Interpreting changes in nascent transcription composite gene profiles using a compartment model

Michael J. Guertin

January 31, 2021

Contents

List of Figures

1 Background

1.1 Transcription Cycle

The transcription field has been interpreting changes in genomic nascent RNA profiles to determine the effects that various treatments have upon steps in the transcription cycle Figure 1. The first interpretation suggested that the Estrogen Receptor (Hah *et al.*, 2011), acts prior to RNA polymerase pausing. Subsequent work has defined the role of other transcription factors, including NFkB, HSF, GAF, and ZNF143 (Danko *et al.*, 2013; Jonkers *et al.*, 2014; Duarte *et al.*, 2016; Sathyan *et al.*, 2019). These studies were first limited to the study of trasncription factors that are rapidly inducible, but the development of rapidly inducible degradation methods has democratized the study of transcription factors by permitting measurements immediately after the factor is inhibited. Advantages include the ability to study essential factors and cleaner results that are not affected by the multitude post-primary effects of TF dysregualtion. Our recent work modeled the effects of changing the rates of any of these steps.

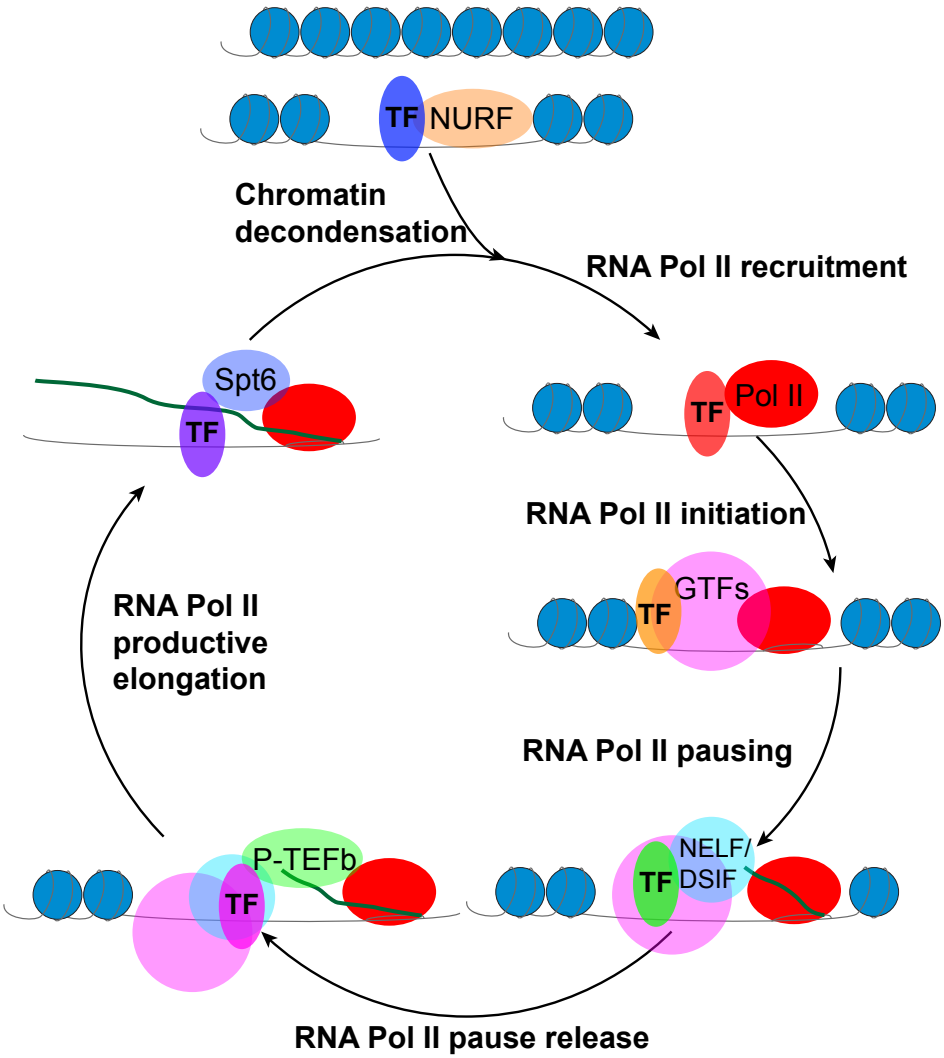

Figure 1: **Transcription Cycle.** Transcription is regulated at many steps and transcription factors tend to specialize in a subset of these steps. Specificity is occurs by selectively interacting with cofactors that are highly specialized in their function. As shown many different sequence-sepecific transcription factors are conferring the specificity of recruitment of various cofactors without DNA binding domains.

1.2 Genome-wide kinetic measurements of chromatin accessibility identifies regulatory transcription factors

We induced adipognesis and measured chromatin accessibility at 0min, 20min, 40min, 60min, 120min, 180min, 240min, and 6 days (Figure 2A). We clustered the kinetic accessibility profiles by their dynamics (Figure 2B) and identified DNA elements that are specifically enriched in the increased and decreased accessibility clusters (Figure 2C). We found 12 transcription factor families and the top six are shown in Figure 2C. Note that many transcription facotr bind the same motif, but we are able to narrow down the effector factors based on relative expression at time 0min and changes in factor expression ove rthe tiem course. Figure 2D&E show that a single transcription factor family is associated with either increases or decreases in accessibility.

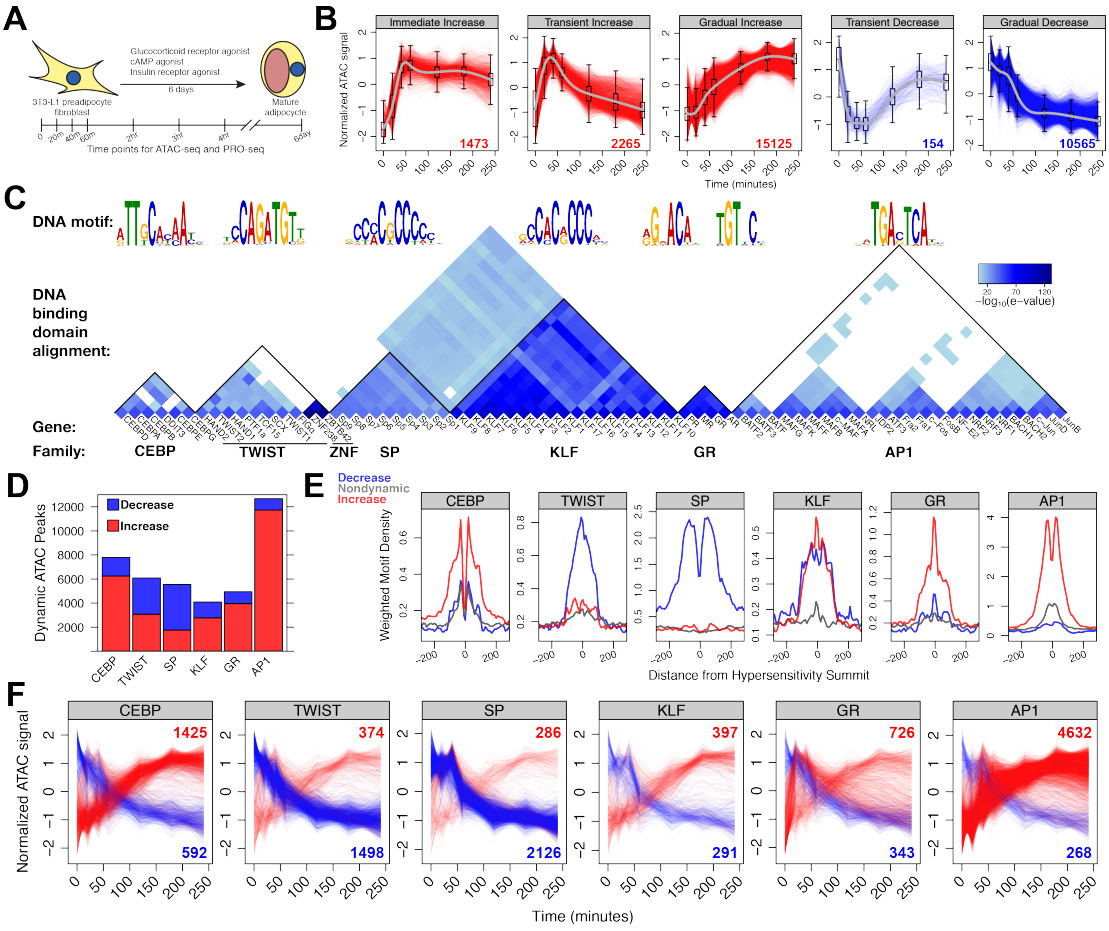

Figure 2: **CEBP, TWIST, SP, KLF, GR, and AP1 transcription factors drive chromatin accessibility dynamics in early adipogenesis.** A) Preadipocyte fibroblast 3T3-L1 cells were treated with an adipgenesis cocktail for the indicated time points: no treatment, 20min, 40min, 60min, 2hr, 3hr, 4hr, and 6day. B) Temporal classification of ATAC peaks revealed five major classes of dynamic peaks. Each red or blue line trace represents a single ATAC peak. C) *De novo* motif analysis (Bailey *et al.*, 2006) identified the six top DNA motifs that are enriched within the dynamic peaks. The transcription factors in the wedge below the seqLogo recognize the respective DNA motifs. The heatmap quantifies the local protein sequence alignment of the DNA binding domains for the genes, as determined by the Smith-Waterman algorithm. Although there are six DNA motifs, the TWIST and ZNF families of DNA binding domains recognize the same motif, despite their lack of evolutionary conservation. D) Dynamic ATAC peaks are classified by the presence of each DNA motif. The red bars represent the number of dynamic ATAC peaks within the *immediate increase*, *transient increase*, and *gradual increase* categories; the blue bars correspond to the *transient decrease* and *gradual decrease* classes. E) ATAC peaks that decrease accessibility are enriched for TWIST and SP motifs; peaks that increase accessibility are enriched for CEBP, KLF, GR, and AP1 motifs.

1.3 Genomic kinetic measurements of nascent transcription identifies regulatory transcription factors

A completely independent way to identify putative regulatory elements is to look for enhancer RNA using dREG (Wang *et al.*, 2019). We found over 200,000 putative regulatory regions with bidirectional signatures ((Figure 3A). These regulatory regions identify overlapping regulatory elements as ATAC, but the elements identified by this method are enriched in promoters and intragenic regions, while ATAC peaks are more intergenic (Figure 3B&C). PRO-seq identifies a subset of motifs compared to ATAC, and one additional motif (Figure 3D&E). The dREG peaks are enriched for motifs that are known to be in promoters (Benner *et al.*, 2013). Further analyses show that bidirectional PRO-seq signatures preferentially identifies promoter-proximal regulatory elements (Figure 3F,G).

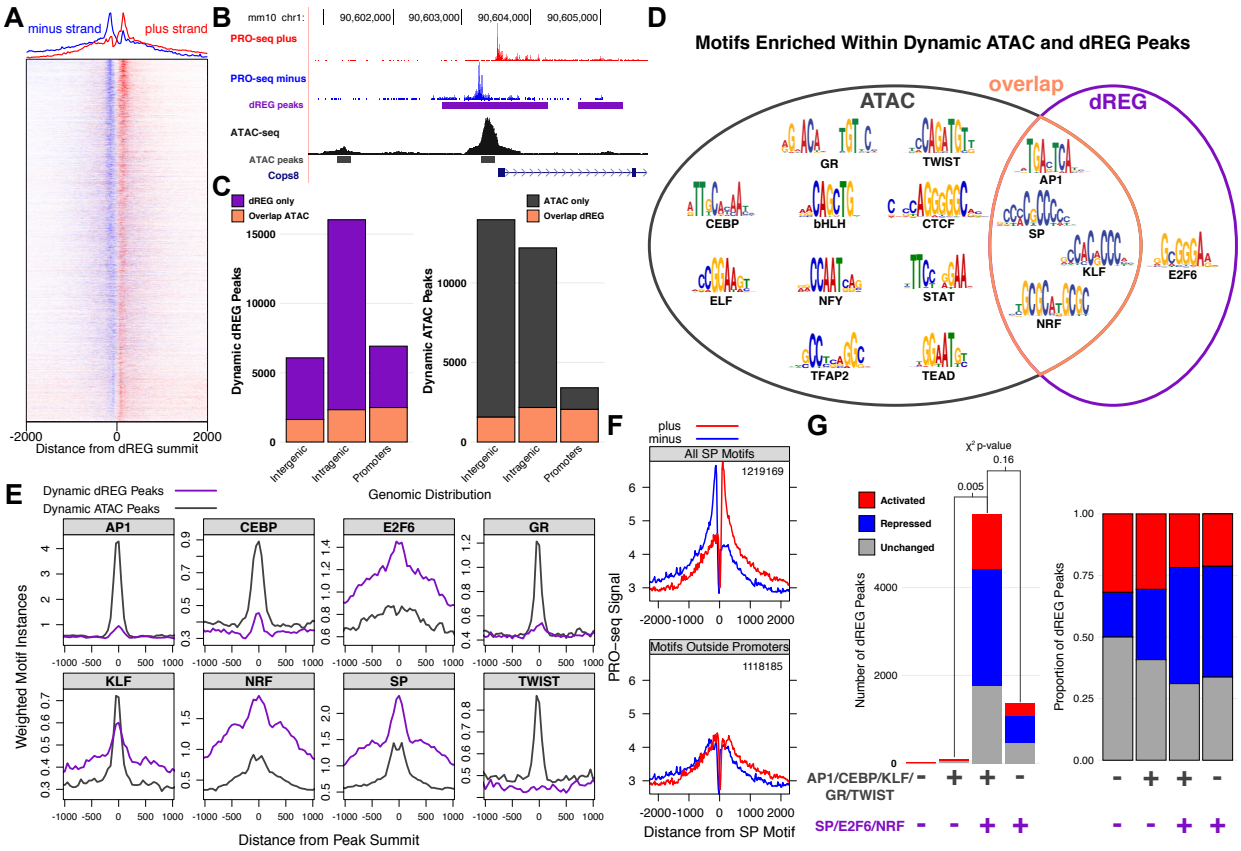

Figure 3: **SP, NRF, and E2F6 transcription factor families drive dynamic bidirectional adipogenic-induced transcription at regulatory regions within gene bodies and promoters.** A) The heatmap illustrates over 200,000 putative regulatory elements identified by a bidirectional trancription signature using discriminative regulatory-element detection (dREG) (Wang *et al.*, 2019). B) Both ATAC and PRO-seq identify a regulatory element within the promoter of Cops8. The upstream intergenic regulatory element is only identified by ATAC, while the intragenic regulatory element within the gene body is only identified by its bidirectional PRO-seq signature (visualized with UCSC browser (Kent *et al.*, 2002)). C) ATAC-seq and PRO-seq identify a distinct set of regulatory regions in the genome. D) Dynamic chromatin accessibility peaks are enriched for a more diverse set of transcription factor motifs. E) ATAC and dREG identify distinct classes of regulatory elements. There is modest CEBP motif enrichment at dREG peaks, despite the CEBP motif not being identified as enriched within dREG peaks. However, PRO-seq and dREG fail to detect a class of regulatory elements in which the dynamics are driven by TWIST and GR factors. F) SP is only associated with bidirectional transcription at promoters rather than distal regulatory elements. The average normalized PRO-seq signal for plus and minus strands around all SP motif instances (top) and all SP motifs excluding those in promoters (bottom). The number of motif instances for each plot is in the top right of the panel. G) DREG-enriched factor motifs are enriched in peaks that decrease bidirectional transcription, which suggests a link between these factors and an early and pervasive decrease in promoter initiation at genes with SP, NRF, and E2F6 motifs. Dynamic bidirectional transcription peaks found in promoters are stratified by the presence or absence of transcription factor motifs. The left plot quantifies the total number of peaks and the right plot scales to the proportion of peaks in each category. The x-axis factor motif categories are defined by the presence or absence of ATAC-enriched factors (AP1, CEBP, GR, KLF, and TWIST) and dREG-enriched factors (SP, E2F6, and NRF).

1.4 Linking transcription factors to their gene targets

Genes are also classified by their kinetics, just as ATAC peaks and enhancer RNA bidirectional peaks (Figure 5A&C). Two predictive features can be used to link regulatory elemetns to genes 1) covariation in regulatory element signal and transcript; and 2) proximity. We can estimate how far a given transcriptino factor can act by plotting the cumulative distribution of proximal dynamic regulatory elements containing the factor's cognate motif (Figure 5B&D). These data indicate that GR can act distally and SP1 factor activity is limited to proximal genes (Figure 5B&D). Activated and repressed genes tend to be proximal to regulatory elements that increase and decrease accessibility, respectively (Figure 5E). We can incorporate additional contrants on the the links between regulatory elements and genes (Figure 5F). Using a set of data-driven rules based on = proximity and covarying transcription and accessibility, we can infer the target genes of regulatory elements. Moreover, if the regulatory element contains a transcription factor motif from Figure 3D, then we link a transcription factor to its target. The simplified network for GR identifies nodes where GR binds to regulate local gene expression (Figure 5G, right path).

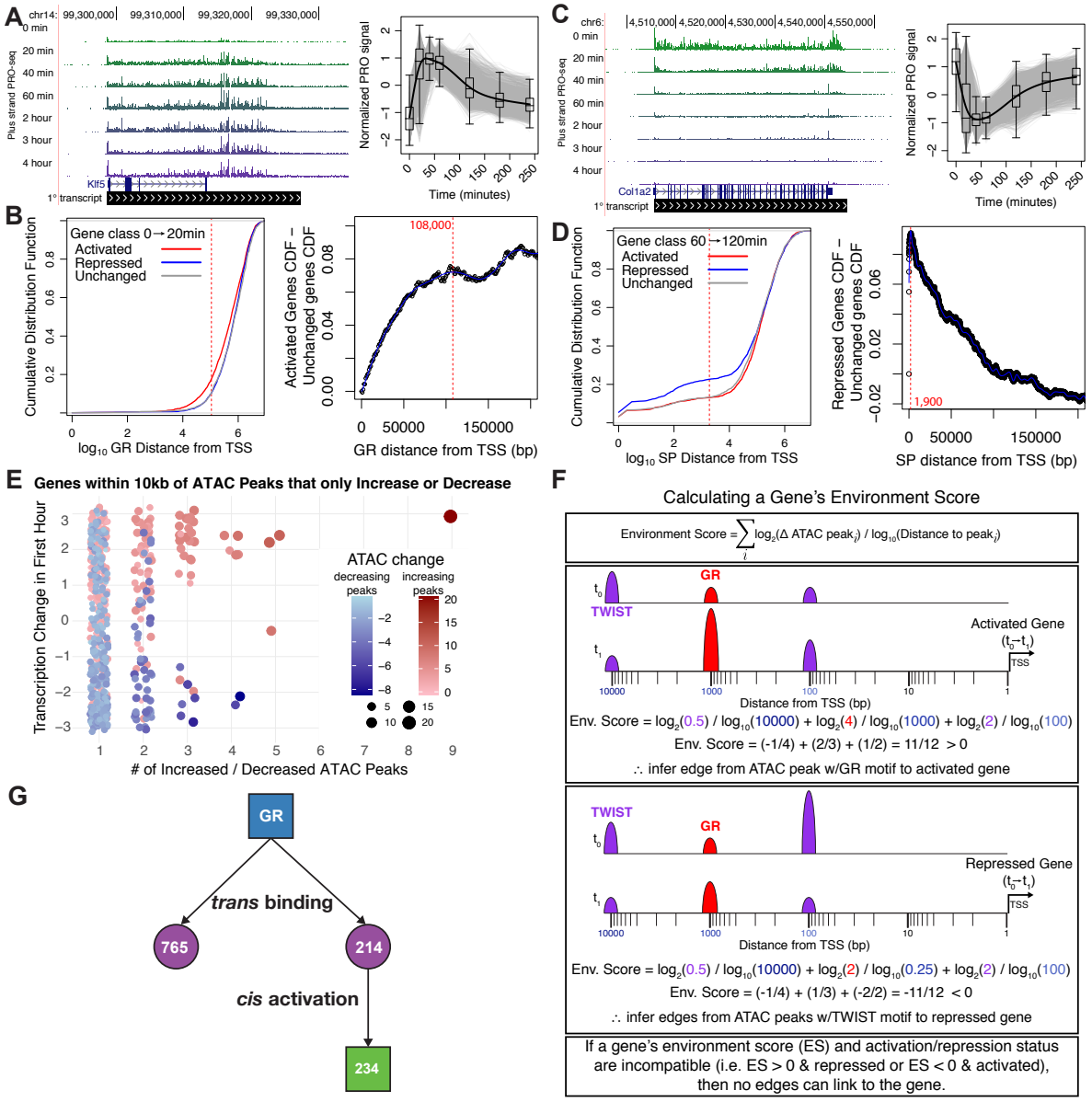

Figure 4: **Correlated chromatin accessibility and transcription dynamics and proximity guide inference of links between regulatory elements and target genes.** A) Klf5 (left) and a cluster of genes (gray traces on the right) are immediately and transiently activated. B) Cumulative distribution plots quantify the relationship between the closest dynamic ATAC peak with a GR motif and the start sites of genes that are activated from 0 - 20min. These ATAC peaks are closer to activated gene class and these traces begin to converge 108,000 bases from the start sites. C) Col1a2 (left) and a cluster of genes (gray traces on the right) are immediately repressed. D) ATAC peaks with SP motifs are closer to the 60 - 120min repressed gene class and these traces begin to converge 1,900 bases from the start sites. E) The number of ATAC peaks that change and their fold change correlate with the changes in expression of proximal genes. Note that genes with only either increasing or decreasing ATAC peaks within 10kb are included. F) The change in accessibility of dynamic ATAC peaks that harbor the specified motif, and none of the other 5 motifs, is consistent with the reciprocal analysis conclusions from panel (E). F) We empirically found that imposing a local chromatin environment score to a gene increased specificity when infering links between regulatory elements and genes. G) The kinetic experiments allow us to link transcription factors to regulatory elements. If the regulatory elements are proximal to genes that are activated or repressed, based on the function of the bound factor, we can link the regulatory elements to their gene targets.

2 Interpreting composite profiles

Now we have a network with candidate target genes for over a dozen transcription factors. We can systematically look at how the distirbution of RNA polymerase changes at these groups of genes to determien what steps in the transcription cycle each factor is liekly to regulate.

We previously generated a two-compartment model to describe alterations in the transcription cycle within the context of composite profiles 5 (Sathyan *et al.*, 2019). The limitations of this model and implementation are that rates are dimentionless, as we only assess how changing the rates qualitatively affected the pause and gene body densities. All considerations and previous analyses assumed steady-state.
To expand on this work the goals for this manuscript are to develop a dynamic model. The long term goals are to 1) incorporate rate estimates from the literature; 2) estimate the contributions of transcription factors cooperating at genes based on kinetic changes in binding (ATAC signal) and nascent transcription (PRO signal); 3) apply mechanistic rules to exisiting kinetic or thermodynamic models (He *et al.*, 2010; Scholes *et al.*, 2017) of transcription output to compute and predict changes in expression over developmental time courses and regulatory cascades.

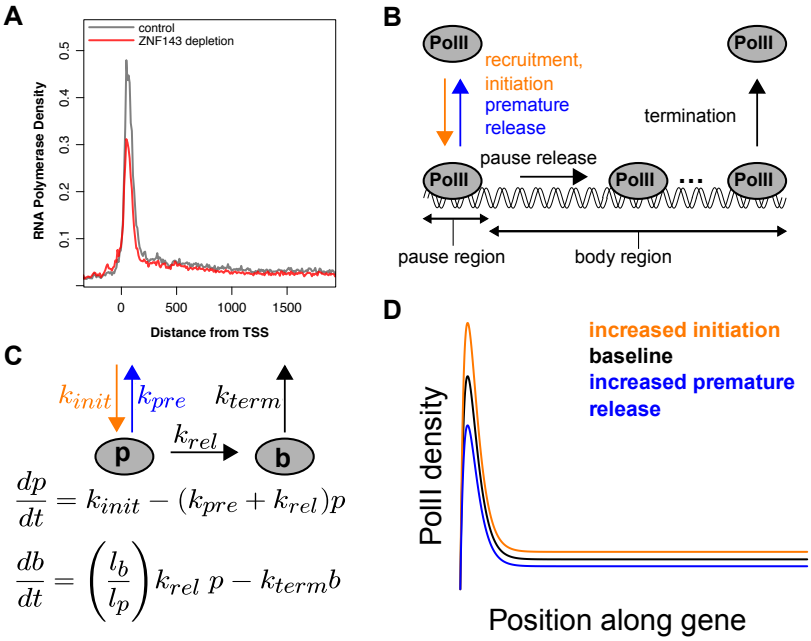

Figure 5: **ZNF143 regulates inititation or premature pause release.** This figure is reproduced from our previous work (Sathyan *et al.*, 2019). A) The composite profile of Pol II density upon ZNF143 indicates that Pol II pause density decreases. B) Model structure and key variables are highlighted in this schematic. C) A mathematical formulation of the two-compartment model, in which p refers to Pol II density at the pause region, and b refers to the density at the gene body region. D) This plot represents a steady-state simulation for a reference model (black), a model in which transcriptional initiation was increased by 25% (orange), and a model in which premature pause release was increased by 60% (blue).

2.1 Developing a dynamic model

The following code chunk contains functions that describe the compartment model and illustrate the plot to resemble conventional genomic composite gene profiles.

The dynamics for the densities of RNA Polymerase at pause site and the gene body (Sathyan *et al.*, 2019), defined as p and b:

$$\frac{\partial p}{\partial t} = k_{init} - (k_{pre} + k_{rel})\ p$$

$$\frac{\partial b}{\partial t} = \left(\frac{l_p}{l_b}\right) k_{rel}\ p - k_{term}\ b$$

Source functions for organizing the data and plotting the composites from: `https://raw.githubusercontent.com/guertinlab/modeling_PRO_composites/main/plotting_composites_lattice.R`. The following code chunk can be retrieved directly from `https://raw.githubusercontent.com/guertinlab/modeling_PRO_composites/main/dynamic_traces.R`

```
library(deSolve)
library(lattice)
#import in misc. plotting and data parsing functions
gitpage = 'https://raw.githubusercontent.com/guertinlab/'
source(paste0(gitpage, 'modeling_PRO_composites/main/plotting_composites_lattice.R'))

#model derived from our G&D paper (doi:10.1101/gad.328237.119):
#Declare differential equations as function
#dP/dt = kinit - (kpre +krel)p
#dB/dt = (lp/lb)*krel * p - kterm * b
#P is the first dependent variable, promoter density; dP is its derivative wrt time
#B is the second dependent variable, body density; dB is its derivative wrt time

density.prime <- function(t, initial.state, params = params)  {
    with(as.list(c(params, initial.state)), {
        dP = kinit - (kpre + krel)*P
        dB = ((lp/lb)*krel)*P - kterm*B
        res = c(dP, dB)
        list(res)
    })
}


#code adapted from our G&D paper (doi:10.1101/gad.328237.119):
#to visualize it in a composite profile form
# function for finding the gene body parameter
# bpeak: desired gene body level
# pausepeak: desired pause region level
# tau: exponential decay constant
# min.pk: minimal level for gene body peak
# max.pk: maximal level for gene body peak
# dpk: resolution for the implicit solution
find.body.param <- function(bpeak=NULL,tau=NULL,
                            pausepeak=NULL,min.pk=0,max.pk=1,
                            dpk=0.001) {
    pk = seq(min.pk,max.pk,dpk) # sequence of peak parameter values
    dat = matrix(0,length(pk),2) # data matrix
    colnames(dat) = c("body_param","body_assymp")
    for (x in 1:length(pk)) {
        bodypeak = pk[x]
        root = tau*(bodypeak+exp(1))*exp(-1)
        peak = (root/tau) * exp(-(root - tau)/tau) + bodypeak *
            (1 - exp(-root/tau))
        body = bodypeak * pausepeak / peak
```

```
        dat[x,1] = bodypeak
        dat[x,2] = body
    }
# look up the ratio of the body parameter over the assymptote
# use linear interpolation
    inter = findInterval(bpeak,dat[,2]) - 1
    m = (dat[(inter+1),1] - dat[inter,1]) /
        (dat[(inter+1),2] - dat[inter,2])
    b = dat[inter,1] - m * dat[inter,2]
    est = m * bpeak + b
    return(est)
}

# function to get the PRO waveform
# bpeak: desired gene body level
# pausepeak: desired pause region level
# bodypeak: body peak value to obtain a level of bpeak
# bp.seq: base pair sequence
# tau: exponential decay constant
get.pro.waveform <- function(bpeak=NULL,pausepeak=NULL,bodypeak=NULL,
                             bp.seq=NULL,tau=NULL){
    root = tau*(bodypeak+exp(1))*exp(-1)
    peak = (root/tau) * exp(-(root - tau)/tau) + bodypeak * (1 - exp(-root/tau))
    vec = sapply(bp.seq,function(x){(pausepeak/peak)*((x/tau) * exp(-(x - tau)/tau) +
                                                      bodypeak * (1 - exp(-x/tau)))})
    vec = unname(vec)
    pars = as.data.frame(rbind(cbind(max(vec),
        vec[bp.seq[length(bp.seq)]]),c(pausepeak,bpeak)))
    names(pars) = c("Peak","Assymp")
    rownames(pars) = c("simulated","requested")
    out = list(vec=vec, pars=pars)
    return(out)
}
```

2.2 Implementing the compartment model

Running thorugh the models to simulate composite profiles and changes in transcription. Note that the rates were chosen based on the observations that RNA Polymerase pausing is rate limiting at most genes and termination is relatively slow, as PolII transcribes approx. 2.5kb/min. For this model initiation rate is twice the pause release rate and the premature (non-productive pause release rate) is 60% the pause release rate. Length of the pause and gene body regions approximate the average in the genome. The raw code can be retieved from `https://raw.githubusercontent.com/guertinlab/modeling_PRO_composites/main/composites_processes.R`

```
#dimensionless rates/lengths
lp = 100 #length promoter
lb = 30000 #length gene body
kinit = 0.1 #initiation rate (pooled in with the unbound polII concentration)
krel = 0.05 #pause release rate (rate-limiting at most genes)
kpre =0.03 #premature release rate (controversial how prevalent this is see David Price H202)
kterm = 0.001 #termination rate relatively slow
times = seq(0,100, by = 5) #time

params = c(kinit=kinit, kpre=kpre, kterm=kterm, krel=krel, lp = lp, lb = lb)
initial.state = c(P = 0.1, B = 0.01)

#Solve series of differential equations
result = ode(y = initial.state,
             times = times,
             func = density.prime,
             parms = params)
```

```
result = data.frame(result)

#plot in lattice
plot.changes.wrt(result, filename = 'dynamic_pro_model_PGBdensities')

#previous plot shows that pause density reaches steady state, but
#gene body density is still linear
plot.changes.wrt(data.frame(ode(y = initial.state,
              times = seq(0,10000, by = 1),
              func = density.prime,
              parms = params)), 'dynamic_pro_model_PGBdensities_ss')


#prepare the simulated composites
dynamic.pro = dynamic.pro.profile(input = result)

#plot the steady state profile as well.
#set to zero and solve
# dP = kinit - (kpre + krel)*P
# dB = ((lp/lb)*krel)*P - kterm*B
pause.peak = kinit/(kpre + krel)
body.peak = (((lp/lb)*krel)*pause.peak)/kterm
steady.state.pro = dynamic.pro.profile(input = data.frame(cbind(1, pause.peak, body.peak)))

#plot composites in lattice
plot.pro.simulation.composites(dynamic.pro,
              filename = 'dynamic_pro_model_density')

plot.pro.simulation.composites(steady.state.pro ,
              filename = 'steady_state_pro_model_density')
```

The most noteworthy observation of the dynamic model is that it takes over an order of magnitude longer to approach steady state density in the gene body as compared to pause region (Figure 6A&B).

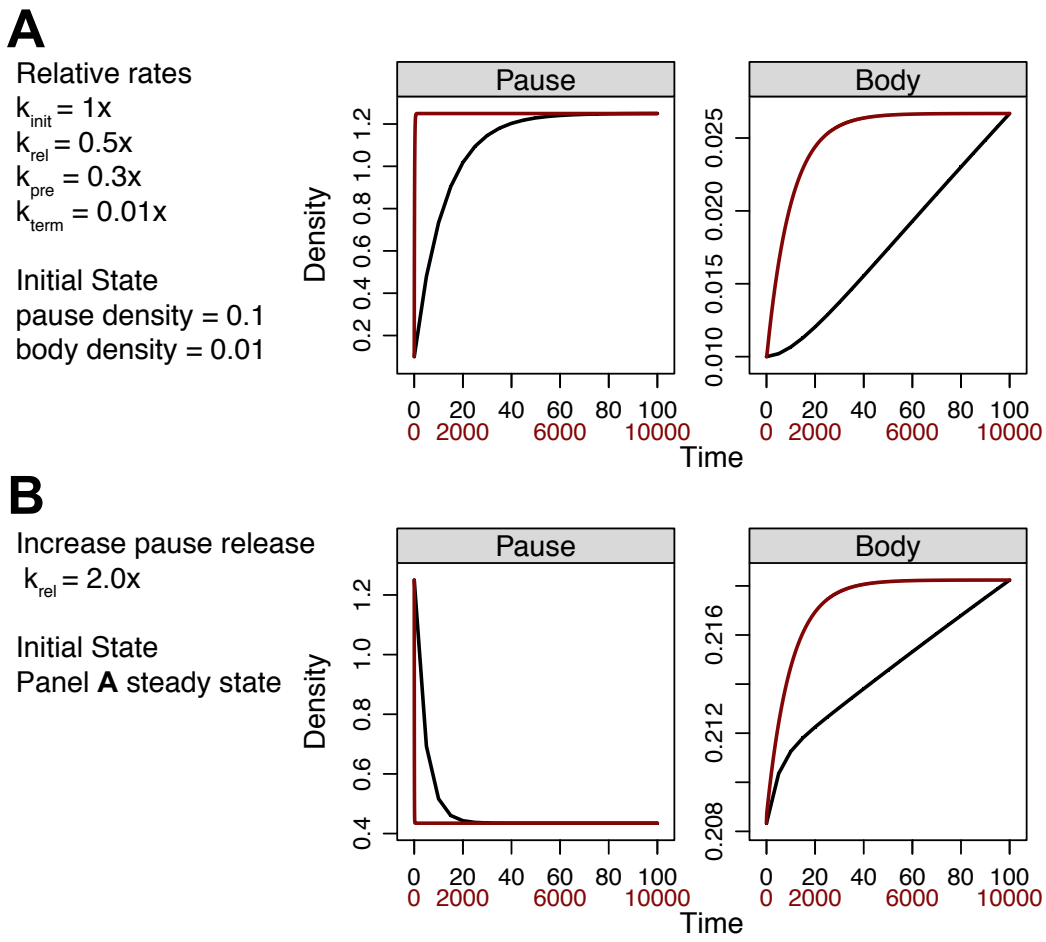

Figure 6: **Time dependent changes in pause and body.** The RNA polymerase density in the pause region reaches a steady state much earlier than the gene body density. Note that the x-axis has a separate scale color keyed to the traces.

2.3 Assessing how changing pause release affects composite profiles over time

By changing the pause release rate by 4x, we observe the change from the previous composites. A goal for the adipogenesis paper is to systematically ascribe function to the 13 factors based on how target genes respond. Isolating genes that are inferred to be exclusively or predominantly regulated by a single factor will be paramount. Much like we established an environment score in Figure 5F, we will use a combination of distance, change in accessiblity, and motif score to establish a relative transcription factor contribution score for each gene.

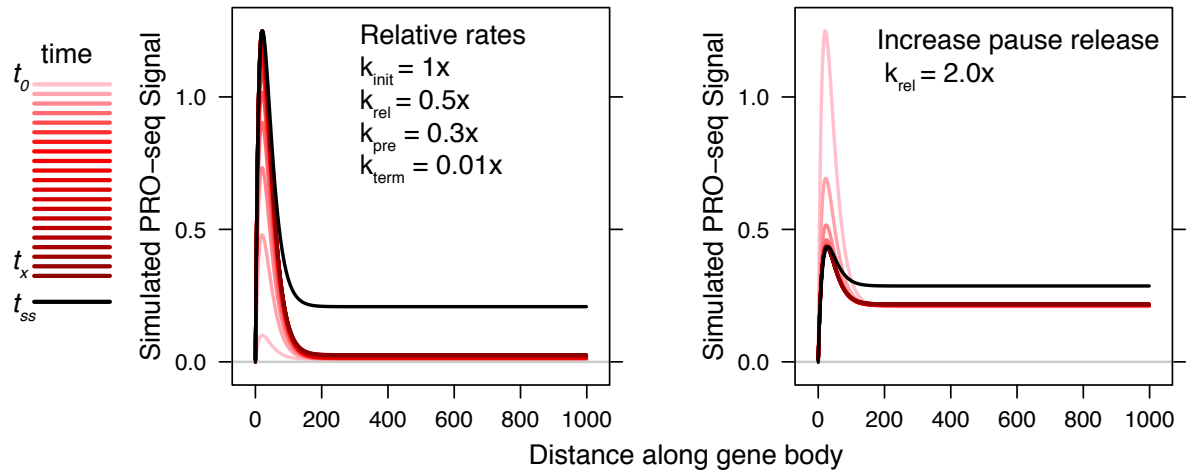

Figure 7: **Time dependent composites track changes in pause and body density.** The rate constants estimates recapitulate the profile ratios that we observe from PRO-seq data. The red traces are for the first 100 units of time and the black trace is the steady state level.

Raw code: `https://raw.githubusercontent.com/guertinlab/modeling_PRO_composites/main/stim_pause_rel.R`

```
#initial conditions are:
#pause.peak
#body.peak
krel.stim=0.2
result.stimulate.pause.release = ode(c(P = pause.peak , B = body.peak),
             times = times,
             func = density.prime,
             parms = c(kinit=kinit,
                       kpre=kpre,
                       kterm=kterm,
                       krel=krel.stim,
                       lp = lp,
                       lb = lb))

result.stimulate.pause.release = data.frame(result.stimulate.pause.release)

#plot in lattice
plot.changes.wrt(result.stimulate.pause.release,
                 filename = 'dynamic_pro_model_stim_pause_release')

plot.changes.wrt(data.frame(ode(y = c(P = pause.peak , B = body.peak),
             times = seq(0,10000, by = 1),
             func = density.prime,
             parms = c(kinit=kinit,
                       kpre=kpre,
                       kterm=kterm,
                       krel=krel.stim,
                       lp = lp,
```

```
                        lb = lb))), 'dynamic_pro_model_stim_pause_PGBdensities_ss')

#plot the steady state profile as well.
pause.peak.stim = kinit/(kpre + krel.stim)
body.peak.stim = (((lp/lb)*krel.stim)*pause.peak.stim)/kterm
steady.state.stim.pro = dynamic.pro.profile(input = data.frame(cbind(22, pause.peak.stim, body.peak.stim)))



#prepare the simulated composites
dynamic.pro.stimulate.pause.release = dynamic.pro.profile(input = result.stimulate.pause.release)

#plot composites in lattice
plot.pro.simulation.composites(dynamic.pro.stimulate.pause.release,
                               filename = 'dynamic_pro_model_density_stim_pause_release')

#steady state composite upon stimuating pause release
plot.pro.simulation.composites(steady.state.stim.pro,
             filename = 'steady_state_pro_model_stim_pause_density')
```

References

Bailey TL, Williams N, Misleh C, Li WW (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." *Nucleic acids research*, **34**(suppl 2), W369–W373.

Benner C, Konovalov S, Mackintosh C, Hutt KR, Stunnenberg R, Garcia-Bassets I (2013). "Decoding a signature-based model of transcription cofactor recruitment dictated by cardinal cis-regulatory elements in proximal promoter regions." *PLoS Genet*, **9**(11), e1003906.

Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, Siepel A, Kraus WL (2013). "Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells." *Molecular cell*, **50**(2), 212–222.

Duarte FM, Fuda NJ, Mahat DB, Core LJ, Guertin MJ, Lis JT (2016). "Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation." *Genes & development*.

Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, Kraus WL (2011). "A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells." *Cell*, **145**(4), 622–634.

He X, Samee MAH, Blatti C, Sinha S (2010). "Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression." *PLoS Comput Biol*, **6**(9), e1000935.

Jonkers I, Kwak H, Lis JT (2014). "Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons." *Elife*, **3**, e02407.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002). "The human genome browser at UCSC." *Genome research*, **12**(6), 996–1006.

Sathyan KM, McKenna BD, Anderson WD, Duarte FM, Core L, Guertin MJ (2019). "An improved auxin-inducible degron system preserves native protein levels and enables rapid and specific protein depletion." *Genes & development*, **33**(19-20), 1441–1455.

Scholes C, DePace AH, Sánchez Á (2017). "Combinatorial gene regulation through kinetic control of the transcription cycle." *Cell systems*, **4**(1), 97–108.

Wang Z, Chu T, Choate LA, Danko CG (2019). "Identification of regulatory elements from nascent transcription using dREG." *Genome research*, **29**(2), 293–303.