and Extended Data Fig. 1). In the absence of DNA, Klf4 forms liquid-like condensates at physiological salt and pH above a concentration of ~1.0 μM, which is above the estimated nuclear concentration of Klf4[28] (Fig. 1c,d, Supplementary Methods and Extended Data Figs. 1b and 2). Untagged Klf4 behaves in a similar way (Extended Data Fig. 2c–e). The addition of λ-DNA triggered the formation of foci below the saturation concentration (C_{SAT}) (Fig. 1e), confirming previous observations that DNA can trigger the foci formation of transcription factors[11].

To further examine the behaviour of Klf4 on DNA, we used dual-trap optical tweezers with confocal microscopy to hold a linearized λ-DNA molecule stretched between two polystyrene beads (Fig. 1f)[29]. We observed many Klf4 foci on the DNA molecule at a Klf4 concentration of 115 nM, which varied with regard to the amount of Klf4 they contained (Fig. 1g, Extended Data Figs. 3 and 4a–d and Supplementary Video 1). Notably, even at concentrations closer to C_{SAT} (250 nM), they grew with time until reaching a finite size with an average of approximately 800 molecules per cluster (Fig. 1h). Furthermore, foci can fuse, they can recover after photobleaching and their position can fluctuate on DNA (Fig. 1i, Extended Data Fig. 5 and Supplementary Video 2). Importantly, because condensates form on DNA well below C_{SAT} (Extended Data Figs. 2 and 3k), DNA does not serve as a classic nucleator for the formation of bulk-phase droplets of the kind depicted in Fig. 1c. This is because after nucleation, bulk-phase droplets can only grow if the solution remains above the saturation concentration[11,30,31]
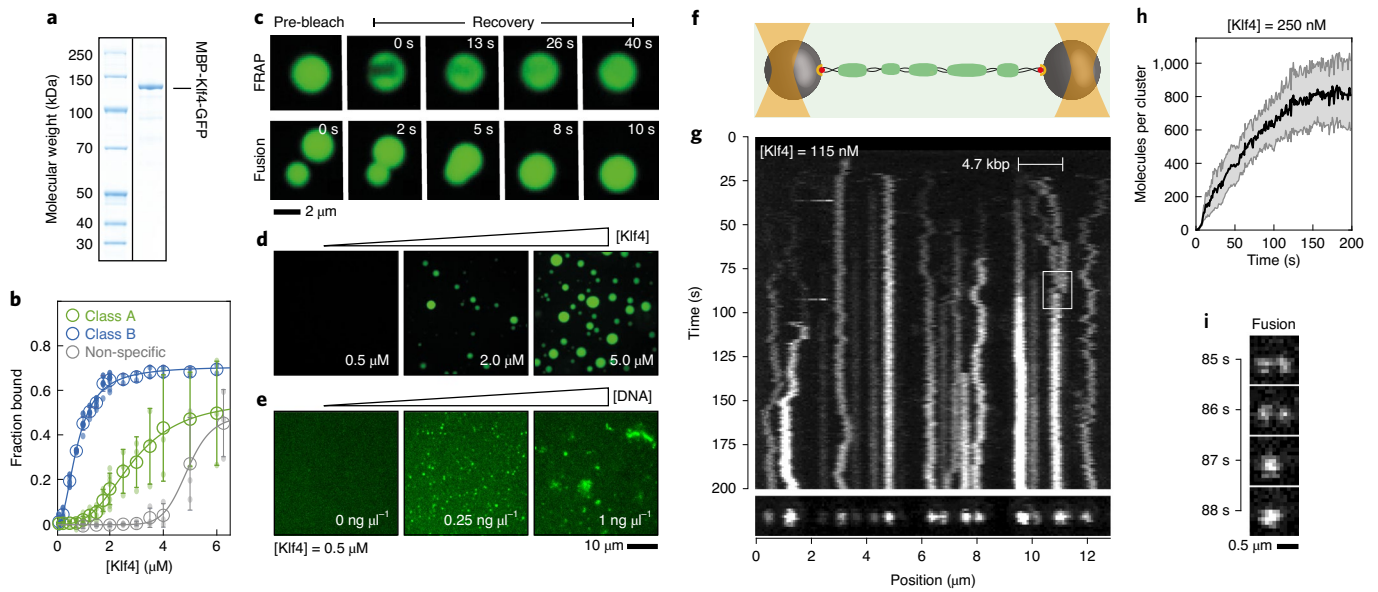
Fig. 1 | DNA drives condensation of Klf4 at concentrations below the saturation concentration for liquid–liquid phase separation. a, Sodium dodecyl sulphate gel showing recombinantly expressed and purified MBP-Klf4-GFP. **b**, Test of the affinity of Klf4-GFP to short dsDNA oligonucleotides using EMSA (Extended Data Fig. 1 and Supplementary Tables 1 and 2). Green, $N=3$: oligonucleotide containing Class B sequence; blue, $N \geq 3$: oligonucleotide containing Class A sequence; grey, $N=3$: oligonucleotide without reported Klf4 consensus sequence. Large circles, mean values; error bars, standard deviation; dots, individual experiments. **c,d**, Bulk-phase separation assay of Klf4-GFP reveals droplet formation above a saturation concentration of ~1.0 μM (Extended Data Fig. 2a). Confocal microscopy images of Klf4-GFP FRAP (Extended Data Fig. 2b) and droplet fusion. The droplets retain their liquid nature for about 40 min (Extended Data Fig. 6b,c). **e**, Addition of λ-DNA to 500 nM Klf4-GFP triggers foci formation. **f**, Schematic of the optical tweezers assay: a single λ-DNA molecule (black) is held between two optically trapped (orange cones) beads via biotin–streptavidin interactions (red and orange) at tension of ~8 pN (Extended Data Fig. 3f,g). The exposure of DNA to a solution of Klf4-GFP (light green) triggers foci formation (dark green). **g**, Top, representative kymograph revealing Klf4-GFP foci formation and dynamics. White horizontal bar, foci displacement on DNA (Extended Data Fig. 5f–h). Bottom, confocal image 200 s after exposure of DNA to Klf4-GFP. **h**, Average number of Klf4-GFP molecules per focus saturates over time. Black line, mean; grey shading, standard error of the mean at 95% confidence (20 foci from 13 experiments). **i**, Foci fusion observed in the indicated region in the kymograph (white box in **g**) (Extended Data Fig. 5a–c and Supplementary Video 1).

(Extended Data Figs. 3i and 6). This suggests that at concentrations below C_{SAT}, a mechanism is at play that is qualitatively different from the standard picture of phase separation in a bulk solution.

To further study the physical nature of Klf4 foci on DNA, we investigated their dependence on protein concentration. Figure 2a shows the representative fluorescent images and corresponding traces of Klf4 intensities at different concentrations, recorded 200 s after the Klf4 solution was introduced to the observation chamber. The number and intensity of Klf4 foci increased with the concentration (Fig. 2a). A probability density histogram of pixel intensities at concentrations above 210 nM reveals a bimodal distribution, indicative of two distinct populations of Klf4 foci (Fig. 2b and Extended Data Fig. 7). The peak of the histogram at a low intensity characterizes Klf4 regions with, on average, less than one molecule per binding site (corresponding to a 10 bp footprint of the Klf4 zinc fingers[32]; Extended Data Fig. 4a,b). We refer to this mode of association as the adsorbed state. The peak at high intensity encompasses Klf4 regions that contain foci with several hundreds and up to a few thousand Klf4 molecules (corresponding to ~2–10 molecular layers of Klf4; Extended Data Fig. 4a,d). We refer to this mode of association as the condensed state. Next, we analysed the intensity histogram as a function of time (Fig. 2c). At low concentrations (below 80 nM), the histogram rapidly forms a peak at intensities that correspond to the adsorbed state that persists over time (Fig. 2c, top). In contrast, at higher concentrations (above 210 nM), the histogram reveals a bifurcation: it first rapidly forms a peak at intensities that correspond to the adsorbed state that subsequently decreases in amplitude, whereas a second peak emerges and quickly moves to higher

intensities that corresponds to the condensed state (Fig. 2c, bottom, Extended Data Fig. 4g and Supplementary Video 3). This bifurcation reveals a switch-like transition from an adsorbed state to a condensed state via a bimodal intensity distribution. The occurrence of these two states and the bimodal intensity distribution depends on the concentration. It is noteworthy that the fraction of DNA that is occupied by the condensed state 200 s after exposure of Klf4 to DNA suddenly increases at a concentration C_{PW} of 86 ± 5 nM (Fig. 2d). We conclude that in our experimental system, Klf4 condensates on DNA are formed in a two-step manner: at low concentrations, the protein merely adsorbs to DNA. At sufficiently high Klf4 concentrations, adsorbed proteins switch to a thick condensate.

How does the DNA surface facilitate the formation of Klf4 condensates? Our data indicate that the formation of Klf4 condensates on DNA is a wetting phenomenon, best described as a process of surface condensation known as prewetting[17–19,33]. Such prewetting transitions occur at a concentration denoted as C_{PW}, which is below the saturation concentration C_{SAT} in the bulk (Fig. 2e,f) and can be understood as follows: Klf4 has an affinity for the DNA surface; hence, the concentration of Klf4 tends to increase in the vicinity of DNA. Therefore, local condensation is facilitated by the surface. However, it cannot extend far away from it, because condensation is not possible in the bulk. Put differently, protein–protein interactions are expected to mediate condensate formation, but the condensate is only stable because of DNA–protein interactions that confine it to the vicinity of the surface.

We next set out to investigate the role of DNA sequence in Klf4 surface condensation. The average intensity profile along the DNA
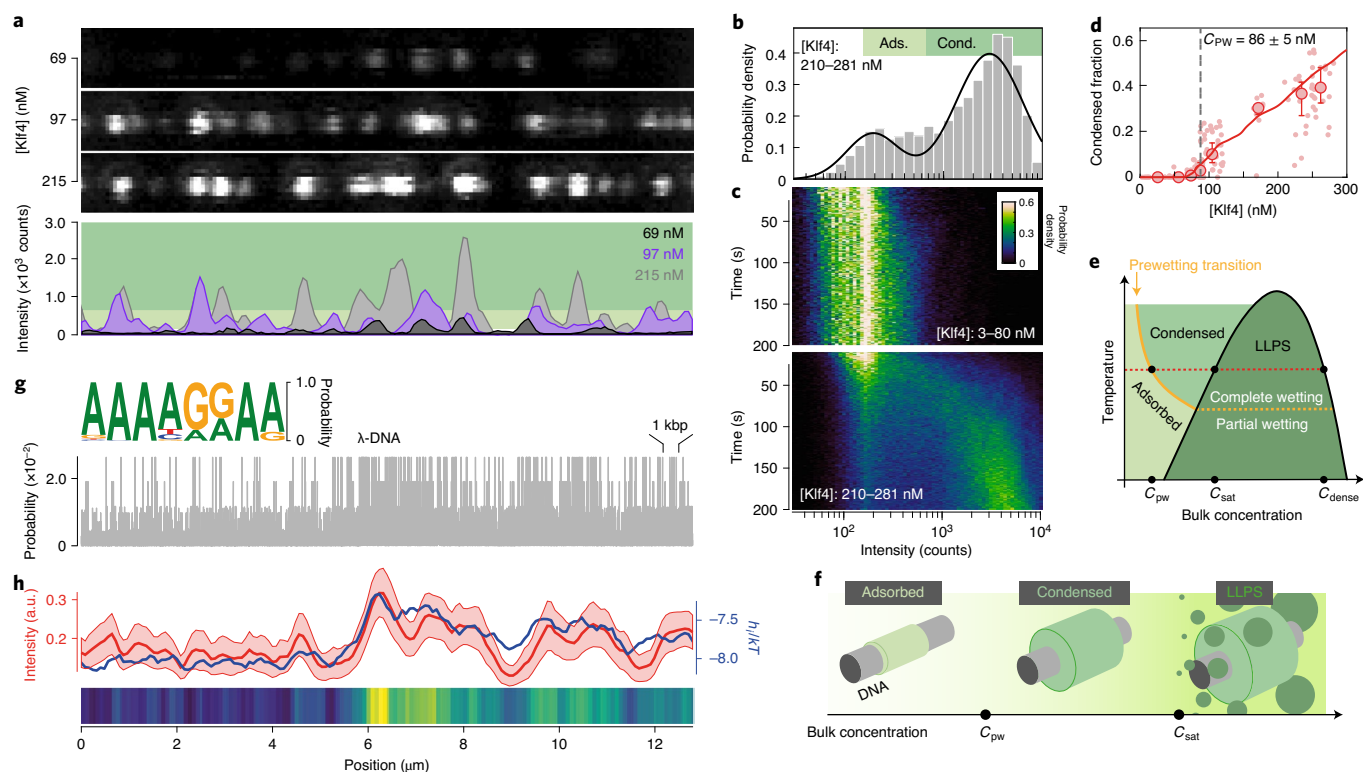
Fig. 2 | Klf4 foci on λ-DNA switch from an adsorbed state to a condensed state via a prewetting-like transition at positions predicted by a Klf4 position weight matrix. a, Confocal images (top) and intensity profiles (bottom) of Klf4-GFP on λ-DNA. Light and dark green, adsorbed and condensed intensity ranges. **b**, Probability density of the logarithm of intensities; $N = 37$ experiments. Black line, fit to the sum of two normal distributions. Intersection (658.5 counts) defines the threshold between adsorbed (light green, Ads.) and condensed (dark green, Cond.) states (Methods and Extended Data Fig. 7). **c**, Same as **b** as a function of time. Top: $N = 60$ (Extended Data Fig. 4g and Supplementary Video 3). **d**, Fraction of DNA occupied by the condensed state. Light dots, individual experiments; red circles, binned medians; error bars, 95% confidence interval (bootstrapping); red line, fit to the heterogeneous two-state model. **e**, Phase diagram for a binary fluid in the presence of a surface[17,54]. In the coexistence regime (dark green, liquid–liquid phase separation (LLPS)) and close to the surface, the dense phase transits from partial to complete wetting when the system crosses a characteristic temperature (yellow dashed line). This first-order transition extends into the single-phase region through the prewetting line (solid yellow line). **f**, Crossing the prewetting line (at C_{PW}) leads to condensation from a thin adsorbed layer (left and middle). Above the saturation concentration C_{SAT}, liquid droplets spontaneously appear in the bulk (right). **g**, Consensus recognition motif of Klf4 in vivo[26] and probability of binding to a given sequence along λ-DNA relative to the consensus motif (Methods and Extended Data Fig. 8). **h**, Blue and colour map, binding energy landscape h_i (kT is the unit of energy, where k is the Boltzmann constant, and T is the temperature), coarse grained over 1 kb (Methods and Extended Data Fig. 9a). Red, average Klf4 intensity along λ-DNA ([Klf4-GFP]: 8–281 nM, $N = 79$). Shaded area, standard error of the mean at 95% confidence.

across all Klf4 concentrations tested reveals enrichment at preferred locations (Fig. 2h). Several Klf4-binding motifs have been reported from in vivo and in vitro studies[24–27]. We chose five of these for further investigation, and used the position weight matrix of these recognition motifs to infer the binding energy landscapes[34] for Klf4 on λ-DNA (Fig. 2g, Extended Data Fig. 8, Supplementary Table 1 and Methods). We find that the measured profile positively correlates with the energy landscapes inferred from four of the five motifs[24–27] (Fig. 2h and Extended Data Fig. 8a–c), with Pearson correlation coefficients of approximately 0.74. We refer to these as class A motifs. The Klf4 intensity profile did not positively correlate with the energy landscape inferred from the fifth motif, which we refer to as a class B motif[24] (Extended Data Fig. 8d). This shows that the position weight matrices of class A, but not class B, motifs provide an accurate parameterization of the binding energy landscape of Klf4 on λ-DNA. Interestingly, electrophoretic mobility shift assay (EMSA) analysis reveals that an oligonucleotide representing the class B consensus motif binds with about 1.3 kT stronger affinity to Klf4 than one representing a class A consensus motif (Extended Data Fig. 1c–f and Supplementary Table 2). Even though Klf4 has higher affinity for class B motifs, these are less well represented on

λ-DNA than class A motifs, explaining the observed binding pattern (Fig. 1b and Extended Data Fig. 8e,f). These results show that in our experiments, the localization of Klf4 condensates on DNA is guided by the underlying DNA sequence. Future work, both in vivo and in vitro, will be required to provide a full parameterization of the sequence-dependent energy landscape of the interaction of Klf4 with DNA.

The data so far indicate that Klf4 condensation on DNA corresponds to a prewetting phenomenon on a heterogeneous substrate, where the heterogeneity is provided by the DNA sequence. Position weight matrices reflect the binding preference of individual molecules to short DNA recognition motifs. However, the condensation phenomenon seen in our experiments is the result of the collective behaviour of many molecules. How can we reconcile single-molecule binding at the length scale of a few base pairs with the sequence dependence of condensation at larger length scales as seen in our experiments (Fig. 2g,h)? We developed a simplified model that considers transitions between a thin adsorbed state and thick condensed state, modulated by a binding energy landscape specified by the DNA sequence (Fig. 3a and Supplementary Note). We represent the stretched DNA polymer by a one-dimensional set
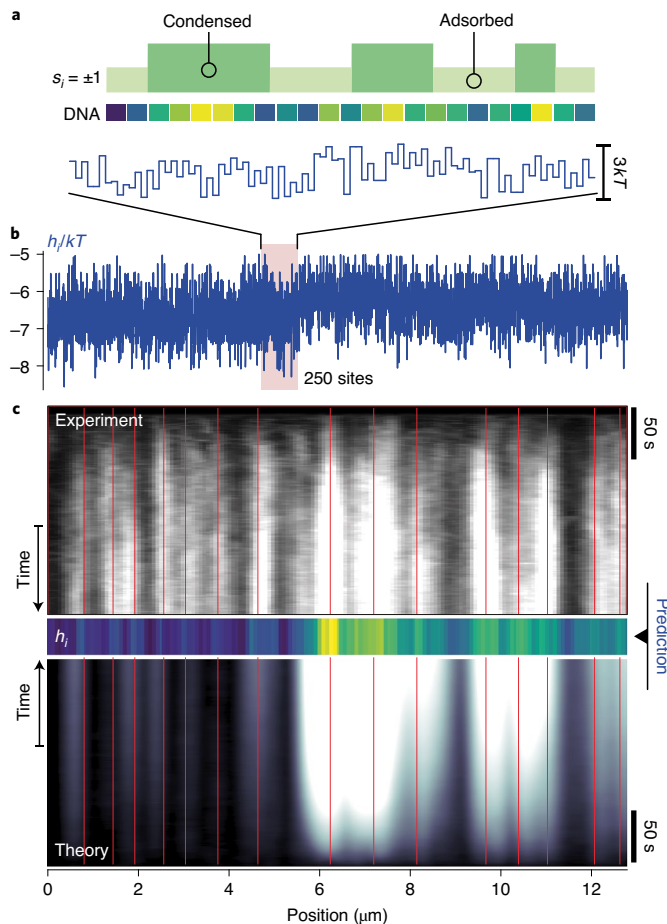
Fig. 3 | Two-state heterogeneous model of prewetting on DNA captures the sequence-dependent switching of Klf4 to a condensed state.
a, Schematic of the two-state heterogeneous model of prewetting (Supplementary Note). The DNA is considered as a one-dimensional lattice of sites that can be in either one of the two states, adsorbed state ($s_i = -1$) or condensed state ($s_i = +1$). **b**, DNA sites have an inhomogeneous propensity for condensation (denoted by h_i; Supplementary Note) given by the binding energy landscape inferred from the position weight matrix (Fig. 2g and Extended Data Fig. 9a, bottom). **c**, Average experimental kymograph ([Klf4-GFP]: 105–281 nM, $N = 34$; top) shows a spatial localization pattern consistent with the coarse-grained binding energy landscape (middle, colour map representation of h_i; Fig. 2h). The average kymograph obtained from the model ($N = 400$; bottom) captures this pattern. The spatial dimension of the model kymograph was convolved with a Gaussian function with a full-width at half-maximum corresponding to that of the point spread function of the microscopy system (Extended Data Fig. 3c). The unconvolved average kymograph and individual kymographs are shown in Extended Data Fig. 9f,g.

of N discrete sites at which Klf4 association can be in either a thin adsorbed state ($s_i = -1$) or a thick condensed state ($s_i = +1$), where $i = 1, ..., N$ (Fig. 3a). This two-state model is formally equivalent to a heterogeneous Ising model[35,36]. Each site i corresponds to a putative binding site of Klf4 corresponding to ten base pairs[32]. At low bulk concentrations, the thin adsorbed state is thermodynamically favoured, whereas at sufficiently high concentrations, the adsorbed molecules collectively switch to form a thick condensed state. This balance is captured by energy h, which is proportional to the bulk chemical potential of Klf4. We capture the DNA sequence by introducing a site-dependent energy bias (h_i for site i) determined using

the position weight matrix corresponding to a binding motif of class A (Fig. 3b and Extended Data Fig. 9a, bottom). The free energy of condensation in the model is given by

$$E = -J \sum_{\langle i,j \rangle} s_i s_j - \sum_i (h + h_i) s_i,$$

where the first sum is over pairs $\langle ij \rangle$ of adjacent sites i and j on the DNA (every pair is counted once). Further, J is an energetic cost related to interfacial tensions. Numerical solutions for the time dependence of condensation as well as for the condensation patterns in the steady state show that the model captures key features seen in experiments: (1) the formation of condensates that coexist with regions that are in the adsorbed state (Figs. 2d and 3c); (2) the dependence of the condensed fraction on protein bulk concentration (Fig. 2d); (3) the sequence dependence of the average spatial condensation pattern in the steady state (Fig. 3c and Extended Data Fig. 9c,f,g); and (4) the time dependence of condensate formation as revealed in kymographs (Fig. 3c and Extended Data Fig. 9f,g). Notably, our model connects the molecular scale with the emerging condensation patterns that involve many molecules, and can be used to predict the condensation pattern for a given sequence of DNA (Fig. 3b,c).

To further analyse the interplay of surface condensation and sequence, we fused a maltose-binding protein (MBP) tag to the disordered N-terminus of Klf4. For this variant, which has an unchanged DNA-binding domain, no bulk-phase separation was observed (data not shown). Notably, MBP-Klf4 forms a thin adsorption layer on λ-DNA following a Hill–Langmuir adsorption isotherm[37], which saturates at a density of less than one molecule per binding site (Fig. 4b, top, Extended Data Fig. 4e,f and Supplementary Note). We next removed the MBP tag from Klf4 after DNA binding (Fig. 4a). Strikingly, the adsorbed layer rapidly rearranged into several condensed foci that localized to positions predicted from the sequence (Fig. 4b,c, Extended Data Fig. 10a–f and Supplementary Video 4). These results show that the properties that drive bulk-phase separation also enable the formation of Klf4 condensates on DNA in a sequence-dependent manner.

We analysed the correlation between the protein localization pattern and the underlying DNA sequence as a function of bulk protein concentration (Fig. 4e and Extended Data Fig. 10g,h). For MBP-Klf4, the correlation coefficient initially increases until it reaches a maximum at ~70 nM ($\rho = 0.66$), followed by a sharp loss of correlation at higher MBP-Klf4 concentrations. This illustrates that the sequence sensitivity of protein localization patterns depends on the protein concentration, and is lost at concentrations beyond the typical binding constant, as expected from the Hill–Langmuir binding kinetics[38]. In the case of Klf4 without the MBP tag, the correlation initially shows a similar increase as a function of concentration ($\rho = 0.76$ at 83 nM), but here the correlation remains high at higher protein concentrations. This reveals that above a certain concentration, the pattern of sequence-dependent localization of condensates is insensitive to bulk protein concentration. We conclude that in contrast to single-molecule binding, surface condensation enables a large dynamic range of bulk concentrations for which the localization pattern remains sequence specific.

Prewetting is an attractive concept for transcription factors because it provides a mechanism for the sequence-dependent formation of small condensates on DNA that are limited in size by interactions with the DNA surface. The transition from an adsorbed layer to a condensed layer serves as a collective amplifier of sequence information that effectively expands the dynamic range at which sequence specificity is achieved. However, how can surface condensation maintain sequence specificity at higher concentrations? Two mechanisms are at play here. First, in the adsorbed state, sequence information is independently used by individual
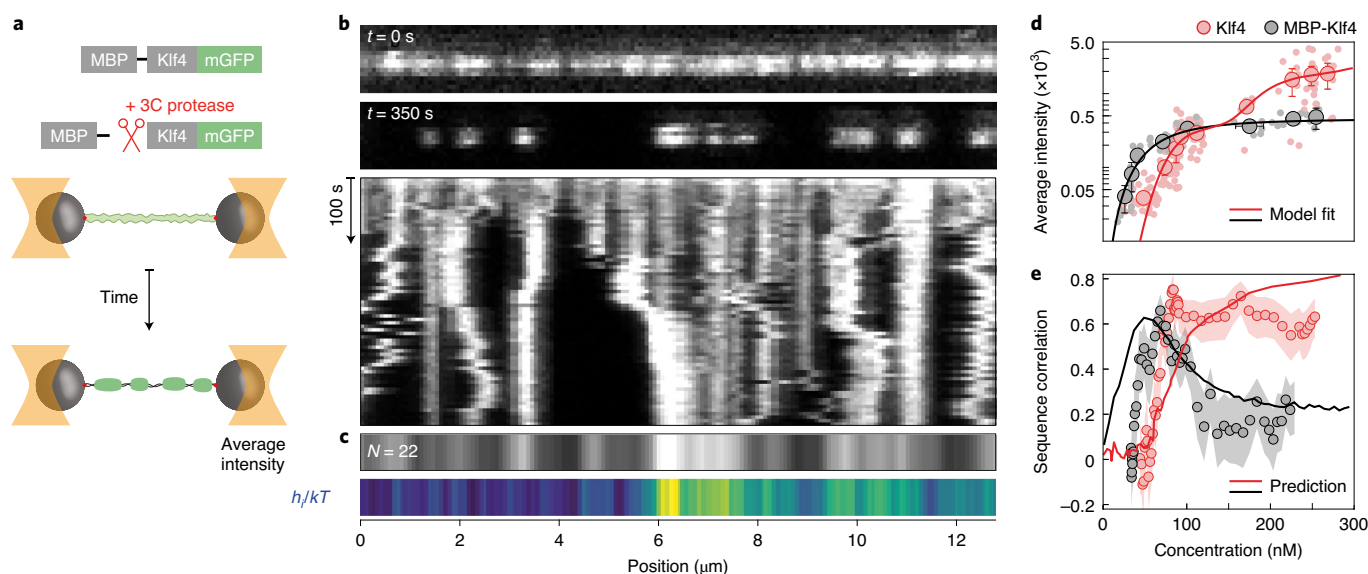
Fig. 4 | Surface condensation enables robust sequence-specific localization of Klf4. a, Schematic depicting the in situ condensation assay: MBP-Klf4-GFP-coated DNA is transferred to a Klf4-free buffer containing 3C protease that can cleave off the MBP tag. **b**, Top, confocal images before (time, $t = 0$ s) and after ($t = 350$ s) transfer to the 3C solution. Bottom, representative kymograph of the 3C-dependent rearrangement of Klf4-GFP on DNA (Extended Data Fig. 10a–f and Supplementary Video 4). **c**, Top, colour map representation of the average intensity profile ($N = 22$ experiments) 350 s after cleavage induction. Bottom, colour map representation of the coarse-grained binding energy landscape (h_i; Fig. 2h). **d**, Average intensity along the DNA for Klf4-GFP (red) and MBP-Klf4-GFP (black) as a function of concentration. Light dots, individual experiments; filled circles, binned medians (Klf4-GFP: 11–34 experiments per bin; MBP-Klf4-GFP: 3–9 experiments per bin); error bars, 95% confidence interval; red line, fit of Klf4-GFP data using the two-state heterogeneous model (Supplementary Note); black line, fit of MBP-Klf4-GFP data to the Hill–Langmuir model (Supplementary Methods). **e**, Correlation between the coarse-grained binding energy landscape (h_i; Fig. 2h,c, bottom) and the average intensity profile (quantified by Pearson's correlation coefficient) for Klf4-GFP (red) and MBP-Klf4-GFP (black) as a function of protein concentration (Extended Data Fig. 10g,h). Circles, correlation resulting from the average intensity profiles selected from a moving window along the concentration axis ($N = 15$ and 12 experiments per bin for Klf4-GFP and MBP-Klf4-GFP, respectively); shaded area, 95% confidence interval (bootstrapping); red line, prediction from the two-state heterogeneous model for condensation using parameter values extracted from the fit in Fig. 2d; black line, prediction from the two-state heterogeneous model for adsorption (Supplementary Note provides the parameter values and Extended Data Fig. 9d,e).

molecules, whereas in the condensed state, sequence information is collectively integrated by the molecules. This is presumably because surface condensation is triggered by a local increase in concentration, which is promoted by the local clustering of binding sites[11] as previously suggested for the formation of transcriptional condensates in the bulk. This might explain how pioneer transcription factors can distinguish recognition sites within enhancers from isolated sites in other regions of the genome[39,40]. Second, when molecules adsorb independently, binding becomes saturated at higher bulk concentrations because each site can only be occupied once[41]. However, a condensate can accommodate a variable number of molecules, even as the concentration increases. Indeed, molecules associated with DNA will be incorporated into existing condensates either directly or via one-dimensional diffusion[34,39,42,43] rather than occupying unfavourable sites. Consequently, a further increase in protein concentration can result in the growth of condensates without altering their localization pattern, rendering the process insensitive to molecular noise[44,45].

Since polymer-surface-mediated condensation leads to the formation of liquid-like compartments, features such as the fusion of transcriptional condensates and recruitment of downstream factors that have been observed previously[7–9,28] can be accounted for here. The limited size of transcription factor condensates provides a possible explanation for the small size of transcriptional foci observed in vivo[8,46,47]. We suggest that polymer-surface-mediated condensation provides a general framework to explain the formation of other nuclear condensates such as heterochromatin or paraspeckles on chromatin or RNA surfaces[48–53].

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/s41567-021-01462-2.

References

1. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169**, 13–23 (2017).
2. Boehning, M. et al. RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat. Struct. Mol. Biol.* **25**, 833–840 (2018).
3. Guo, Y. E. et al. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* **572**, 543–548 (2019).
4. Lu, H. et al. Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **558**, 318–323 (2018).
5. Wei, M.-T. et al. Nucleated transcriptional condensates amplify gene expression. *Nat. Cell Biol.* **22**, 1187–1196 (2020).
6. Basu, S. et al. Unblending of transcriptional condensates in human repeat expansion disease. *Cell* **181**, 1062–1079.e30 (2020).
7. Sabari, B. R. et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, eaar3958 (2018).
8. Cho, W.-K. et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412–415 (2018).
9. Chong, S. et al. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* **361**, eaar2555 (2018).

10. Nair, S. J. et al. Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. *Nat. Struct. Mol. Biol.* **26**, 193–203 (2019).

11. Shrinivas, K. et al. Enhancer features that drive formation of transcriptional condensates. *Mol. Cell* **75**, 549–561.e7 (2019).

12. Boija, A. et al. Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175**, 1842–1855.e16 (2018).

13. Weber, S. C, A. P. Evidence for and against liquid-liquid phase separation in the nucleus. *Non-Coding RNA* **5**, 50 (2019).

14. McSwiggen, D. T. et al. Evidence for DNA-mediated nuclear compartmentalization distinct from phase separation. *Elife* **8**, e47098 (2019).

15. Shin, Y. et al. Liquid nuclear condensates mechanically sense and restructure the genome. *Cell* **175**, 1481–1491.e13 (2018).

16. McSwiggen, D. T., Mir, M., Darzacq, X. & Tjian, R. Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Genes Dev.* **33**, 1619–1634 (2019).

17. Cahn, J. W. Critical point wetting. *J. Chem. Phys.* **66**, 3667–3672 (1977).

18. De Gennes, P. G. Wetting: statics and dynamics. *Rev. Mod. Phys.* **57**, 827–863 (1985).

19. Quéré, D. Wetting and roughness. *Annu. Rev. Mater. Res.* **38**, 71–99 (2008).

20. Zaret, K. S. Pioneer transcription factors initiating gene network changes. *Annu. Rev. Genet.* **54**, 367–385 (2020).

21. Ghaleb, A. M. & Yang, V. W. Krüppel-like factor 4 (KLF4): what we currently know. *Gene* **611**, 27–37 (2017).

22. Yamanaka, S., Takahashi, K., Okita, K. & Nakagawa, M. Induction of pluripotent stem cells from fibroblast cultures. *Nat. Protoc.* **2**, 3081–3089 (2007).

23. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).

24. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

25. Shields, J. M. & Yang, V. W. Identification of the DNA sequence that interacts with the gut-enriched Krüppel-like factor. *Nucleic Acids Res.* **26**, 796–802 (1998).

26. Wan, J. et al. Methylated *cis*-regulatory elements mediate KLF4-dependent gene transactivation and cell migration. *Elife* **6**, e20068 (2017).

27. Hu, S. et al. Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell* **139**, 610–622 (2009).

28. Sharma, R. et al. Liquid condensation of reprogramming factor KLF4 with DNA provides a mechanism for chromatin organization. *Nat. Commun.* **12**, 5579 (2021).

29. Candelli, A., Wuite, G. J. L. & Peterman, E. J. G. Combining optical trapping, fluorescence microscopy and micro-fluidics for single molecule studies of DNA–protein interactions. *Phys. Chem. Chem. Phys.* **13**, 7263–7272 (2011).

30. Kashchiev, D. *Nucleation: Basic Theory with Applications* (Butterworth Heinemann, 2000).

31. Pruppacher, H. R., Klett, J. D. & Wang, P. K. Microphysics of clouds and precipitation. *Aerosol Sci. Technol.* **28**, 381–382 (1998).

32. Schuetz, A. et al. The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation. *Cell. Mol. Life Sci.* **68**, 3121–3131 (2011).

33. Rouches, M., Veatch, S. L. & Machta, B. B. Surface densities prewet a near-critical membrane. *Proc. Natl Acad. Sci. USA* **118**, e2103401118 (2021).

34. Phillips, R., Kondev, J., Theriot, J., Garcia, H. & Nigel, O. *Physical Biology of the Cell* (Garland Science, 2012).

35. Imry, Y. & Ma, S. Random-field instability of the ordered state of continuous symmetry. *Phys. Rev. Lett.* **35**, 1399–1401 (1975).

36. Blossey, R., Kinoshita, T. & Dupont-Roc, J. Random-field Ising model for the hysteresis of the prewetting transition on a disordered substrate. *Phys. A: Stat. Mech. Appl.* **248**, 247–272 (1998).

37. Langmuir, I. The adsorption of gases on plane surfaces of glass, mica and platinum. *J. Am. Chem. Soc.* **40**, 1361–1403 (1918).

38. Berg, O. G. & von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–743 (1987).

39. Dror, I., Rohs, R. & Mandel-Gutfreund, Y. How motif environment influences transcription factor search dynamics: finding a needle in a haystack. *BioEssays* **38**, 605–612 (2016).

40. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. *PLoS Genet.* **10**, e1004226 (2014).

41. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annu. Rev. Cell Dev. Biol.* **35**, 357–379 (2019).

42. Elf, J., Li, G.-W. W. & Xie, X. S. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316**, 1191–1194 (2007).

43. Marklund, E. et al. DNA surface exploration and operator bypassing during target search. *Nature* **583**, 858–861 (2020).

44. Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010–2013 (2005).

45. Grah, R., Zoller, B. & Tkačik, G. Nonequilibrium models of optimal enhancer function. *Proc. Natl Acad. Sci. USA* **117**, 31614–31622 (2020).

46. Jackson, D., Hassan, A., Errington, R. & Cook, P. Visualization of focal sites of transcription within human nuclei. *EMBO J.* **12**, 1059–1065 (1993).

47. Pancholi, A. et al. RNA polymerase II clusters form in line with surface condensation on regulatory chromatin. *Mol. Syst. Biol.* **17**, e10272 (2021).

48. Fox, A. H., Nakagawa, S., Hirose, T. & Bond, C. S. Paraspeckles: where long noncoding RNA meets phase separation. *Trends Biochem. Sci.* **43**, 124–135 (2018).

49. Larson, A. G. et al. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* **547**, 236–240 (2017).

50. Gibson, B. A. et al. Organization of chromatin by intrinsic and regulated phase separation. *Cell* **179**, 1–15 (2019).

51. Fei, J. et al. Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J. Cell Sci.* **130**, 4180–4192 (2017).

52. Strom, A. R. et al. Phase separation drives heterochromatin domain formation. *Nature* **547**, 241–245 (2017).

53. Erdel, F. et al. Mouse heterochromatin adopts digital compaction states without showing hallmarks of HP1-driven liquid-liquid phase separation. *Mol. Cell* **78**, 236–249.e7 (2020).

54. Levinson, P., Jouffroy, J. & Brochard, F. Wetting transition for a thin cylinder. *J. Phys. Lett.* **46**, 21–26 (1985).

Methods

Protein expression and purification. Proteins were expressed in Sf9 cells (Expression Systems, 94-001F) for 72 h using the baculovirus system[55]. For all the Klf4 constructs, the preparation was done on ice using pre-cooled solutions. The cells were resuspended in lysis buffer (50 mM bis-tris propane (pH 9.0), 500 mM KCl, 500 mM arginine-HCl, 6.25 μM ZnCl$_2$, 5% glycerol, EDTA (ethylenediaminetetraacetic acid)-free protease inhibitor cocktail set III (Calbiochem) and 0.25 U ml^{-1} benzonase (in-house)) and lysed by sonication. The lysate was cleared by centrifugation for 1 h at 13,000 × g and 4 °C. The supernatant was incubated with amylose resin (NEB) for at least 30 min at 4 °C. After washing with wash buffer I (50 mM bis-tris propane (pH 9.0), 500 mM KCl, 500 mM arginine-HCl and 5% glycerol), the beads were transferred into Econo-Pac gravity columns (Bio-Rad) and washed with wash buffer II (50 mM bis-tris propane (pH 9.0), 1 M KCl, 500 mM arginine-HCl and 5% glycerol) followed by wash buffer I. MBP-Klf4 was eluted using elution buffer (50 mM bis-tris propane (pH 9.0), 500 mM KCl, 500 mM arginine-HCl, 10 mM maltose and 5% glycerol). The eluate was concentrated using Vivaspin 50,000 MWCO concentrators (GE Healthcare or Sartorius) and subjected to size-exclusion chromatography at 4 °C using a Superdex 200 column (GE Healthcare) and size-exclusion chromatography buffer (50 mM bis-tris propane (pH 9.0), 500 mM KCl, 5% glycerol and 1 mM DTT (dithiothreitol)).

After concentrating the sample as described above, the proteins were stored at 4 °C for no longer than 2 weeks. MBP-Klf4 was buffer exchanged using Zeba Spin Desalting Columns (Thermo Scientific) into Klf4 buffer (25 mM tris (pH 7.4), 500 mM KCl, 1 mM DTT and 0.1 mg ml^{-1} BSA (bovine serum albumin)). For MBP-Klf4-GFP (plasmid TH1528) and MBP-Klf4-mCherry (plasmid TH1529), unless stated otherwise, the MBP moiety was cleaved off with 10% (v/v) 3C protease (in-house; 1 U μl^{-1}) for at least 1 h on ice (Extended Data Fig. 1a). For MBP-Klf4-MBP (plasmid TH1696), both MBP tags were cleaved off with 10% v/v 3C protease and 10% (v/v) TEV (tobacco etch virus) protease (both in-house) for at least 2 h on ice (Extended Data Fig. 2c). In both cases, the sample was spun for 10 min at 20,000 × g and 4 °C and the concentration was remeasured using either adsorption at 280 nm or GFP fluorescence.

Phase separation assays. Klf4-GFP was kept on ice and diluted with a pre-cooled solution to prevent premature phase separation at higher temperatures (Extended Data Fig. 2f). The protein was pre-diluted with cold Klf4 buffer to four times the final concentration and then mixed in a ratio of 1:4 with cold dilution buffer (25 mM tris (pH 7.4), 1 mM DTT and 0.1 mg ml^{-1} BSA) to obtain the Klf4 assay buffer (25 mM tris (pH 7.4), 125 mM KCl, 1 mM DTT and 0.1 mg ml^{-1} BSA) in a total volume of 20 μl. For assays containing DNA, the dilution buffer would also contain appropriate amounts of DNA. The samples were mixed by pipetting, and 18 μl was transferred into 384-well medium-binding microplates (Greiner Bio-One). The samples were incubated at room temperature for 20 min before imaging. Samples that contained DNA were additionally spun at 3,200 × g for 2 min. The images were taken using an Andor Eclipse Ti inverted spinning-disc microscope with an Andor iXon 897 electron-multiplying charge-coupled device camera and a UPLSAPO ×40/0.95 numerical aperture (NA) air objective or ×60/1.20 NA water-immersion objective (Nikon). Data from at least three independent experiments were averaged. Data analysis was performed as described elsewhere[56]. For untagged Klf4, differential interference contrast microscopy was done using a Zeiss LSM 880 inverted single-photon point-scanning confocal system utilizing a transmitted-light detector and ×40/1.2 NA C-Apochromat water-immersion objective (Zeiss), which is suitable for differential interference contrast.

Determining the concentration of the dilute phase. The Klf4 samples were set up as those for the phase separation assays, but instead of transfer to a microplate, the samples were incubated for 20 min at room temperature in 1.5 ml Eppendorf tubes. To obtain a standard curve, samples with a final KCl concentration of 500 mM were also prepared and treated in parallel. After incubation, the samples were spun in a temperature-controlled centrifuge at 20,000 × g for 15 min at 21 °C. Here 5 μl supernatant was added to 15 μl Klf4 buffer or, in the case of the control samples, a corresponding buffer, to reach the same final KCl concentration. These samples were transferred to 384-well non-binding microplates (Greiner Bio-One) and imaged with a wide-field fluorescence microscope (DeltaVision Elite, Applied Precision) using ×10/0.4 NA dry objective and a Photometrics electron-multiplying charge-coupled device camera. The median fluorescence values for each field of view were obtained using the Fiji software (https://fiji.sc/). The control samples were used to generate a standard curve that correlates fluorescence intensity with protein concentration. This curve was then used to calculate the original protein concentration in the sample supernatant. To determine C_{SAT}, we fitted a two-component piecewise linear function to the curve of dilute phase concentration versus the total concentration (Extended Data Fig. 2a, left). The number of points considered on each side was varied and the two optimal lines were selected as those having the maximum difference between their slopes. The C_{SAT} value for a given dataset was determined as the intercept of the lowest slope curve (Extended Data Fig. 2a, left). Fitting was done in MATLAB (version R2018b).

EMSAs. Reactions were setup at 4 °C at the indicated final protein concentrations. The Klf4 samples contained 25 mM tris (pH 7.4), 125 mM KCl, 6% glycerol, 1 mM DTT, 0.1 mg ml^{-1} BSA, 7.5 nM Cy5-dsDNA and 37.5 ng poly-d(IC) (poly-(5'-phosphono-3'-deoxy-cytidine compound with 5'-phosphono-2'-deoxy-inosine)). The oligonucleotides used in this study are listed in Supplementary Table 2. The absence of condensed material under these conditions was confirmed by fluorescence microscopy (data not shown). The samples were incubated for 20 min at 4 °C before they were loaded onto a pre-run 4–20% Novex TBE gel (Invitrogen). Electrophoresis was performed at 250 V for 45 min in TBE buffer (89 mM tris, 89 mM boric acid and 2 mM EDTA). The gels were then imaged using a Typhoon FLA 9500 fluorescence imager (GE Healthcare). Band intensities were determined using the Fiji software, and data plotting and fitting was done in MATLAB. The following expression was used to fit the data[57]:

$$f = b + \left[\frac{m - b}{1 + \left(\frac{K_d}{P_t} \right)^n} \right],$$

where P_t is the total protein concentration and K_d is the dissociation constant; m and b are normalization factors for the upper and lower asymptotes of the DNA titration curve, respectively; and n is the Hill coefficient.

Optical tweezers with confocal microscopy. Experiments involving optical tweezers were performed on a Lumicks C-trap instrument with integrated confocal microscopy and microfluidics. Bacteriophage λ-DNA was biotinylated on both ends as described elsewhere[29]. Attachment of the λ-DNA-dCas9 complex to 4.42 μm Spherotech streptavidin-coated polystyrene beads was done using the laminar flow. For all the experiments, the trap position was kept constant to render an average force of 8.22 ± 2.65 pN (Extended Data Fig. 3f).

The protein stock was centrifuged for 10 min at 20,000 × g. The supernatant concentration was measured and diluted in Klf4 assay buffer following a dilution series: the solution containing the maximum concentration of a given series was flushed into the flow cell. After recording for 10–15 experiments, the remaining volume in the syringe was removed, and the protein was diluted and reloaded into the syringe. The flow chamber was flushed before each experiment and sealed during the course of it.

For confocal imaging, a 488 nm laser was used for excitation, with emission detected in the channel with a blue filter (525/25 nm). After a λ-DNA molecule was tethered between the beads, an image of the dCas9-EGFP (enhanced green fluorescent protein) probe was acquired in the buffer channel with 10% excitation intensity (this imaging setting is referred to as high excitation). We then started continuous acquisition with 5% excitation intensity as the beads–DNA system was transferred to the channel of the microfluidics chip containing the protein of interest. The interaction process was monitored for 200 s at a frame rate of ~1 s^{-1} with a low pixel integration time of 0.08 ms (Fig. 1g and Extended Data Fig. 3). After 200 s, an image was acquired using the high-excitation imaging conditions. Analysis of the intensity distributions and quantification of number of molecules (Fig. 2a,b and Extended Data Fig. 4) was done for the high-excitation settings. To determine the number of molecules per cluster over time (Fig. 1h), time series were acquired using a pixel integration time of 1 ms (referred to as low excitation; Extended Data Fig. 3), conditions in which the dCas9-EGFP probe was detectable for the first few frames, before the beads–DNA system reached the protein solution.

For the in situ condensation assay (Fig. 4b and Extended Data Fig. 10), after the binding process was recorded, the beads–DNA system was transferred back to the buffer channel containing either the assay buffer or the assay buffer with 2% (v/v) 3C protease (in-house; 1 U μl^{-1}). This process was recorded for more than 500 s under the low-excitation settings at a frame rate of 0.2 s^{-1}.

Fluorescence recovery after photobleaching (FRAP) experiments were performed as follows: after a binding experiment (low-excitation settings), the chamber was gently flushed. A pre-bleach time series was acquired for 20 s. A smaller region of interest (ROI) for the FRAP experiment was imaged with a high-excitation laser intensity (90% excitation). To capture recovery, a 200-s-long time series was then acquired at a frame rate of ~1 s^{-1} (Extended Data Fig. 5).

Analysis of tweezers data. *Intensity emission per EGFP.* For each experiment, confocal images of the dCas9-sgRNA-λ-DNA complex were acquired under the high-excitation settings. For the time series (low-excitation settings), the dCas9 probe was detectable for the first few frames, before the beads–DNA complex reached the protein solution. To confirm the position of the dCas9 probe, intensity profiles along the DNA were aligned using the beads centre as the reference and flipped when required. To confirm the position of the target sequence, the target locations were superimposed with the average profile (Extended Data Fig. 3). This alignment criteria was then used to analyse all the Klf4 and MBP-Klf4 intensity profiles shown throughout this work. The sequence information was converted to spatial units by taking into account the extension per base pair (xbp = 0.32 nm bp^{-1}) at the average experimental force. The integration of the total intensity in an ROI of 21 pixels × 21 pixels around the detected probe rendered the total number of counts under the given imaging conditions. The probability distribution of integrated counts for several experiments exhibits a multi-mode Gaussian distribution,

consistent with having four sites for dCas9 binding in λ-DNA (Supplementary Table 3). A fit to a Gaussian mixture model rendered the mean and standard deviation of each mode. The emission intensity per EGFP was then calculated as follows:

$$I_{GFP} = \frac{\sum_{j=1}^{N-1} \left(I_{j+1} - I_J \right)}{N - 1},$$

where I_j is the mean of mode j and N is the number of modes (Extended Data Fig. 3 and Supplementary Table 5).

Intensity distributions. The pixel values used in the calculation of intensity distributions were obtained as follows: after background subtraction (to remove the contributions from the protein in solution), the maximum projection intensity profile along the DNA was determined in a region of 20 pixels around the DNA axis (Extended Data Fig. 7). We next filtered the profiles with a spatial mask. In brief, using the 'findpeaks' function in MATLAB, we detected the peaks above a threshold corresponding to the background value of the background subtracted image (Extended Data Fig. 7). Data points in a five-pixel window, along the horizontal direction and centred at the position of each peak, were selected. The window was displaced from left to right and accepted if there was no overlap. From the histograms of the obtained pixel intensity values, we computed the probability density of the logarithm of pixel intensities[58]. The probability density versus logarithm of intensities was fitted to either one- or two-component Gaussian mixture model in the linear scale. To compare the intensity distributions (Fig. 2b) with the intensity distributions over time (Fig. 2c, Extended Data Fig. 4g and Supplementary Video 3), time-series images were multiplied by a factor (13.4 ± 2.9) to compensate for the intensity-value differences between the low- and high-excitation imaging conditions. From the last frame of each time series and the corresponding high-excitation image acquired immediately thereafter, we computed the mean intensity in an ROI of 30 pixels × 100 pixels in the centre of the confocal image. Time-series intensities were multiplied by the ratio of these means.

Classification of pixels into adsorbed or condensed. Pixels were classified into adsorbed or condensed based on their intensity. An intensity above the background and below the layer threshold resulted in a classification as adsorbed, whereas an intensity above the layer threshold was classified as condensed. To determine the background threshold, we extracted the background values (after background subtraction) along a line away from the DNA and pulled together all the experiments corresponding to Klf4-GFP. The probability density of the logarithm of pixel intensities was fitted to a normal probability density function. The background threshold was defined as the mean plus three times the standard deviation of this distribution (Extended Data Fig. 7). To determine the layer threshold, we computed the probability density of the logarithm of pixel intensities along the masked maximum projection profiles pulling together 60 Klf4-GFP experiments recorded at low concentrations ([Klf4]: 3–80 nM). We extracted the mean value of this distribution by fitting the data to a normal probability density function. We next computed the same quantity for 37 experiments recorded at higher concentrations ([Klf4]: 210–281 nM). Here the probability density shows bimodality, and we fitted this distribution to a two-component Gaussian mixture model, constraining the mean of the low-intensity mode to the value obtained at low concentrations. Fitting was done in MATLAB using the 'nonlinear least squares' method and weights of $1/(\text{probability density}) + w$, where $w = 10$ sets the strength of the weights.

Condensed fraction. The condensed fraction (Fig. 2d) was determined for each experiment as the number of pixels falling into the condensed category divided by the length of the considered ROI (161 pixels). For this calculation, we considered the pixels obtained by the masking procedure (as discussed above; Extended Data Fig. 7). Binned medians with error bars (95% confidence interval) were obtained by bootstrapping ('bootci' function in MATLAB) using 10,000 bootstrap samples. Binned medians contain 11–36 individual experiments.

Analysis of Klf4 sequence-specific binding to λ-DNA. To assess the sequence specificity of Klf4 localization on λ-DNA, we used a Klf4 sequence motif reported elsewhere[26]. The reported sequence logo was converted to a position weight matrix using the Logo2PWM tool[59]. A position weight matrix contains one row for each of the four DNA bases and a column for each position of the motif (Fig. 2g, Extended Data Fig. 8 and Supplementary Table 1). The values of the matrix represent the relative frequency to find a certain base at a given position within the motif. Since these matrices are generally derived from the genome-wide in vivo data of protein binding to DNA, they inform us of how likely it is to find a protein bound to a given sequence of bases.

We denote the matrix by M_{nb}^{F}, where M_{nb}^{F} is the relative frequency to find base b at nucleotide position n on the forward strand when the protein is bound to DNA. M_{nb}^{R} is defined analogously for the reverse strand. Here b can be any of the four possible nucleotides, namely, A, T, G or C. The position weight matrix describing the frequency of bases on the complementary strand is denoted as M_{nb}^{C}. Although position weight matrices give a fairly accurate estimate of the consensus

sequences, any sequence further away from the consensus (represented by small values in the position weight matrix) is not well represented[60]. To account for this limitation, any element in the position weight matrix where $M_{nb} \leq e$ is replaced by a minimal value e. To discuss Klf4 binding to either strand of DNA, we, therefore, use the following average position weight matrix:

$$\bar{M}_{nb} = \frac{M_{nb}^{F} + M_{nb}^{R}}{2}. \tag{1}$$

For a sequence that differs by a single base n from the consensus sequence, the probability ratio $\lambda_{n,b}$ of Klf4 binding to the two sequences is related to the position weight matrix by

$$\bar{M}_{nb} = \frac{\lambda_{n,b}}{\sum_b \lambda_{n,b}}. \tag{2}$$

Using this equation, we can obtain the probability ratio of Klf4 binding to a given sequence $\bar{B}$ (where $\bar{B} = (b_1, ..., b_L)$; b_n is the base at position n along the sequence) relative to the consensus motif, which reads

$$P\left(\bar{B} \right) = \prod_{n=1}^{L} \lambda_{n,b}, \tag{3}$$

where L is the number of bases along the sequence. Here P lies between 0 and 1; for consensus sequence $\bar{B}^*$, $P(\bar{B}^*) = 1$. We use equation (3) to infer the landscape of Klf4-λ-DNA-binding probability, as shown in Fig. 2h. To infer the binding energy landscape along λ-DNA from the position weight matrix, we use the following equation:

$$\epsilon_{b,n} - \epsilon_{b^*,n} = kTln\left(\frac{\bar{M}_{nb}}{\bar{M}_{nb^*}} \right), \tag{4}$$

where $\epsilon_{b,n}$ is the binding energy contribution for nucleotide position n for the corresponding case b_n. Further, $\epsilon_{b^*,n}$ is the binding energy contribution corresponding to consensus base (b_n^*) at nucleotide position n. Hence, the binding energy difference for a given sequence $\bar{B}$ with respect to binding to the consensus sequence is given by

$$\delta E\left(\bar{B} \right) = \sum_n \left(\epsilon_{b,n} - \epsilon_{b^*,n} \right). \tag{5}$$

Equations (4) and (5) allow us to infer the binding energy landscape for Klf4 binding to λ-DNA (Fig. 2h and Extended Data Figs. 8 and 9a). For details of how we obtain these equations, see Supplementary Note.

Correlation as a function of concentration. Independent experiments were sorted based on the experimental protein concentration. Intensity profiles were normalized to their maximum intensity and then averaged in groups selected from a moving window along the concentration axis (Fig. 4e) or in specific concentration bins (Extended Data Fig. 10g,h). For each average intensity profile, the correlation with the coarse-grained binding energy profile was quantified as the Pearson's correlation coefficient. The coarse-grained binding energy profile was first interpolated into the dimensions of the experimental profile.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
All data generated or analysed in this study are available from the corresponding authors upon reasonable request.

Materials
All materials will be made available upon request after the completion of a Material Transfer Agreement.

Code availability
The code used to analyse the data and perform numerical simulations is available from the corresponding authors upon reasonable request.

References
55. Lemaitre, R. P., Bogdanova, A., Borgonovo, B., Woodruff, J. B. & Drechsel, D. N. FlexiBAC: a versatile, open-source baculovirus vector system for protein expression, secretion, and proteolytic processing. *BMC Biotechnol.* **19**, 20 (2019).
56. 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).
57. Ryder, S. P., Recht, M. I. & Williamson, J. R. Quantitative analysis of protein-RNA interactions by gel mobility shift. *Methods Mol. Biol.* **488**, 99–115 (2008).

58. van Kampen, N. G. Chapter I—Stochastic Variables. in *Stochastic Processes in Physics and Chemistry* 3rd edn (ed van Kampen, N. G.) 1–29 (Elsevier, 2007).

59. Gao, Z., Liu, L. & Ruan, J. Logo2PWM: a tool to convert sequence logo to position weight matrix. *BMC Genomics* **18**, 709 (2017).

60. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).

61. Soufi, A. et al. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).

62. Good, N. E. et al. Hydrogen ion buffers for biological research. *Biochemistry* **5**, 467–477 (1966).

63. Wang, M. D., Yin, H., Landick, R., Gelles, J. & Block, S. M. Stretching DNA with optical tweezers. *Biophys. J.* **72**, 1335–1346 (1997).

Author contributions

J.A.M., S.W. and A.K. performed the experiments. S.G. provided reagents. S.C. and F.J. developed the theory with input from all the authors. J.A.M., S.W. and S.C. analysed the data with input from all the authors. J.A.M., S.W., S.C., A.K., A.A.H., F.J. and S.W.G. wrote the manuscript.

Competing interests

A.A.H. is a founder of Dewpoint Therapeutics and a member of the board as well as a shareholder in Caraway Therapeutics. All other authors have no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41567-021-01462-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41567-021-01462-2.

Correspondence and requests for materials should be addressed to Anthony A. Hyman, Frank Jülicher or Stephan W. Grill.

Peer review information *Nature Physics* thanks Stephanie Weber and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.
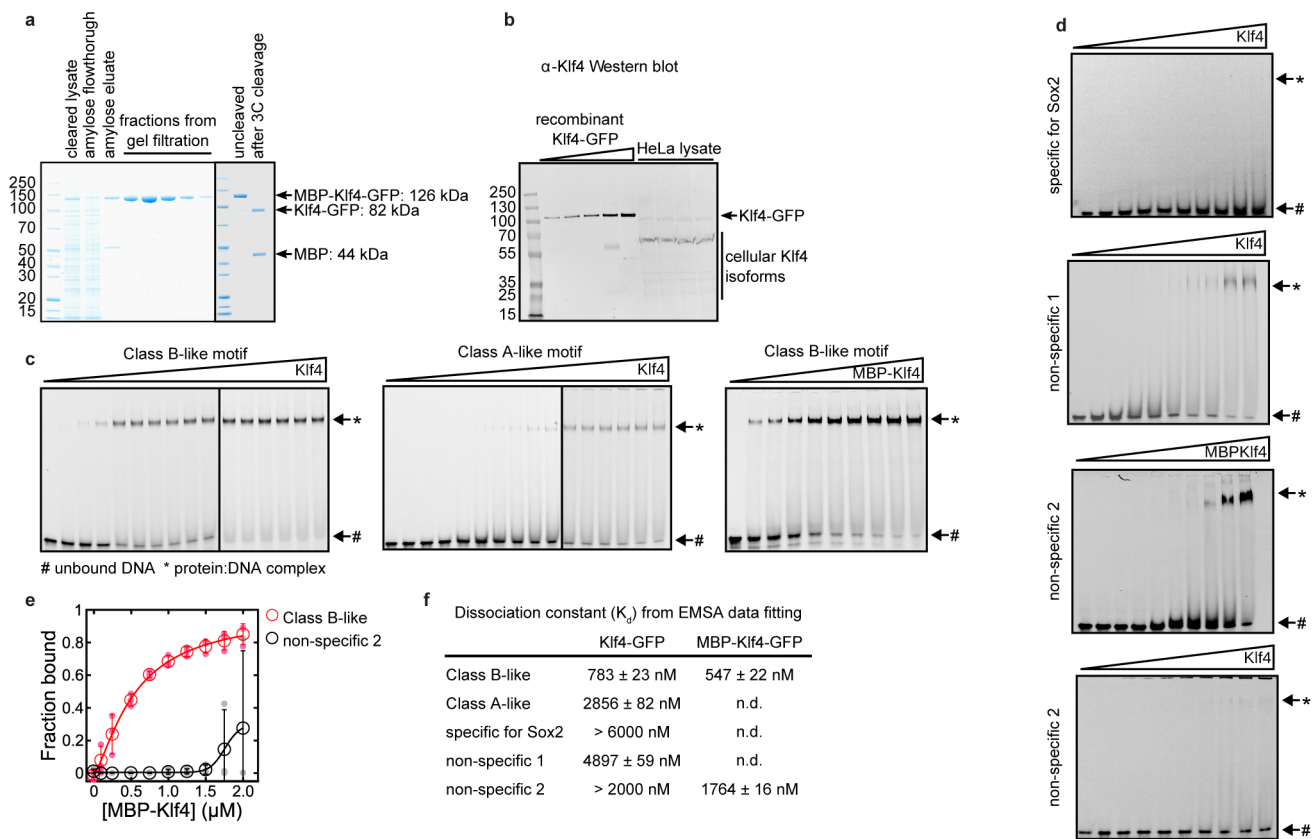
Extended Data Fig. 1 | Klf4 can be recombinantly purified and binds DNA in a sequence-specific manner. a, SDS gel showing a representative purification of MBP-Klf4-GFP. First, the Sf9 insect cell lysate is cleared by centrifugation (cleared lysate) before it is subjected to amylose resin (amylose flowthrough and amylose eluate). The eluate is then concentrated and further purified by size exclusion chromatography (fractions from gel filtration). Right before the assay, the concentrated MBP-Klf4-GFP sample (uncleaved) is treated with 3 C PreScission protease (see Methods) to remove the MBP-tag from Klf4-GFP (after 3 C cleavage). **b**, The concentration of intracellular Klf4 was estimated using HeLa lysates and Western blotting with fluorescent secondary antibodies (see Supplementary Methods). A representative example blot is shown. Recombinantly expressed and purified Klf4 was used to generate a standard curve on each Western blot. With this, the amount of Klf4 in HeLa lysates was determined. **c, d**, Representative examples of TBE gel images visualising Cy5-labelled dsDNA oligonucleotides in electrophoretic mobility shift assays (EMSAs) are depicted. Klf4-GFP and MBP-Klf4-GFP bind to dsDNA in a sequence-specific manner and cause an up-shift of the oligonucleotide in the gel. The oligonucleotides used can be found in Supplementary Table 2[61], sequence motifs are depicted in Supplementary Table 1. For assays using the DNA 'non-specific 2', concentrations up to 2 μM are shown, for all others up to 6 μM. **e**, EMSAs were performed as shown in **c, d** to test the affinity of MBP-Klf4-GFP to short dsDNA oligonucleotides with (red) or without (black) specific binding sites for the protein. Large circles, mean of N = 3 experiments; error bars, standard deviation; dots, individual experiments. **f**, Data fitting of the graphs shown in **e** and Fig. 1b allows determination of the dissociation constant (K_d, see Methods). The error margin indicates the 95% confidence value.
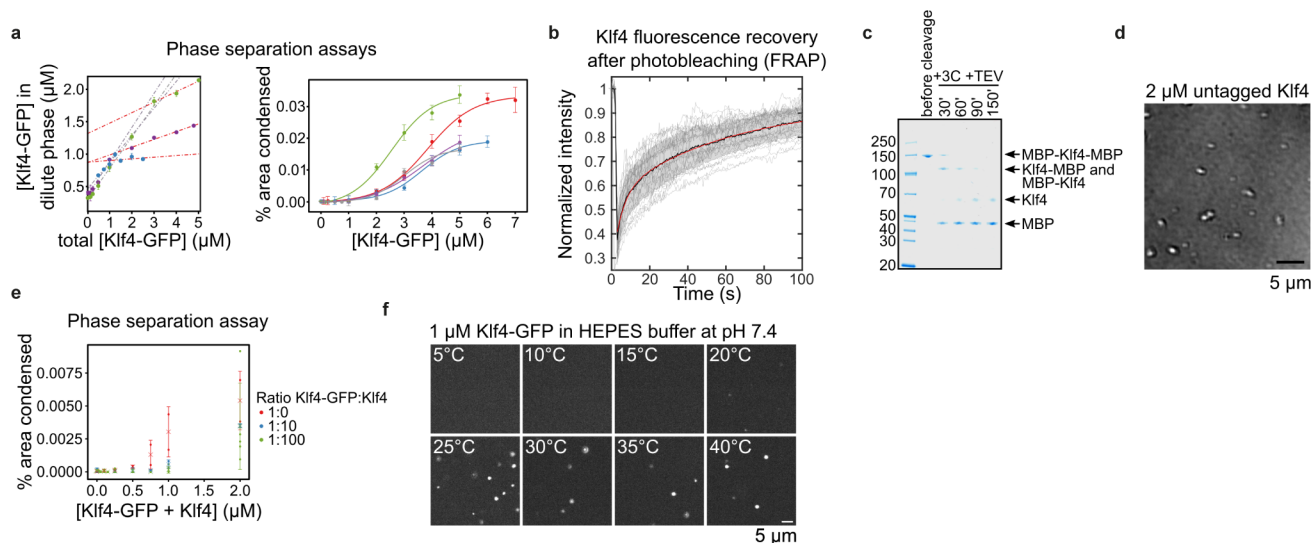
Extended Data Fig. 2 | Klf4 forms liquid droplets in vitro. a, Different phase separation assays were performed to estimate the saturation concentration (C_{SAT}) of Klf4-GFP. Left, the concentration of the dilute phase was measured after phase separating Klf4-GFP *in vitro*. For fitting, we used the two components piecewise linear function (gray and red lines) that shows maximum difference between slopes (see Supplementary Methods). The average $C_{SAT} = 1.0 \pm 0.3 \, \mu M$ (mean ± standard deviation) was determined from the intercepts of the lines with the lowest slope (red lines). Three independent repeats each from a different Klf4-GFP preparation are shown. Right, quantification of the fraction of the condensed area in confocal fluorescence microscopy images show droplets at concentrations above 1 µM. Five repeats are shown. Symbols, mean of 25 fields of view; error bars, standard deviation. **b**, FRAP of Klf4-GFP (grey, N = 28), black, mean, shaded area, 95% confidence interval. Red line, fit to the mean (93% mobile fraction, see Supplementary Methods). **c**, Untagged Klf4 was obtained by cleaving purified MBP-Klf4-MBP with TEV and 3C PreScission protease (see Supplementary Methods). SDS gel depicting a typical purification and cleavage time course. The protein was only used after complete removal of both MBP tags. **d**, Example DIC (differential interference contrast) microscopy image of untagged Klf4 droplets (same conditions as in Fig. 1c). **e**, Phase separation assay for different ratios of untagged Klf4 and Klf4-GFP (as in **a**, right). For a ratio of 1:0 (100% Klf4-GFP), the same five examples shown in **a** (right) are used. For the 1:10 (10% Klf4-GFP) and 1:100 (1% Klf4-GFP), two repeats are shown. Cross, mean; dots, individual experiments; error bars, standard deviation. **f**, Confocal images of 1 µM Klf4-GFP at different temperatures (same field of view) using a self-built temperature stage[62]. 25 mM HEPES pH 7.4 was used as buffering component (ΔpKa/10 °C = –0.14 for HEPES and -0.31 for Tris).
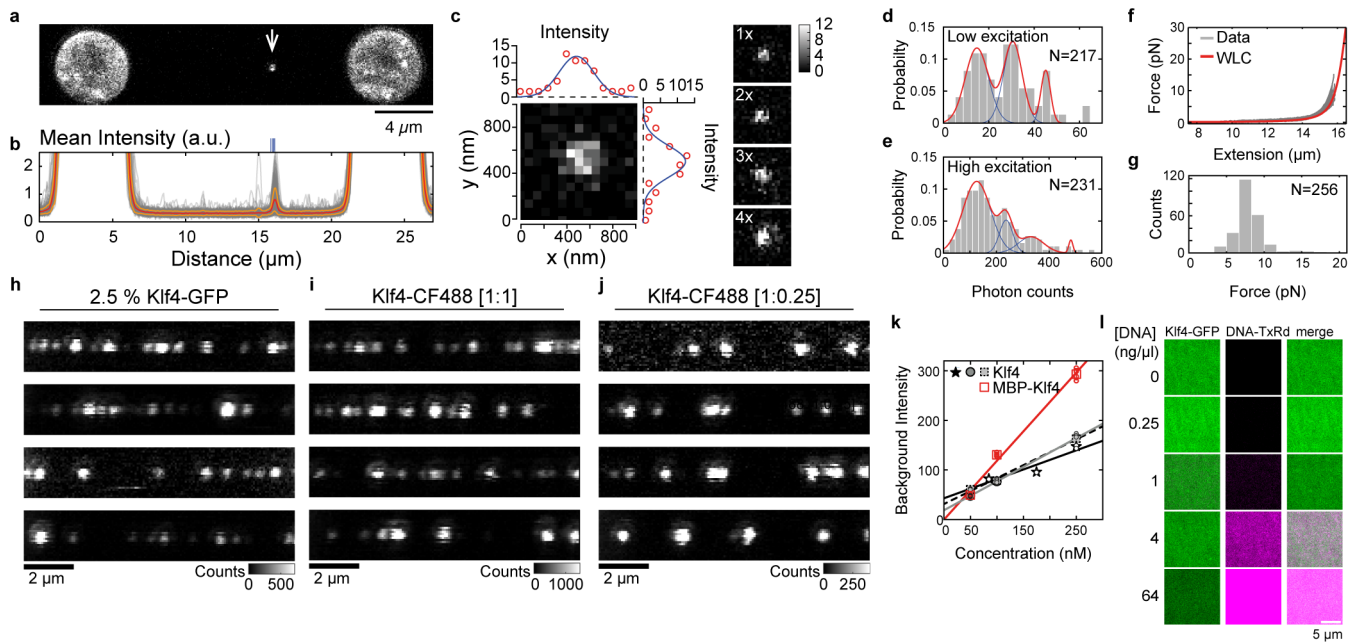
Extended Data Fig. 3 | Calibration of the optical tweezers assay and control for a nucleation mechanism. a, Example confocal image of a dCas9-EGFP (white arrow) labelled λ-DNA molecule held between two 4.42 μm diameter polystyrene beads. **b**, Mean Intensity profile along the DNA. Gray lines, N = 270 experiments. Red, orange, mean and standard deviation. Blue lines, position of the four dCas9 target sites (Supplementary Table 3 and Supplementary Methods). **c**, Left, example of an individual dCas9-EGFP molecule. Intensity profile in x and y directions (red circles) fitted to a gaussian function (solid blue line, FWHM = 355.2 and 319.5 nm respectively). Right, examples of 1, 2, 3 and 4 dCas9-EGFP molecules with integrated intensity values of: 156.1, 237.0, 344.2 and 488.8 photon counts. **d**, **e**, Probability distribution of integrated intensity values of dCas9 on λ-DNA (N, number of experiments). Red line, fit to a gaussian mixture model (see Methods and Supplementary Table 5). **f**, Representative force extension curves. Grey lines (N = 30), red line, Worm-like chain model[63] (contour length, L_C = 16.49 μm (48.514 kbp), persistence length, L_P = 50 nm, stretch modulus of K = 1200 pN). **g**, Histogram of experimental forces. F_{exp} = 8.22 ± 2.65 pN (mean ± standard deviation). At these forces, the contributions from an intrinsic globular state of the DNA (entropic regime) or the deformation of the double helix structure (enthalpic regime) can be avoided. **h–j**, Representative confocal images of a mixture of 2.5% Klf4-GFP and 97.5% untagged Klf4 and untagged Klf4 labelled with a small molecule dye (CF488A) at the specified ratios of Klf4:Klf4-dye (Supplementary Methods). [Klf4] = 200 nM. **k**, Background intensity as a function of the loaded concentration. Large symbols, mean, error bars, standard deviaiton. Lines, linear fit (Supplementary Methods and Supplementary Table 4). **i**, Phase separation assay with 500 nM Klf4-GFP and increasing amounts of Sulforhodamine 101-X (TxRd)-labelled dsDNA oligonucleotides containing 9 Klf4 binding sites. The short oligonucleotides fail to induce Klf4 condensation, indicating that Klf4 condensates do not form on DNA via a nucleation mechanism. Images of the same color have the same contrast settings.
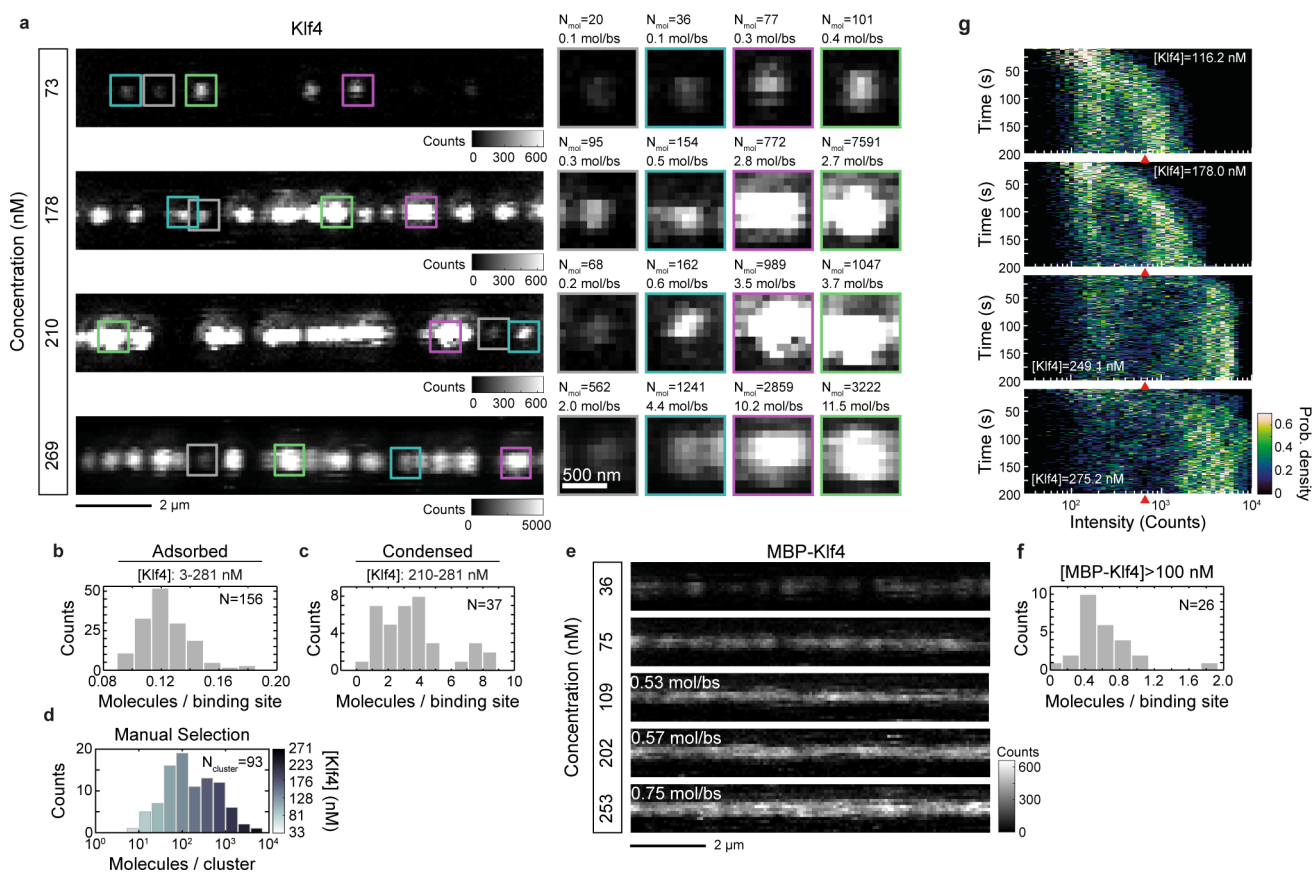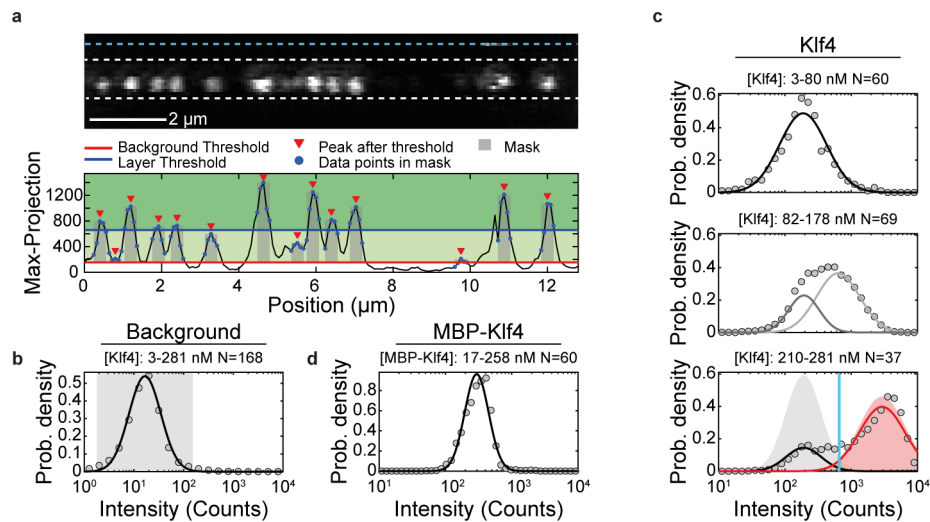
Extended Data Fig. 4 | Quantification of the number of molecules and transition between the adsorbed and condensed states. a, Left: Representative confocal images of Klf4-GFP on λ-DNA. Right: Representative clusters, the number of molecules per cluster (N_{mol}) and the number of molecules per binding site (number of molecules divided by the length of a binding site, 10 bp, see Supplementary Methods) are shown. **b**, **c**, Number of molecules per binding site for the adsorbed and condensed states (quantified as the sum of the integrated intensities corresponding to the pixels classified as adsorbed or condensed divided by the intensity per GFP). N, number of experiments. **d**, Histogram of number of molecules per cluster selected manually. **e**, Representative confocal images of MBP-Klf4-GFP coated DNA. **f**, Number of molecules per binding site computed as in **a**. Only experiments that exhibit full coverage were considered for this calculation. **g**, Probability density of the logarithm of intensities as a function of time after exposure of DNA to Klf4-GFP for individual experiments. Intensity distributions were computed for each frame as described in Extended Data Fig. 7. Red triangle, position of the adsorption threshold used in pixel classification (Fig. 2b and Extended Data Fig. 7). After an initial binding step (first ~ 50 s in top) the intensity distribution bifurcates into the adsorbed and condensed states (low and high intensity branches respectively). Over time, these states coexist and the condensed one increases its brightness (the high intensity branch migrates toward higher intensity values) until it reaches a stable value (~10^3 photon counts in top). The way these states are populated is suggestive of a switch-like behaviour: the condensed state becomes more populated over time at the expense of the adsorbed one. Increasing the bulk concentration reduces the time required for the bifurcation. At higher bulk concentrations, the condensed state gets brighter (~$5*10^3$ in bottom) and overall, more populated.
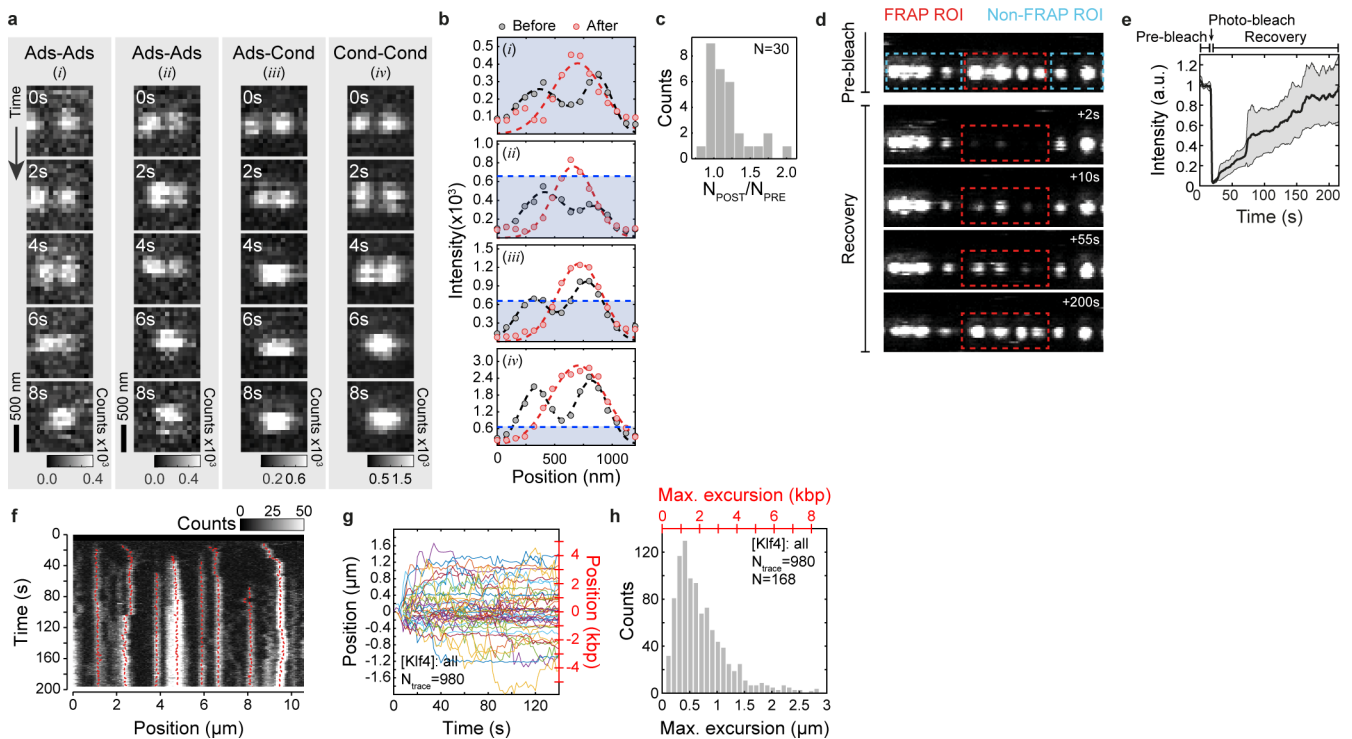
Extended Data Fig. 5 | Properties of Klf4 clusters on DNA. a, Confocal images of representative time lapses of the four types of fusion events observed (*i-iv*). (i) corresponds to Fig. 1i. **b**, Intensity profiles along the DNA before (black) and after (red) fusion. The events were classified into adsorbed-adsorbed (*i, ii*, N = 9); adsorbed-condensed (*iii*, N = 9) and condensed-condensed (*iv*, N = 12). Blue shade and dotted line, intensity region corresponding to the adsorbed state. **c**, Histogram of the ratio of number of molecules before and after fusion. **d**, Representative confocal microscopy images of the time course of a FRAP experiment. Time stamps are referenced to the time of induced photo-bleaching. **e**, Mean intensity across the FRAP ROI corrected for photo-bleaching. Individual traces of mean intensity in the FRAP ROI were divided by the mean intensity in the non-FRAP ROI (cyan squares in **d**) and normalized to the value prior to the photo-bleaching step. Black line shows the average trace for N = 13 experiments, grey shade, standard error of the mean at 95% confidence. **f**, Kymograph showing the displacement of Klf4-GFP clusters along the DNA. In red, the trajectories of segmented clusters are shown. Segmentation was done using the python package Trackpy (see Supplementary Methods) with initial search parameters of $P_R = 15$ pixels, $S_R = 6$ pixels, $t_{memory} = 30$ s and $t_{min} = 100$ s. **g**, Position of foci over time along the DNA relative to the initial point of each trace (Position = x(t) -x(t = 0), where x(t) is the position over time). A random selection of 50 traces is shown. **h**, Histogram of the maximum excursion length (defined as the farthest point a cluster moves from the initial position. The number of traces (N_{trace}) is indicated. N = 168 experiments were considered for [Klf4] = 3–281 nM.
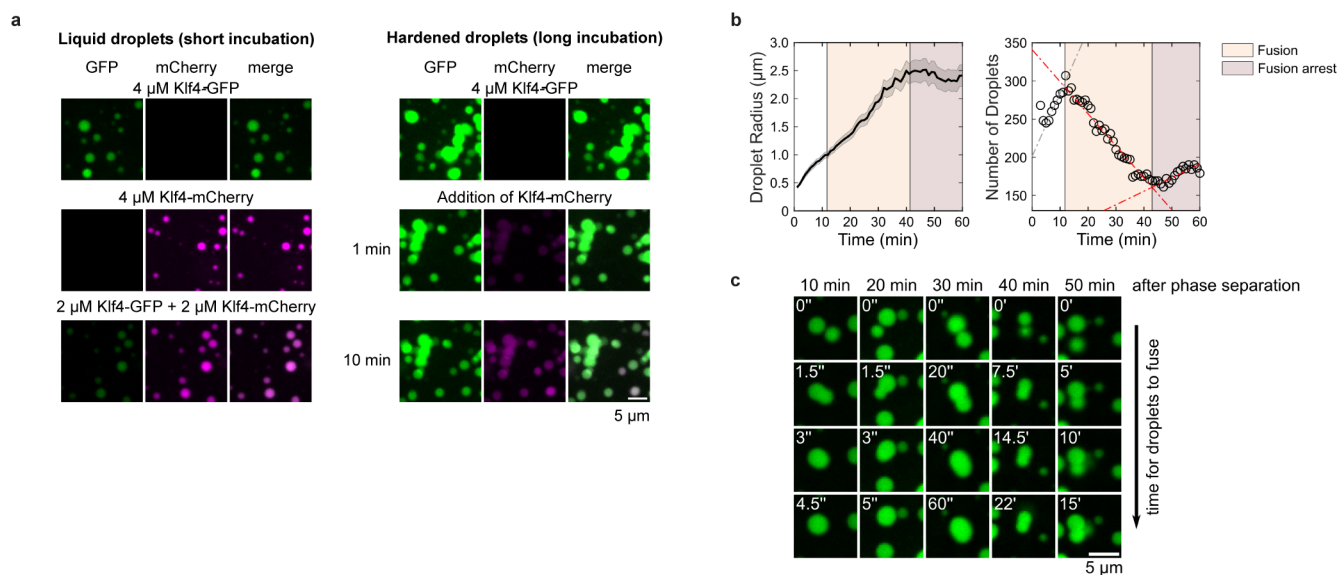
Extended Data Fig. 6 | Hardened Klf4 droplets can incorporate new material and Klf4 droplets harden within 40 min. a, In order to exclude the possibility that Klf4 droplets stopped growing on DNA because of hardening, we tested whether hardened droplets retain the ability to incorporate new material. Right, for this, Klf4-GFP droplets were formed and incubated until hardened (top row). After 50 min, a fresh solution of Klf4-mCherry was added (middle row). After only 10 min (bottom row), the green Klf4-GFP droplets enriched red signal of Klf4-mCherry. Left, as a control, to test whether both Klf4-GFP and Klf4-mCherry individually (top and middle row) or in combination (bottom row) are able to form droplets under these conditions, the standard incubation time for droplet assays was used (short incubation: 20 min). See Supplementary Methods for details. All confocal microscopy images of the same colour in the same panel have the same contrast settings. **b**, The droplet radius (left) and number of droplets in the field of view (right) of a solution with 20 μM Klf4-GFP was measured every minute after induction of phase separation for one hour. Many fusion events could be observed initially (white and yellow areas). However, after 43 min the droplets stopped fusing (brown area). This is one order of magnitude larger than the observation time in the optical tweezers assay (200 s). The onset of hardening was determined following a similar procedure than in Extended Data Fig. 2a (see Methods). **c**, Individual examples of droplet fusion events are depicted at different times after phase separation was induced, as indicated at the top. Note how the fusion time increases with time until fusion stops and droplets only stick to each other when hardened (" seconds, ' minutes). In order to compare fusion times of droplets that are roughly of the same size, the examples are at different Klf4-GFP concentrations ranging from 5 to 20 μM.

Extended Data Fig. 7 | Determination of the intensity thresholds for pixel classification into adsorbed and condensed states. a, Top, example confocal image of Klf4-GFP on λ-DNA. Bottom, intensity profile and data processing steps (see Methods). **b**, Probability density of the logarithm of pixel intensities of background values (after background subtraction) along a line away from the DNA (cyan dashed line in **a**, top). Black line, fit to a normal probability density function. The background threshold ($I_{th-bg}=152.9$, red line in **a**, bottom) was defined as the mean plus 3 times the standard deviation of the distribution (grey area, upper boundary). **c**, Probability density of the logarithm of pixel intensities, along the intensity profiles pulling together experiments in the concentration range indicated. For low protein concentrations (top), the probability density can be fitted to a normal probability density function (black line, mean $\mu_1=187.96$ and standard deviation $\sigma_1=237.47$). For high protein concentrations (bottom), the probability density was fitted to a two components gaussian mixture model constraining the mean of the first mode to the value extracted from the low concentration distribution in the top panel (black line $\sigma_1=162.50$, area $a_1=0.22$ and red line, $\mu_2=2.93*10^3$, $\sigma_2=3.52*10^3$, $a_2=0.78$ respectively). We define the adsorption layer upper boundary ($I_{th-ads}=658.5$, cyan line) as the crossing point between the first and second modes, normalized to the same area independently (grey and red areas, rescaled here for representation purposes: maximum value=0.65 and 0.50 respectively). In the intermediate concentration range, an unconstrained fit to a two components gaussian mixture model rendered a low intensity component with a mean similar to the one observed at low and high concentrations (dark grey line, $\mu_1=193.63$, $\sigma_1=136.37$, $a_1=0.31$) and a high intensity component (light grey line, $\mu_2=634.66$ and $\sigma_2=708.06$, $a_2=0.69$)). **d**, The probability density pulling together all MBP-Klf4 experiments, can be fitted to a normal probability density function (black line, $\mu=295.40$ and $\sigma=151.74$). **b**–**d**, The number of experiments considered in each case (N) is indicated.
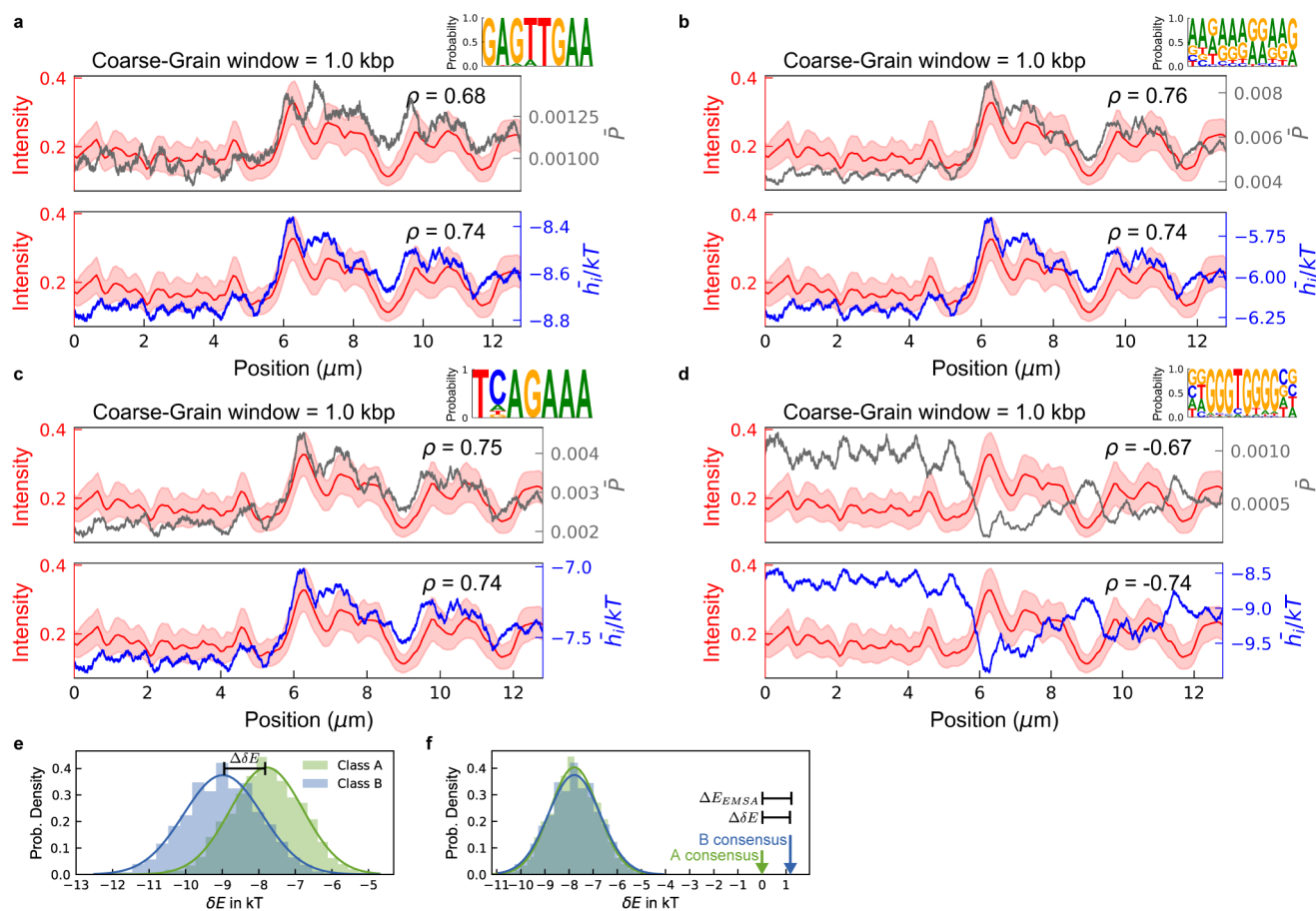
Extended Data Fig. 8 | Sequence dependence of the Klf4 localization pattern. a–d, Top right, consensus recognition motifs considered in this work. Red line, average intensity profile along λ-DNA for N = 79 experiments in the concentration range 8–281 nM (red). Shaded area, standard error of the mean at 95% confidence. Top: grey, coarse-grained binding probability profile. Bottom: blue, coarse-grained binding energy profile (see Methods). The Pearson's correlation (ρ) between the average intensity profile and the corresponding calculated profile is indicated. **a,** Class A binding motif 2 (*in vivo*[26]). **b,** Class A binding motif 3 (*in vitro*[25]). **c,** Class A binding motif 4 (*in vitro*[27]). **d,** Class B binding motif (*in vivo*[24]). **e,** Similarity of classes A[26] and B[24] motifs to λ-DNA sequence composition quantified by the histograms of the inferred binding energy difference (δE) to the respective class consensus sequences. Histograms reveal a gaussian-like peak (solid lines are fits to gaussian distributions) of similar width and height at negative values of δE, shifted relative to each other. The more a peak is positioned towards the right, the higher the similarity with the sequence composition of λ-DNA. The value of at the peak of each histogram corresponds to the difference in binding energy of the respective consensus sequence to Klf4 and the binding energy of Klf4 to a typical sequence on λ-DNA. The distance between the two peaks ($\Delta\delta E$) should correspond to the difference in binding energy of the classes A and B consensus sequences to Klf4, and class A consensus sequences should therefore have a weaker affinity for Klf4 than the class B consensus sequence. **f,** Shifting the two histograms along the δE axis such that the peaks overlap, reveals a binding energy difference of $\Delta\delta E = 1.20$ kT, which can be compared with that obtained by EMSA $\Delta E_{EMSA} = 1.28$ kT (Fig. 1b, Extended Data Fig. 1).
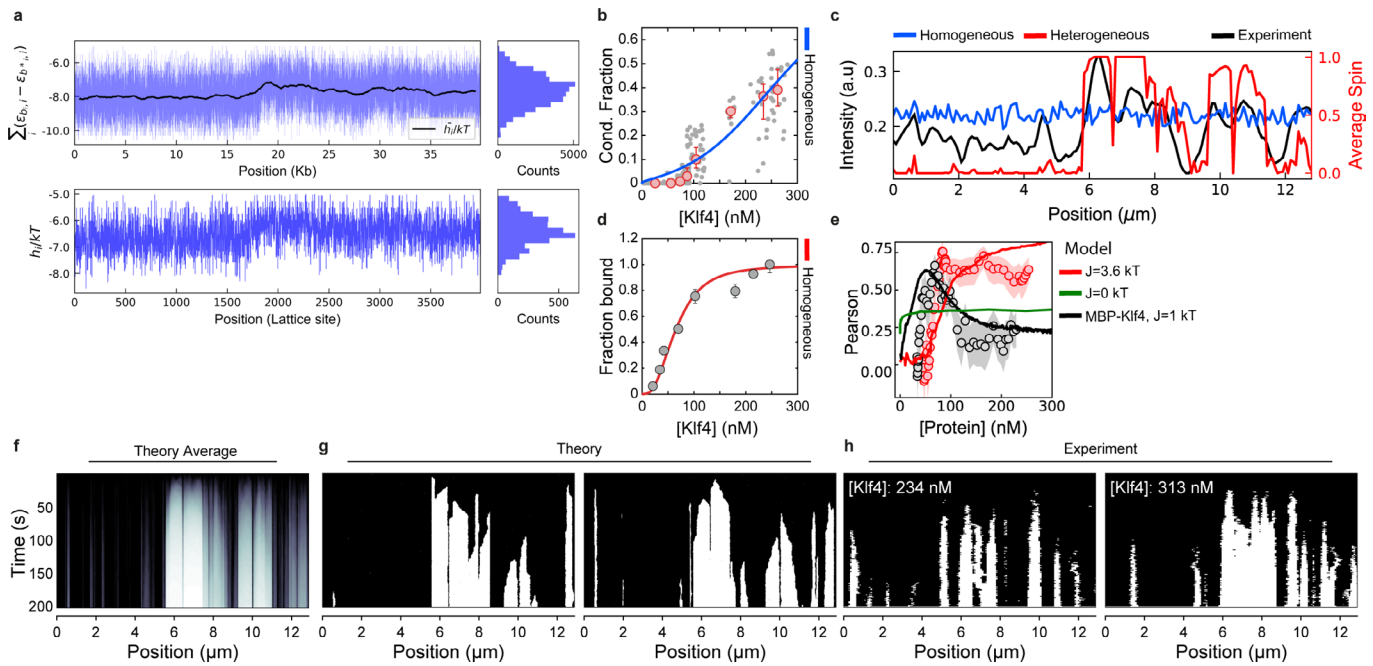
Extended Data Fig. 9 | A two-state model with heterogeneous binding energies can successfully account for experimental observations. a, Top: Binding energy landscape along λ-DNA as inferred from the position weight matrix (*in vivo*[26], Class A binding motif 1, Fig. 2g). Black line, coarse-grained binding energy landscape ($\bar{h}_i$, 1 kb window moving average). Right, corresponding histogram of binding energies. Bottom: Inhomogeneous propensity for condensation for the model (h_i, Eq. 17 and 18 Supplementary Note). Right, corresponding histogram of h_is. **b**, Condensed fraction for Klf4-GFP at different concentrations (as in Fig. 2d). Blue line, fit to the model with homogeneous binding energies (parameter values: J = 3.37 kT, C_0 = 292 nM, α = 0.0016 kT). **c**, Average steady-state spatial profiles of Klf4 condensation along λ-DNA. Blue line, steady-state profile (obtained after 200 iterations) for homogeneous binding energies (average over 100 individual kymographs). This average can vary from 0 to 1 (0 and 1 correspond to the -1 and +1 states in the model). Red line, average steady-state profile for heterogeneous binding energies. Black line, average pixel intensity of Klf-GFP along λ-DNA. **d**, Fraction bound of MBP-Klf4-GFP at different concentrations. Red line, fit to the model with heterogeneous binding energies (parameter values: J = 1.02 kT, C_0 = 64.95 nM, α = 0.0016 kT). **e**, Solid lines, Pearson correlation coefficient between the spatial condensation profile obtained from the model and $\bar{h}_i$ as a function of bulk protein concentration. Red and green lines, model predictions with heterogeneous binding energies for J = 3.16 kT and J = 0 kT. Black line, model prediction, for the model fitted to the data in **d**. Circles, experimental correlations between intensity and $\bar{h}_i$ profiles (same as in Fig. 4e). **f**, Average kymograph (400 realizations) predicted from the model at a concentration of 250 nM. **g**, Representative kymograph realizations at 234 nM and 313 nM concentrations (see Supplementary Note). **h**, Experimental kymographs at the indicated concentrations (thresholded to display only the condensed state in white).
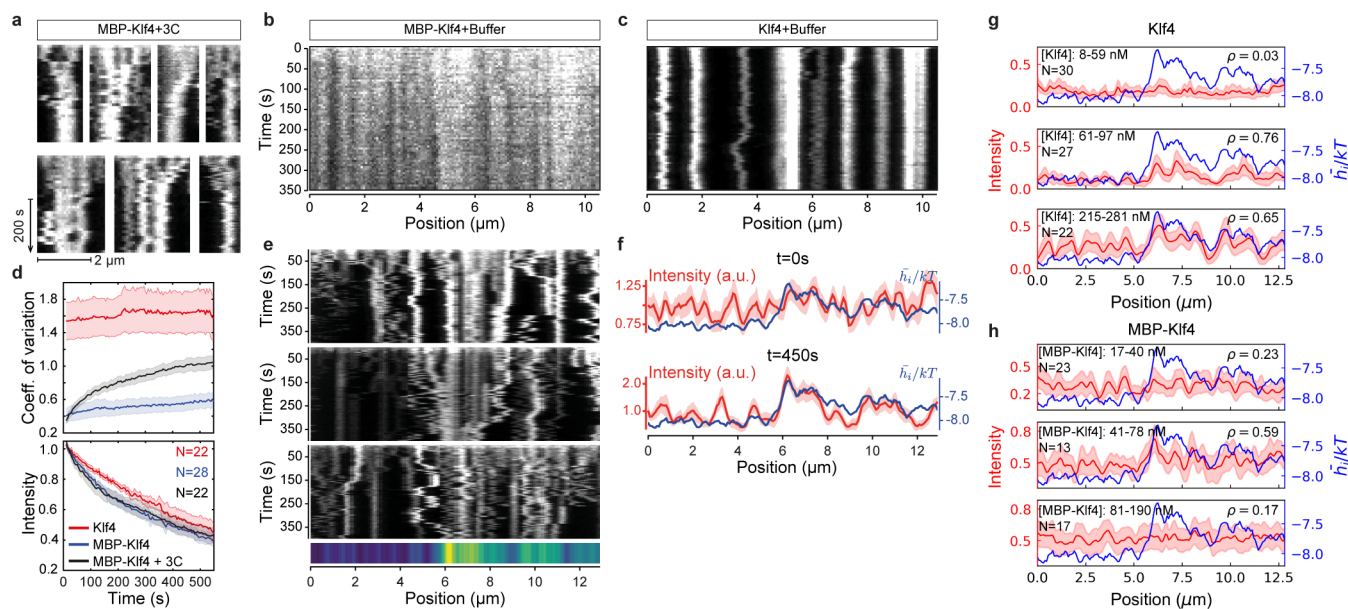
Extended Data Fig. 10 | Klf4 transition from adsorbed to condensed layer and from low to high sequence specificity on DNA upon MBP cleavage.
a, Representative examples of kymograph ROIs corresponding to the 3 C dependent MBP-Klf4 condensation. Each example corresponds to a different experimental realization. **b**, **c**, Example kymographs of MBP-Klf4 and Klf4 coated DNA when transferred to the assay buffer (which does not contain the 3 C protease). **d**, Coefficient of variation (CV, top) and mean (bottom) of the distribution of intensity values along the DNA over time after cleavage. An increase in the CV while the mean intensity decays in a similar way for all conditions indicates a rearrangement of material on the DNA that leads to local enrichment of cleaved Klf4-GFP. Solid lines, mean, shaded area, standard error of the mean at 95% confidence. The number of experiments (N) is indicated. **e**, Representative kymographs corresponding to the 3 C dependent MBP-Klf4 condensation process. The coarse-grained binding energy profile is shown at the bottom as a guide to the eye (see Fig. 2h and Methods). **f**, Red, average intensity profile along λ-DNA for N = 22 experiments binned at the indicated time after transfer of the bead-DNA system to a solution containing the 3 C protease. Shaded area, standard error of the mean at 95% confidence. The coarse-grained binding energy profile is shown in blue (see Fig. 2h and Methods). The Pearson's correlation between the average intensity and the binding energy profiles, increases from ρ=0.51 at t = 0 s to ρ=0.69 at t = 450 s. **g**, **h**, Average intensity profile along λ-DNA for N experiments binned in the indicated concentration range (red) for Klf4 (**g**) and MPB-Klf4 (**h**). Shaded area, standard error of the mean at 95% confidence. The coarse-grained binding energy profile is shown in blue (see Fig. 2h and Methods). The correlation between the average intensity profile and the binding energy profile, quantified by Pearson's correlation coefficient (ρ), is indicated in each case.

nature research

Corresponding author(s):   Stephan W. Grill
Frank Jülicher
Anthony A. Hyman

Last updated by author(s): Sep 11, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	ImageStudio v4.0, Typhoon FLA 9500 v1.1, Andor iQ3.3, Andor iQ3.6, DeltaVision softWoRx v7.2.0, OLYMPUS cellSens Dimension v3.1, Carl Zeiss Zen 2.3 SP1 FP3 (black) v14.0.0.0, Bluelake v1.4.2- 1.6.13
Data analysis	ImageStudio v4.0, Matlab R2017b, Matlab R2018b, Fiji running with ImageJ v1.53c and Java v1.8.0_66, RStudio v1.0.143 running with R v3.4.4, Python 3.7.6

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated or analysed in this study are available form the corresponding authors upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We based the sample sizes on the variability within the ensemble and chose the sample sizes empirically.
Data exclusions	One data set was excluded from Supplementary Fig. 3a as some of the calculated concentrations were below 0. No other data was excluded. However, different analyses were performed using different sub-sections of the data as specified in the figure legends and Methods.
Replication	Variability was observed between individual Klf4 preparations. This variability is included in the reported errors.
Randomization	Not relevant for this study.
Blinding	Not relevant for this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	α-Klf4 antibodies: abeomics, Cat-No: 11-12004; goat-α-rabbit IRDye 800CW: Li-Cor, Cat-No: 926-3221
Validation	As shown in Supplementary Fig. 1b, the α-Klf4 antibody recognises recombinant Klf4 on a Western blot. The identitiy of the recombinant protein was confirmed by mass spectrometry. Further, the detected bands in cell lysates are in accordance with the Klf4 isoforms specified on Uniprot. The antibody was not used for any other application than Western blot.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Jan Ellenberg Group, EMBL Heidelberg
Authentication	The HeLa cell line used in this study was authenticated by STR analysis (Eurofins).
Mycoplasma contamination	Mycoplasma tests were performed regularly and only cells that tested negative were used for the experiments.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.