RNA Pol II Length and Disorder Enable Cooperative Scaling of Transcriptional Bursting

Porfirio Quintero-Cadena[a], Tineke L. Lenstra[b], Paul W. Sternberg[a,*]

[a]*California Institute of Technology, Division of Biology and Biological Engineering, Pasadena, CA*
[b]*The Netherlands Cancer Institute, Oncode Institute, Division of Gene Regulation, Amsterdam, the Netherlands*

Abstract

RNA Polymerase II contains a disordered C-terminal domain (CTD) whose length enigmatically correlates with genome size. The CTD is crucial to eukaryotic transcription, yet the functional and evolutionary relevance of this variation remains unclear. Here, we use smFISH, live imaging, and RNA-seq to investigate how CTD length and disorder influence transcription. We find that length modulates the size and frequency of transcriptional bursting. Disorder is highly conserved and mediates CTD-CTD interactions, an ability we show is separable from protein sequence and necessary for efficient transcription. We build a data-driven quantitative model, simulations of which recapitulate experiments and support CTD length promotes initial polymerase recruitment to the promoter but slows down its release from it, and that CTD-CTD interactions enable promoter recruitment of multiple polymerases. Our results reveal how these tunable parameters provide access to a range of transcriptional activity, offering a new perspective for the mechanistic significance of CTD length and disorder in transcription across eukaryotes.

Keywords:
Transcriptional bursting, RNA Pol II, CTD length, Phase separation

[*] Correspondence: pws@caltech.edu

1. Introduction

In eukaryotes, the RNA Polymerase II complex that transcribes protein-coding genes is typically composed of 12 subunits (Hantsche and Cramer, 2017). The largest and catalytic subunit RPB1 contains a repetitive and unstructured C-terminal Domain (CTD) that is a major factor for establishing critical protein-protein interactions throughout transcription and downstream processes (Harlen and Churchman, 2017).

Each of the heptad amino acid repeats in the CTD, whose number ranges from 5 in *Plasmodium yoelii* to 60 in *Hydra* (Chapman et al., 2008; Yang and Stiller, 2014), can be subject to post-translational modifications that regulate its physical interactions and consequently RNA Pol II function (Eick and Geyer, 2013; Harlen and Churchman, 2017). The CTD's repetitive nature most likely arose in the last eukaryotic common ancestor (Yang and Stiller, 2014), and its length appears to enigmatically correlate with genome size (Chapman et al., 2008; Yang and Stiller, 2014). What is the role of CTD length variation in transcription?

Truncating the number of CTD repeats impacts cell growth and animal development, with a minimal number required for viability (Nonet et al., 1987; Bartolomei et al., 1988; West and Corden, 1995; Gibbs et al., 2017; Lu et al., 2019), and reduces the transcriptional output from enhancer responsive genes (Allison and Ingles, 1989; Scafe, C; Young, 1990; Gerber et al., 1995; Aristizabal et al., 2013). Enhancers physically interact with promoters via protein-protein interactions to activate transcription in bursts of activity (Chubb et al., 2006; Bartman et al., 2016; Chen et al., 2018). Given that mRNA output decays rapidly with increasing separation between enhancers and promoters (Dobi and Winston, 2007; Quintero-Cadena and Sternberg, 2016), an intriguing possibility is that CTD expansion facilitates enhancer function over physical distances to promoters that scale with genome size (Allen and Taatjes, 2015).

Increasingly relevant for the understanding of biological phenomena, liquid-liquid phase separation is an emerging signature of proteins with disordered, repetitive domains (Banani et al., 2017; Shin and Brangwynne, 2017). Low complexity domains that exhibit this property appear to be abundant in nuclear proteins, including the CTD and major transcription factors (Cho et al., 2018; Chong et al., 2018; Qiu et al., 2018; Shin et al., 2018). CTD length has also been implicated in its ability to form (Boehning et al., 2018) and bind phase-separated droplets, with a minimum threshold that parallels its viability requirement (Kwon et al., 2013). In addition, the interaction of the CTD with these droplets can be dynamically modulated by phosphorylation (Kwon et al., 2013; Chong et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019), a major post-translational modification that precedes transcription initiation (Payne et al., 1989; Svejstrup et al., 1997). In light of these observations, phase separation provides an appealing framework to explain certain transcriptional phenomena. From this perspective, the CTD could provide a bridge for the RNA polymerase to dynamically participate in multi-molecular assemblies of transcription factors and DNA loci that facilitate the function of highly active enhancers (Hnisz et al., 2017).

Extensive investigations have revealed many roles of CTD sequence and post-translational modifications (Eick and Geyer, 2013; Harlen and Churchman, 2017). On the other hand, the functional and evolutionary relevance of CTD length and the mechanism by which it influences transcription have not been systematically investigated.

Here, by quantitatively analyzing snapshots and dynamics of transcription in budding yeast, we show that CTD length can modulate transcription burst size and frequency. We strengthen the evolutionary relevance of the CTD's long disorder, and provide evidence that its role in transcription is separable from amino acid sequence. Specifically, we demonstrate that the function of the CTD's long disorder can be supplemented by similarly unstructured protein domains. These proteins can interact with and recruit others of their kind, an ability that is necessary for efficient transcription *in vivo*. We use these features, together with known CTD protein-protein interactions, to construct an integrative and quantitative model that explains how CTD length influences the dynamics of eukaryotic transcription.

2. Results

2.1. CTD is enriched in disordered amino acids and its length inversely correlates with gene density across eukaryotes

The CTD of representative species has been shown to be a random coil (Portz et al., 2017), suggesting this structural feature is relevant for its function and could itself be used to identify it. We sought to learn whether this was a common signature in all known protein sequences. We searched for RPB1 protein sequence homologs, the CTD-bearing catalytic subunit of RNA Pol II (Figure 1A, top). We recovered 542 unique sequences from 539 species and 338 genera, whose length varies from 1374 to 3055 amino acids (Figure 1B), with some evident bias that likely stems from the limited availability of genome sequences.

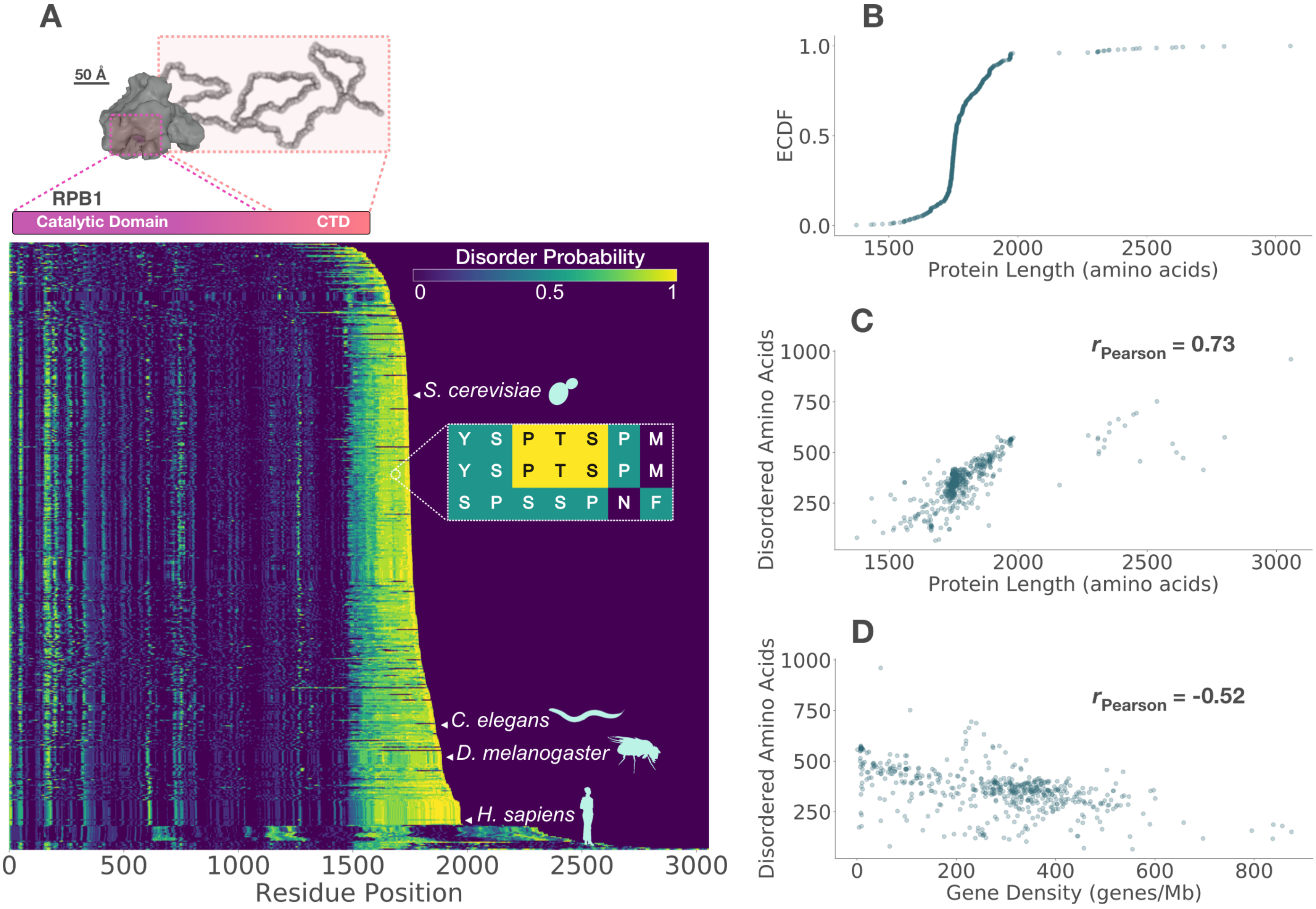


Figure 1: **RPB1, the catalytic subunit of RNA Polymerase II, contains an unstructured C-terminal Domain (CTD) whose length correlates with gene density across eukaryotes.** (A) Top: Cartoon of 12 subunit RNA Pol II complex with an unstructured CTD drawn to scale, adapted from 1Y1W.PBD (Kettenberger et al., 2004; Portz et al., 2017). RPB1 is highlighted in pink and its CTD in orange. Bottom: distribution of disorder probability along RPB1 sequence homologs sorted by length. Representative species are highlighted. Inset illustrates residue coloring by disorder probability. (B) Empirical cumulative distribution function (ECDF) of RPB1 lengths. (C) RPB1 length positively correlates with number of predicted disordered amino acids and inversely with gene density (D). Pearson correlation coefficient is shown for each pair of variables.

We computed the disorder probability per amino acid along each protein sequence. We found that with few exceptions, the C-termini of RPB1 sequences is enriched in disordered amino acids (Figure 1A, bottom).

Protein length is positively correlated with the number of disordered amino acids (Figure 1C). As noted in previous reports (Chapman et al., 2008; Yang and Stiller, 2014), these C-terminal sequences are enriched in amino acids from the heptad repeat YSPTSPS (Figure S1A). Like amino acid content, overall charge, aromaticity and hydrophobicity have a compact distribution (Figure S1B-D).

CTD length has been shown to correlate with genome size for a few representative species that span a wide range of genome sizes (Chapman et al., 2008; Yang and Stiller, 2014), from $1x10^7$ bp in yeast to $3x10^9$ bp in human. To systematically investigate the generality of this phenomenon, we compiled a list of genome sizes and their estimated gene number. We then computed the gene density (genes per megabase of DNA) for each species, in order to account for long stretches of non-coding DNA that are more common in large genomes. The number of disordered amino acids in RPB1 homologs inversely correlated with gene density (Figure 1D): sparse genomes tend to have polymerases with longer CTDs.

This systematic characterization of RPB1 sequences builds on previous extensive reports (Chapman et al., 2008; Yang and Stiller, 2014) and highlights three important features of the CTD. First, protein disorder is a highly conserved and likely functionally relevant feature of the CTD. Second, RPB1 length variation mostly originates from the number of disordered amino acids in its C-terminus. Third, CTD length is inversely correlated with gene density.

2.2. *CTD length modulates transcription burst size and frequency*

We sought to understand the role of CTD length using *Saccharomyces cerevisiae* as our model. Wild-type yeast CTD contains 26 heptad repeats (CTDr). We generated strains in which the genomic copy of RPB1 was engineered to have 14, 12, 10, 9, and 8 CTDr, the minimum required for viability in yeast (West and Corden, 1995). Consistent with previous reports (Nonet et al., 1987; West and Corden, 1995), the growth rate of these strains was compromised by CTD truncation. The magnitude of the decrease in growth rate increased with decreasing CTD length (Figure 2A); while 14 and 12 CTDr strains have only a subtle growth phenotype, the decrease in growth rate becomes evident with 10 CTDr and progressively larger with 9 and 8 CTDr (Figure 2A, inset).

RNA Pol II is mostly present in the nucleus, imported as a fully assembled complex (Boulon et al., 2010; Czeko et al., 2011). We hypothesized that nuclear polymerase becoming rate-limiting could explain the observed phenotypes, given that the CTD has been linked to nuclear import (Carre and Shiekhattar, 2011). We fused the fluorescent protein mScarlet to RPB1, and unexpectedly found that CTD truncation increased its nuclear levels (Figure 2B). This effect is presumably explained by the requirement of the CTD for ubiquitination (Huibregtse et al., 1997; Somesh et al., 2005) and the resulting accumulation of the protein complex. We were unable to fuse mScarlet to shorter CTD strains; however, this trend suggested the polymerase does not become rate-limiting upon CTD truncation.

We then asked how CTD length influences transcription. We focused on the galactose transcriptional response, because like other inducible pathways it has been shown to be sensitive to CTD truncations (Allison and Ingles, 1989; Scafe, C; Young, 1990) and involves the differential expression of over 2000 transcripts (Figure 2E), about a third of the yeast's transcriptome. We measured the transcriptional phenotype of CTD truncation using RNA-seq, comparing wild-type with the 10CTDr strain with and without galactose (Figure 2C). We fitted these data to a linear model that allowed us to estimate the individual contributions, for each measured transcript, of four components in the experiment: batch effects, galactose induction, CTD truncation, and its interaction with galactose induction (Figure 2C, right).

We identified over 2000 transcripts affected by CTD truncation at a q-value threshold of 0.1, from a distribution of q-values that indicates a strong transcriptional phenotype (Figure 2D,E). A significant proportion of the galactose responding transcripts exhibited a statistical interaction with CTD truncation. This effect revealed a surprisingly specific, globally antagonistic relationship between two components: 1) galactose

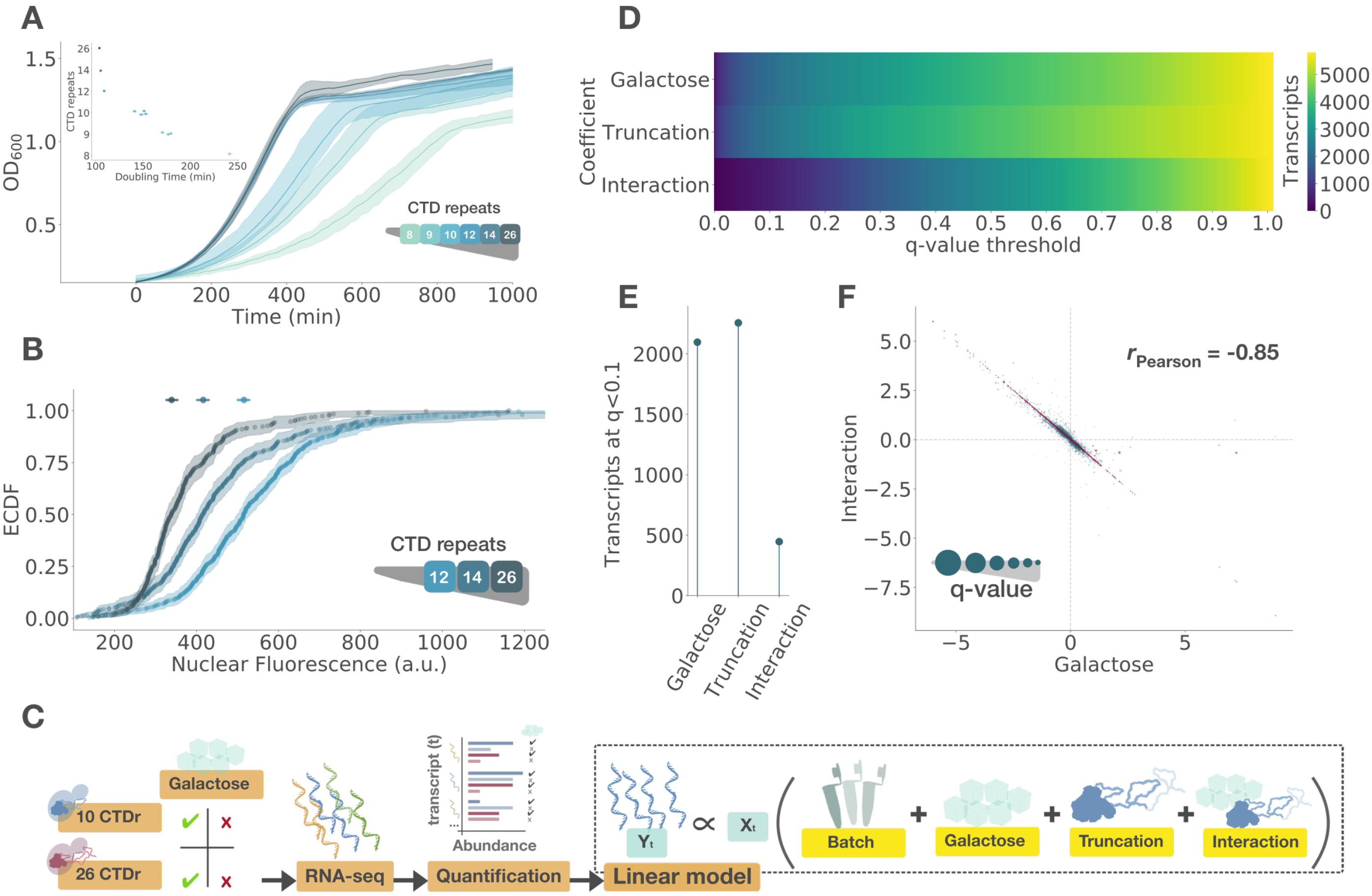


Figure 2: **CTD truncation reduces growth rate, increases nuclear Pol II concentration and antagonizes transcription activation genome-wide.** (A) Mean optical density (OD) over time of *S. cerevisiae* strains with 26 (wild-type), 14, 12, 10, 9, and 8 CTD repeats (CTDr). Shaded area shows the range of measured ODs at each time point from three biological replicates per line. Inset shows mean doubling times (DT) with standard error by line. (B) Empirical cumulative distribution function (ECDF) of mScarlet-RPB1 nuclear fluorescence in strains with 26, 14, and 12 CTDr from three biological replicates. Shaded area is bootstrapped 99% confidence interval (CI) and top markers show median with 99% CI. (C) Experimental design to measure the transcriptomic phenotype of CTD truncation and its influence on the transcriptional response after 2 hours of galactose induction via RNA-seq from three biological replicates. A linear model was used to fit the RNA-seq data, where each coefficient estimates the influence on each measured transcript of batch effects, galactose induction, CTD truncation and its interaction with induction. For each coefficient, (D) cumulative distribution, in number of transcripts, of false discovery rates (q-values) and (E) number of differentially expressed transcripts detected at a q-value threshold of 0.1. (F) Comparison of the log fold-change of each transcript resulting from galactose induction and its interaction with CTD truncation. Red points show the positions on the diagonal $x = y$. Marker size of each point is inversely proportional to the q-value of the interaction ($ms = -log(q_{int})$); dotted lines reference no change at zero and the Pearson correlation is indicated. Direct targets of GAL4 listed in Lesurf et al. (2016) are plotted in orange.

induction and 2) the interaction of truncation with galactose induction (Figure 2F). In other words, CTD truncation reduced the magnitude of change in abundance of most transcripts upon galactose induction.

We sought to understand the source of this antagonism by visualizing transcription dynamics in living cells. We introduced 14 copies of the bacteriophage sequence PP7 in the 5' UTR of GAL10, a strongly galactose responsive gene. Each of the PP7 repeats forms an RNA hairpin that can be bound by a pair of PP7 coat proteins fused to GFP (Coulon et al. (2014); Lenstra et al. (2015), Figure 3A, top). This system allowed us to visualize the dynamics of GAL10 transcription upon galactose induction as fluorescence bursts arising from the transcription site (TS; Figure 3A, bottom). We found that expressing TS intensity as the ratio of spot to mean nuclear fluorescence could reliably account for the differences in PP7-GFP levels observed between strains (Figure S2D-H).

Given the burstiness of transcription, we hypothesized that two parameters could play a role in the diminished transcriptional output, namely burst size and frequency. We measured transcription fluorescence traces for 26 (wild-type), 14, and 12 CTDr strains (Figure 3B). CTD truncation decreased the intensity of fluorescence bursts (Figure 3C), suggesting a decrease in burst size. Also, truncation increased the time interval between bursts (Figure 3D), which is closely related to burst frequency. We could similarly observe this frequency decay by looking at whether the average cell was active or inactive over time for each strain (Figure S3A). These average traces also show that burst frequency remained roughly constant after activation, only declining towards the end likely due to photobleaching and mRNA-bound PP7-GFP nuclear export. The autocorrelation of normalized intensity traces increased in amplitude with CTD truncation, similarly supporting a decrease in burst frequency (Figure S3B-D). Differences in burst duration measured from this analysis were more subtle. This potential ambiguity could mean that burst duration is independent of size, or that the observed decay in TS intensity was influenced by the decay in burst frequency. Overall, the live transcription measurements suggested CTD length can simultaneously modulate burst size and frequency.

The transcriptional activity in strains with shorter CTDs was too weak to be visualized in these movies without incurring in phototoxic illumination. To circumvent this problem, we took a single snapshot per field of view with maxium laser intensity during a 20 minute window after 30 minutes of galactose induction. This approach additionally allowed us to obtain a better estimate of TS fluorescence. The fraction of active cells per field of view was impacted by CTD truncation (Figure 3E). We observed a transition similar to the growth phenotype in this assay, where the magnitude of the effect progressively increased with CTD truncation. We also observed a consistently moderate shift in the distributions of TS fluorescence with CTD truncation (Figure 3F). The comparatively small magnitude of this decrease suggested the decay in fraction of active cells is primarily driven by burst frequency. From these measurements, we conclude that CTD length can modulate both the size and the frequency of transcriptional bursting.

2.3. Fusion to disordered proteins can rescue the function of a CTD-truncated RNA Pol II

Given the conservation of protein disorder (Figure 1A) and recent evidence that the CTD can form and interact with phase separated droplets (Kwon et al., 2013; Chong et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019), we hypothesized that the function of the CTD's long disorder could be supplied by other proteins of similar chemical and structural features. We tested this idea by fusing the low complexity domains (LCD) of the human proteins FUS and TAF15, which are not present in the yeast genome, to the C-terminus of a 10CTDr truncated RPB1. These LCDs contain neither a known nuclear localization sequence (Gal et al., 2011; Marko et al., 2012) nor ubiquitination sites (Mertins et al., 2013) that could supplement CTD function in a predictable manner; they are similar in amino acid composition, particularly in the frequency of tyrosines, but share little sequence similarity with the CTD (Figure S4A-C).

Astonishingly, strains carrying either protein fusion showed an improved growth rate over the 10CTDr strain. Fusion to FUS LCD progressively rescued growth rates of strains with 9 and 8 CTDr. Furthermore, strains with 7 and 6 CTDr remained viable when fused to FUS, surpassing the minimum requirement of 8 CTDr alone (Figure 4A). This supression was particularly striking because the CTD has been extensively mutated, typically with detrimental effects to transcription or downstream processes. This new minimal length is also noticeably close to the four heptad repeats that directly contact Mediator in an assembled preinitiation complex (Robinson et al., 2012, 2016).

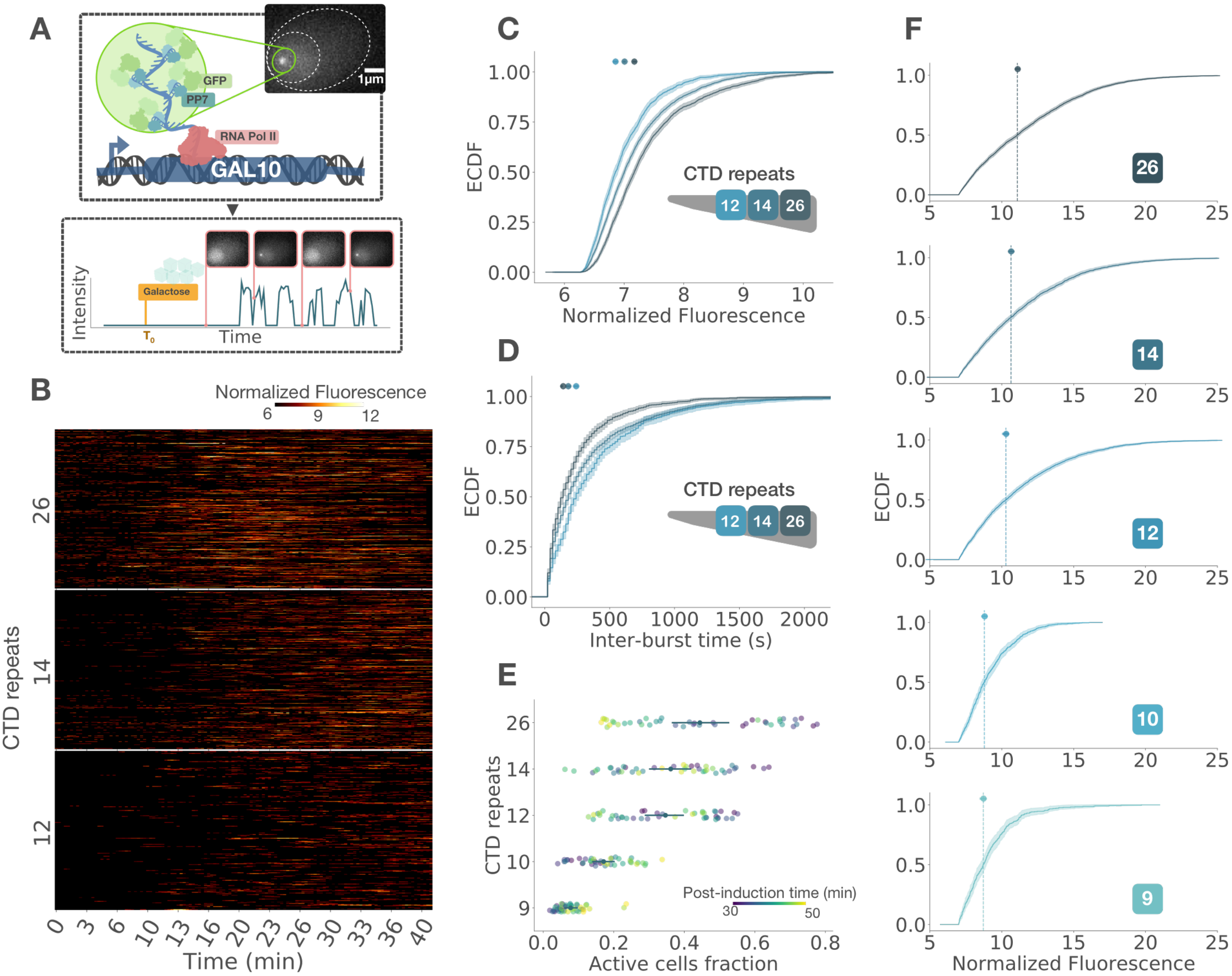


Figure 3: **CTD length modulates transcription burst size and frequency.** (A) Experimental strategy used to observe transcription dynamics in live cells. The 5' end of a single allele of galactose-inducible GAL10 is tagged with RNA hairpins that are bound by nuclear expressed GFP-PP7. This protein fusion results in a fluorescent spot in the cell nucleus upon transcription activation, whose fluorescence through time is recorded to investigate transcription dynamics. (B) Transcriptional traces by cell of strains with 26 (wild-type), 14, and 12 CTD repeats (CTDr) from three biological replicates. These are related to burst size and frequency through (C) the empirical cumulative distribution function (ECDF) of transcription site fluorescence intensities and the ECDF of inter-burst times, in seconds (D). Shaded area is bootstrapped 99% confidence interval (CI) and top markers show median with 99% CI. (E) Fraction of active cells per field of view from two biological replicates measured from high-laser-power snapshots of strains with 26, 14, 12, 10, and 9 CTDr. Middle points indicate mean with bootstrapped 99% CI. Color indicates time after galactose induction. (F) ECDFs of normalized fluorescence intensities of transcription bursts from these snapshots. Vertical dotted lines indicate median and shaded area bootstrapped 99% CI. Number of CTD repeats is indicated in the lower right corner of each plot.

We next probed whether the improved growth phenotype originated from a transcriptional rescue. We used RNA-seq to compare the transcriptomes of the FUS and TAF15 rescued strains and their response to

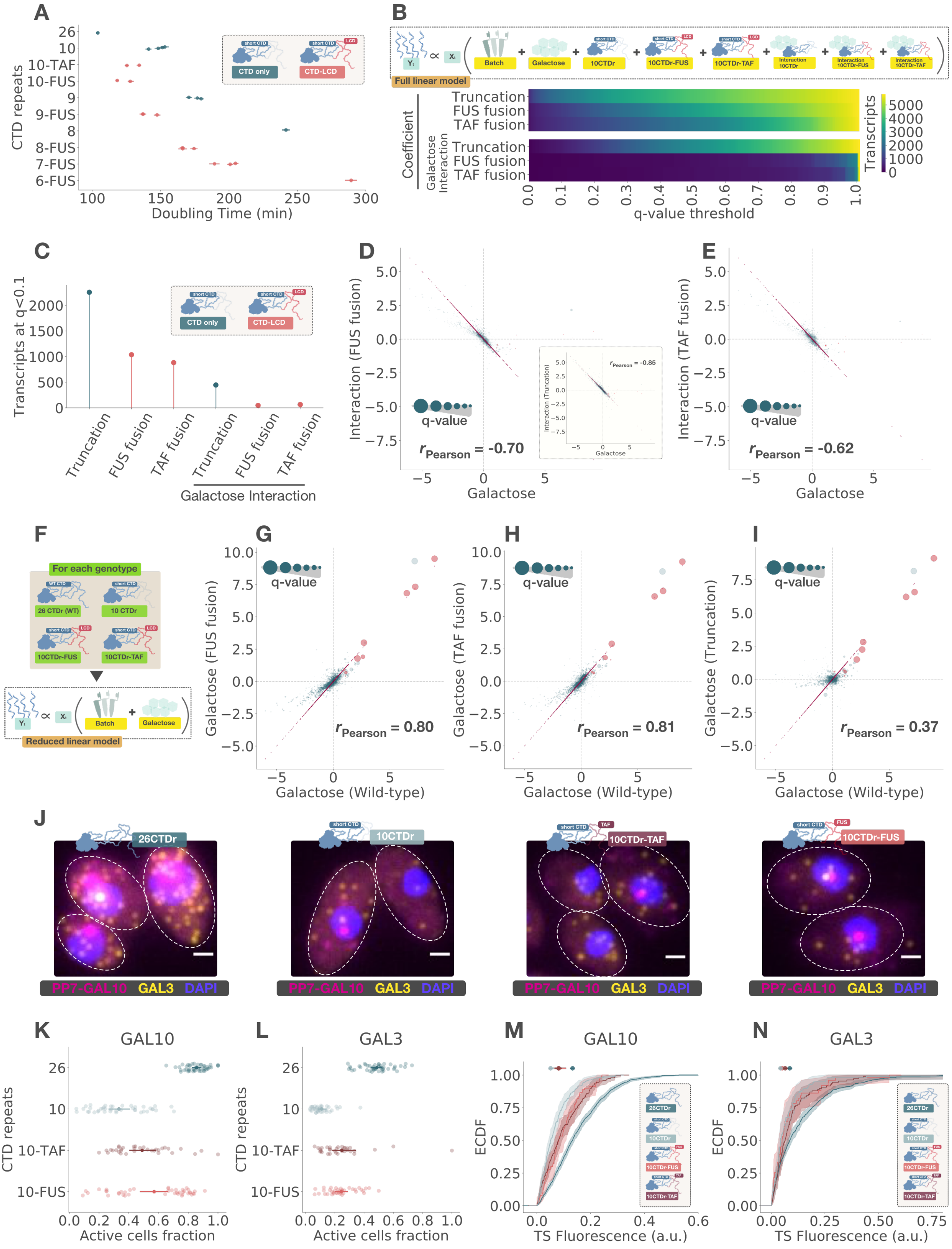

A
CTD repeats
26
10
10-TAF
10-FUS
9
9-FUS
8
8-FUS
7-FUS
6-FUS
Doubling Time (min)
100
150
200
250
300
short CTD
CTD only
short CTD
LCD
CTD-LCD
B
Yᵢ
∝
Xᵢ
Batch
Galactose
10CTDr
10CTDr-FUS
10CTDr-TAF
Interaction 10CTDr
Interaction 10CTDr-FUS
Interaction 10CTDr-TAF
Full linear model
Coefficient
Galactose Interaction
Truncation
FUS fusion
TAF fusion
Transcripts
5000
4000
3000
2000
1000
0
q-value threshold
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
C
Transcripts at q<0.1
2000
1500
1000
500
0
Truncation
FUS fusion
TAF fusion
Galactose Interaction
D
Interaction (FUS fusion)
q-value
$r_{Pearson}$ = -0.70
Galactose
Interaction (Truncation)
$r_{Pearson}$ = -0.85
E
Interaction (TAF fusion)
$r_{Pearson}$ = -0.62
F
For each genotype
WT CTD
26 CTDr (WT)
10 CTDr
10CTDr-FUS
10CTDr-TAF
Reduced linear model
G
Galactose (FUS fusion)
$r_{Pearson}$ = 0.80
Galactose (Wild-type)
H
Galactose (TAF fusion)
$r_{Pearson}$ = 0.81
I
Galactose (Truncation)
$r_{Pearson}$ = 0.37
J
26CTDr
10CTDr
TAF
10CTDr-TAF
FUS
10CTDr-FUS
PP7-GAL10 GAL3 DAPI
K
GAL10
L
GAL3
Active cells fraction
M
N
ECDF
TS Fluorescence (a.u.)
26CTDr
10CTDr
10CTDr-FUS
10CTDr-TAF

Figure 4:

Figure 4: **Fusion of the low-complexity domain (LCD) of FUS or TAF15 to a truncated polymerase can rescue its function and reduce the CTD length required to support cell growth.** (A) Comparison of doubling times (DT) of strains with wild-type and decreasing number of CTD repeats (CTDr) with and without fusion to the LCD of FUS or TAF15. Individual points come from independent lines when available and indicate mean DTs with standard error estimated from three biological replicates per line. (B) Top: Linear model used to estimate the effect of truncation to 10CTDr, subsequent LCD fusion, and the interaction of each of these components with galactose induction. Bottom: Resulting cumulative distributions of q-values, in number of transcripts, for each coefficient, excluding galactose and batch effects from three biological replicates. (C) Number differentially expressed transcripts detected at a q-value threshold of 0.1. Comparison of the log fold-change of each transcript induced by galactose and its interaction with 10CTDr fused with FUS (D) and TAF15 (E) LCDs. Inset shows comparison with 10CTDr interaction alone. (F) Reduced linear model used to measure galactose induction in each strain individually. Comparisons of log fold-change of each transcript induced by galactose in wild-type and 10CTDr (G), 10CTDr-FUS (H), and 10CTDr-TAF15 (I). Red points show the positions on the diagonal $x = y$. Marker size of each point is inversely proportional to the q-value of the coefficient in the y-axis ($ms = -log(q_y)$); dotted lines reference no change at zero and the Pearson correlation is indicated. Direct targets of GAL4 listed in Lesurf et al. (2016) are plotted in orange. (J) Representative images of smFISH with probes for a single allele of PP7-GAL10 and both alleles of GAL3 after two hours of galactose induction for 26CTDr (wild-type), 10CTDr, 10CTDr-TAF, and 10CTDr-FUS strains as indicated on top of each image. White dotted contours mark cell outlines. Scale bar is 1 μm. Fraction of active cells per field of view for GAL10 (K) and GAL3 (L) measured from two biological replicates of smFISH. Mean with 99% bootstrapped confidence interval (CI) is shown on top of each group. Corresponding empirical cumulative distribution functions (ECDF) with 99% bootrsapped CI of transcription site intensities of GAL10 (M) and GAL3 (N). Medians with 99% CI are shown on top.

galactose induction with that of the wild-type and 10CTDr strains. Using principal component analysis, we observed LCD fusion results in transcriptomes in-between that of the wild-type and truncated strains, under induced and uninduced conditions (Figure S5A). As described in Figure 2B, we fitted the data to a linear model to identify the contributions to each transcript of galactose induction, CTD truncation, FUS or TAF15 LCD fusion to a 10CTDr truncated polymerase, and their interaction with galactose (Figure 4B, top). We found the number of differentially expressed transcripts resulting from CTD truncation decreased from 2256 to 1037 and 883 transcripts at a q-value threshold of 0.1 upon fusion to FUS or TAF15 LCDs, respectively (Figure 4C). More generally, the distribution of q-values resulting from CTD truncation shifted towards less significant values (Figure 4B, middle), a sign of considerable amelioration in the transcriptional phenotype. These LCD fusions additionally shifted the distribution of q-values of the interaction between CTD truncation and galactose induction (Figure 4B, bottom), abolishing the measurable effect at a q-value of less than 0.1 (Figure 4C). Moreover, the global antagonism of this interaction term with galactose induction nearly vanished (Figure 4D,E).

We interrogated these data further using other linear models, the simplest of which consists of independently measuring the galactose induction in each strain (Figure 4F). In this comparison, the galactose response of most transcripts in the rescued strains closely resembles that of the wild-type (Figure 4G,H), more than the 10CTDr alone (Figure 4I). The FUS and TAF15 transcriptional phenotypes, as measured using the full model (Figure 4B, top), were highly correlated (Figure S5B). Based on this observation, we fitted the data to a model in which we consider the two rescued strains as a single LCD group by pooling their transcriptomes together (Figure S5C). This model increased the number of transcripts that we can confidently call differentially expressed at a q-value threshold of 0.1 to 1392 (Figure S5D). This effect supports the transcriptional rescue occurs through a single pathway, and that there is a CTD sequence-dependent signature that remains shared among the three 10CTDr strains. This signature was evident from the high similarity between the truncation and LCD fusion coefficients (Figure S5E). From this experiment, we conclude that the sequence and long disorder of the CTD have separable roles in transcription, the latter of which can be supplemented by the similarly disordered LCDs of FUS or TAF15.

For unknown reasons, our live imaging system did not work with the FUS rescued strains. We circumvented this issue by using two-color Single-Molecule Fluorescence *in situ* Hybridization (smFISH). We used probes against the PP7 repeats, allowing us to detect mRNA from a single allele of GAL10, and against GAL3, which could detect RNA from both of its alleles (Figure 4J). The fraction of active cells consistently increased for both the single allele of GAL10 and the two alleles of GAL3 (Figure 4K,L). In addition, we observed an increase in TS fluorescence intensity of both rescued strains over the 10CTDr strain (Figure 4L,M), suggesting the fraction of active cells increased specifically because of an increased burst size.

Together, these results show the CTD's long disorder can influence transcription in a way that depends on its chemical and structural properties rather than its precise amino acid sequence.

2.4. CTD, FUS and TAF15 LCDs can self-interact and this ability is necessary for efficient transcription

The CTD can bind the LCD of FUS and more strongly that of TAF15 (Kwon et al., 2013). TAF15 can also interact with components of the Mediator complex (Takahashi et al., 2011). However, the avidity of these interactions does not appear to correlate with the extent of rescue (Figure 4A). Other experiments have shown these LCDs are able to phase-separate (Kwon et al., 2013). These phases are thought to form as a result of intermolecular interactions that collectively drive droplet formation (Banani et al., 2017; Shin and Brangwynne, 2017). We hypothesized that FUS, TAF15, and the CTD could be involved in the recruitment of RNA Pol II to the TS via these self-interactions.

We first tested whether FUS and TAF15 variants, with tyrosine to serine misense mutations that make them significantly less able to bind phase-separated droplets *in vitro* (Kwon et al. (2013); data reproduced in Figure S4D), would fail to rescue the growth phenotype of a 10CTDr strain.

We observed a correlation between droplet binding rates and the extent of growth rescue upon fusion of these protein variants to the truncated RNA polymerase (Figure 5A,B). This rescue also correlated with the compromised ability of these proteins to function as transcription factors when fused to a DNA binding domain (Kwon et al., 2013).

We investigated these mutants more closely by using an assay designed to measure self-interactions *in vivo*, which we define as the ability of a protein to interact with and recruit others of its kind. We speculated that fusing a self-interacting protein to PP7-GFP would lead to brighter spots in our live transcription assay (Figure 5C, left). Heterologously expressed FUS and TAF15 can form punctate structures in yeast (Couthouis et al., 2011; Ju et al., 2011). However, in this assay wild-type FUS and TAF15 LCD fusions distributed uniformly across the cell nucleus, presumably due to lower protein concentrations. On the other hand, the spots that formed after galactose induction became brighter with either LCD compared to PP7-GFP alone. Moreover, more than a single spot became visible soon after transcription activation (Figure 5C, right).

We characterized this phenomenon via smFISH of galactose induced strains carrying PP7-GFP with and without FUS and TAF15 LCD fusion (Figure 5D). This experiment suggested that the increase in the number of spots and their brightness could be a result of 1) indirect recruitment of PP7-GFP-LCD to each mRNA scaffold but mostly 2) LCD-mediated physical interactions between mRNA molecules outside of the TS. We proceeded to use this assay to determine whether a protein can self-interact at physiological concentrations *in vivo*.

We counted the number of bright spots per cell arising 30 minutes after galactose induction in snapshots taken during a 20 minute window with a maximum intensity laser. This number is almost always 1 for PP7-GFP alone, corresponding to the TS, except for when a cell had just duplicated the GAL10 locus during cell division. A 10CTDr-PP7-GFP protein fusion mostly produced a single spot, similar to the non-self-interacting

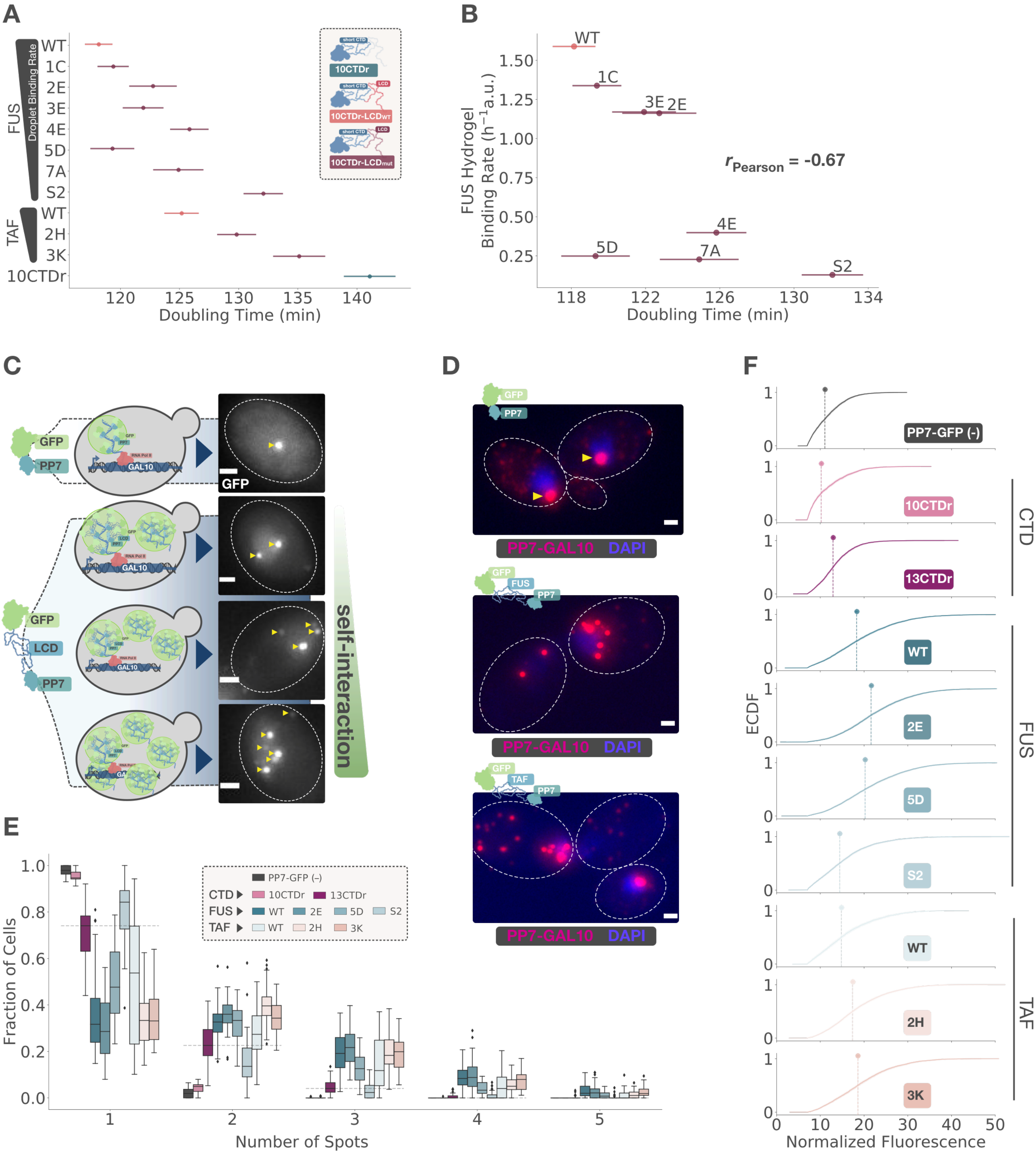


Figure 5: **CTD and the low complexity domains (LCD) of FUS and TAF15 can self-interact *in vivo* and this ability is necessary for the function of RNA Pol II.** (A) Doubling times (DT) of 10 CTD repeat (CTDr) strains with and without fusion to FUS or TAF15 wild-type and mutated LCDs sorted by droplet binding rate as reported in Kwon et al. (2013) and comparison of FUS mutants with these rates (B). Pearson correlation is indicated. Color indicates whether fused LCD is present, wild-type or mutant. Individual points come from independent lines when available and indicate mean DTs with standard error estimated from three biological replicates per line.

Figure 5: (C) Diagram of self-interaction assay. An LCD is fused to GFP and PP7 coat proteins; GFP-PP7 fusion with no LCD is used as control (top). Upon induction of transcription, these fusion proteins bind mRNA scaffolds. If LCDs can self interact, they increase the brightness of spots by recruiting more proteins to the scaffold, and by bringing more than one mRNA together outside of the active transcription site via LCD-LCD interactions (bottom). Representative snapshots of cells with increasing number of GFP fluorescent spots are shown in the right. (D) smFISH after 30 minutes of galactose induction with probes that bind PP7 hairpins in GAL10 mRNA on strains constitutively expressing PP7-GFP with and without LCD fusion as indicated in the upper left corner of each image. Cells without an LCD have a single nuclear RNA complex per cell corresponding to the transcription site (top, yellow arrowheads). Individual mRNA molecules are visible as dimmer spots. Fusion to FUS (middle) or TAF15 (bottom) LCD leads to the formation of multiple RNA complexes, visible as spots brighter than a single mRNA, in individual cells as a result of intermolecular LCD-LCD interactions. White dotted contours mark cell outlines; scale bar is 1μm. (E) Fraction of cells per field of view that contain each number of GFP spots by strain from three biological replicates. The horizontal blue dotted lines indicate the number of spots in the 13CTDr GFP-PP7 fusion. Protein fused to PP7-GFP is indicated by color. (F) Corresponding empirical cumulative distribution functions (ECDF) of GFP fluorescence intensities from the brightest spot in each cell, typically corresponding to the transcription site. Protein fused to PP7-GFP is indicated by color and in the lower right of each plot.

control. On the other hand, a 13CTDr and all of the FUS and TAF15 variants resulted in a higher fraction of cells with more than a single bright spot (Figure 5E).

We also compared the fluorescence intensity of the brightest spot per cell, presumably the TS, to the control expressing PP7-GFP only. Proteins that formed multiple bright spots also increased the intensity of the brightest spot (Figure 5F). The 10CTDr construct resembled the negative control more than 13 CTDr. Generally, these measurements support that 13CTDr, FUS, and TAF LCDs can self-interact, and the extent of self-interaction qualitatively recapitulates the transitions observed in our growth and transcription assays.

These results suggest that self-interactions are necessary for the transcriptional rescue of CTD truncation, and support the idea that this ability is a key attribute that the CTD's long disorder contributes to transcription.

2.5. An integrative transcription model explains the influence of CTD length

In light of our evidence, we sought to devise a quantitative model for transcription that captured the effects of perturbing CTD length and illuminated the role of disorder-mediated self-interactions. We built upon a model that includes an active and an inactive state, which enables it to produce transcriptional bursting, and specifies that an RNA polymerase molecule can only be recruited during the active state, in agreement with experimental data (Bartman et al., 2019).

The CTD is known to physically interact with Mediator (Thompson et al., 1993; Kim et al., 1994), a prevalent component of the preinitiation complex (PIC) (Allen and Taatjes, 2015). Given the repetitive nature of the CTD, we postulated that the number of repeats and hence CTD length could modulate the affinity with which the polymerase binds a Mediator-bearing PIC. Following this logic, we chose to explicitly refer to the active state as the assembled PIC, primed for RNA Pol II binding.

Polymerase release from the PIC is preceded by CTD phosphorylation (Payne et al., 1989; Svejstrup et al., 1997), which disrupts their physical interaction (Jeronimo and Robert, 2014; Wong et al., 2014). Assuming each of the CTD repeats contributes to this interaction, we reasoned that their number should correlate with the rate of polymerase release. Our rationale is that given the ratio of unphosphorylated to phosphorylated repeats determines the physicochemical state of the CTD and its interaction with Mediator (Robinson et al., 2016), the number of phosphorylation sites, or CTD length, should be proportional to the time it takes for CTD kinases to reach this threshold.

To recapitulate, we postulated that CTD length influences transcription by enhancing polymerase recruitment rate to the PIC (β) and by decreasing polymerase release rate (ϕ) from the PIC (Figure 6A, blue).

An important feature of the current model for transcription is that only a single polymerase molecule can bind the PIC at a given time (Bartman et al. (2019); Figure 6A, blue). To incorporate the ability of the CTD to self-interact, we postulated an additional molecular state that allows for the recruitment of more than single polymerase molecule to the extant PIC-Pol II complex via CTD-CTD self-interactions (Figure 6A, pink).

Finally, we used this model to assess the transcriptional rescue observed upon fusion of FUS or TAF15 LCDs to a CTD-truncated RNA polymerase. We reasoned that these LCDs would contribute the ability to self-interact, but make it more difficult for the polymerase to be released into the gene because their phosphorylation may not be as efficient. Specifically, we postulated that fusing a self-interacting LCD to a truncated CTD would fix self-recruitment rate ϵ and release rate ϕ, while the rate of initial recruitment β would still be determined by CTD length.

We interrogated the consistency of our models with experimental data using stochastic simulations. For simplicity, we set CTD length (CTD_L) to be a number in the range (0-1), directly proportional to both polymerase recruitment rates β and ϵ, whose complement ($1 - CTD_L$) is proportional to polymerase release rate ϕ. We visualized these simulations as transcriptional traces for each model, including states of PIC assembly, numbers of PIC-bound and phosphorylated polymerases (Figure 6B and Figure S6), akin to our live transcription imaging data (Figure 3B). From these simulations we computed the distributions of burst sizes, inter-burst times and fraction of active cells, and asked how CTD length would affect these parameters. Importantly, to be consistent with prior literature, we define the start and end of a burst to be concomitant with PIC assembly and disassembly, respectively, and only consider those that yield at least one mRNA molecule.

We compared the outcomes of the one-polymerase, the many-polymerases and the rescue models. The identity of the distributions of burst sizes and inter-burst times resembled geometric and exponential, respectively, for all the models (Figure S7). In addition, because the difference between models lies in the states after the start of a burst and hence does not influence burst frequency, the model choice did not affect the resulting distributions of inter-burst times (Figure 6C, S7C).

The three models predicted that shortening the CTD would progressively increase the average inter-burst time (Figure 6C). This in turn translated into a progressive decrease in the fraction of active cells (Figure 6D). Importantly, the magnitude of change in these numbers increased with decreasing CTD length. This effect qualitatively reproduced the observed transitions in growth and transcription phenotypes resulting from CTD truncation (Figure 2A, 3E).

Although dominated by burst frequency in this regime, the fraction of active cells was also influenced by burst size (Figure S8A). Our simulations predicted that such influence would lead to a modest but consistent increase in the fraction of active cells across a range of CTD lengths when comparing rescued with truncated CTD lengths (Figure 6D). In contrast to frequency, burst size was significantly influenced by model choice. The rescue model predicted a consistently higher mean burst size at almost every CTD length, except at the extreme of a long CTD (Figure 6E). These results are strikingly consistent with our experiments, in that rescued strains show a moderate increase in the fraction of active cells compared to truncated CTDs (Figure 4K,L), and TS intensity increased upon LCD fusion (Figure 4M,N). These comparisons of simulations with experiments support the existence of a state with more than one polymerase and that LCD fusion specifically rescues burst size via self-interactions.

The one-polymerase and the many-polymerases models produced very similar mean burst sizes with short CTDs. However, burst sizes resulting from each model deviated significantly with longer CTDs (Figure 6E). In particular, they increased exponentially when many polymerases were allowed to bind the PIC, but only linearly with a single polymerase. These predictions suggest the impact of binding more than one polymerase may become increasingly relevant as the CTD grows longer.

We also sought to elucidate the relationship between burst size, duration, and TS intensity in our exper-

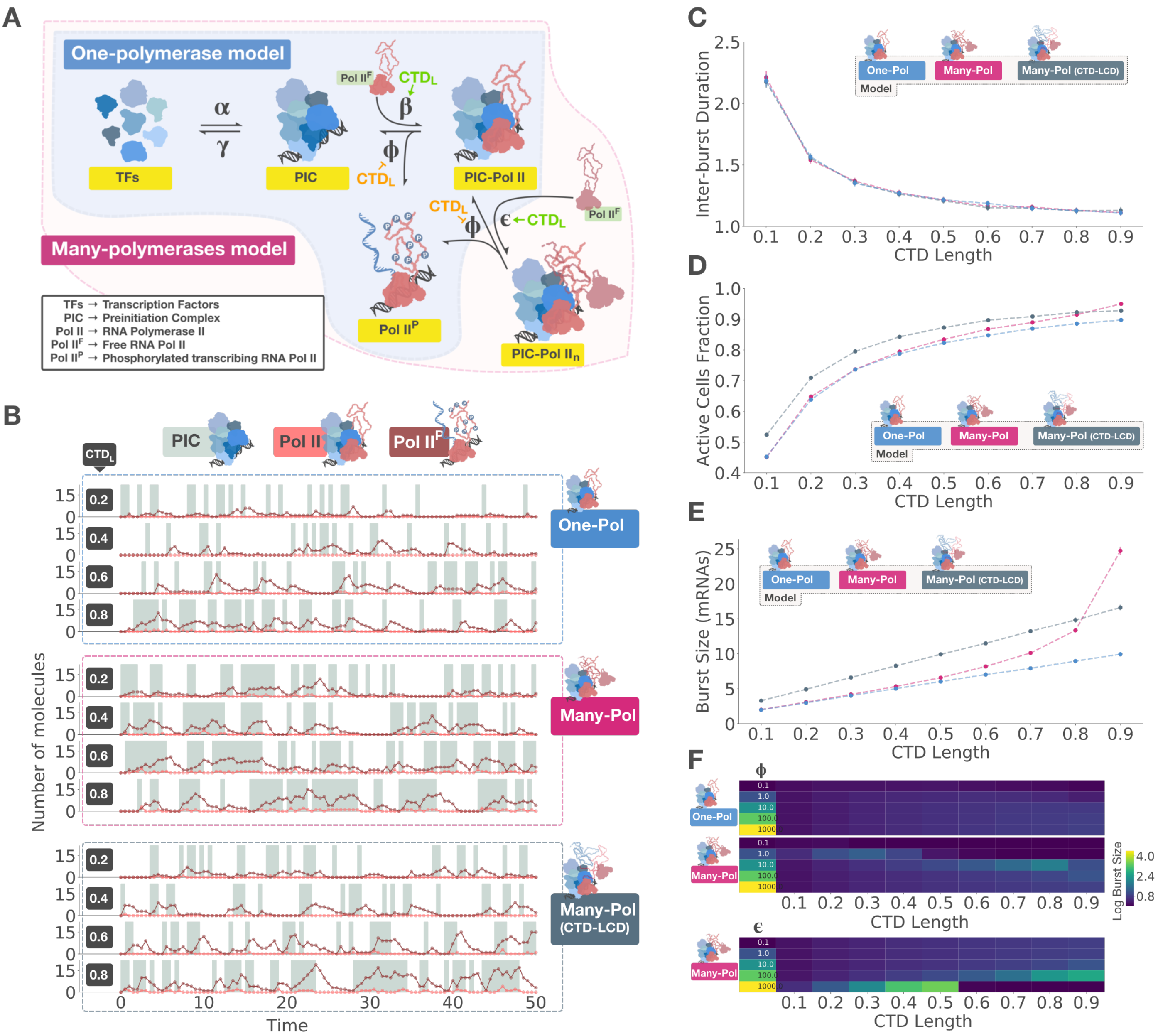


Figure 6: **An integrative model for transcription activation explains the role of CTD length.** (A) CTD-centric model for transcription activation. Transcription factors (TFs) assemble to form a preinitiation complex (PIC) to which an RNA polymerase binds. In the one-polymerase model (blue), only a single polymerase is allowed to bind the complex at a time. In the many-polymerases model (pink), more than one polymerase can bind the complex via CTD-CTD interactions. Fusion of a truncated CTD to a self-interacting protein is simulated using the many-polymerases model with fixed ϵ and ϕ. The positive or negative influence of CTD length (CTD_L) on each rate is indicated in green and orange, respectively. (B) Representative traces from stochastic simulations by model, indicated to the right, as a function of CTD_L, indicated in the left side of each plot. Background shows PIC assembly state; number of PIC bound and phosphorylated (transcribing) polymerases are shown in color as indicated in the legend on top. Corresponding mean inter-burst durations (C), mean fractions of active cells (D) and mean burst sizes (E) with 99% bootstrapped confidence interval. (F) Effect of varying ϕ (top) and ϵ (bottom) on the log burst size.

iments by comparing these three values in our simulations. Because CTD length inhibits release rate in the one-polymerase and the many-polymerases models, burst duration increased faster than burst size (Figure S8B); these parameters were linearly proportional only when release rate was fixed in the case of the rescue model. These results potentially explain why burst duration decays more subtly (Figure S3B-D) than TS intensity (Figure 3D,F). A potential source of uncertainty in measuring burst size is that the decay in TS intensity could be a result of decreased burst frequency, given that GAL10 is transcribed in highly frequent bursts that could overlap in time and inflate the measured intensity of a burst. We simulated how this intensity signal would vary across CTD lengths and under different elongation rates δ, which determines how long a given mRNA stays and contributes to the TS signal. We compared the fraction of active cells with the difference between this observed TS intensity, influenced by δ and measured as the distributions of peak intensities in the final simulated traces, and the true burst size, which we kept track of as we generated the traces and is independent of δ (Figure S8C). The trend in this analysis is that if δ is low, the fraction of active cells would remain high across CTD lengths and TS intensity would overestimate the true burst size because sequential bursts would overlap; if δ is large, the fraction of active cells would remain low, and TS intensity would underestimate the true burst size because intensity would decrease before the end of a burst. Both of these effects were enhanced with longer CTDs. The experimental range of active cells fractions (Figure 3E) suggests a scenario where the estimate from TS intensity lies between a slight overestimation to an underestimation of the true burst size; in the latter situation, inferring burst size from these data would be a conservative estimate. In the context of the canonical transcription model (Figure 6A), the polymerase binding rate β intrinsically links burst size and frequency. We thus conclude that the simplest explanation consistent with simulations, experiments and previous literature is that burst size and frequency both decrease with CTD truncation.

Varying each of the model parameters individually (Figure S8E,F) offered an additional insight. By modulating the rate of polymerase phosphorylation ϕ –conceivably in local nuclear environments that limit CTD kinase activity– or by increasing the rate of self-interaction ϵ, it is possible to dramatically increase burst size under the many-polymerases model (Figure 6F). This effect would be a direct consequence of increased concentrations of unphosphorylated Pol II, akin to recently reported droplets in cells (Chong et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019). Although not specified in our model, liquid-liquid phase separation may thus naturally emerge from this transcription logic.

Our CTD-centric models helped us understand our experimental observations and integrate them with prior knowledge. By simulating them, we were able to capture the behavior of transcription upon CTD truncation and subsequent fusion to LCDs, illuminating the role of CTD length and providing support for a novel molecular state where more than a single polymerase can bind the PIC.

3. Discussion

3.1. A CTD-centric model offers mechanistic insights into transcriptional bursting

RNA Polymerase II is essential and extremely conserved across eukaryotes, yet the amino acid length of its catalytic subunit varies dramatically (Figure 1A,B). As discussed above, the number of disordered amino acids in the CTD closely follows this length variation, increasing with genome size as coding sequences become more scattered (Figure 1C,D).

The influence of CTD length on transcription is consistent with a simple quantitative model based on known protein-protein interactions with Mediator, CTD phosphorylation and disorder-mediated self-interactions (Figure 6A). We explicitly assume that at this level of abstraction, the major, Poissonian source of stochasticity in transcription comes from the multimolecular assembly of the preinitiation complex (PIC) at the enhancer and the promoter, ocurring at the start of each burst and preceding Pol II recruitment. This assumption is motivated by the CTD's influence on both burst size and frequency (Figure 3C-F) and by extensive previous experimental evidence. PIC formation is a rate-limiting step in transcription (Kuras and Struhl, 1999;

Li et al., 1999) and transcription bursts are concomitant with enhancer-promoter interactions (Bartman et al., 2016; Chen et al., 2018). Roughly, the likely order of protein recruitment events upon activation is enhancer specific transcription factors, a single Mediator complex (Petrenko et al., 2016) and general transcription factors, and finally RNA polymerase followed by CTD kinases (Bryant and Ptashne, 2003; Krishnamurthy and Hampsey, 2009). In specifying our model, we put forward a view in which bursts are a result of recurrent Pol II binding to the assembled PIC, and inactivity periods a consequence of PIC disassembly (Figure 6A). This reductionist framework thus offers an intelligible perspective of the mechanism of eukaryotic transcriptional bursting.

3.2. CTD length cooperatively scales transcription

The probability of interaction between two genomic loci is highly dependent on the physical distance between them (Lieberman-Aiden et al., 2009). As a result, order-of-magnitude variation in genome sizes and in the physical spacing between enhancers and promoters represents a challenge: how does the transcription machinery overcome an increasingly infrequent event? Our simulations suggest increasing CTD length can reduce the number of times an assembled PIC fails to recruit RNA Pol II and produce mRNA before disassembly (Figure S8E), and that self-interactions can considerably increase burst size (Figure 6E,F). By exploiting each rare assembly event, CTD length-enhanced recruitment and self-interactions could contribute to resolve the transcription scaling paradox. This compensatory mechanism would complement changes in genome organization (Szabo et al., 2019), without which enhancer-promoter interactions may never occur in the first place.

In this context, CTD length bears an important distinction with the strength of self-interaction and polymerase recruitment rate, determined by interactions with the PIC. While the naive expectation is that these parameters should be correlated, devations from the consensus repeat YSPTSPS may provide a way to modulate them independently. This hypothesis could explain why fruit flies with a CTD of wild-type length that is made up entirely of consensus repeats do not survive, but animals with a yeast CTD remain viable (Lu et al., 2019). Based on this rationale, we speculate CTD length and sequence in eukaryotes coevolves with the physical spacing between enhancers and promoters, primarily determined by genome organization (Lieberman-Aiden et al., 2009; Szabo et al., 2019) and genome size.

3.3. CTD-CTD self-interactions link transcription activation to phase-separation

It is possible that FUS and TAF15 LCDs rescue CTD truncation through an alternative recruitment mechanism that does not involve self-interactions, given they can also function as transcription factors when fused to a DNA binding domain (Kwon et al., 2013). This hypothesis would nonetheless be consistent with our inference that CTD length modulates polymerase binding rate to the PIC, but it would not support a many-polymerases state. On the other hand, our data do not suggest rescue is driven by enhanced direct recruitment to the PIC, given the increase in fraction of active cells seems to be predominantly driven by burst size and not frequency (Figure 4K-N), while direct recruitment would enhance both parameters. Additionally, we find a correlation between LCD ability to bind liquid droplets (Figure S4D) and self-interact with the extent of phenotypic rescue upon fusing them to a truncated RNA Pol II (Figure 5A,C,D).

Self-interactions additionally offer a logical connection between the mechanism of transcription activation and liquid-liquid phase separation (LLPS). Our model predicts that when self-interaction strength is large or the rate of polymerase release is small, large transcription bursts could emerge (Figure 6F), implying a high local concentration of unphosphorylated polymerases. This environment has been observed in LLPS droplets at super-enhancers of live cells (Chong et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019). A corollary of this idea is that the average gene and the super-enhancer gene can both be transcribed using the same mechanisms, but only the latter would manifest LLPS droplets as an epiphenomenon of enhanced polymerase recruitment or kinase exclusion. Super-enhancers would then be at the extreme of the distribution of burst sizes, which is consistent with the observation of only a few droplets per cell whose number does not nearly match the total number of transcribed genes. In this scenario, LLPS could result in emergent behaviors whose understanding would require a different quantitave framework; our model may not apply to these CTD-lengths but could provide a useful expectation to compare them with. In other words, CTD length variation may result in regimes of transcription activation governed by different dynamics.

3.4. Self-interactions support a multi-polymerase complex

The key proposition of our model that allows the incorporation of self-interactions is the existence of a molecular complex that can bind more than one RNA Polymerase molecule (Figure 6A, pink). Short-lived Pol II clusters observed in mammalian cells that overlap with active transcription sites and whose duration correlates with mRNA output (Cisse et al., 2013; Cho et al., 2016) could be a direct observation of this event. On the other hand, Pol II pausing appears to negatively correlate with transcription initiation (Shao and Zeitlinger, 2017; Gressel et al., 2017), which could suggest that new polymerases may not be able to bind an occupied promoter. Distinguishing the perhaps differential ability of PIC-bound and paused Pol II's CTD to self-interact would be helpful to understand the relationship of this observation with a many-polymerases state.

Pol II is released from the promoter upon CTD phosphorylation (Jeronimo and Robert, 2014; Wong et al., 2014), based on which we argue that CTD length influences release rate. Along this line, depletion of yeast CTD-kinase Kin28 causes an upstream shift in Pol II occupancy along genes (Wong et al., 2014), with a pattern that resembles proximal-promoter accumulation in metazoans (Adelman and Lis, 2012) and is consistent with a defective promoter escape. A conspicuously similar shift was observed upon mutating CTD's serine 5 (Collin et al., 2019), the specific CTD residue phosphorylated for transcription initiation (Eick and Geyer, 2013; Harlen and Churchman, 2017). We find self-interactions correlate with the efficiency of transcription (Figure 5). A sensible interpretation of these experiments is that decreasing CTD-kinases or their activity on the CTD lead to increased RNA Pol II at the promoter by extending the time window for self-interaction mediated recruitment. These observations raise the hypothesis that promoter accumulation of Pol II in metazoans (Adelman and Lis, 2012), congruently not observed in yeast (Steinmetz et al., 2006), could be contingent on a higher phosphorylation release-threshold linked to a long CTD and a multi-polymerase complex. Experiments that directly measure the number of polymerases that can bind the PIC, and how CTD length influences RNA Pol II occupancy profile would be highly informative in this regard.

In summary, our study integrates experimental results and simulations to explain how CTD length influences transcription activation. We revise the current model of transcription by providing evidence that self-interactions are a key feature in this process, intrinsically linked to a state in which multiple polymerases can bind the PIC. This line of reasoning offers a sound connection between a reductionist, concrete transcriptional logic and the emerging perspective of phase-separation, generating testable hypotheses that will further clarify the functional and evolutionary relevance of CTD length variation.

4. Acknowledgements

We thank Mitchell Guttman, Matt Thomson and members of the Sternberg laboratory for helpful discussions, Heun J. Lee, Andres Collazo and Giada Spigolon for imaging assistance, Igor Antoshechkin and Vijaya Kumar for RNA-seq experiments, and Steven Mcknight and Masato Kato for reagents. This work was supported by the Howard Hughes Medical Institute with which PWS was an investigator, by the Gordon Ross Medical Foundation and the Benjamin M. Rosen Graduate Fellowships, by the Biological Imaging Center at the Caltech Beckman Institute, and by the Millard and Muriel Jacobs Genetics and Genomics Laboratory.

5. Author Contributions

Conceptualization, P.Q.C., P.W.S; Methodology, P.Q.C, P.W.S., T.L.L.; Software, P.Q.C.; Formal analysis, P.Q.C., T.L.L.; Investigation, P.Q.C.; Data Curation, P.Q.C.; Writing – Original Draft, P.Q.C., P.W.S.; Writing – Review & Editing, P.Q.C., P.W.S., T.L.L.; Visualization, P.Q.C.; Supervision, P.W.S.; Funding Acquisition, P.W.S.

6. Declaration of interests

The authors declare no competing interests.

7. Methods

7.1. Data analysis

Except when indicated, all programming, data extraction, wrangling, calculations and plotting were done using Python 3.7 with standard scientific libraries (Oliphant, 2007; Jones et al.; Millman and Aivazis, 2011). All scripts used in this paper are available in the following github repository: https://github.com/WormLabCaltech

7.2. Image analysis

Maximum-intensity projections were used for all z-stack images, sometimes generated and often visualized using Fiji (Schindelin et al., 2012).

Cells were segmented using local thresholds and the Watershed algorithm. Candidate 2D fluorescent peaks were detected and tracked using Trackpy (Allan et al., 2018) with minor adaptations.

For PP7 transcription dynamics imaged with low laser intensity, only the brightest peak per cell per frame was kept. A Gaussian-Process Classifier (GPC) trained with a set of manually classified images was then used to distinguish transcription sites from spurious peaks, only keeping those with a GPC probability of at least 0.5 (Figure S2A-C). Transcription intensity was expressed as the fold-change of peak over mean nuclear fluorescence. This metric yielded overlapping intensity distributions of the same strain imaged with different settings (Figure S2D-H). Autocorrelation analysis was carried out as previously described (Lenstra and Larson, 2016). Missing timepoints where no peak was detected were imputed using the intensity at the position of the previous spot. For snapshots and smFISH images taken with maximum laser intensity, in which signal-to-noise ratio was greater, manually determined intensity thresholds were used.

7.3. RPB1 bioinformatic analysis

RPB1 homologs were retrieved by searching Ensembl database (Zerbino et al., 2018) using the HMMER online tool (Finn et al., 2011) with default settings, starting with the yeast RPB1 protein sequence. Amino acid sequences were analyzed for disorder locally using MobiDB-lite (Necci et al., 2017), which provides a consensus score derived from eight disorder predictors, in a machine running Unix Debian 4.9 and Python 2.7. Genome sizes and gene numbers were scraped from Ensembl websites using a custom script.

7.4. Genetic constructs

All constructs used in this paper were built using PCR amplification, Gibson (Gibson et al., 2009) or golden gate (Engler et al., 2008) assembly methods and verified by Sanger sequencing. Plasmids are listed in Table S1.

Wild-type LCDs of FUS (residues 1-214) and TAF15 (residues 1-208) were as previously defined (Kwon et al., 2013) and obtained by PCR amplification from human cell line 293T cDNA. Plasmids with coding sequences of previously reported FUS and TAF15 LCD tyrosine-to-serine mutants (Kwon et al., 2013) were a gift from Steven Mcknight.

The DNA sequence for CTD truncation repair templates was redesigned to facilitate PCR amplification, and together with yeast codon-optimized mScarlet coding sequence, synthesized as an Integrated DNA Technologies (IDT) gBlock and cloned into their respective vectors. sgRNAs were purchased as individual oligos, hybridized and cloned into pWS082 using golden gate assembly.

7.5. Strain Engineering

All transformations were carried out using the LiAc/SS Carrier DNA/PEG method (Daniel Gietz and Woods, 2002). Strains are listed in Table S2.

Strain YTL047A (Donovan et al., 2019) was generated by transforming diploid *S. cerevisiae* BY4743 with a PCR product containing the PP7 loop cassette and a loxP-kanMX-loxP marker, which was subsequently removed with Cre recombinase. A single allele of GAL10 was tagged. All strains used in this study are derivatives of YTL047A.

RPB1 modifications were engineered in both alleles using CRISPR-Cas9 with gRNAs of improved stability (Ryan et al., 2014), antibiotic-mediated selection of cells proficient in gap repair (Horwitz et al., 2015), and plamids from the Yeast MoClo Toolkit (Lee et al., 2015) modified by Tom Ellis's lab. sgRNA sequences are listed in Table S3.

CTD truncations, mScarlet and LCD RPB1 strains were generated by transforming YTL047A with 100 ng of BsmBI linearized and gel-purified Cas9-kanR plasmid (pWS173), 200 ng of each EcoRV linearized sgRNA vector (pWS082 derivatives), and 2-5 ug linearized repair template, selected for with G418. Correctly modified strains were identified using PCR of zymolyase digested colony scrapes followed by Sanger sequencing.

Strains for live transcription imaging were generated by integrating a single copy of GFP-Envy (Slubowski et al., 2015) fused to PP7 coat protein under an rpl15A promoter and a functional ura3 gene into the ura3Δ0 locus by transforming PacI linearized pTL174 and selected for with plates lacking uracil. Strains for self-recruitment assays (Figure 5) were constructed in the same way, except integrating PP7-LCD-GFP fusions. smFISH of self-recruitment assays was done on YTL047A transformed with plasmids pTL092, pQC075 or pQC076 (Figure 5D).

7.6. Cell growth measurements

Optical density (OD) was measured at an absorbance wavelength of 600 nm for 16 to 24 hours every 15 minutes using 1:100 dilutions of overnight cultures in 150 uL of YPD in a Falcon flat-bottom 96-Well Clear Assay Plate with lid on a Biotek Cytation 3 microplate reader with 1000 rpm shaking at 30C.

Doubling times were estimated non-parametrically from time derivatives of OD measurements with Gaussian processes using previously described software (Swain et al., 2016).

7.7. Live fluorescence microscopy

All microscopy experiments were done using early to mid-log cultures (typically 5e6 to 1e7 cells/mL) growing at 30C with 250 RPM shaking.

mScarlet-RPB1 strains were imaged on 2% agarose pads on coverslips at room temperature immediatly after spinning down at 3600 RCF cultures growing in Synthetic Complete (SC) 2% Glucose media on a Zeiss Imager Z2 microscope with an Axiocam 506 Mono camera, 63x oil objective, 150ms exposure time and 25% laser intensity.

Live transcription and self-interaction imaging was done on concanavalin-A-coated MatTek dishes at 30C as previously described (Lenstra and Larson, 2016) using an Leica DMI 6000 wide-field fluorescence microscope with an Andor Zyla 5.5 or a Hamamatsu Flash 4.0 v3 camera with a 100x oil objective. Cells were induced by adding galactose dissolved in 2 mL SC to 1 mL SC 2% raffinose for a final 3 mL SC 2% galactose and imaged immediately every 20 sec for around 1 hour (live transcription) or after 30 min for 20 min (self-interaction). Live movies were taken with 150 ms exposure, 9 z-stacks every 0.5 μm, and minimal laser intensity to avoid photo-toxicity. Snapshots were imaged once per field-of-view with maximum laser power, 150 ms exposure, and 9-15 manually set z-stacks every 0.5 μm.

7.8. RNA-seq

RNA was extracted from mid-log cultures growing in SC 2% raffinose after 2h of 2% galactose or blank induction using Zymo Quick-RNA Fungal/Bacterial Microprep Kit (Catalog # R2010) lysed in an MP Biomedicals FastPrep-24 machine.

RNA integrity was assessed using RNA 6000 Pico Kit for Bioanalyzer (Agilent Technologies #5067-1513) and mRNA was isolated using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB #E7490). RNA-seq libraries were constructed using NEBNext Ultra II RNA Library Prep Kit for Illumina (NEB #E7770) following manufacturer's instructions. Briefly, mRNA isolated from 1 μg of total RNA was fragmented to the average size of 200 nt by incubating at 94C for 15 min in first strand buffer, cDNA was synthesized using random primers and ProtoScript II Reverse Transcriptase followed by second strand synthesis using NEB Second Strand Synthesis Enzyme Mix. Resulting DNA fragments were end-repaired, dA tailed and ligated to NEBNext hairpin adaptors (NEB #E7335). After ligation, adaptors were converted to the 'Y' shape by treating with USER enzyme and DNA fragments were size selected using Agencourt AMPure XP beads (Beckman Coulter #A63880) to generate fragment sizes between 250 and 350 bp. Adaptor-ligated DNA was PCR amplified followed by AMPure XP bead clean up. Libraries were quantified with Qubit dsDNA HS Kit (ThermoFisher Scientific #Q32854) and the size distribution was confirmed with High Sensitivity DNA Kit for Bioanalyzer (Agilent Technologies #5067- 4626). Libraries were sequenced on Illumina HiSeq2500 in single read mode with the read length of 50 nt following manufacturer's instructions. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4.

RNA-seq quantification was performed using Kallisto (Bray et al., 2016) with 200 bootstraps in sigle-end mode with average length of 300 bp and standard deviation of 20 bp. Differential expression analysis was done with Sleuth (Pimentel et al., 2017) using the linear models described in the results and supplementary sections.

7.9. smFISH

smFISH experiments were carried out as described previously (Trcek et al., 2012; Lenstra et al., 2015). TYE665 labeled PP7 probes were purchased from IDT. A set of 48 Quasar 570 labeled probes were designed to target the coding sequence of Gal3 and purchased from Biosearch Technologies. Probe sequences are listed in Table S3.

Mid-log yeast cultures were fixed with paraformaldehyde and permeabilized with lyticase. Hybridization solution with 0.1 uM probes, 10% dextran sulfate, 10% formamide, and 2x Sodium Saline Citrate (SSC) was used to hybridize probes in fixed cells for 4 hours at 37 C. Coverslips were washed twice for 30 min with 10% formamide, 2x SSC at 37C, followed by rinses with 2x SSC, and 1x PBS for 5 minutes. PLL-coated 18 mm diameter #1.5 thickness coverslips were purchased from Neuvitro, mounted on microscope slides using ProLong Gold or Glass Antifade Mountant with DAPI (Life Technologies).

smFISH samples were imaged at room temperature with maximum laser power, 300 ms exposure and 9-15 manually set z-stacks every 0.2 μm on the Leica microscope with 100x objective described above.

7.10. Stochastic simulations

Stochastic simulations were performed using software described in Bois and Elowitz (2019) with minor modifications to extract burst start, end and size while generating Gillespie samples. Rates were chosen according to Bartman et al. (2019), with $\alpha = 1$, $\gamma = 3$, $\beta = 30$, $\epsilon = 10$ and $\phi = 100$. For trace visualization purposes, a rate of phosphorylated Pol II removal (elongation rate) $\delta = 1$ was used.

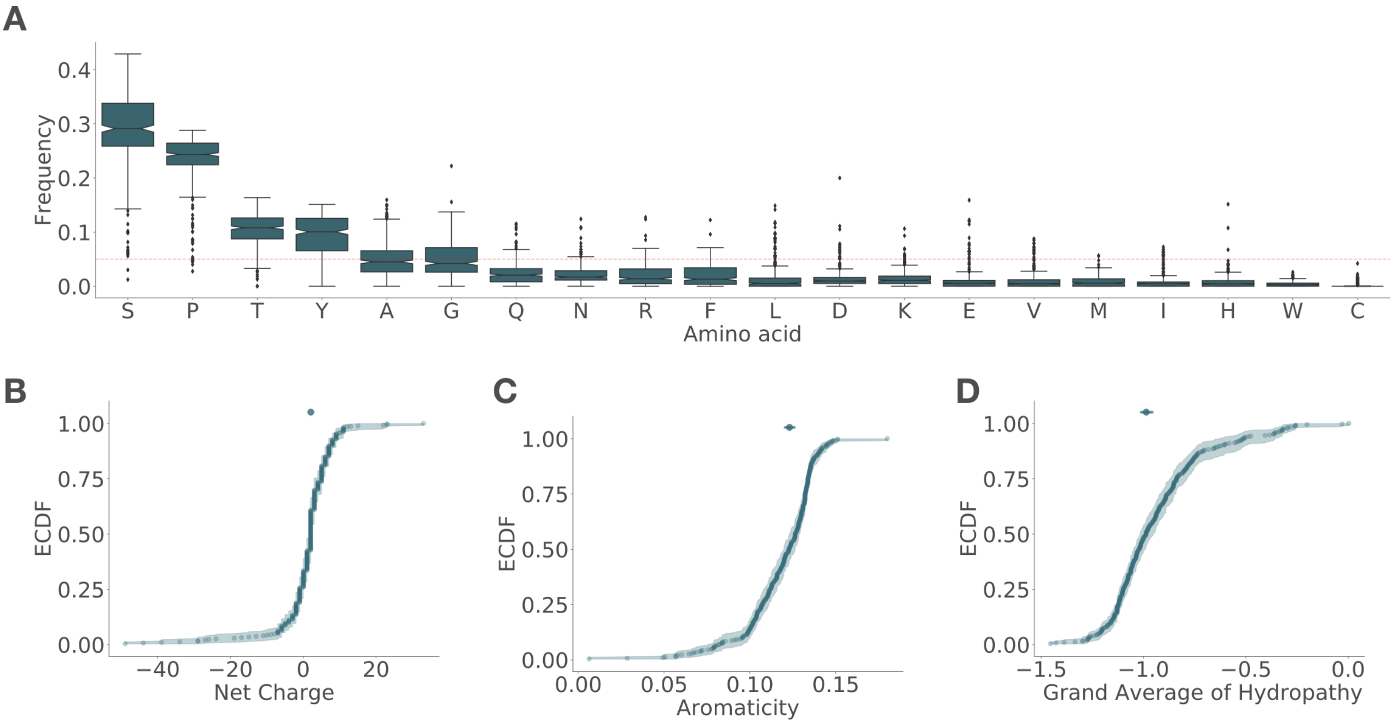


Figure S1: **CTDs share amino acid composition. Related to Figure 1.** CTDs were identified as the longest contiguous disordered region in RPB1 sequences. Only the longest protein per genus was considered. (A) Amino acid frequency sorted by mean abundance. Red dotted horizontal line indicates a uniform amino acid frequency of 1/20. Empirical cumulative distributions (ECDF) of net charges (B), aromaticity (C) and hydrophobicities (D) based on the grand average of hydropathy score (Kyte and Doolittle, 1982). Shaded area is bootstrapped 99% confidence interval (CI) and top markers show median with 99% CI.

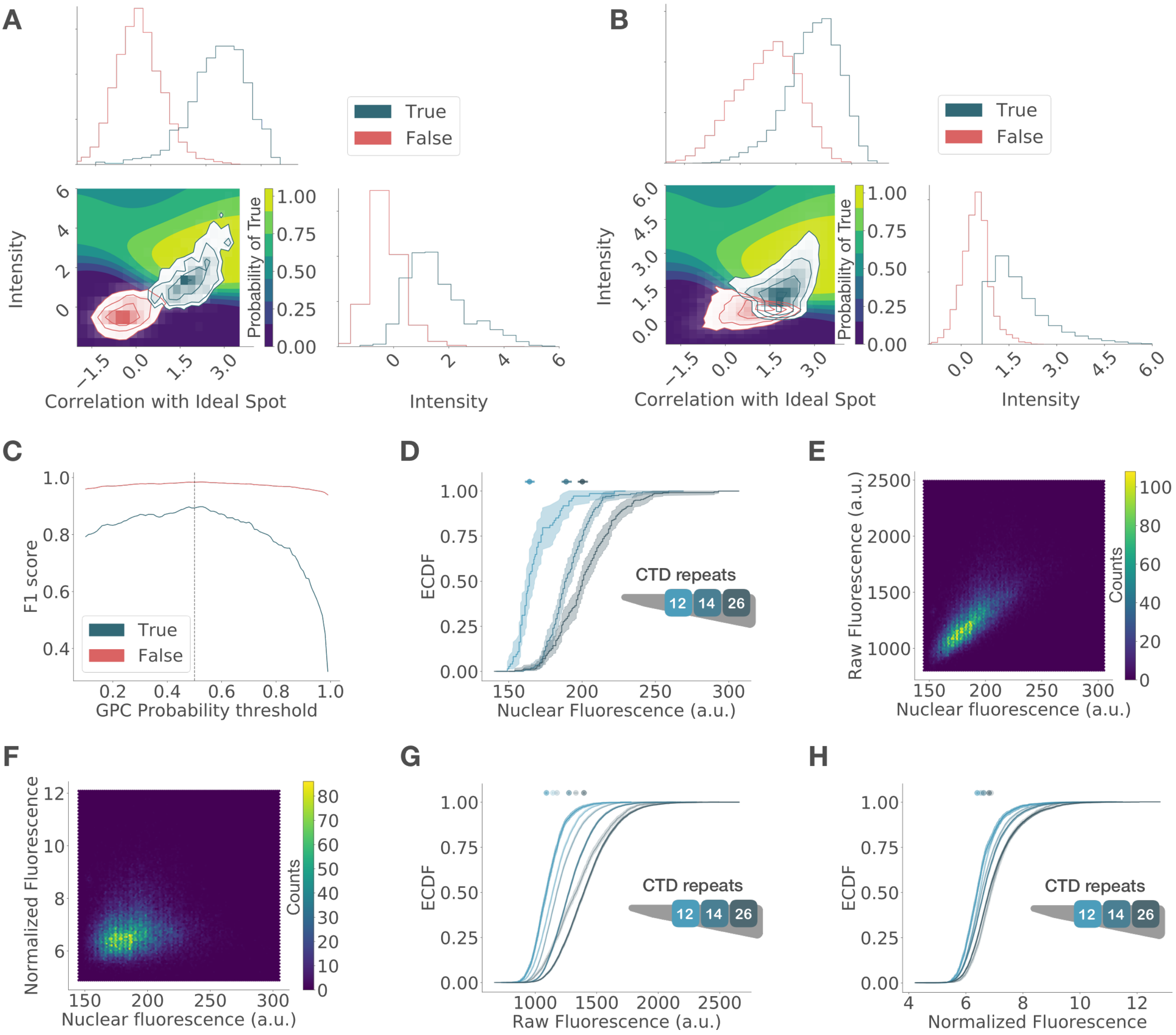


Figure S2: **Classification and normalization of PP7-GFP spots enables quantification of transcription dynamics and cross-strain comparisons. Related to figure 3.** Candidate spot images were obtained automatically using Trackpy's (Allan et al., 2018) peak detection algorithm. A sample of this image set was manually classified as True or False. For classification, spot images were represented using two features: correlation with an ideal spot (a single light point source blurred with a 2D Gaussian function) and intensity. (A) Histograms show the distribution of correlations (top) and intensities (right) of manually labeled spots. Left corner plot shows the joint distributions. This 2D data set was used to train a Gaussian-Process Classifier (GPC), resulting in the decision surface shown underneath, whose color indicates the probability of being a true spot. Candidate spots with a GPC probability above 0.5 were classified as True (B). This threshold was determined based on the change in the accuracy of classification (C), measured using the F1 score on a test set. The vertical dotted line indicates this probability threshold. Mutant strains show different PP7-GFP expression levels, as seen in the empirical cumulative distribution functions (ECDF) of mean nuclear fluorescence by strain (D). These differences result in a correlation observed in the hexagonal bin plot comparing mean nuclear fluorescence with raw spot fluorescence (E), which is removed after normalization (F). Normalized fluorescence is the ratio of spot fluorescence over mean nuclear fluorescence.

Figure S2: The efficacy of normalization can also be seen in the ECDFs of raw burst fluorescence by strain imaged with two laser intensities that artificially shift the intensity distributions of the same strains (G), which overlap after normalization (H). Transparency is used to indicate a different laser intensity. Shaded area is bootstrapped 99% confidence interval (CI) and top markers show median with 99% CI.

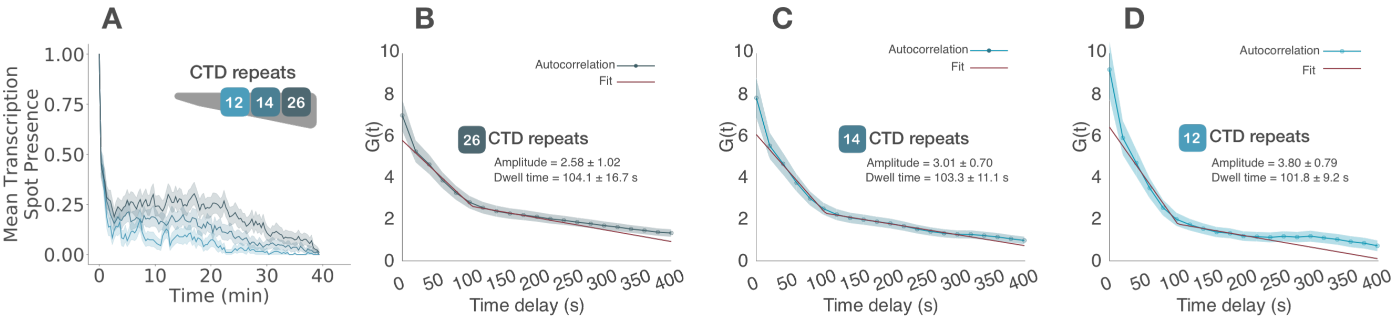
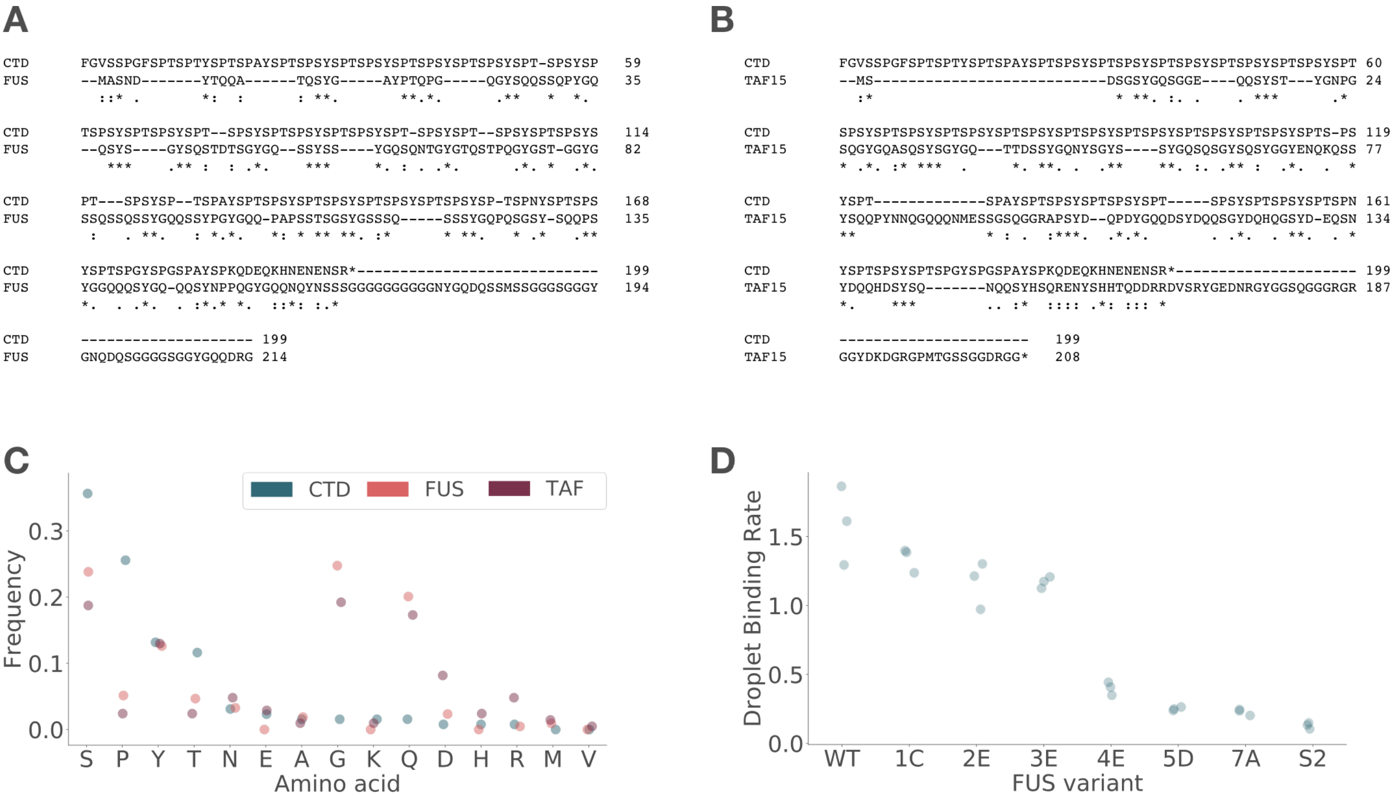


Figure S3: **Transcription burst frequency remains constant after activation and decreases with CTD truncation. Related to Figure 3.** (A) Mean aligned GAL10-PP7 boolean transcription traces. Boolean traces were obtained by marking with 1 and 0 the presence or absence of a transcription spot (TS), respectively. These traces were aligned and trimmed to begin with the first appearance of a TS and averaged over time, only considering cells that were active during the movie. These traces show the average frequency remains mostly constant over time and decreases with CTD length. Shaded area is bootstrapped 95% mean confidence interval. Frequency decay is also evident from an increase in amplitude, inversely related to frequency, in the autocorrelation of intensity traces corrected for non-steady-state effects in wild-type (B), 14 (C) and 12 (D) CTDr strains. Shaded area indicates standard error of the mean.



```
CTD      FGVSSPGFSPTSPTYSPTSPAYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPT-SPSYSP   59
FUS      --MASND-------YTQQA------TQSYG-----AYPTQPG-----QGYSQQSSQPYGQ   35
           ::* .       *:  :      : **.       **.*.      .**   *  *.

CTD      TSPSYSPTSPSYSPT--SPSYSPTSPSYSPTSPSYSPT-SPSYSPT--SPSYSPTSPSYS   114
FUS      --QSYS----GYSQSTDTSGYGQ--SSYSS----YGQSQNTGYGTQSTPQGYGST-GGYG   82
            ***    .** :  : .*.    ***     *. : . .*.      .*. *  .*.

CTD      PT---SPSYSP--TSPAYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSP-TSPNYSPTSPS   168
FUS      SSQSSQSSYGQQSSYPGYGQQ-PAPSSTSGSYGSSSQ-----SSSYGQPQSGSY-SQQPS   135
          :   . **.   : *.*.   *: * ** **. :*      * **.   * .*   .**

CTD      YSPTSPGYSPGSPAYSPKQDEQKHNENENSR*----------------------------   199
FUS      YGGQQQSYGQ-QQSYNPPQGYGQQNQYNSSSGGGGGGGGGGNYGQDQSSMSSGGGSGGGY   194
         *.  . .*.  . :*.* *.  ::*: :.*

CTD      -------------------- 199
FUS      GNQDQSGGGGSGGYGQQDRG 214
```

```
CTD        FGVSSPGFSPTSPTYSPTSPAYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPT 60
TAF15      --MS---------------------------DSGSYGQSGGE----QQSYST---YGNPG 24
             :*                            * **. :. .    . ***      .*

CTD        SPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTS-PS 119
TAF15      SQGYGQASQSYSGYGQ---TTDSSYGQNYSGYS----SYGQSQSGYSQSYGGYENQKQSS 77
           * .*. :* ***  .     *. **. .  .**    **. :. .** :  .*.  .  *

CTD        YSPT-------------SPAYSPTSPSYSPTSPSYSPT-----SPSYSPTSPSYSPTSPN 161
TAF15      YSQQPYNNQGQQQNMESSGSQGGRAPSYD--QPDYGQQDSYDQQSGYDQHQGSYD-EQSN 134
           **               * : .  :***.  .*.*.       . .*.  . **.  . *

CTD        YSPTSPSYSPTSPGYSPGSPAYSPKQDEQKHNENENSR*--------------------- 199
TAF15      YDQQHDSYSQ-------NQQSYHSQRENYSHHTQDDRRDVSRYGEDNRGYGGSQGGGRGR 187
           *.    ***        .. :*  :::: .*: ::: *

CTD        ----------------------   199
TAF15      GGYDKDGRGPMTGSSGGDRGG*   208
```

Figure S4: **FUS and TAF15 low complexity domains (LCD) are different in sequence but similar in amino acid composition to the CTD. Related to figures 4 and 5**. Protein alignments of FUS (A) and TAF15 (B) LCDs with yeast CTD. (C) Amino acid frequency in each of these proteins, sorted by CTD frequency. Only amino acids present in at least one protein are shown. (D) *In vitro* droplet binding rates of FUS variants used in this study. These numbers are the slopes obtained from a linear regression of LCD-GFP binding to wild-type FUS LCD droplets, measured as droplet fluorescence intensity over time. Each point is an experimental replicate; data are from Kwon et al. (2013).

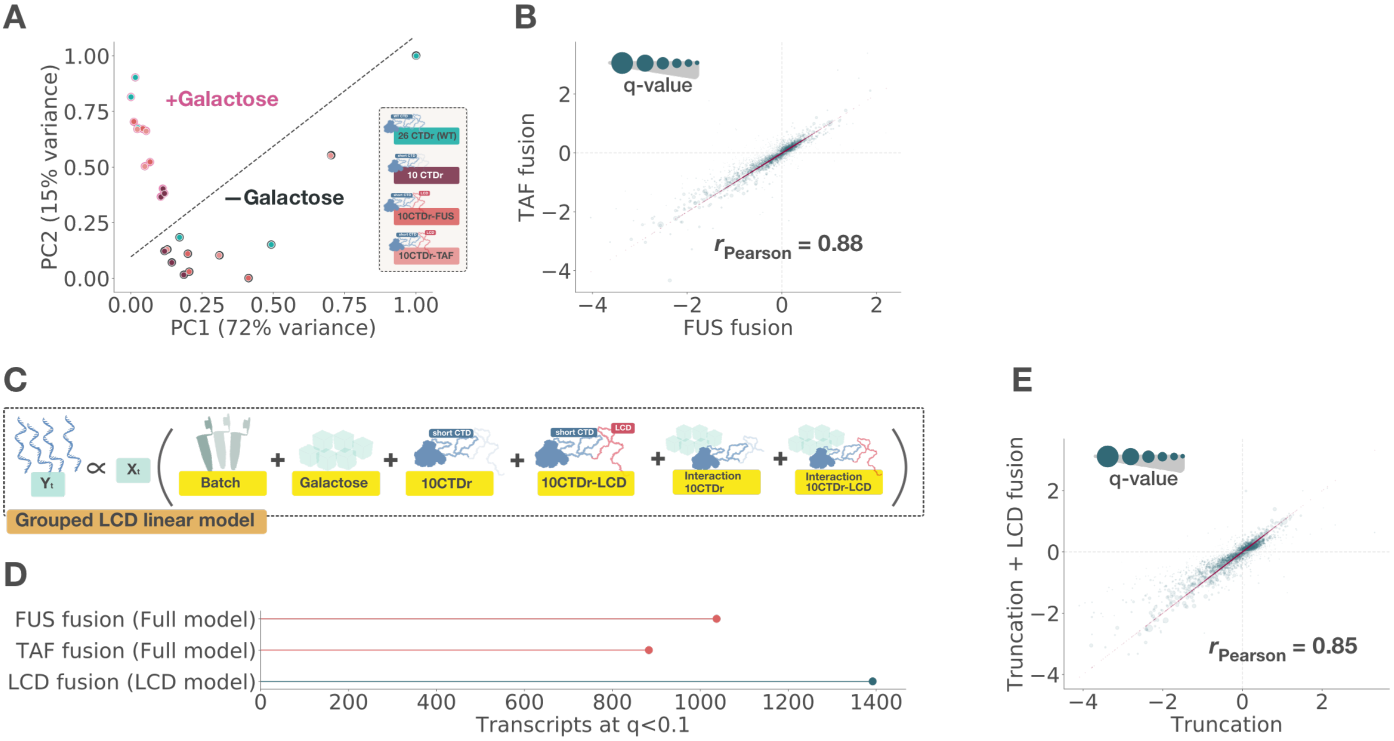


Figure S5: **Fusion of a CTD-truncated polymerase to FUS or TAF15 low complexity domains (LCD) results in convergent transcriptomes. Related to figures 2 and 4.** (A) Principal compoment analysis (PCA) with the first two PCs scaled to the range [0,1], which together explain 87% of the variance. Each strain has three biological replicates and two conditions. Marker edge color indicates the presence (pink) or absence (black) of galactose in the media; these groups are also divided by the dotted diagonal line. (B) Comparison of the log fold-change of each transcript resulting from FUS and TAF LCD fusion to 10CTDr truncated RNA Pol II under the full linear model shown in Figure 4B. Red points show the positions on the diagonal $x = y$. Marker size of each point is inversely proportional to the q-value of the interaction ($ms = -log(q_{int})$); dotted lines reference no change at zero and the Pearson correlation is indicated. (C) Alternative linear model where FUS and TAF rescued strains are grouped together. This grouping results in a higher number of genes identified for LCD fusion under a q-value threshold of 0.1 than for individual coefficients (D). Using this model, (E) comparison of the log fold-change of each transcript resulting from truncation with and without LCD fusion.

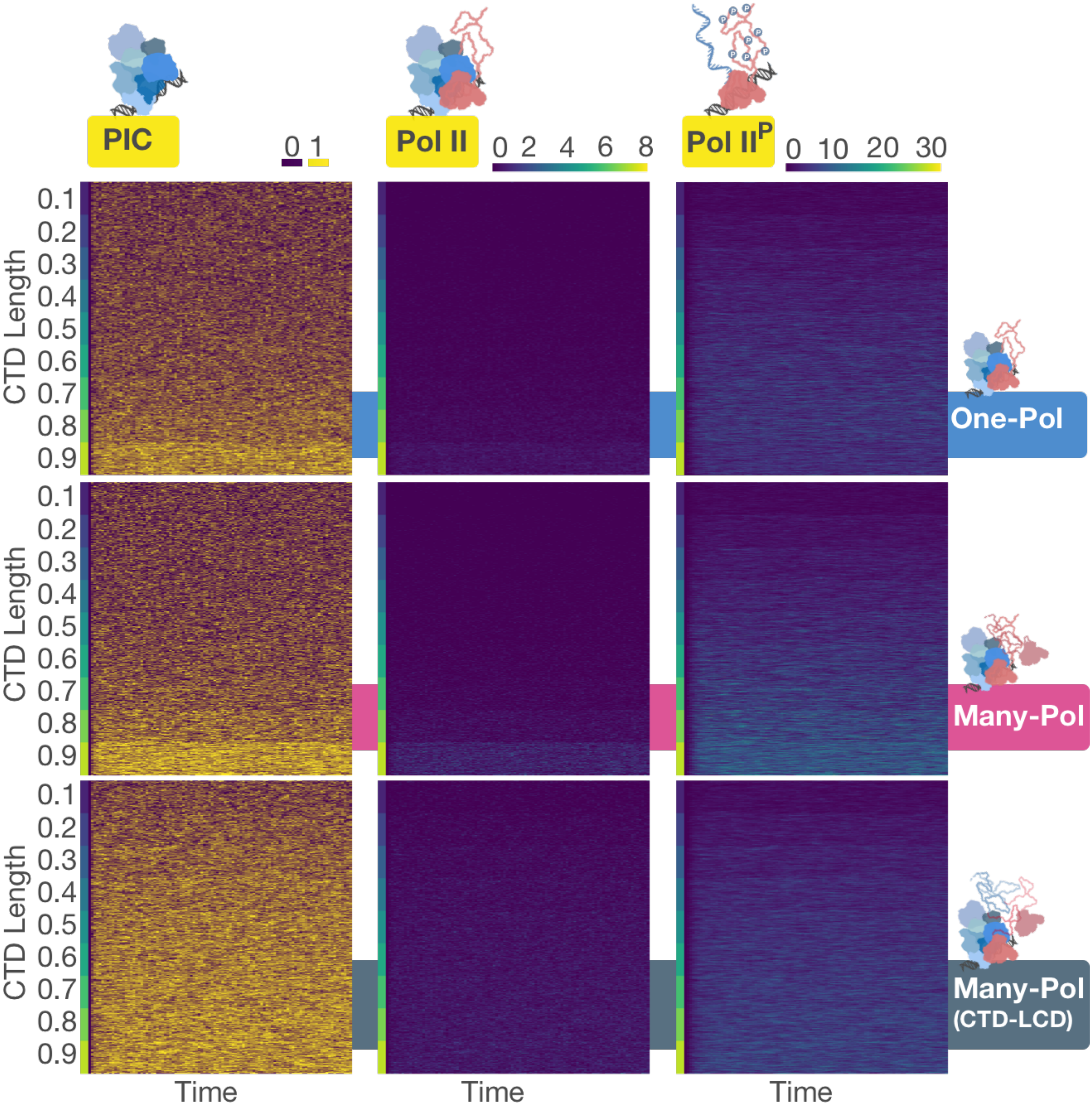


Figure S6: **Gillespie simulations yield traces akin to live transcription imaging. Related to figure 6.** Traces from stochastic simulations of PIC assembly states, number of PIC bound and phosphorylated (transcribing) polymerases for each model as a function of CTD_L, indicated with a colorbar to the left of each panel.

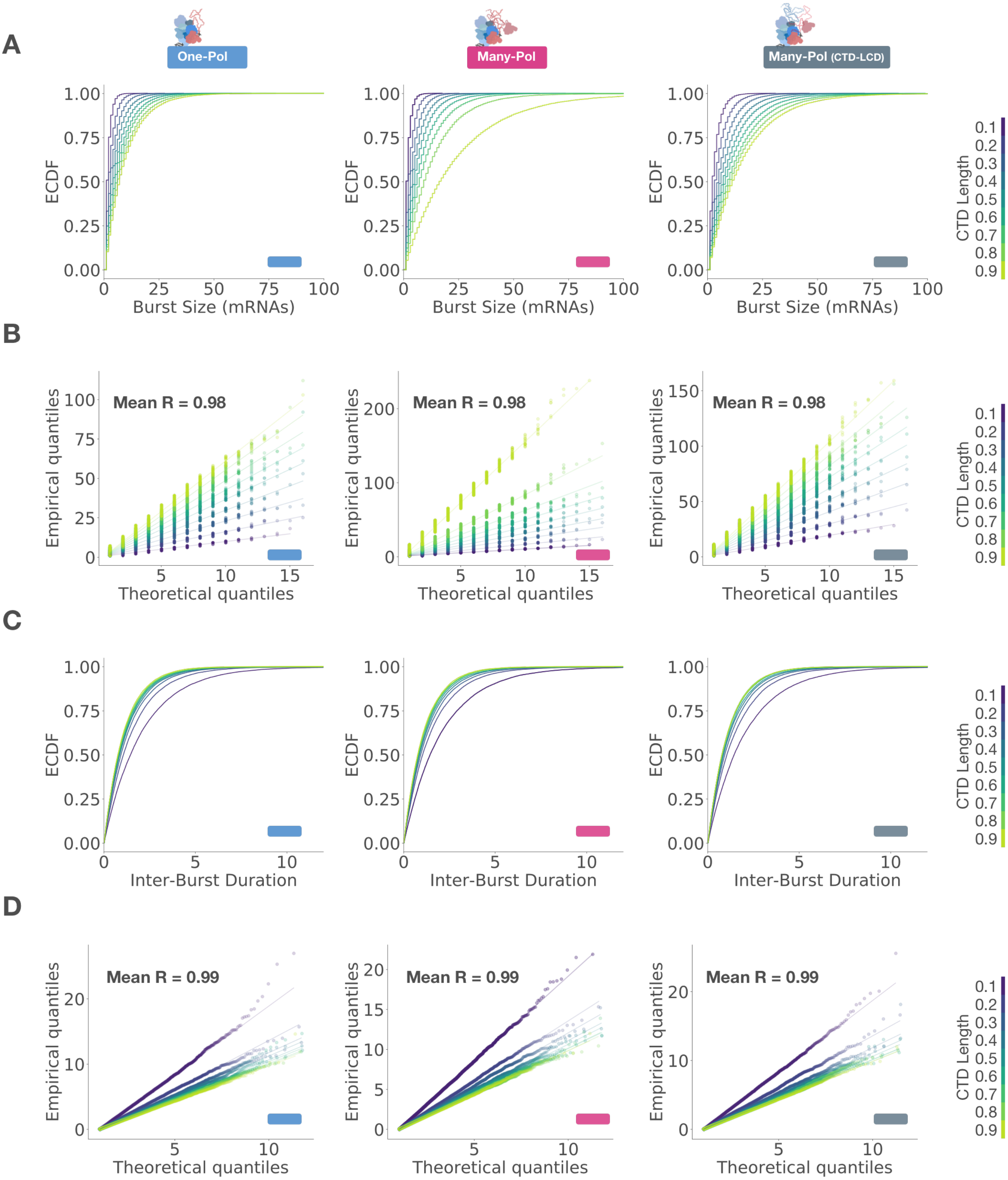

A
One-Pol
Many-Pol
Many-Pol (CTD-LCD)
ECDF
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Burst Size (mRNAs)
CTD Length
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
B
Mean R = 0.98
Empirical quantiles
Theoretical quantiles
C
Inter-Burst Duration
D
Mean R = 0.99

Figure S7:

Figure S7: **Transcription models produce geometric and exponential distributions of burst sizes and inter-burst durations, respectively. Related to figure 6.** (A) Empirical cumulative distribution functions (ECDF) of burst sizes by CTD length for each model. (B) Q-Q plots comparing quantiles from simulated distributions and a geometric distribution. Similarly, (C) ECDFs of inter-burst durations and (D) Q-Q plots comparing their quantiles with an exponential distribution. Each column comes from the model indicated by the color on top and the lower right corner in each plot. The mean of the square root of the coefficient of determination (R) by model is indicated in each quantile comparison. CTD length is indicated by the color shown to the right of each row.

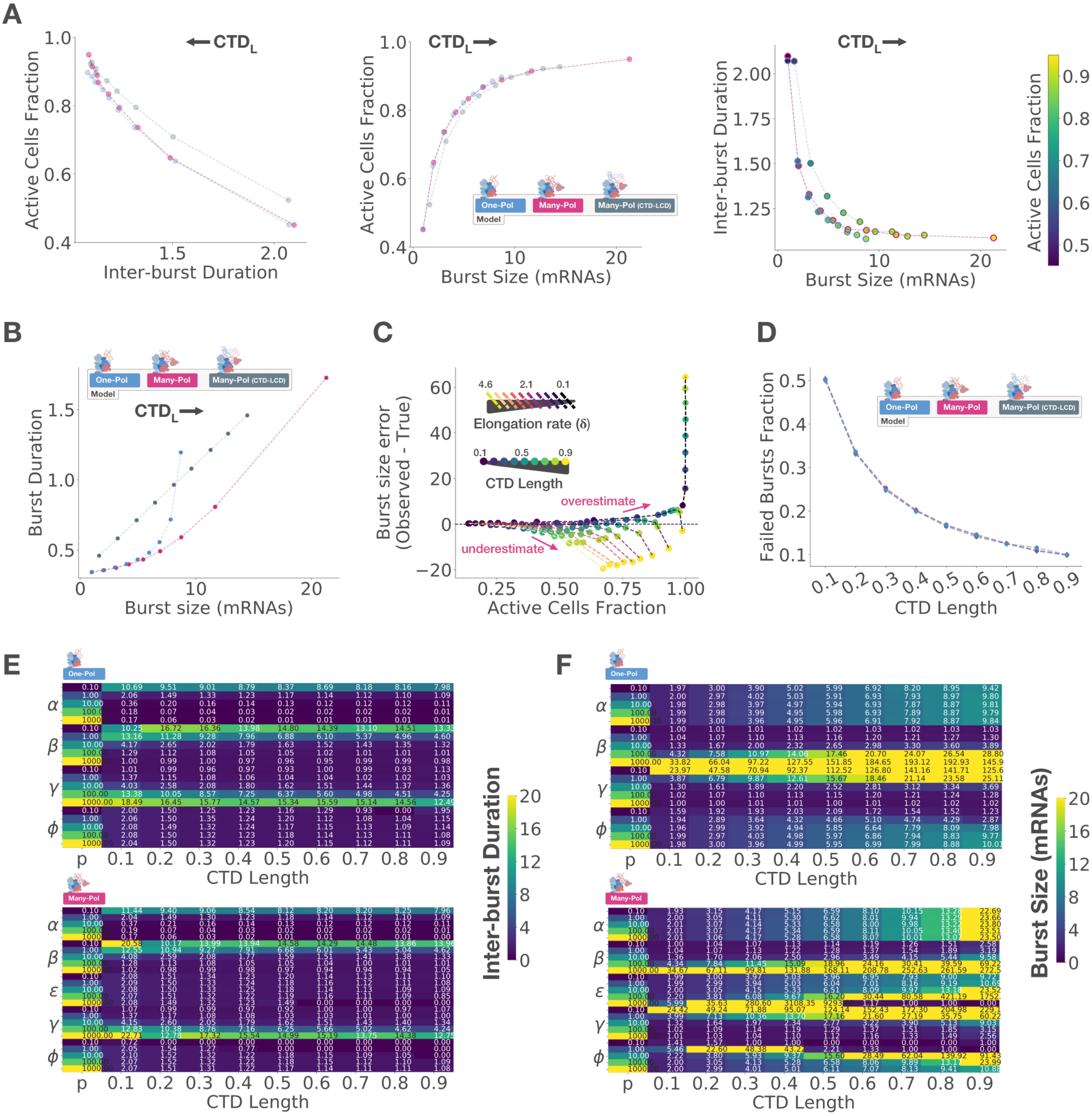


Figure S8:

Figure S8: **Parameter exploration with stochastic simulations provides insights into experimental observations. Related to figure 6.** (A) Comparison of the mean active cells fraction with means of inter-burst duration (left), burst size (middle) and both of these numbers (right) with increasing CTD length (CTD_L) by model, indicated with color. Direction of CTD increase is indicated with an arrow on top of each plot. (B) Comparison of mean burst duration with mean burst size with increasing CTD length by model. (C) Comparison of the error in burst size estimate, computed as the difference between the means of the observed transcription site intensity and the true burst size, with the fraction of active cells as a function of CTD_L under the many-polymerases model. The elongation rate (δ) determines the time that a given mRNA spends bound to the transcription site and contributes to the observed intensity, thus influencing the fraction of active cells at a given time. (D) Comparison of fraction of failed bursts, where an assembled preinitiation complex produced zero mRNAs before disassembly, as a function of CTD_L by model. Error bars indicate 99% bootstrapped confidence interval. Mean inter-burst duration (E) and burst size (F) as a function of CTD_L and individually varying parameter values, while the others are held constant, as indicated in the first left column of each heatmap. Colormap is artificially fixed to the range [0-20] for visualization purposes and actual numbers are shown in each cell. Model is indicated in the top left corner.

References

Adelman, K., Lis, J.T., 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nature reviews. Genetics 13, 720–731.

Allan, D.B., Caswell, T., Keim, N.C., van der Wel, C.M., 2018. trackpy: Trackpy v0.4.1 (Version v0.4.1).

Allen, B.L., Taatjes, D.J., 2015. The Mediator complex: a central integrator of transcription. Nature reviews. Molecular cell biology 16, 155–166.

Allison, L.A., Ingles, C.J., 1989. Mutations in RNA polymerase II enhance or suppress mutations in GAL4. Proceedings of the National Academy of Sciences 86, 2794–2798.

Aristizabal, M.J., Negri, G.L., Benschop, J.J., Holstege, F.C.P., Krogan, N.J., Kobor, M.S., 2013. High-Throughput Genetic and Gene Expression Analysis of the RNAPII-CTD Reveals Unexpected Connections to SRB10/CDK8. PLoS Genetics 9, e1003758.

Banani, S.F., Lee, H.O., Hyman, A.A., Rosen, M.K., 2017. Biomolecular condensates: Organizers of cellular biochemistry. Nature Reviews Molecular Cell Biology 18, 285–298.

Bartman, C.R., Hamagami, N., Keller, C.A., Giardine, B., Hardison, R.C., Blobel, G.A., Raj, A., 2019. Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. Molecular Cell 73, 519–532.e4.

Bartman, C.R., Hsu, S.C., Hsiung, C.C.S., Raj, A., Blobel, G.A., 2016. Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. Molecular Cell 62, 237–247.

Bartolomei, M.S., Halden, N.F., Cullen, C.R., Corden, J.L., 1988. Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II. Molecular and cellular biology 8, 330–9.

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A.S., Yu, T., Marie-Nelly, H., McSwiggen, D.T., Kokic, G., Dailey, G.M., Cramer, P., Darzacq, X., Zweckstetter, M., 2018. RNA polymerase II clustering through carboxy-terminal domain phase separation. Nature Structural and Molecular Biology 25, 833–840.

Bois, J.S., Elowitz, M.B., 2019. Stochastic simulation of biological circuits. Caltech Library , https://doi.org/10.7907/V8SD–Q741.

Boulon, S., Pradet-Balade, B., Verheggen, C., Molle, D., Boireau, S., Georgieva, M., Azzag, K., Robert, M.C., Ahmad, Y., Neel, H., Lamond, A.I., Bertrand, E., 2010. HSP90 and its R2TP/Prefoldin-like cochaperone are involved in the cytoplasmic assembly of RNA polymerase II. Molecular Cell 39, 912–924.

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34, 525–527.

Bryant, G.O., Ptashne, M., 2003. Independent Recruitment In Vivo by Gal4 of Two Complexes Required for Transcription. Molecular Cell 11, 1301–1309.

Carre, C., Shiekhattar, R., 2011. Human GTPases Associate with RNA Polymerase II To Mediate Its Nuclear Import. Molecular and Cellular Biology 31, 3953–3962.

Chapman, R.D., Heidemann, M., Hintermair, C., Eick, D., 2008. Molecular evolution of the RNA polymerase II CTD. Trends in Genetics 24, 289–296.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., Gregor, T., 2018. Dynamic interplay between enhancer–promoter topology and gene activity. Nature Genetics 50, 1296–1303.

Cho, W.K., Jayanth, N., English, B.P., Inoue, T., Andrews, J.O., Conway, W., Grimm, J.B., Spille, J.H., Lavis, L.D., Lionnet, T., Cisse, I.I., 2016. RNA Polymerase II cluster dynamics predict mRNA output in living cells. eLife 5, e13617.

Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V., Cisse, I.I., 2018. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science 361, 412–415.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., Tjian, R., 2018. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science 361, eaar2555.

Chubb, J.R., Trcek, T., Shenoy, S.M., Singer, R.H., 2006. Transcriptional Pulsing of a Developmental Gene. Current Biology 16, 1018–1025.

Cisse, I.I., Izeddin, I., Causse, S.Z., Boudarene, L., Senecal, A., Muresan, L., Dugast-Darzacq, C., Hajj, B., Dahan, M., Darzacq, X., 2013. Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells. Science 341, 664–667.

Collin, P., Jeronimo, C., Poitras, C., Robert, F., 2019. RNA Polymerase II CTD Tyrosine 1 Is Required for Efficient Termination by the Nrd1-Nab3-Sen1 Pathway. Molecular Cell 73, 655–669.e7.

Coulon, A., Ferguson, M.L., de Turris, V., Palangat, M., Chow, C.C., Larson, D.R., 2014. Kinetic competition during the transcription cycle results in stochastic RNA processing. eLife 3, e03939.

Couthouis, J., Hart, M.P., Shorter, J., DeJesus-Hernandez, M., Erion, R., Oristano, R., Liu, A.X., Ramos, D., Jethava, N., Hosangadi, D., Epstein, J., Chiang, A., Diaz, Z., Nakaya, T., Ibrahim, F., Kim, H.J., Solski, J.A., Williams, K.L., Mojsilovic-Petrovic, J., Ingre, C., Boylan, K., Graff-Radford, N.R., Dickson, D.W., Clay-Falcone, D., Elman, L., McCluskey, L., Greene, R., Kalb, R.G., Lee, V.M.Y., Trojanowski, J.Q., Ludolph, A., Robberecht, W., Andersen, P.M., Nicholson, G.A., Blair, I.P., King, O.D., Bonini, N.M., Van Deerlin, V., Rademakers, R., Mourelatos, Z., Gitler, A.D., 2011. A yeast functional screen predicts new candidate ALS disease genes. Proceedings of the National Academy of Sciences 108, 20881–20890.

Czeko, E., Seizl, M., Augsberger, C., Mielke, T., Cramer, P., 2011. Iwr1 Directs RNA Polymerase II Nuclear Import. Molecular Cell 42, 261–266.

Daniel Gietz, R., Woods, R.A., 2002. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method, in: Methods in Enzymology, pp. 87–96.

Dobi, K.C., Winston, F., 2007. Analysis of transcriptional activation at a distance in Saccharomyces cerevisiae. Molecular and cellular biology 27, 5575–5586.

Donovan, B.T., Huynh, A., Ball, D.A., Patel, H.P., Poirier, M.G., Larson, D.R., Ferguson, M.L., Lenstra, T.L., 2019. Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. The EMBO Journal 38, e100809.

Eick, D., Geyer, M., 2013. The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code. Chemical Reviews 113, 8456–8490.

Engler, C., Kandzia, R., Marillonnet, S., 2008. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. PLoS ONE 3, e3647.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 39, W29–W37.

Gal, J., Zhang, J., Kwinter, D.M., Zhai, J., Jia, H., Jia, J., Zhu, H., 2011. Nuclear localization sequence of FUS and induction of stress granules by ALS mutants. Neurobiology of Aging 32, 2323.e27–2323.e40.

Gerber, H.P., Hagmann, M., Seipel, K., Georgiev, O., West, M.a., Litingtung, Y., Schaffner, W., Corden, J.L., 1995. RNA polymerase II C-terminal domain required for enhancer-driven transcription. Nature 374, 660–2.

Gibbs, E.B., Lu, F., Portz, B., Fisher, M.J., Medellin, B.P., Laremore, T.N., Zhang, Y.J., Gilmour, D.S., Showalter, S.A., 2017. Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II C-terminal domain. Nature Communications 8, 15233.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., Smith, H.O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature Methods 6, 343–345.

Gressel, S., Schwalb, B., Decker, T.M., Qin, W., Leonhardt, H., Eick, D., Cramer, P., 2017. CDK9-dependent RNA polymerase II pausing controls transcription initiation. eLife 6, 1–24.

Hantsche, M., Cramer, P., 2017. Conserved RNA polymerase II initiation complex structure. Current Opinion in Structural Biology 47, 17–22.

Harlen, K.M., Churchman, L.S., 2017. The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. Nature Reviews Molecular Cell Biology 18, 263–273.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., Sharp, P.A., 2017. A Phase Separation Model for Transcriptional Control. Cell 169, 13–23.

Horwitz, A.A., Walter, J.M., Schubert, M.G., Kung, S.H., Hawkins, K., Platt, D.M., Hernday, A.D., Mahatdejkul-Meadows, T., Szeto, W., Chandran, S.S., Newman, J.D., 2015. Efficient Multiplexed Integration of Synergistic Alleles and Metabolic Pathways in Yeasts via CRISPR-Cas. Cell Systems 1, 88–96.

Huibregtse, J.M., Yang, J.C., Beaudenon, S.L., 1997. The large subunit of RNA polymerase II is a substrate of the Rsp5 ubiquitin-protein ligase. Proceedings of the National Academy of Sciences of the United States of America 94, 3656–61.

Jeronimo, C., Robert, F., 2014. Kin28 regulates the transient association of Mediator with core promoters. Nature Structural & Molecular Biology 21, 449–455.

Jones, E., Oliphant, T., Peterson, P., Others, . SciPy: Open source scientific tools for Python , http://www.scipy.org/ [Online; accessed 2019-09-30.

Ju, S., Tardiff, D.F., Han, H., Divya, K., Zhong, Q., Maquat, L.E., Bosco, D.A., Hayward, L.J., Brown, R.H., Lindquist, S., Ringe, D., Petsko, G.A., 2011. A Yeast Model of FUS/TLS-Dependent Cytotoxicity. PLoS Biology 9, e1001052.

Kettenberger, H., Armache, K.J., Cramer, P., 2004. Complete RNA Polymerase II Elongation Complex Structure and Its Interactions with NTP and TFIIS. Molecular Cell 16, 955–965.

Kim, Y.J., Björklund, S., Li, Y., Sayre, M.H., Kornberg, R.D., 1994. A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. Cell 77, 599–608.

Krishnamurthy, S., Hampsey, M., 2009. Eukaryotic transcription initiation. Current Biology 19, R153–R156.

Kuras, L., Struhl, K., 1999. Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. Nature 399, 609–613.

Kwon, I., Kato, M., Xiang, S., Wu, L., Theodoropoulos, P., Mirzaei, H., Han, T., Xie, S., Corden, J.L., McKnight, S.L., 2013. Phosphorylation-Regulated Binding of RNA Polymerase II to Fibrous Polymers of Low-Complexity Domains. Cell 155, 1049–1060.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology 157, 105–132.

Lee, M.E., DeLoache, W.C., Cervantes, B., Dueber, J.E., 2015. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. ACS Synthetic Biology 4, 975–986.

Lenstra, T.L., Coulon, A., Chow, C.C., Larson, D.R., 2015. Single-Molecule Imaging Reveals a Switch between Spurious and

Functional ncRNA Transcription. Molecular Cell 60, 597–610.

Lenstra, T.L., Larson, D.R., 2016. Single-Molecule mRNA Detection in Live Yeast, in: Current Protocols in Molecular Biology. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 14.24.1–14.24.15.

Lesurf, R., Cotto, K.C., Wang, G., Griffith, M., Kasaian, K., Jones, S.J.M., Montgomery, S.B., Griffith, O.L., 2016. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. Nucleic Acids Research 44, D126–D132.

Li, X.Y., Virbasius, A., Zhu, X., Green, M.R., 1999. Enhancement of TBP binding by activators and general transcription factors. Nature 399, 605–609.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J., 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293.

Lu, F., Portz, B., Gilmour, D.S., 2019. The C-Terminal Domain of RNA Polymerase II Is a Multivalent Targeting Sequence that Supports Drosophila Development with Only Consensus Heptads. Molecular Cell 73, 1232–1242.e4.

Marko, M., Vlassis, A., Guialis, A., Leichter, M., 2012. Domains involved in TAF15 subcellular localisation: Dependence on cell type and ongoing transcription. Gene 506, 331–338.

Mertins, P., Qiao, J.W., Patel, J., Udeshi, N.D., Clauser, K.R., Mani, D.R., Burgess, M.W., Gillette, M.A., Jaffe, J.D., Carr, S.A., 2013. Integrated proteomic analysis of post-translational modifications by serial enrichment. Nature Methods 10, 634–637.

Millman, K.J., Aivazis, M., 2011. Python for Scientists and Engineers. Computing in Science & Engineering 13, 9–12.

Nair, S.J., Yang, L., Meluzzi, D., Oh, S., Yang, F., Friedman, M.J., Wang, S., Suter, T., Alshareedah, I., Gamliel, A., Ma, Q., Zhang, J., Hu, Y., Tan, Y., Ohgi, K.A., Jayani, R.S., Banerjee, P.R., Aggarwal, A.K., Rosenfeld, M.G., 2019. Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. Nature Structural & Molecular Biology 26, 193–203.

Necci, M., Piovesan, D., Dosztányi, Z., Tosatto, S.C., 2017. MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics 33, 1402–1404.

Nonet, M., Sweetser, D., Young, R.A., 1987. Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. Cell 50, 909–915.

Oliphant, T.E., 2007. Python for Scientific Computing. Computing in Science & Engineering 9, 10–20.

Payne, J.M., Laybourn, P.J., Dahmus, M.E., 1989. The transition of RNA polymerase II from initiation to elongation is associated with phosphorylation of the carboxyl-terminal domain of subunit IIa. Journal of Biological Chemistry 264, 19621–19629.

Petrenko, N., Jin, Y., Petrenko, N., Jin, Y., Wong, K.H., Struhl, K., 2016. Mediator Undergoes a Compositional Change during Transcriptional Activation Article Mediator Undergoes a Compositional Change during Transcriptional Activation. Molecular Cell 64, 443–454.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., Pachter, L., 2017. Differential analysis of RNA-seq incorporating quantification uncertainty. Nature Methods 14, 687–690.

Portz, B., Lu, F., Gibbs, E.B., Mayfield, J.E., Rachel Mehaffey, M., Zhang, Y.J., Brodbelt, J.S., Showalter, S.A., Gilmour, D.S., 2017. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Nature Communications 8, 15231.

Qiu, C., Liu, F., Xu, L., Deng, B., Xiao, M., Si, J., Lin, L., Zhang, Z., Wang, J., Guo, H., Peng, H., Peng, L.M., 2018. Dirac-source field-effect transistors as energy-efficient, high-performance electronic switches. Science 361, 387–392.

Quintero-Cadena, P., Sternberg, P.W., 2016. Enhancer Sharing Promotes Neighborhoods of Transcriptional Regulation Across Eukaryotes. G3 (Bethesda, Md.) 6, 4167–4174.

Robinson, P.J., Trnka, M.J., Bushnell, D.A., Davis, R.E., Mattei, P.J., Burlingame, A.L., Kornberg, R.D., 2016. Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. Cell 166, 1411–1422.e16.

Robinson, P.J.J., Bushnell, D.A., Trnka, M.J., Burlingame, A.L., Kornberg, R.D., 2012. Structure of the Mediator Head module bound to the carboxy-terminal domain of RNA polymerase II. Proceedings of the National Academy of Sciences 109, 17931–17935.

Ryan, O.W., Skerker, J.M., Maurer, M.J., Li, X., Tsai, J.C., Poddar, S., Lee, M.E., DeLoache, W., Dueber, J.E., Arkin, A.P., Cate, J.H., 2014. Selection of chromosomal DNA libraries using a multiplex CRISPR system. eLife 3, e03703.

Scafe, C; Young, R., 1990. RNA polymerase II C-terminal repeat influences response to transcriptional enhancer signals. Nature 374, 685–689. NIHMS150003.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nature Methods 9, 676–682.

Shao, W., Zeitlinger, J., 2017. Paused RNA polymerase II inhibits new transcriptional initiation. Nature genetics 49, 1045–1051.

Shin, Y., Brangwynne, C.P., 2017. Liquid phase condensation in cell physiology and disease. Science 357, eaaf4382.

Shin, Y., Chang, Y.C., Lee, D.S., Berry, J., Sanders, D.W., Ronceray, P., Wingreen, N.S., Haataja, M., Brangwynne, C.P., 2018.

Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome. Cell 175, 1481–1491.e13.

Slubowski, C.J., Funk, A.D., Roesner, J.M., Paulissen, S.M., Huang, L.S., 2015. Plasmids for C-terminal tagging in Saccharomyces cerevisiae that contain improved GFP proteins, Envy and Ivy. Yeast 32, 379–387.

Somesh, B.P., Reid, J., Liu, W.F., Søgaard, T.M.M., Erdjument-Bromage, H., Tempst, P., Svejstrup, J.Q., 2005. Multiple mechanisms confining RNA polymerase II ubiquitylation to polymerases undergoing transcriptional arrest. Cell 121, 913–923.

Steinmetz, E.J., Warren, C.L., Kuehner, J.N., Panbehi, B., Ansari, A.Z., Brow, D.A., 2006. Genome-Wide Distribution of Yeast RNA Polymerase II and Its Control by Sen1 Helicase. Molecular Cell 24, 735–746.

Svejstrup, J.Q., Li, Y., Fellows, J., Gnatt, A., Bjorklund, S., Kornberg, R.D., 1997. Evidence for a mediator cycle at the initiation of transcription. Proceedings of the National Academy of Sciences 94, 6075–6078.

Swain, P.S., Stevenson, K., Leary, A., Montano-Gutierrez, L.F., Clark, I.B., Vogel, J., Pilizota, T., 2016. Inferring time derivatives including cell growth rates using Gaussian processes. Nature Communications 7, 13766.

Szabo, Q., Bantignies, F., Cavalli, G., 2019. Principles of genome folding into topologically associating domains. Science Advances 5, eaaw1668.

Takahashi, H., Parmely, T.J., Sato, S., Tomomori-Sato, C., Banks, C.A., Kong, S.E., Szutorisz, H., Swanson, S.K., Martin-Brown, S., Washburn, M.P., Florens, L., Seidel, C.W., Lin, C., Smith, E.R., Shilatifard, A., Conaway, R.C., Conaway, J.W., 2011. Human Mediator Subunit MED26 Functions as a Docking Site for Transcription Elongation Factors. Cell 146, 92–104.

Thompson, C.M., Koleske, A.J., Chao, D.M., Young, R.A., 1993. A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. Cell 73, 1361–1375.

Trcek, T., Chao, J.A., Larson, D.R., Park, H.Y., Zenklusen, D., Shenoy, S.M., Singer, R.H., 2012. Single-mRNA counting using fluorescent in situ hybridization in budding yeast. Nature Protocols 7, 408–419.

West, M.L., Corden, J.L., 1995. Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. Genetics 140, 1223–1233.

Wong, K.H., Jin, Y., Struhl, K., 2014. TFIIH Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. Molecular Cell 54, 601–612.

Yang, C., Stiller, J.W., 2014. Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proceedings of the National Academy of Sciences of the United States of America 111, 5920–5.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O.G., Janacek, S.H., Juettemann, T., To, J.K., Laird, M.R., Lavidas, I., Liu, Z., Loveland, J.E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D.N., Newman, V., Nuhn, M., Ogeh, D., Ong, C.K., Parker, A., Patricio, M., Riat, H.S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S.E., Kostadima, M., Langridge, N., Martin, F.J., Muffato, M., Perry, E., Ruffier, M., Staines, D.M., Trevanion, S.J., Aken, B.L., Cunningham, F., Yates, A., Flicek, P., 2018. Ensembl 2018. Nucleic Acids Research 46, D754–D761.