TRANSPARENT PROCESS

OPEN ACCESS

molecular systems biology

Unique features of transcription termination and initiation at closely spaced tandem human genes

Noa Nissani & Igor Ulitsky*

Abstract

The synthesis of RNA polymerase II (Pol2) products, which include messenger RNAs or long noncoding RNAs, culminates in transcription termination. How the transcriptional termination of a gene impacts the activity of promoters found immediately downstream of it, and which can be subject to potential transcriptional interference, remains largely unknown. We examined in an unbiased manner the features of the intergenic regions between pairs of 'tandem genes'—closely spaced (< 2 kb) human genes found on the same strand. Intergenic regions separating tandem genes are enriched with guanines and are characterized by binding of several proteins, including AGO1 and AGO2 of the RNA interference pathway. Additionally, we found that Pol2 is particularly enriched in this region, and it is lost upon perturbations affecting splicing or transcriptional elongation. Perturbations of genes involved in Pol2 pausing and R loop biology preferentially affect expression of downstream genes in tandem gene pairs. Overall, we find that features associated with Pol2 pausing and accumulation rather than those associated with avoidance of transcriptional interference are the predominant driving force shaping short tandem intergenic regions.

Keywords Argonaute proteins; genome organization; Pol2 transcription; tandem genes; transcriptional interference

Subject Categories Chromatin, Transcription & Genomics

DOI 10.15252/msb.202110682 | Received 9 September 2021 | Revised 9 March 2022 | Accepted 10 March 2022

Mol Syst Biol. (2022) 18: e10682

Introduction

Transcription of messenger RNAs (mRNAs) and long noncoding RNAs (lncRNAs) begins with the heavily regulated transcription initiation where RNA polymerase 2 (Pol2) complex is assembled on promoters. It proceeds with transcriptional elongation, which is often coordinated with processing of the nascent RNA (capping its 5′ and splicing out the introns), and culminates with 3′ end formation, which almost universally involves cleavage and polyadenylation (CPA) of the nascent RNA transcript following a polyadenylation signal (PAS) and the addition of a poly(A) tail. As opposed to transcription initiation, the mechanisms of transcription termination are relatively less understood (Gruber & Zavolan, 2019).

It is important to distinguish between the site of CPA, which defines the end of the transcript and the site of Pol2 release from the DNA (or 'transcription termination site'), since Pol2 dissociates from the DNA between 100 bp to several kb downstream of the polyadenylation site (Hagenbüchle *et al*, 1984; Ashfield *et al*, 1991a; Tantravahi *et al*, 1993; Dye & Proudfoot, 2001). The disengagement of Pol2 from the DNA is thought to be important for both Pol2 recycling into the cellular pool and for insulating elongating Pol2 from downstream promoters to allow proper initiation at the downstream gene (Greger & Proudfoot, 1998; Gromak *et al*, 2006).

There are two main models for the dissociation of Pol2 from the DNA: the allosteric model and the torpedo model. The allosteric model proposes that once the elongating Pol2 passes over a functional PAS, it undergoes a conformational change (Zhang *et al*, 2015). This conformational change is thought to be mediated by the association of CPA factors with Pol2 CTD, which results in a Pol2 pause and its eventual release. The torpedo model connects between nascent RNA cleavage step and Pol2 release. In this model, XRN2, as part of the 5′-to-3′ RNA degradation machinery, engages with the free 5′ formed on the nascent transcript that is still being synthesized by Pol2 following CPA (Connelly & Manley, 1988; West *et al*, 2004). XRN2 digests this RNA faster than the speed of Pol2 elongation and thus acts as a torpedo until it reaches Pol2 and triggers its release from the DNA (Proudfoot, 2016). Indeed, the depletion of Xrn2 in mammalian cells or of its ortholog Rat1 in the yeast *Saccharomyces cerevisiae* has a strong effect on transcriptional termination (Kim *et al*, 2004; West *et al*, 2004).

Having a paused, or slowed-down, Pol2 downstream of termination sites is not a feature uniquely ascribed to the allosteric model, but it may also enhance XRN2-mediated termination and promote Pol2 recycling. Paused Pol2 may be giving an advantage to XRN2 in this race. G-rich elements are thought to enhance Pol2 pausing, possibly through promoting the formation of R-loop structures (DNA: RNA hybrids) (Skourti-Stathaki *et al*, 2011). Potentially related, a G-rich sequence motif, the MAZ element (G_5AG_5), bound by MAZ transcription factor (TF), was also shown to be required for efficient termination between closely spaced human complement genes C2 and factor B, which are separated by a mere 421 bp (Ashfield *et al*, 1991b, 1994). This sequence element was also shown to be sufficient for polyadenylation *in vitro*. In contrast, similarly G-rich Sp1

Departments of Biological Regulation and Molecular Neuroscience, Weizmann Institute of Science, Rehovot, Israel
*Corresponding author. Tel: +972-8-9346421; E-mail: igor.ulitsky@weizmann.ac.il

binding sites G_5CG_5 were not effective in this system. Pausing at such sequences was shown to be an intrinsic property of Pol2, at least *in vitro* (Yonaha & Proudfoot, 1999).

The different transcription steps are orchestrated by the modification of heptapeptide repeats of the free CTD of Pol2 largest subunit, RPB1 (Harlen & Churchman, 2017). At the promoter, Pol2 is mostly unphosphorylated. During initial elongation, Serine-5 and Tyrosine-1 gradually become phosphorylated (S5P and Y1P), which might aid with the recruitment of the RNA capping machinery. Later elongation stages are correlated with a decrease in S5P and an increase in S2P. Furthermore, S2P is associated with 3′ ends of genes and interacts with components of the CPA machinery. In addition, phosphorylated threonine-4 (T4P) has also been correlated with termination regions (McCracken *et al*, 1997; Hsin & Manley, 2012; Heidemann *et al*, 2013; Schlackow *et al*, 2017; Nojima *et al*, 2018), and phosphorylated Tyrosine-1 (Y1P) has been linked to termination in yeast (Mayer *et al*, 2012) and mammals (Shah *et al*, 2018).

While features of closely positioned promoters have been studied systematically (Chen *et al*, 2016), features of closely spaced termination and initiation regions have been less explored. The mammalian genome consists of two predominant types of regions—gene-rich and gene-poor—and those are associated with overall differences in sequence composition and other features, such as intron length (Amit *et al*, 2012). Gene-rich regions are usually characterized by more euchromatic regions, which are thought to facilitate wide expression. A subset of these clustered genes are found in particularly close proximity to each other (< 2 kb distance). A further subset of these precede each other on the same strand, we refer to such gene pairs as “tandem genes”. The precision of the termination process is thought to have significant importance especially in the case of tandem genes, as it ensures the production of two stable transcripts, avoiding a readthrough transcript starting at a non-properly terminated upstream gene and continuing at the promoter of the downstream gene (Shearwin *et al*, 2005; Proudfoot, 2016). There are examples of such transcriptional interference between adjacent genes having regulatory roles in model organisms (Nguyen *et al*, 2014), and specific proteins as well as appropriate Pol2 speed were shown to ensure efficient transcriptional termination (Krzyszton *et al*, 2018; Yu *et al*, 2019; Leng *et al*, 2020), suggesting that uncontrolled termination can cause substantial crosstalk between adjacent transcription units. In cases where both genes are well-expressed in the same cells, we expect that efficiency of termination is particularly important to allow both transcripts to be expressed at the right levels, and we hypothesize that this efficiency is encoded in the intergenic sequence. Notably, this intergenic sequence may optimize both avoidance of interference between the two transcripts, as well as potential recycling of the terminating Pol2 to be re-used in transcription of the downstream gene, although experimental evidence of such recycling is currently lacking and is difficult to obtain.

We became interested in the potential crosstalk between adjacent transcriptional units following our observation that in mouse embryonic fibroblasts, loss of the *Chaserr* lncRNA and the consequent increased dosage of CHD2 chromatin remodeler leads to repression of promoters found within 2 kb of highly transcribed genes on the same strand (Rom *et al*, 2019). This observation suggested that there is potential for substantial transcriptional crosstalk between closely spaced tandem genes in mammalian cells, echoing studies in yeast (Martens *et al*, 2004; Hainer *et al*, 2011; Pruneski *et al*, 2011; Thebault *et al*, 2011) and in other species (Nguyen *et al*, 2014; Shuman, 2020). This motivated us to look more broadly at features and perturbation sensitivities shared by such transcriptional units, which we describe below.

Results

Prevalence of closely spaced and co-expressed tandem gene pairs in the human genome

We first analyzed the overall prevalence of adjacent closely spaced genes found on the same strand. We grouped pairs of adjacent human protein-coding genes, considering genes > 5 kb and < 800 kb in length, and removing all pairs of overlapping genes (see Materials and Methods). We then split them into groups based on their genomic orientation relative to each other: divergent genes, situated on different strands with transcription facing in opposite directions (2,737 pairs, 23.2%); convergent, pairs on different strands, with transcription directed towards each other (2,866 pairs, 24.3%); or tandem genes, transcribed from the same strand (6,177 pairs, 52.4%). The distribution of the relative orientations of genes in the human genome is thus roughly as expected by chance. Adjacent gene pairs were further separated into groups based on their distance from one another (defined as the minimal distance between transcribed bases). Interestingly, the proportional share of closely spaced genes (< 2 kb and 2–5 kb) was generally higher than expected, with the exception of divergent genes separated by 2–5 kb, which were relatively depleted, whereas divergent genes in the < 2 kb category were relatively more common, as expected for genes that can share a common promoter (Fig 1A and Dataset EV1). Notably, 222 (49%) of the close tandem pairs, 285 (51%) of the close divergent pairs, and 315 (60%) of the close convergent pairs have homologs that meet the same criteria in the mouse genome (see Materials and Methods), suggesting that close tandem spacing is not strongly avoided. We then focused on the set of tandem genes, and analyzed in detail 457 pairs of adjacent, closely spaced (< 2 kb) genes (with the constraints imposed above). We further classified tandem pairs as co-expressed (188 pairs, ~41%) based on HepG2 cell line expression data from the ENCODE project (Djebali *et al*, 2012) (we obtained very similar results when using ENCODE K562 data, and so focused on HepG2 in the rest of the analysis, unless indicated otherwise) (see Materials and Methods and Dataset EV2–EV10). We analyzed the co-expression proportion in the other subgroups divided by orientations and distances and found that co-expression was most common in the closely spaced pairs of genes, regardless of their orientation (Fig 1A). Furthermore, this trend remained if the distance between the genes was defined as the distance between their promoters (Fig 1B). We conclude that compared to the convergent orientation, there are fewer tandem gene pairs separated by < 2 kb, perhaps because the A/T-rich polyadenylation signals are less likely to co-occur with G/C-rich promoters (see below). Nevertheless, when the distance between adjacent genes is short, there is no evidence that tandem genes are less likely to be co-expressed, which could be expected if transcriptional interference was posing a substantial obstruction to transcription from downstream promoters (in which case we would expect to see selection for avoidance of co-expression, specifically for tandem pairs), although such interference can still come from unannotated ncRNAs that we do not consider here. As expected, there was a mild yet significant correlation between the expression levels of the two tandem genes (Fig EV1A, Spearman R = 0.24 $P = 0.007$, ENCODE HepG2 RNA-seq data), but there was no preference for the upstream or downstream transcripts to be more abundant ($P = 0.49$ for Wilcoxon paired test comparing expression of the upstream and the downstream genes in HepG2 cells, Bonferroni-adjusted $P > 0.1$ for each of seven other cell lines examined).

Sequence features of intergenic regions separating tandem co-expressed gene pairs

We reasoned that co-expressed pairs of closely spaced tandem genes, that need to be produced at similar conditions in many tissues, would require, at least in some tissues where they are highly expressed, a more precise regulation over the termination process. Additionally, we considered the possibility that the vicinity of the CPA site of one gene to the promoter of another may facilitate the recycling of Pol2 machinery via some sequence or transcriptional features within the "short tandem intergenic region" (STIR). As a group, closely spaced co-expressed tandem genes tend to have substantially shorter introns than other genes (Fig 1C), consistent with previous observations about differences between gene-rich and gene-poor regions of the human genome (Amit *et al*, 2012). Markedly, introns of co-expressed tandem genes are also shorter than those of co-expressed closely spaced divergent or convergent genes. These differences dictated our strategy for selecting control genomic regions. As controls for the set of co-expressed adjacent tandem gene pairs, we matched for each pair of a gene A found upstream of B five "promoter-control" genes for B—these genes had a similar intronal length and expression levels as B (based on ENCODE expression data mentioned above), but no close upstream gene within at least 5 kb. Similarly, we matched five "3′ control" genes for A—these genes had a similar intronal length and expression pattern to A, but no close downstream neighbor (Fig EV1B and Materials and Methods). Because the set of expressed genes differs between cell lines, the set of controls was also cell-line–specific. Importantly, in the following analyses, the lengths of the control regions we use are the same as those of the STIRs. For example, as controls for the 789 bp STIR between the genes *TULP1* and *TEAD3* we used 789 bp upstream of the TSS of five genes expressed similarly and with similar intronal length as TEAD3, and 789 bp regions downstream of the CPA sites of five genes expressed similarly and with similar intronal length as TULP1. These controls were used in all subsequent analyses.

We considered the possibility that read-through transcription through the STIR might occasionally generate fusion transcripts which may confound our analysis. Such read-through has been observed upon different perturbations in human cells (Vilborg *et al*, 2017; Arnold *et al*, 2021). Several lines of evidence suggest that these events are exceedingly rare in unperturbed cells. First, the examination of ChRNA-seq data of nascent transcripts showed that read coverage immediately upstream of the TSS of the downstream gene is reduced to almost baseline levels (Fig EV1C). Second, splicing efficiency of the downstream gene (which might be reduced if parts of it correspond to a long 3′ UTR of the upstream gene) was rather indistinguishable from that of controls, including in the first intron, where splicing efficiency was even greater in tandem genes compared to controls (Fig EV1D and E). Third, when considering long-read nano-COP data (Drexler *et al*, 2020), of the 20,925 reads mapping to either the upstream or the downstream gene in a tandem pair, only 122 (0.6%) overlapped exons of both genes, and none contained spliced-out introns in both genes. For only 1.8% of the 131 tandem gene pairs for which both genes had at least 10 nano-COP reads, there were at least 2 reads overlapping both genes. In contrast, 90% of the reads that overlapped the CPA site of the upstream gene overlapped at least 50 nt of the STIR, and 57% of the upstream genes had at least 2 reads aligning to both the gene and the STIR. These results suggest that whereas Pol2 that transcribes the upstream gene continues to transcribe after the CPA site, as expected, it very rarely extends the upstream transcript beyond the TSS position of the downstream gene.

Short tandem intergenic regions were generally slightly more GC-rich than both types of control regions (58% G/C on average for STIRs compared to 46% and 55% at 3′ control regions and promoter regions in HepG2 cell line, respectively) (Fig 1D). The nucleotide composition at STIRs was more similar to the promoter control

Figure 1. Genomic architecture and nucleotide composition of tandem genes.

A Donut chart displaying the distribution of different gene orientations (inner circle, dataset of genes defined in the Materials and Methods section), distances between the genes (middle circle), and portion of co-expressed genes within each distance group (outer circle) defined in HepG2 cell line, using ENCODE RNA-seq data.

B Density plot of the $\log_{10}$-transformed distance between the promoters of tandem (top), divergent (middle), or convergent (bottom) co-expressed (light red) or non-co-expressed (yellow) pairs of genes. Co-expression was tested in the HepG2 cell line.

C Boxplot of distribution of the sums of introns lengths for co-expressed genes in the various genomic orientations. Co-expression was defined based on HepG2 data. Respective plotted group sizes are 188 upstream or downstream tandem genes, 570 divergent genes, 404 convergent genes, and 8,733 non-tandem genes. The thickened line represents the median intronal length of the genes, the lower and upper boxplot hinges correspond to first and third quartiles of the data, respectively. The whiskers represent the minimal/maximal existing values within 1.5 × inter-quartile range. Outliers were removed from the analysis. ****$P \leq 0.0001$; Wilcoxon rank-sum test.

D Overall nucleotide frequency within HepG2 co-expressed STIRs and within their promoter or 3′ controls. Shown are the standard deviation and paired Wilcoxon rank-sum test *P*-values between STIRs and either control. ****$P \leq 0.0001$.

E Metagene analysis showing binned nucleotide ratio within co-expressed STIRs (purple), promoter controls (beige) or 3′ controls (orange), of HepG2 cells. Heatmap shows the Bonferroni corrected *P*-value of paired Wilcoxon rank-sum test between STIRs and their respective controls within each bin.

F Metagene analysis of transposable element occupancy within HepG2 co-expressed STIRs or their respective controls. Bottom heatmap as in E.

G Average PhyloP conservation scores within HepG2 co-expressed STIRs between the CPA site of the upstream gene and the TSS of the downstream gene (purple), and in 1 kb flanking sequence. For the control genes, the TSSes of promoter control genes are aligned to the TSSes of the downstream genes (beige), and the CPA sites of 3′ control genes are aligned to the CPA sites of the upstream genes (orange).

▶

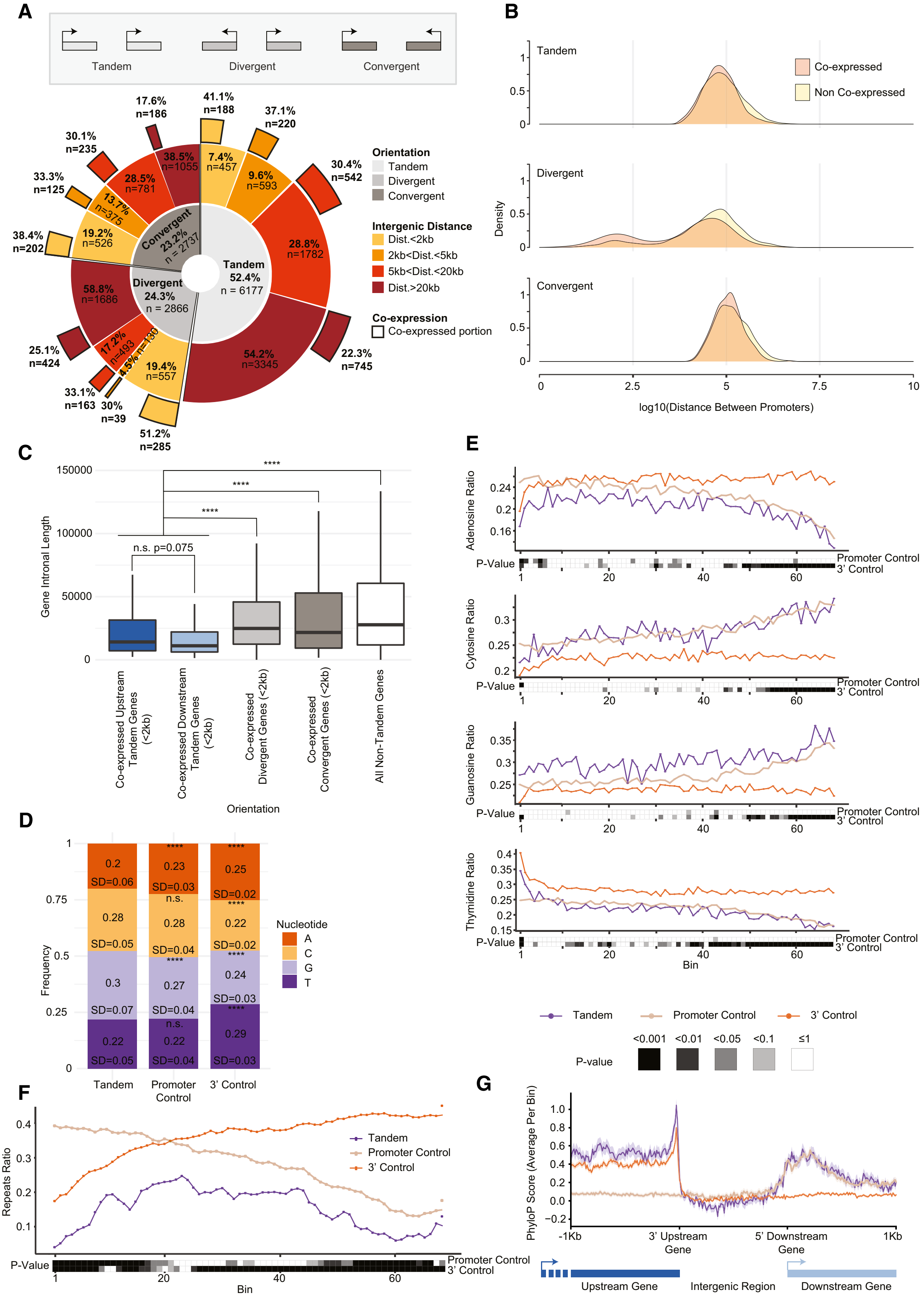
A
Tandem
Divergent
Convergent
Orientation
Tandem
Divergent
Convergent
Intergenic Distance
Dist.<2kb
2kb<Dist.<5kb
5kb<Dist.<20kb
Dist.>20kb
Co-expression
Co-expressed portion
Tandem 52.4% n = 6177
Divergent 24.3% n = 2866
Convergent 23.2% n = 2737
7.4% n=457
9.6% n=593
28.8% n=1782
54.2% n=3345
58.8% n=1686
17.2% n=493
4.5% n=130
19.4% n=557
19.2% n=526
13.7% n=375
28.5% n=781
38.5% n=1055
41.1% n=188
37.1% n=220
30.4% n=542
22.3% n=745
51.2% n=285
30% n=39
33.1% n=163
25.1% n=424
38.4% n=202
33.3% n=125
30.1% n=235
17.6% n=186
B
Tandem
Divergent
Convergent
Co-expressed
Non Co-expressed
Density
0
0.5
1
0
2.5
5
7.5
10
log10(Distance Between Promoters)
C
Gene Intronal Length
0
50000
100000
150000
****
****
****
n.s. p=0.075
Co-expressed Upstream Tandem Genes (<2kb)
Co-expressed Downstream Tandem Genes (<2kb)
Co-expressed Divergent Genes (<2kb)
Co-expressed Convergent Genes (<2kb)
All Non-Tandem Genes
Orientation
D
Frequency
0
0.25
0.5
0.75
1
0.2 SD=0.06
0.28 SD=0.05
0.3 SD=0.07
0.22 SD=0.05
****
0.23 SD=0.03
n.s.
0.28 SD=0.04
****
0.27 SD=0.04
n.s.
0.22 SD=0.04
****
0.25 SD=0.02
****
0.22 SD=0.02
****
0.24 SD=0.03
****
0.29 SD=0.03
Tandem
Promoter Control
3' Control
Nucleotide
A
C
G
T
E
Adenosine Ratio
0.16
0.2
0.24
Cytosine Ratio
0.2
0.25
0.3
Guanosine Ratio
0.25
0.3
0.35
Thymidine Ratio
0.15
0.2
0.25
0.3
0.35
0.4
P-Value
Promoter Control
3' Control
1
20
40
60
Bin
Tandem
Promoter Control
3' Control
P-value
<0.001
<0.01
<0.05
<0.1
≤1
F
Repeats Ratio
0.1
0.2
0.3
0.4
Tandem
Promoter Control
3' Control
P-Value
Promoter Control
3' Control
1
20
40
60
Bin
G
PhyloP Score (Average Per Bin)
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
-1Kb
3' Upstream Gene
5' Downstream Gene
1Kb
Upstream Gene
Intergenic Region
Downstream Gene

Figure 1.

regions than to the 3′ control regions (Fig 1E). Notably, when compared to the promoter controls, there was an enrichment of Gs, spanning the full length of the STIR, and the depletion of As in the beginning of the STIRs (Fig 1E). Furthermore, STIRs were depleted of transposable elements compared to both controls (Fig 1F). This may point to the functional importance of STIR sequences, but interestingly, when considering sequence conservation during vertebrate evolution, we observed slightly lower conservation levels in STIRs compared to control sequences. Notably, the 3′ region of the upstream co-expressed tandem gene shows some preferential conservation compared to its 3′ control gene set (Fig 1G). The proximity of a CPA site upstream of a promoter on the same strand thus seems to have a mild effect on the sequence composition in the STIR and on evolutionary conservation at the upstream CPA site.

G-rich sequence motifs are enriched in STIRs

The observation that STIRs have a biased sequence composition prompted us to look for specific enriched motifs within these regions. To this end, we used STREME (Sensitive, Thorough, Rapid, Enriched Motif Elicitation) from the MEME suite package (Bailey, 2011, 2021; Bailey *et al*, 2015) and filtered for motifs both enriched within STIRs compared to the background model (as defined by STREME) and more prevalent in STIRs relative to promoter and 3′ control regions using FIMO (Grant *et al*, 2011) (see Materials and Methods). We performed the enrichment analysis using both an "RNA" mode that seeks motifs only in the strand where both genes are transcribed, and a "DNA" mode, in which motifs are sought on both strands. Our analysis highlighted 'GGGGCGGG' and 'GGGGCGGGGSC' found in RNA (STREME $P = 0.023$ and $P = 0.035$, respectively) and 'CCTTCCC' found in DNA ($P = 0.02$) modes (Fig 2A and Appendix Fig S1A) as the lead motifs found in ~33%, ~49% and ~43% of the STIRs, as opposed to ~22%, ~33% and ~30% of control promoter regions and ~3%, ~15% and 23% of the 3′ controls, and with ~0.53, ~0.97 and ~0.43 average motif occurrences in STIRs as opposed to averages of ~0.36, ~0.74 and ~0.3 or ~0.04, ~0.21 and ~0.23 in the promoter and 3′ controls, respectively. Notably, most enriched motifs in this analysis are particularly G/C-rich (Fig 2B and Appendix Fig S1A and B). We verified that these motifs were also enriched relative to random dinucleotide-preserving shuffled STIR sequences, suggesting their presence does not merely reflect the G-richness of STIRs (See Materials and Methods, Appendix Fig S1C).

Moreover, the examination of specifically the GGGGNGGGG motifs across the STIRs of co-expressed pairs in HepG2 and K562 showed enrichment in STIRs over both types of controls. As expected, we found enrichment of the 'GGGGCGGGG' variant, which is very similar to the 'GGGGCGGG' motif identified *de novo* by STREME. 'GGGGTGGGG' and to some extent 'GGGGAGGGG' were also enriched in STIRs. Intriguingly, this was not the case for 'GGGGGGGGG' which was relatively depleted from all sequences inspected, and further depleted in STIRs. Interestingly, the enrichments are evident when we consider the G-rich motifs but not necessarily their C-rich reverse complements, pointing to the specific importance of Gs on the transcribed strand (Fig 2C and D and Appendix Fig S1D), and consistent with the overall enrichment of Gs mentioned above. Intersection of the enriched motifs with the JASPAR database (Castro-Mondragon *et al*, 2022) using Tomtom (Bailey *et al*, 2009) (see Materials and Methods) pointed at three TFs as potentially binding these motifs, MAZ, Sp5 (a member of the Sp1 family, discussed below), and ZNF148, which is a relatively poorly studied TF.

MAZ elements are G-rich motifs that were previously associated with efficient termination and are being used as termination signals *in vitro* and *in vivo* at individual genes (Ashfield *et al*, 1991a, 1994; Yonaha & Proudfoot, 1999). Since our candidate motifs resemble the 'MAZ' element G_5AG_5, and due to the general enrichment of G-rich motifs, we checked whether there is preferential binding of MAZ in STIRs. Notably, ENCODE ChIP-seq data of MAZ in HepG2 cells displayed a slightly higher binding signal in STIRs compared to both types of controls, with no significant difference between STIRs and promoter controls (Fig 2E). Notably, the majority of the signal stemmed from the promoter of the downstream gene and not the 3′ of the upstream gene or the control. A similar known G-rich element is the Sp1 binding motif (G/T)GGGCGG(G/A)(G/A)(C/T) (Nagaoka *et al*, 2001), and it has been previously proposed that MAZ but not Sp1 contribute to Pol2 pausing downstream of the CPA site (Yonaha & Proudfoot, 1999). Interestingly, Sp1 also showed a slightly but not significantly higher binding signal in STIRs compared to promoter controls in HepG2, closely resembling the binding pattern of MAZ (Fig 2F).

Short tandem intergenic regions therefore preferentially harbor G-rich motifs, and those are typically located proximally to the downstream promoter, and are associated with a slightly increased binding of MAZ and to a lesser extent SP1 near the downstream promoter.

Specific proteins preferentially bind STIRs

These findings led us to seek genome-wide evidence for preferential binding of other proteins within STIRs. Using ENCODE ChIP-seq data, we considered 338 factors profiled with ChIP-seq by the ENCODE project in HepG2 and/or K562 cells. When considering just the control regions, the factors were generally much more likely to bind upstream of control TSSs (in regions length-matched to STIRs) than to regions downstream of control CPA sites (Fig EV2A and B). When comparing the fraction of STIRs bound by each factor to the fractions of the respective control regions, we found enriched binding of several factors in STIRs, including AGO2, AGO1, RBM22, BCL3, MYNN and NFATC1 (Figs 3A–G and EV2C–F).

Interestingly, G-rich stretches and AGO proteins were previously implicated in transcriptional termination at specific genes (Skourti-Stathaki *et al*, 2014). The model of Pol2 pausing following PAS transcription suggests that G-rich terminator elements found tens of bp downstream of the CPA site further enhance Pol2 pausing. This pausing was suggested to be facilitated by R-loop structures reported to be particularly enriched downstream of the CPA sites of some genes. A study by (Skourti-Stathaki *et al*, 2011) suggested that R-loops at G-rich regions in termination regions of the β-actin gene induce antisense transcription, which leads to the generation of dsRNA and the recruitment of the RNA-interference (RNAi) factors such as AGO1, AGO2, DICER, and the G9a histone lysine methyltransferase. This recruitment was reported to lead to deposition of the H3K9me2 repressive mark and to the recruitment of heterochromatin protein 1γ (HP1γ), which in turn may reinforce Pol2 pausing and promote transcription termination. Since we observed

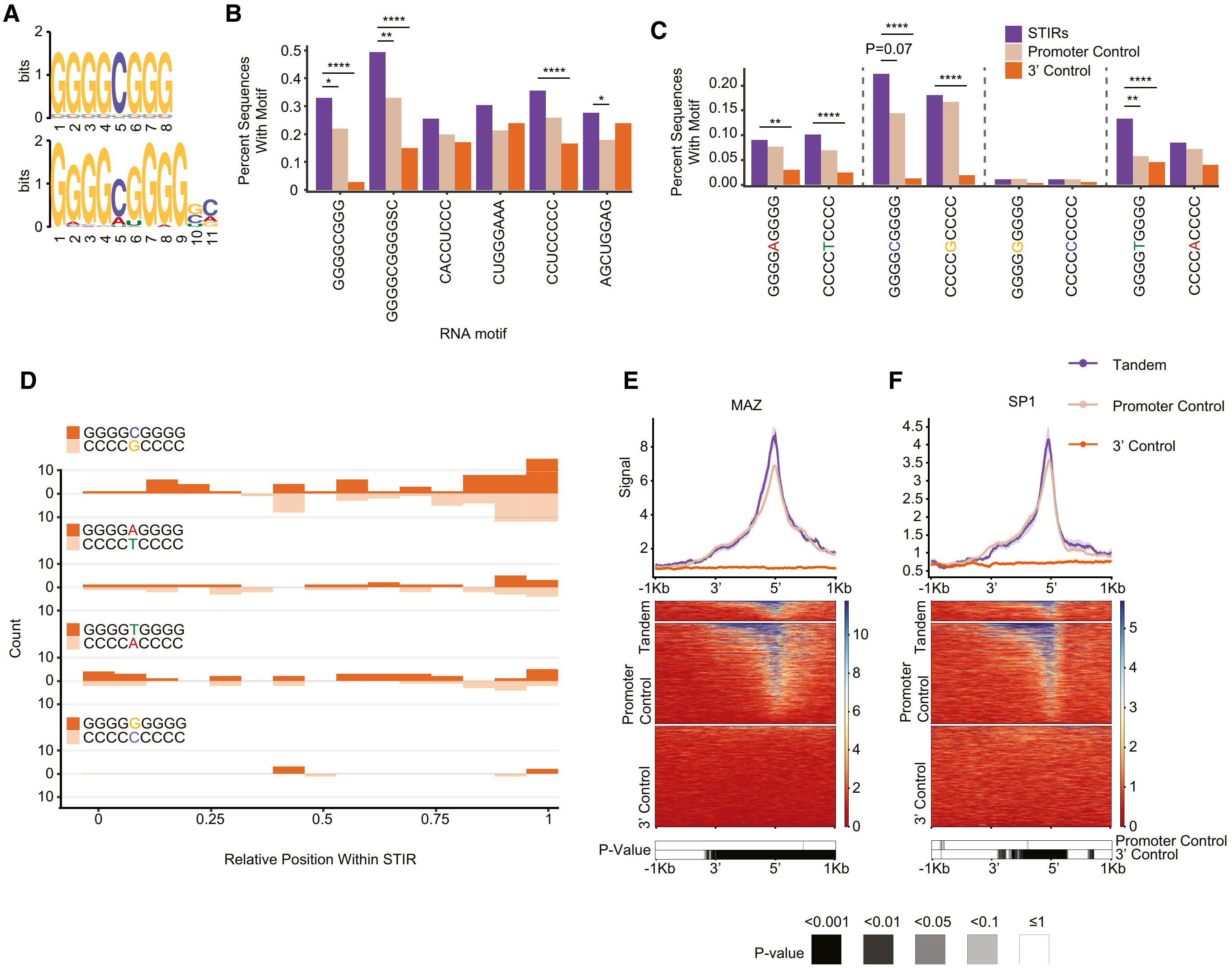

Figure 2. Enriched motifs in STIRs.

A Logo representation of motifs identified by STREME in the "RNA" mode as enriched in either HepG2 (top) or K562 (bottom) co-expressed STIRs.

B Barplots showing the proportion of STIRs (purple) or control sequences (beige and orange) that carry the RNA-mode STREME-discovered motif.

C Barplots showing the proportion of co-expressed STIRs (purple) or control sequences (beige and orange) for the G_4NG_4 motifs (or their reverse complement), for co-expression in HepG2 cells.

D Histogram showing the binned distribution of G-rich motifs (or their reverse complement) across the co-expressed STIRs of HepG2 cells.

E Metagene analysis (top) and corresponding binding heatmap (center) of MAZ in ENCODE ChIP-seq data in HepG2 co-expressed STIRs and flanking 5′ and 3′ regions and at the control regions (same controls as in Fig 1 and throughout the manuscript). Top graph shows median and standard error and the bottom heatmap shows the binned Bonferroni-corrected paired Wilcoxon rank-sum test *P*-value heatmap.

F As in (E), but for the SP1 ChIP-seq data in HepG2 cells.

Data information: (B, C) Shown are Bonferroni corrected proportion test *P*-values (**P* ≤ 0.05, ***P* ≤ 0.01, *****P* ≤ 0.0001).

significant binding of both AGO1 and AGO2—components of the RNAi machinery in STIRs, we sought to further examine the other factors tied with the same pathway in these regions. Analyzing data of RNA:DNA hybrids profiled using the S9.6 antibody in K562 cell line by (Sanz *et al*, 2016), we indeed observed a substantial enrichment of R-loops in STIRs (Fig 4A). To further examine the model suggested by Skourti-Stathaki *et al*, we examined the binding of the other factors using available ENCODE data, including G9a (EHMT2)—an H3K9 methyltransferase, HP1γ (CBX3), and the H3K9me2 chromatin mark. All these factors were not enriched in STIRs (Fig EV3A–C). Conversely, G9a showed ~30% depletion within STIRs, peaking at the promoter, along with slightly lower levels of H3K9me2 binding overall, precluding the promoter region, which was similarly depleted of H3K9me2 (Figs EV3A and C). Notably, G9a has additional substrates, such as H3K9 (Shinkai & Tachibana, 2011), therefore it might be recruited there under other circumstances. Alternatively, we hypothesized that the relatively low levels of H3K9me2 and its histone lysine methyltransferase within STIRs

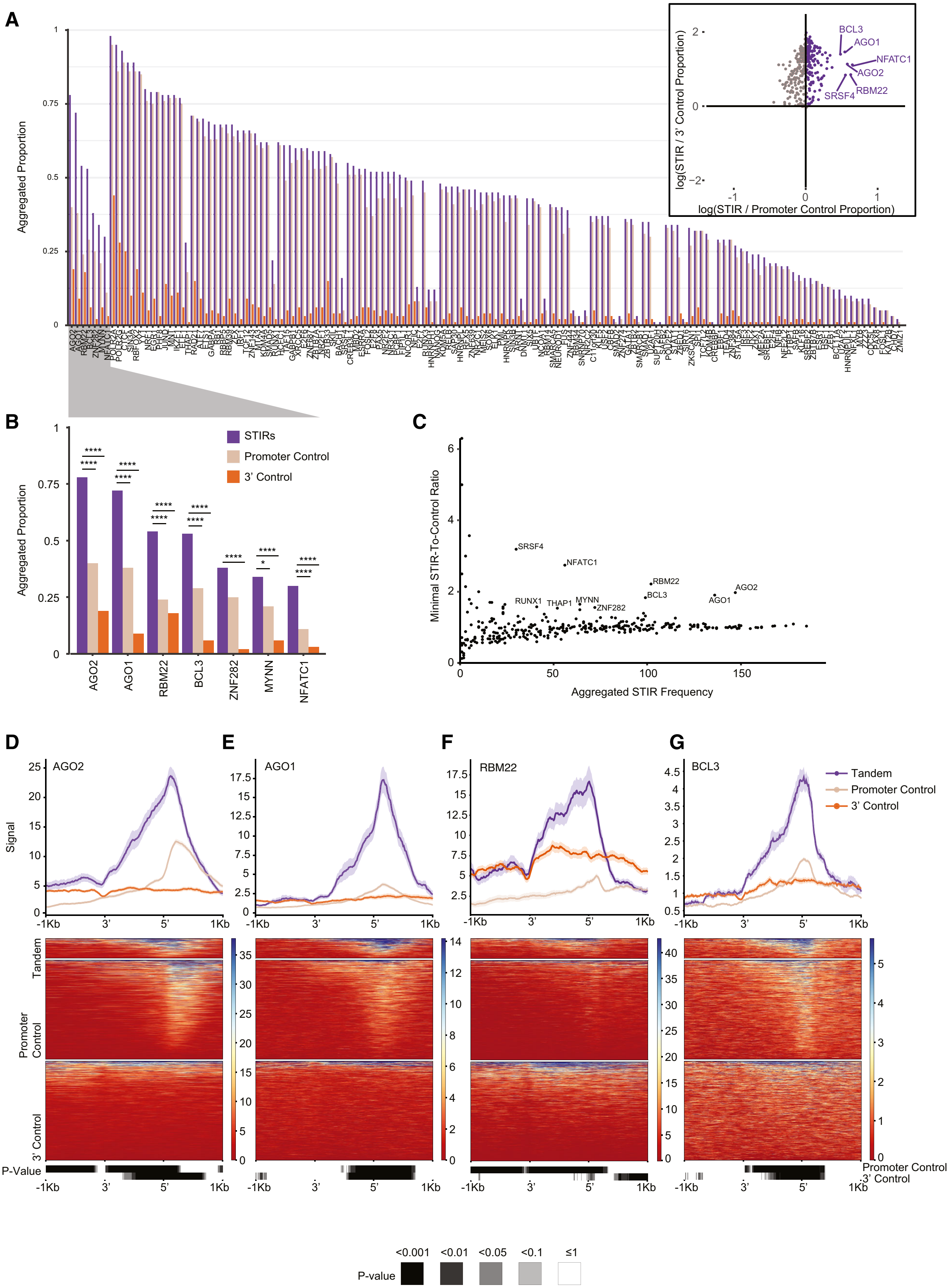
A
Aggregated Proportion
1
0.75
0.5
0.25
0
AGO2
AGO1
RBM22
BCL3
ZNF282
MYNN
NFATC1
log(STIR / 3' Control Proportion)
log(STIR / Promoter Control Proportion)
2
0
−2
-1
0
1
BCL3
AGO1
NFATC1
AGO2
RBM22
SRSF4
B
Aggregated Proportion
STIRs
Promoter Control
3' Control
****
*
AGO2
AGO1
RBM22
BCL3
ZNF282
MYNN
NFATC1
C
Minimal STIR-To-Control Ratio
Aggregated STIR Frequency
0
50
100
150
2
4
6
SRSF4
NFATC1
RBM22
BCL3
AGO2
AGO1
RUNX1
THAP1
MYNN
ZNF282
D
E
F
G
AGO2
AGO1
RBM22
BCL3
Signal
Tandem
Promoter Control
3' Control
-1Kb
3'
5'
1Kb
Tandem
Promoter Control
3' Control
P-Value
Promoter Control
3' Control
P-value
<0.001
<0.01
<0.05
<0.1
≤1

Figure 3.

◀ **Figure 3. Enriched binding of proteins at STIRs.**

A Barplot showing the proportion of co-expressed STIRs (purple), promoter control- (beige), and 3′ control- (orange) sequences bound by the different proteins (analyzed in HepG2 cells). Bound sequences were aggregated and counted once when multiple binding peaks per sequence were observed. Shown are only proteins with higher binding frequency at STIRs over both controls. Proteins are ordered by ranked frequency in STIRs and ranked descending order of calculated minimal ratio of frequencies between STIR and each of the two controls. Inset scatter plot shows the log-transformed proportion ratio between binding sites at STIRs or at either control, in purple are proteins enriched in STIRs over both types of controls. Indicated are several notably enriched proteins.

B Zoom-in on the top STIRs protein binding candidates in HepG2 cells. Shown are Bonferroni corrected proportion test *P*-values (**P* ≤ 0.05, *****P* ≤ 0.0001).

C Scatter plot showing the number of HepG2 co-expressed STIRs bound by each protein (as in (A)) versus the minimal tandem-to-control ratio calculated for each control. Indicated are the top proteins enriched in STIRs. Y-axis values > 6 were filtered out.

D–G Metagene analysis (top) and corresponding binding heatmap (center) of median ChIP-seq signal of selected STIR-binding protein candidates AGO2 (D), AGO1 (E), RBM22 (F) and BCL3 (G). Bottom heatmap corresponds to binned paired Wilcoxon rank-sum tests (Bonferroni corrected).

can be explained by the tendency of tandem genes to reside within relatively gene-rich regions, which are generally less associated with heterochromatic marks (Gilbert *et al*, 2004; Sanz *et al*, 2016). Therefore, we next examined the binding of PHF8, an H3K9me2 demethylase (Zhu *et al*, 2010). Interestingly, PHF8 levels were somewhat higher in control promoter regions than in STIRs, perhaps because it is recruited by H3K9me2, which was also higher in control promoter regions (Fig EV3C and D). We conclude that whereas AGO binding and R-loops are prevalent in STIRs genome-wide, there is no genome-wide evidence for preferential activity of the H3K9me2-associated machinery in these regions.

AGO2 binding promotes the expression of tandem gene pairs

Notable enrichment of AGO1 and AGO2 binding in STIRs led us to further look for possible consequences of the binding. We examined the changes in the ENCODE gene expression dataset of polyadenylated RNA following AGO1 and AGO2 knockdown (KD) in K562 and HepG2 cell lines. For AGO2 KD we observed a significant yet mild decrease in the expression of the tandem co-expressed genes compared to their respective controls (Fig 4B). For AGO1, we observed smaller changes in the same direction, which were significant only in HepG2 cells, and only for the downstream gene in the tandem pair (Figs 4C and EV3E). Examination of the correlation between the downstream and upstream tandem genes expression changes following the KD showed little to no correlation between the tandem pairs (Figs 4D and EV3F–H), except the notable scarcity of tandem gene pairs where both genes were up-regulated following AGO2 KD (Fig EV3H).

While there was no significant change in expression when considering all the tandem pairs, when we integrated ENCODE K562 AGO1 ChIP-seq data, we observed increased binding of AGO1 to chromatin in the STIRs that were associated with relative downregulation of gene expression of both the upstream and downstream genes following the KD (Wilcoxon test $P < 0.05$). Furthermore, we found a significant negative correlation between AGO1 binding and the change in the expression of the upstream gene following KD (Spearman's correlation coefficient: −0.295, $P = 1.2 \times 10^{-4}$) and a similar yet non-significant (Spearman's correlation coefficient: −0.134, $P = 0.088$) effect for AGO1 binding and the change in expression of the downstream gene (Fig 4D). In addition, examining the connection between R-loop signal and AGO1 KD in K562 cells showed significant correlation between changes in the downstream co-expressed tandem gene expression and R-loop signal ($P = 3.1 \times 10^{-4}$) (Fig 4E). However, R-loop signal intensity did not correlate with gene expression changes of co-expressed tandem genes following AGO2 KD in K562 cells (Fig EV3H).

Pol2 accumulates in STIRs

Due to the short distances between the tandem gene pairs that we considered, and the continued association of Pol2 with DNA downstream of the CPA site, with the addition of our constraints for picking only pairs of genes that co-express in the relevant cell type, we reasoned that STIRs might be associated with distinct patterns of Pol2 occupancy. To examine that, we used mammalian Native Elongation Transcript sequencing (mNET-seq) data from (Schlackow *et al*, 2017). The mNET-seq strategy uses antibodies for several different Pol2 CTD marks, to obtain segments of nascent-RNA bound by Pol2 in its different states. Importantly, in contrast to ChIP-seq, mNET-seq data are strand-specific, allowing to consider Pol2 traveling strictly on the same strand as the considered tandem genes. We used the same definitions described above to define a set of 159 tandem genes co-expressed in HeLa cells and their controls. Overall, ~1.25 fold enrichment over the promoter control of total Pol2 (phosphorylated and non-phosphorylated, tested using CMA601 antibody) was observed at the peak just downstream of the promoter in the downstream tandem gene (Fig 5A and Appendix Fig S2A). There was also substantial Pol2 presence within the STIR whereas it was absent from both controls (Fig 5A and Appendix Fig S2A), resulting in an overall 6.8 and 8.4-fold enrichment of Pol2 occupancy across the STIR, for the 3′ control and promoter control, respectively (see Materials and Methods). Interestingly, all modifications of Pol2 showed enrichment at tandem intergenic regions and at the downstream promoters with slightly different patterns. For example, T4P modification (using 6D7 anti-CTD Thr4-P modification antibody) showed a stable signal for the 3′ control, downstream of the CPA site, and a gradually increasing signal that was enriched by ~7 fold compared to the 3′ control in STIRs (Fig 5B). Surprisingly, the signal peaked at approximately the promoter of the downstream tandem gene followed by a decreasing signal in the following 0.5 kb that eventually reached background levels. As expected (Schlackow *et al*, 2017), T4P signal was generally depleted at the promoter controls (Fig 5B and Appendix Fig S2G). Another notable example is the S2P modification (tested with CMA602 Ser2-P CTD antibody), which showed an enrichment along the STIR that peaked downstream of the transcription start site and was then almost completely absent after 1 kb. Notably, in both controls, S2P seemed to be depleted, with a low signal at the promoter control just downstream of the promoter (Fig 5C and Appendix Fig S2B). Y1P modification was reported to be essential for the function of Positive Transcription Elongation Factor b (P-TEFb) in phosphorylating Ser2 and transitioning from transcription initiation to elongation (Mayfield *et al*, 2019), S5P modification was suggested to facilitate the recruitment

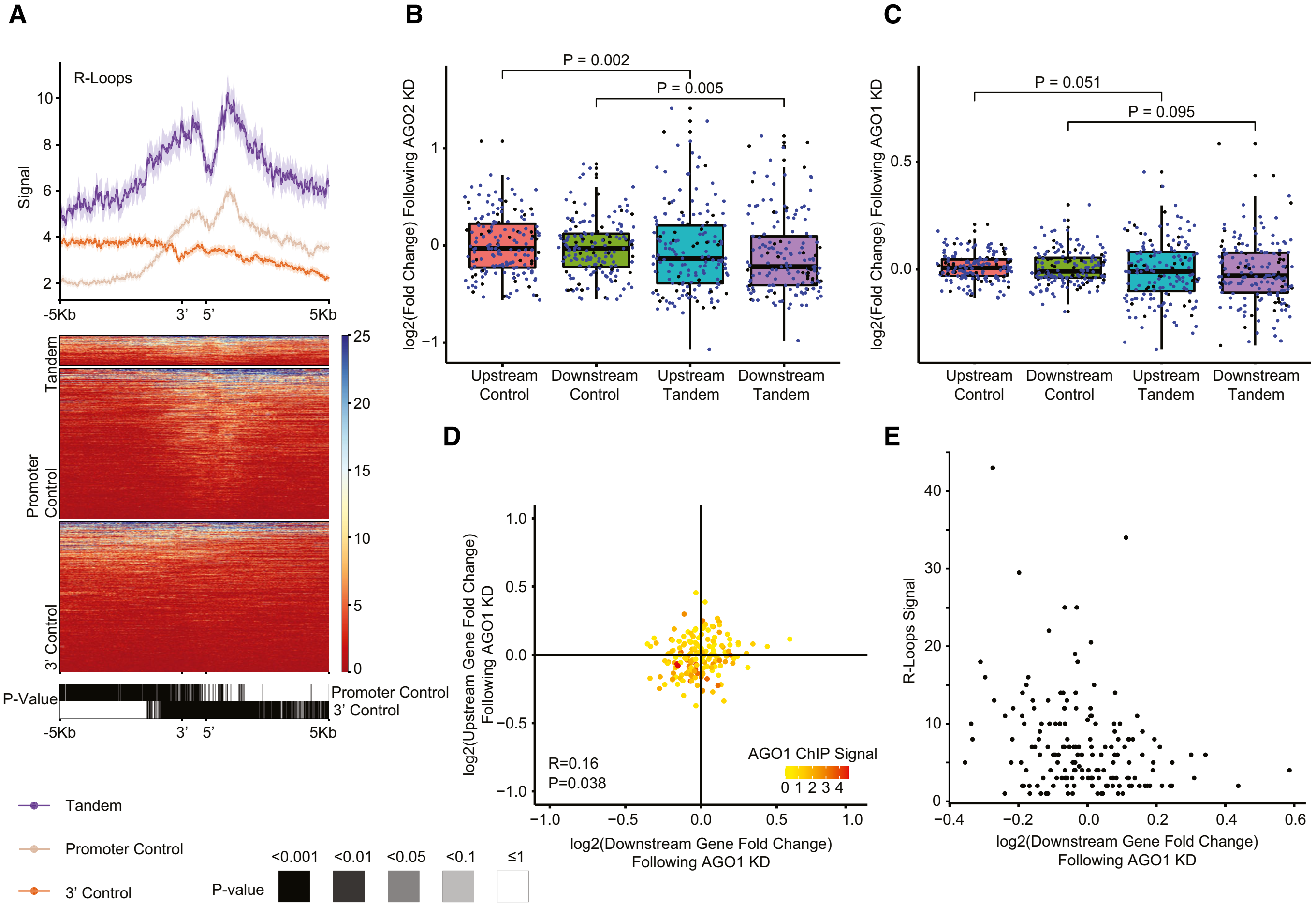

Figure 4. Features promoting the proper expression of tandem gene pairs.

A Metagene analysis (top) and corresponding heatmap (center) showing S9.6 DNA:RNA antibody DRIP-seq median signal in K562 co-expressed STIRs and flanking regions and in their controls. Bottom heatmap shows the corrected *P*-value of binned paired Wilcoxon rank-sum test. Data from Sanz *et al* (2016).

B Boxplot of expression changes in upstream or downstream co-expressed tandem genes (164 genes in each group) following AGO2 KD in K562 cells (Data from ENCODE). Five genes were used as controls per tandem gene (see Materials and Methods, Fig EV1B) and aggregated using the average $\log_2$-transformed fold change value of each quintet. Blue dots correspond to tandem pairs co-expressed in both K562 and HepG2 cell lines (or their respective control). Black dots are tandem genes co-expressed only in the respective cell line. The thickened line represents the median $\log_2$ fold change following AGO2 KD, the lower and upper boxplot hinges correspond to first and third quartiles of the data, respectively. The whiskers represent the minimal/maximal existing values within 1.5 × inter-quartile range. Outliers were removed from the analysis. *P*-values were obtained using paired Wilcoxon rank-sum tests.

C As in (B), for AGO1 KD in K562 cells.

D Scatter plot showing the changes in expression following AGO1 KD in K562 cells. Each dot represents the change in expression of a single tandem pair. Colors indicate median AGO1 ChIP-seq signal per STIR. Calculated Pearson's correlation coefficients between upstream- and downstream- tandem genes changes in expression and *P*-value are indicated. Spearman correlation between the downstream- or upstream- gene expression changes following KD and median AGO1 ChIP-seq signal at STIR was tested, with coefficients of −0.134 or −0.295 and $P = 0.088$ or $P = 1.2 \times 10^{-4}$, respectively. Wilcoxon test of STIR ChIP-seq signal tested between the third quadrant and all other quadrants: $P = 1.7 \times 10^{-3}$.

E Scatter plot showing median R-loop signal at STIRs as a function of the changes in expression of the downstream gene in the co-expressed tandem pairs following AGO1 KD in K562 cell line. Spearman correlation coefficient of −0.278, $P = 3.1 \times 10^{-4}$.

of the stabilizing capping enzyme complex (Ho & Shuman, 1999), and S7P is enriched at promoters and gene bodies, and is thought to regulate snRNA biogenesis (Egloff *et al*, 2007; Harlen & Churchman, 2017). For these three modifications, both the tandem genes and the promoter controls show a peak of signal enrichment just downstream of the promoter, with ~1.25-, ~4-, or ~2.5- fold higher signal at tandem downstream promoter over the promoter control genes, for Y1P, S5P or S7P, respectively (tested using 3D12 Tyr1-P CTD antibody, CMA603 Ser5-P CTD antibody, and 4E12 Ser7-P antibody, respectively). Interestingly, these signals exist also within the STIR, whereas they are depleted upstream of the promoter controls (Fig 5D–F and Appendix Fig S2C,D,F). Intriguingly, treatment with Pladienolide B (Pla-B), a splicing inhibitor that binds SF3B1 spliceosome subunit, made the pattern of S5P Pol2 binding almost indistinguishable between the tandem genes and the promoter controls, suggesting that splicing substantially contributes to the Pol2 accumulation in the intergenic region and around the promoter of the downstream gene. Potentially, the inhibition of splicing affects

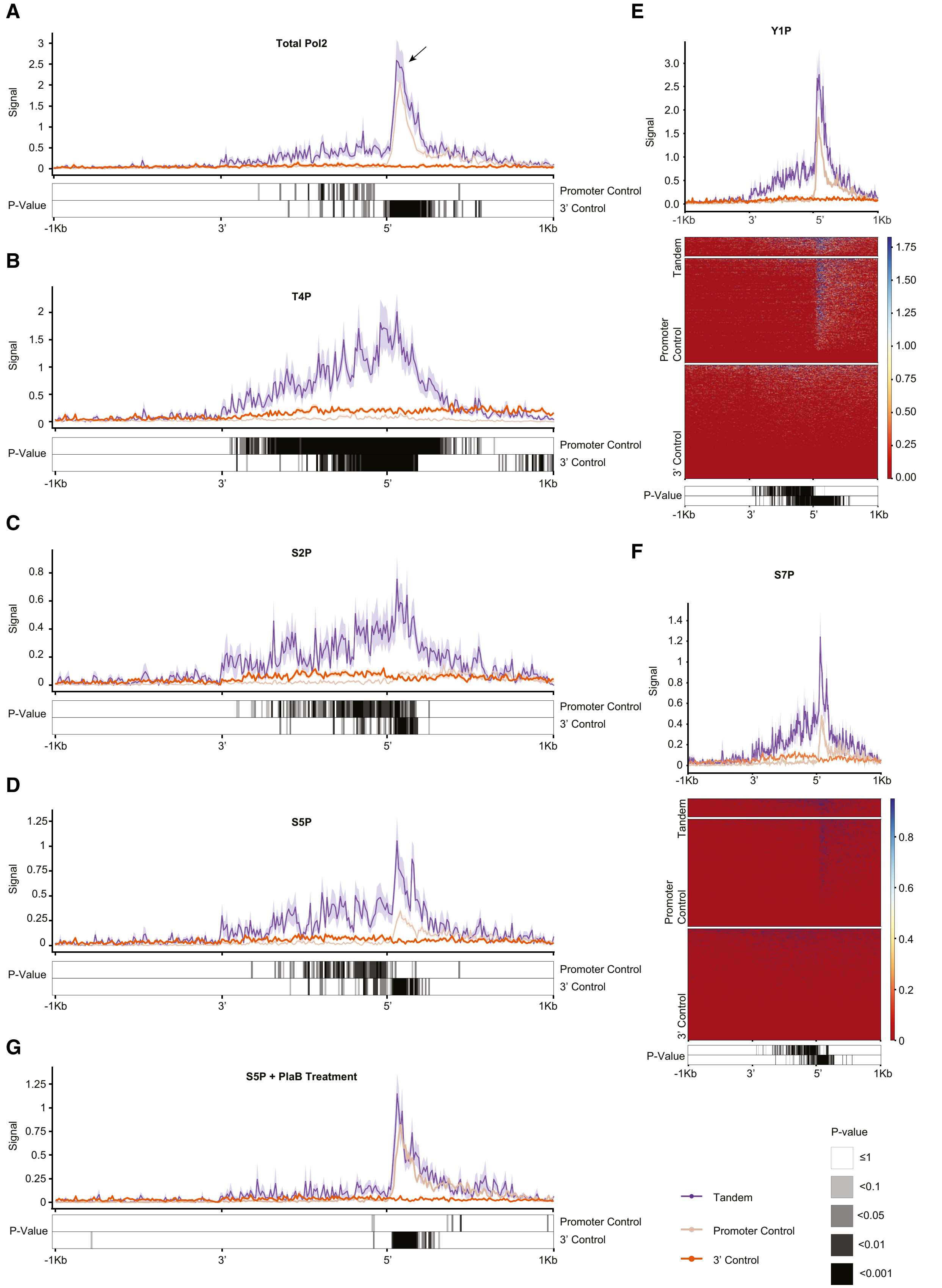
A
Total Pol2
Signal
3
2.5
2
1.5
1
0.5
0
P-Value
Promoter Control
3' Control
-1Kb
3'
5'
1Kb
B
T4P
C
S2P
D
S5P
G
S5P + PlaB Treatment
E
Y1P
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Tandem
Promoter Control
3' Control
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
F
S7P
1.4
1.2
0.8
0.6
0.4
0.2
P-value
≤1
<0.1
<0.05
<0.01
<0.001
Tandem
Promoter Control
3' Control

Figure 5.

◀ **Figure 5. Pol2 accumulates in STIRs.**

A–D Metagene analysis (top) showing the different median Pol2 carboxy-terminal domain modifications (phosphorylated and non-phosphorylated, T4P, S2P, and S5P, respectively) occupancy signal at the sense strand of STIRs or controls and their flanking regions. Bottom are the respective corrected paired Wilcoxon rank-sum test *P*-values. Data from Schlackow *et al* (2017). Shown is a representative replicate.

E, F Metagene analysis (top) and corresponding heatmap (center) of median Pol2 CTD modifications (Y1P and S7P, respectively) mNET-seq signal and standard error at STIRs and controls, and their flanking regions. Bottom are the respective corrected Wilcoxon-paired test *P*-values. Data from Schlackow *et al* (2017).

G As in D, but showing the occupancy signal of S5P modification for Pla-B treated cells.

elongation of Pol2 through the upstream gene, leading to paucity of Pol2 around the gene end. Yet, it is difficult to measure this scarcity for 3′ control regions, since Pol2 levels near gene ends are generally very low. Alternatively, Pol2 that is stalled at the intergenic region is associated with the transcripts produced at the transcription round that just terminated but are still engaged in the midst of splicing, and so splicing inhibition causes disengagement of such immobilized Pol2 (Fig 5G and Appendix Fig S2E). Interestingly, the overall presence of Pol2 within STIRs is dramatically higher than in the 3′ parts of transcribed genes, where the occupancy of Pol2, including all the different modifications, is barely over background levels (Fig 5A–F).

As further evaluation of the role of upstream gene elongation in the accumulation of Pol2 in STIRs, we also examined the consequences of loss of Spt6 and Rtf1, two factors reported to separately control Pol2 progression. We used mNET-seq data from K562 following Spt6 and Rtf1 depletion from (Žumer *et al*, 2021) and examined Pol2 distribution within STIRs (Fig EV4). Spt6 was shown to assist progression of elongating Pol2 through nucleosomes, and its depletion was shown to have little to no effect on Pol2 accumulated at promoters (Žumer *et al*, 2021). Indeed, data in K562 recapitulated the pattern we observed in HeLa cells, showing strong accumulation of Pol2 in STIRs in control conditions (DMSO-treated cells). This accumulation was strongly reduced upon targeted degradation of Spt6 (Fig EV4A and B), with a smaller effect on Pol2 in the downstream promoter. When examining Pol2 pausing under control conditions, tandem downstream genes had lower Pol2 pausing index than promoter controls. Spt6 degradation mildly reduced pausing at both tandem genes and their controls (see Materials and Methods and Fig EV4E–G). Interestingly, no effect was seen for Rtf1 depletion (Fig EV4C and D).

In order to study further the nature of the Pol2 accumulating at STIRs and at the downstream promoter, we analyzed data on Pol2 occupancy in HCT116 cells treated with conditions resulting in a high salt concentration (Erickson *et al*, 2018). Pol2 in STIRs and at the downstream promoter in a tandem pair was more resistant to NaCl or glucose treatments, that increase salt concentrations within the cells, than Pol2 pausing at the control promoters (Appendix Fig S3). These results suggest that at least some of the Pol2 complexes in the aforementioned regions are in an "elongation" rather than in a "pre-initiation/poised" state, as the latter state is more sensitive to high salt concentrations (Erickson *et al*, 2018).

NELF-E KD suggests involvement in transcription regulation of downstream tandem genes

We next tested whether the expression of tandem genes is particularly sensitive to the loss of specific protein factors. We examined this by analyzing KD data of 245 different protein factors, in 440 experiments (in HepG2 or K562 cell lines). Importantly, these data are derived from RNA-seq of poly(A)-selected transcripts, and so informative only for expression of mature RNA products. For most factors, we found concordant changes in expression of the upstream and downstream genes (Fig 6A), possibly related to their overall shared features (see Discussion), and so we were particularly interested in factors whose loss preferentially affected just the upstream or the downstream gene. This analysis highlighted Negative Elongation Factor Complex Member E (NELF-E), the knockdown of which led to an increased expression of the downstream gene in a tandem pair, while not affecting the upstream gene (Fig 6 and Dataset EV11). NELF-E is a part of the NELF complex (composed of units A, B,C/D and E). Based on current knowledge, release of Pol2 from the proximal-promoter region and its conversion to elongating state is dependent on the displacement of the negative elongation factor (NELF) complex from the nascent transcript. This process is thought to include the phosphorylation of several factors including Ser2 of the CTD, NELF-E, and the Spt5 subunit of the DSIF (DRB Sensitivity Inducing Factor) complex by positive transcription elongation factor b (P-TEFb). These phosphorylations are followed by NELF dissociation and Pol2 transition from promoter-proximal pausing to elongation (Lu *et al*, 2016). NELF-E KD in HepG2 cells led to the greatest median fold change in expression of the downstream tandem gene versus its controls accompanied by a negligible median KD effect over the upstream tandem gene compared to its controls (Fig 6B and Dataset EV11). The trend was similar in K562 cells yet the effect was less robust (Appendix Fig S4A). We next examined data from DLD-1 cells where NELF-C or NELF-E were tagged with an AID domain enabling inducible degradation (Aoi *et al*, 2020), and chromatin occupancy of various factors was examined before and after degradation (Appendix Fig S4B–E). Occupancy of NELF-C and NELF-E was similar between STIRs and control promoters, with slightly lower occupancy in STIRs (Appendix Fig S4B and C). Pol2 with both S2P and S5P modifications was strongly enriched in STIRs compared to controls, fitting the observations from the mNET-seq data (Appendix Fig S4D and E) and S2P-modified Pol2 signal in STIRs was substantially more sensitive to loss of NELF-C than control promoters. We conclude that reduction in NELF complex levels leads to reduction of S2P-modified Pol2 which accumulates at STIRs and to increase in the expression of the downstream gene. Future studies may elucidate the mechanistic connection between these changes.

In addition to NELF-E, we found SMN1 and XRN2 KD as having similar, yet less robust effects of upregulation of the downstream and not the upstream tandem gene (Fig 6A and Appendix Fig S4 F and G). Intriguingly, SMN was previously suggested to play a role in R-loops resolution (Zhao *et al*, 2016; Jangi *et al*, 2017). Fewer KD experiments negatively affected just the downstream gene. Notably, KD of U2AF1 had this effect in both K562 and HepG2 cells (Fig 6A), and mutations in U2AF1 in myelodysplastic syndromes are

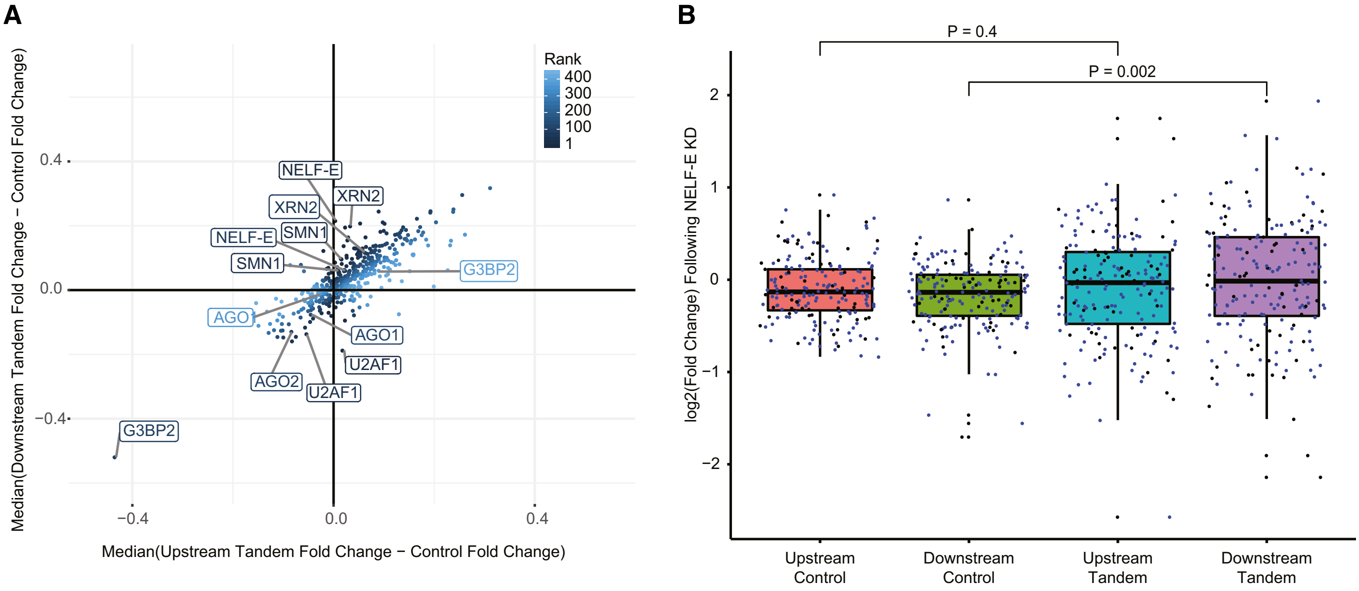

Figure 6. NELF-E involvement in transcription regulation of downstream tandem genes.

A Scatter plot showing the median difference in expression changes, in poly(A)-selected RNA-seq datasets, between downstream tandem (y-axis) or upstream tandem (x-axis) genes and the averaged expression changes in control. Candidate genes with high KD effect on expression compared to controls in the downstream tandem gene but not in the upstream tandem gene are ranked lower and appear darker and vice versa. Shown are results from 440 KD experiments done in either HepG2 or K562. KDs with the lowest ranking, and other proteins of interest are marked.

B Boxplot of expression changes in upstream or downstream 188 pairs of co-expressed tandem genes following NELF-E KD in HepG2 cells (data from ENCODE) or their averaged aggregated controls (5 controls per tandem gene). Blue dots correspond to tandem pairs co-expressed in both K562 and HepG2 cell lines (or their respective control). Black dots are tandem genes co-expressed only in the respective cell line. The thickened line represents the median $\log_2$-transformed fold change following NELF-E KD, the lower and upper boxplot hinges correspond to first and third quartiles of the data, respectively. The whiskers represent the minimal/maximal existing values within 1.5 × inter-quartile range. Outliers were removed from the analysis. *P*-values were obtained using paired Wilcoxon rank-sum tests.

associated with reduction of Pol2 pause release and R loop accumulation (Chen *et al*, 2018; Nguyen *et al*, 2018).

Discussion

Our results show that STIRs differ from regions flanking 5′ or 3′ ends of other genes in sequence composition, protein associations and in the strong accumulation of Pol2. These regions primarily resemble promoters, which implies that sequences required for regulated transcription initiation are longer and/or experience a stronger selective pressure compared to sequences required for efficient termination, and yet differ from other promoters, in particular in strong accumulation of Pol2 bearing CTD modifications traditionally associated with gene ends.

In our analysis, we focused on regions separating well-expressed protein-coding genes, but pervasive transcription of mammalian genomes leads to many additional transcripts produced near most human promoters, mainly in an antisense orientation to the protein-coding genes (Chen *et al*, 2016). Many of these transcripts are rapidly degraded, which results in a wide range of expression levels. Inspection of RNA-seq read coverage on both strands shows that antisense transcription is also present at our STIRs, but at much lower levels than the transcription of the sense strands, and, within the first half of the STIR, at levels lower than that of control promoters (Fig EV1C). Transcription at downstream promoters in tandem pairs thus appears to be more unidirectional than in the controls. In some cases, there could also be additional transcripts on the sense strand that we are not considering, which can affect the boundaries of the STIRs and potentially invalidate some of our control genes, but inspection of ChRNA-seq coverage suggests that such transcripts are either rare or very lowly expressed (Fig EV1C). Full-transcriptome segmentation into gene units that are agnostic to the protein-coding potential, as performed in some species (Ivanov *et al*, 2021), and which would benefit from long-reads, would potentially allow a more accurate analysis of the cross-talk between closely spaced transcriptional units. In any case, the wide range of expression levels of the different transcripts will retain the challenge of how to define “intergenic regions”.

Our observation of a slightly elevated guanine content at the tandem intergenic regions was accompanied by enrichment of specific G-rich motifs in these regions compared to the control regions and to shuffled controls. As mentioned above, these findings echo studies of individual genes showing several possible functions for G-rich sequences. These past reports include demonstrations of links between the G-rich regions and MAZ, and SP1 (Arhin *et al*, 2002; Oberg *et al*, 2005; Dalziel *et al*, 2007), and with the formation of R-loops. GGGGAGGGG “MAZ” elements positioned in proximity to the polyA site was reported to contribute to Pol2 pausing and thus promote more efficient transcription termination facilitated by the XRN2 exonuclease. Interestingly, MAZ itself does not seem to have an effect on the termination activity, as RNAi-mediated depletion of

MAZ did not affect termination (Gromak *et al*, 2006). This observation fits our analysis showing lack of enrichment of MAZ binding at either 3′ control CPA sites or the CPA sites at STIRs (Fig 2E) despite the overall G-richness of the tandem intergenic regions (Fig 1D and E). These findings together may point towards the importance of the G-rich structure of the sequences themselves, e.g., for regulation of Pol2 speed, or to binding of other protein(s), rather than MAZ. Gromak *et al*. reported that the proximity of MAZ elements to the polyA site is important for the termination process, as a construct containing MAZ element at a 2 kb distance from the polyA site, rather than a proximal one, significantly reduced the termination efficiency. Our analysis only partly supports this notion, since we observed no bias for the MAZ elements to appear in proximity to the CPA site (Fig 2E). MAZ elements and not GGGGCGGGG "SP1 elements" were proposed to specifically promote Pol2 pausing and polyadenylation (Yonaha & Proudfoot, 1999). Interestingly, the tandem intergenic regions we analyzed were enriched with 'GGGGCGGG' or 'GGGGCGGGGSC' motif that resembles the SP1 element rather than the MAZ element (Fig 2A,C,D). Furthermore, all types of G-rich motifs (with the notable exception of $(G)_9$) were much more prevalent upstream of the promoter controls, rather than at 3′ controls, and indeed we observed binding of both SP1 and MAZ centered at the TSS rather than at the CPA site (Fig 2E and F). Our findings thus do not support the notion of the importance of specifically an Adenosine flanked by runs of Guanines in facilitating Pol2 pausing at termination sites (Fig 2D).

Dense genes are associated with shorter introns (Amit *et al*, 2012), which fits our observation of shorter total intronal length in tandem genes as a group (Fig 1C). Other features associated with shorter introns include lower levels of low-complexity sequences, higher GC-content, proximity to another transcription unit and slower transcription rate (Veloso *et al*, 2014). Intriguingly, most of these features are exhibited to some extent at the tandem genes we analyzed, in some cases within the intergenic region rather than in the gene bodies, such as the lower rate of transposable elements (Fig 1F) and the higher GC-content (Fig 1D). Transposable elements are the main source for intronal expansion (Wu *et al*, 2013). The lower rate of transposable elements at STIRs is potentially related to the enrichment of Pol2 (Fig 5), which may provide less opportunity for their introduction to these regions. Interestingly, we also find an overall higher level of sequence evolution in STIRs compared to control regions (Fig 1G). This increase is possibly related to the potential of the G-rich motifs to form G-quadruplex structures, which are associated with genome instability and increased mutation rate (Bochman *et al*, 2012).

Based on the current model of transcription termination, transcription continues for a few hundreds of bps to several kbs post the CPA site before Pol2 is released from the DNA. This is thought to be accompanied by the T4P CTD modification of Pol2. Considering this model, we expect that the Pol2 T4P signal, other Pol2 modifications associated with termination, and the general Pol2 signal to be similar at the termination sites of all genes, irrespective of their proximity to another transcriptional unit. Therefore, it is unexpected that we find extensive Pol2 signal in STIRs, compared to the control genes, in the "total Pol2" mNET-seq data (Fig 5A, using anti-CTD CMA601 antibody, recognizing both phosphorylated and unphosphorylated CTD; Stasevich *et al*, 2014; Nojima *et al*, 2015). Further, this trend is seen when examining Pol2 with all the different CTD modifications (Fig 5B–F and Appendix Fig S2), including high levels of T4P signal downstream of the downstream TSS while being almost at background levels in the corresponding regions in both controls. Particularly striking is the prominent peak of T4P-modified Pol2 at the downstream promoter, whereas such Pol2 is rarely found at control promoters (Fig 5B and Appendix Fig S2G). One potential explanation for this observation is that there is a unique regulatory regime taking place when Pol2 is transcribing tandem transcriptional units. In that case, it is possible that the STIR may constitute a "preparation area" for a new cycle of transcription of the downstream tandem genes. This would include slowing down of Pol2 at the intergenic region, which may be supported by its enriched signal at 3′ ends of upstream tandem genes over control genes. Potentially, the 3′ control genes represent cases where Pol2 is not necessarily recycled or where recycling is distributed over a substantially longer genomic sequence, or where fast release of Pol2 might be favored, to maintain its nuclear pool. In tandem genes, the Pol2 that finished transcribing the upstream gene can potentially be used to transcribe the downstream one, perhaps before its CTD marks "reset" from their termination-associated state. At the moment, we note that there is no experimental support for the ideas described here, and that obtaining such support is challenging using the available methodologies. Specifically, since we analyze bulk data from a large number of cells, it is unclear to what extent the transcription events of tandem genes occur concurrently or in short temporal succession. The combination of methods enabling single-cell metabolic labeling with those for single-cell chromatin occupancy can be particularly useful for addressing this question in the future (Erhard *et al,* 2019; Bartosovic *et al*, 2021).

In order to examine which region of the STIR constitutes the primary accumulation site of Pol2, we examined Pol2 occupancy in tandem intergenic regions of increasing sizes (while keeping the other criteria the same, Appendix Fig S5). We observed that with increasing intergenic distances there was a less pronounced Pol2 accumulation in the intergenic region, and that Pol2 accumulation was predominantly found near the downstream TSS. This suggests that Pol2 that finishes the transcription of the upstream gene continues to be associated with chromatin and predominantly pauses near the downstream promoter (while carrying the termination-associated marks T4P and S2P). Notably, this pausing might be facilitated by the G-rich sequences that are also predominantly found near the downstream promoter (Fig 2D). Furthermore, our finding that NELF-E depletion preferentially leads to an increase in expression of the downstream genes in the tandem pairs suggests that the Pol2 pausing may have a functional role in restricting expression from the downstream promoter.

As mentioned above, additional features of tandem genes may also support slower Pol2 dynamics within STIRs. Alternatively, it is also plausible that although we only considered tandem genes that are co-expressed in bulk RNA-seq data, the actual transcription cycles of the two tandem genes are disjoint events. This may be supported by the T4P signal we see at the promoter region. The signal may stem from Pol2 which has not yet dissociated from the DNA after transcribing the upstream gene, and is not going to be involved in the transcription of the downstream gene. If this is the case, since we controlled for expression, we would expect the STIR profile to resemble the superposition of the signals of both types of controls. Notably, this does not appear to be the case, at least when considering the median Pol2 occupancy signal (Fig 5).

When considering chromatin marks in STIRs, as mentioned above, we did not observe a notable enrichment of H3K9me2 in STIRs, where it was depleted similarly to other promoters. Other histone marks also showed largely unremarkable patterns within STIRs (Fig EV5), which were largely superpositions of the patterns of the control regions, with a notable exception of H3K79me2, which showed a reduced pattern in the gene body of the downstream gene. For H3K36me3, the reduction to background levels was faster in STIRs compared to the 3′ controls, likely a consequence of the nucleosome-depleted region at the downstream promoter. Similarly, for H3K4me3, we found a smaller and narrower peak centered at the −1 nucleosome, likely reflecting reduced levels of divergent transcription from the downstream gene promoter.

Finally, we considered the possibility that different gene subsets are responsible for the enriched binding patterns of the various factors that we found enriched within STIRs, and/or that different factors tend to preferentially co-bind the same regions. To test this, we clustered the binding data (Appendix Fig S6 and S7 and Dataset EV12–EV13). However, clustering of both the STIRs and the factors did not point toward a specific regulatory pathway or subsets of tandem genes with the same protein binding patterns.

An immense amount of research has been dedicated so far into understanding transcription initiation in mammalian cells, and relatively less attention has been dedicated to transcriptional elongation and termination. Still, these events were usually studied in isolation, e.g., by considering separately promoters and termination regions. Our results suggest that the presence of an upstream termination region within up to 2 kb can have a dramatic effect on the protein occupancy at promoters, and those proteins lead to a significant effect on the transcriptional activity of the downstream genes. We further found that some of the features previously associated with efficient termination at individual genes (e.g., G-rich elements and MAZ binding) are likely mostly related to downstream promoters and not the upstream genes. Together with the emerging importance and understanding of architectural chromatin domains, this further suggests that integrative analysis of gene regulation on the genome rather than on the single-gene level will most likely be required for detailed and accurate models of gene regulation.

Materials and Methods

Extraction of tandem genes

46,012 Human hg19 Refseq coding transcript annotations were downloaded from UCSC Genome Browser (GB). Unique start and end data per gene were kept to remove different inner splicing variants. Non-coding transcripts of these genes were integrated and multiple transcripts of the same gene were flattened, choosing the minimal start coordinate and maximal end coordinate. Genes from the 23 aligned chromosomes were kept leaving 19,464 flattened transcripts. Extraction of the five closest downstream (for the definition of tandem or convergent genes) or upstream (for the definition of divergent genes) genes per gene was done using the "closest" method of the bedtools-gnu/2.25.0 package. Pairs of genes were filtered to keep only the non-overlapping ones, and only the ones where both genes are longer than 5 kb and shorter than 800 kb. Genes of the different orientations were then divided into four groups based on their minimal distance from one another, as reported by bedtools.

Mouse homologs analysis

Mouse Refseq curated mm10 genes were downloaded from GB and were processed in a similar manner to the human genes to obtain 497 mouse tandem genes. Mouse genes with homology to the human tandem gene were obtained using the Ensembl database (Kinsella *et al*, 2011), and mouse pairs were kept if both were tandem in both mouse and human and marked as having "one2one" orthology type.

Determining co-expressed tandem gene set

ENCODE expression data for multiple cell lines was quantified using RSEM as described (Zuckerman & Ulitsky, 2019). The tandem pairs were first filtered for expressed pairs (requiring both genes with expression > 2 TPM). Differences between the upstream and downstream tandem genes expression level were calculated, transformed to absolute values and empirical cumulative distribution function was operated over the values. The 75th percentile was chosen as the maximal allowed expression difference threshold between the two co-expressed genes (for example 66.37 and 82.12 for HepG2 and K562, for the set of tandem genes), leading to 188 and 164 co-expressed tandem pairs in HepG2 and K562, respectively.

Creating the control set for the co-expressed tandem genes

A set of 8,861 non-tandem genes was defined as genes > 5 kb and < 800 kb with minimal distance > 5 kb from another gene on any strand. Total intronal length was calculated for both tandem and non-tandem genes based on the average total intronal length of all the isoforms of that gene annotated by Refseq. Each co-expressed tandem gene was paired with five control genes with the closest weighted resemblance to it both in expression levels and in total intronic length (For example, for the HepG2 co-expressed pair MRM2 and MAD1L1 which have TPMs of 12.68 and 14.02 and total mean intronal length of 6,307 and ~366 kb, respectively, respective control genes were RRP7A with TPM of 12.33 and intronal length of 6,034 and ASAP1 with TPM of 12.94 and intronal length of ~385 kb. For both types of controls, sequences with the same length as the length of the intergenic region of the original tandem pair were extracted either upstream of the promoter region ("promoter controls", controlling for the downstream co-expressed tandem gene) and downstream of the 3′ end ("3′ control", controlling for the upstream co-expressed tandem gene) and were set as the control set for the intergenic region of the tandem pair.

Splicing efficiency analysis

Splicing efficiency analysis was done as described in (Zuckerman & Ulitsky, 2019), using RefSeq introns annotations. Overall and first-intron splicing score distributions were used for further analysis.

Pausing index calculation

Signal was calculated over tandem genes or controls using kentUtils bigWigAverageOverBed. Pausing index was defined as the ratio between the summation of the reads signal over the first 200 bases following the TSS and the reads signal over the whole gene normalized to the gene length. Datasets used for this calculation are GSM4836456, GSM4836452, GSM4836445 and GSM4836447.

Nano-COP reads analysis

Nano-COP reads (Drexler *et al*, 2020) with accessions GSM4073916, GSM4073917, GSM4073918, GSM3498218, and GSM3498220 were aligned to the hg19 human genome assembly using minimap2 (Li, 2018). Using bedtools, aligned reads were intersected with the set of upstream- or downstream genes of the tandem pairs, or with the STIRs. Sense reads from all experiments were combined.

Sequence analysis

FASTA format sequences of the intergenic region of HepG2 and K562 co-expressed tandem pairs and their control sequences were extracted. The sequences were divided into ~70 bins (based on the minimal length of the co-expressed tandem intergenic regions), and the nucleotide composition ratio was calculated. Control sequence compositions were aggregated using the mean of each set of five controls. Significance was tested using paired Wilcoxon rank-sum tests. Genomic repeats overlapping individual positions within the tandem intergenic regions and control sequences were obtained using RepeatMasker tracks of GB and were filtered to keep only non-simple repeat annotations. Bedtools' maskfasta function was used for soft masking and FASTA sequences for the tandem co-expressed genes and controls were extracted and the sequences were binned as described above. PhyloP evolutionary conservation scores were obtained from GB, and plots were created using the deepTools package (Ramírez *et al*, 2014).

Enriched motifs in co-expressed tandem intergenic regions

188 and 164 non-masked tandem intergenic region sequences of HepG2 and K562 co-expressed tandem genes were used as input for STREME (with the arguments --kmer 1 --minw 6 --maxw 12, and in either "DNA" or "RNA" mode), which yielded 37 motifs with $P < 0.05$. Motif occurrences were counted in the tandem intergenic region and control sequences of the cell line where they were originally detected at using FIMO with STREME output. Scores were summarized and motifs were filtered to keep ones with higher prevalence in the tandem intergenic regions both in the measure of total number of motif occurrences per sequence and in the relative number of genes carrying the motif relative to the controls. Filtered motifs were ranked by the minimal tandem-to-control scores ratio and ordered by the sum of ranks of both measures. Proportion tests were applied over the filtered motifs and *P*-values were corrected using Bonferroni correction. For the G-rich enriched RNA motifs ('GGGGCGGG' and 'GGGGCGGGGSC'), 1,000 di-nucleotide preserved, randomly shuffled sequences of HepG2 or K562 STIRs were generated and scanned for the aforementioned motifs using FIMO. Enriched motifs were compared to JASPAR 2022 vertebrates motif database (Castro-Mondragon *et al*, 2022) using Tomtom (Gupta *et al*, 2007) from the MEME suite package (motifs were first manually converted to DNA alphabet to match the JASPAR motifs alphabet). Significant JASPAR motifs (E-value and *q*-value < 0.05) mutual to both G-rich motifs were kept.

TF binding and metagene analysis

ChIP-seq binding cluster data across the human genome were obtained for 338 proteins profiled by the ENCODE project from the GB (Dataset EV14). Binding sites were intersected with the co-expressed tandem intergenic regions and control sequences of HepG2 or K562 cells using the intersect function of bedtools. Enriched TF binding at co-expressed tandem intergenic regions over controls was calculated by taking the minimal ratio value between the prevalence of TF binding at the tandem regions fractionated by either the promoter- or 3′- control regions normalized binding prevalence. For example: AGO1 binding sites intersected at least once with 136 out of 188 (~72%) of co-expressed tandem intergenic region sequences, but only 357 or 89 of the 940 (~38% and ~9%) of the promoter and 3′ control sequences in HepG2, respectively. The analysis produced a set of 6 TFs with the highest minimal ratio between co-expressed tandem intergenic region- and control- binding which were found in both HepG2 and K562 cell lines. To further examine the binding pattern throughout the tandem intergenic- or control- regions, ChIP-seq data of the 6 candidates were obtained from the ENCODE project and were visualized as a metagene plot using deepTools (Ramírez *et al*, 2014), averaging the bins using the median value and showing the standard error. In addition, each ChIP-seq experiment was adjusted with its own set of co-expressed tandem genes and control genes based on the specific cell type used for the ChIP-seq experiment as explained previously. Metagene plots for EHMT2, H3K9me2, PHF8 and HP1γ ChIP-seq data (from the ENCODE project) and R-loop data (from the GEO database, accession GSE70189) were drawn in a similar manner.

ENCODE shRNA-seq data analysis

ENCODE shRNA experiments of 245 genes (440 experiments) done in K562 and/or HepG2 were analyzed using DESeq2 (Dataset EV15). For each factor, data of the co-expressed tandem genes and their controls within the respective cell line was extracted and the distribution and $\log_2$-transformed fold change of gene expression was plotted and tested using Wilcoxon paired rank-sum test for the set of upstream or downstream genes within the co-expressed tandem pairs and their averaged controls. Scatterplot of the changes in expression following KD of AGO1 and AGO2 for the downstream and upstream co-expressed tandem genes were plotted and tested using Pearson's correlation. Additional correlation tests for ChIP-seq or R-loops median signal and co-expressed tandem genes expression change following KD were done using Spearman's correlation. Median difference in expression changes following KD between the upstream or downstream gene and the controls were calculated. KD experiments were ranked based on minimal-to-maximal effect over the upstream gene compared to controls, maximal-to-minimal effect over the downstream gene compared to controls, and overall absolute difference between the effect on the upstream and downstream gene. Ranking was done so that KD targets with low ranks are those affecting mostly the downstream tandem genes (and not the upstream tandem genes) and vice versa.

Pol2 modifications analysis

Data of Pol2 binding was downloaded from the GEO database (accession GSE81662) in bigWig format. Matrices for the plus and minus strand of the co-expressed tandem intergenic regions in HeLaS3 and their respective controls were built using deepTools and binning was done based on the median value. Matrices were then filtered according to strand and were combined. Extreme values were filtered to remove rows containing outliers with values above 99.99% and under 0.01% of the total combined matrix. To calculate the overall enrichment of total Pol2 at STIRs over control 3′ or promoter regions, raw output sense of "total Pol2" (GSM2357382) matrices were filtered to keep only the intergenic region, and overall signal was summed at STIRs, or averaged per control quintet of each STIR and then summed.

Clustering of tandem genes and binding experiments

Maximal value per gene per binding experiment was calculated over the intergenic region and flanking regions (1 kb on each side, unless stated otherwise) and over the associated control regions, using the deepTools matrix output. For the control experiment, each quintet controlling for a single tandem gene was first aggregated using mean value to create a mean matrix per experiment. Tandem max matrix was divided by either control matrix, genes with over 17 missing values across experiments were removed from the cluster analysis, matrix values exceeding 10 were converted to 10. Clustering of the columns was done using Pearson's correlation using the "complete.obs" option to handle NA values and complete-link measure. Genes were clustered using complete-link measure and euclidean distances method. Log_2-transformed FPKM values for gene expression annotations were computed using the ENCODE RNA-seq data.

Data availability

This study includes no data deposited in external repositories.

Expanded View for this article is available online.

Acknowledgements
We thank members of the Ulitsky lab for useful discussions and comments on the manuscript.

Author contributions

Noa Nissani: Conceptualization; Data curation; Software; Formal analysis; Validation; Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. **Igor Ulitsky:** Conceptualization; Data curation; Formal analysis; Supervision; Funding acquisition; Investigation; Writing—original draft; Project administration; Writing—review and editing.

In addition to the CRediT author contributions listed above, the contributions in detail are:

NN and IU conceived and designed the study. NN developed the computational pipelines and analyzed data under supervision of IU. NN and IU wrote the manuscript.

Disclosure and competing interests statement

The authors declare that they have no conflict of interest.

References

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B *et al* (2012) Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep* 1: 543 – 556

Aoi Y, Smith ER, Shah AP, Rendleman EJ, Marshall SA, Woodfin AR, Chen FX, Shiekhattar R, Shilatifard A (2020) NELF regulates a promoter-proximal step distinct from RNA Pol II pause-release. *Mol Cell* 78: 261 – 274

Arhin GK, Boots M, Bagga PS, Milcarek C, Wilusz J (2002) Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals. *Nucleic Acids Res* 30: 1842 – 1850

Arnold M, Bressin A, Jasnovidova O, Meierhofer D, Mayer A (2021) A BRD4-mediated elongation control point primes transcribing RNA polymerase II for 3′-processing and termination. *Mol Cell* 81: 3589 – 3603

Ashfield R, Enriquez-Harris P, Proudfoot NJ (1991a) Transcriptional termination between the closely linked human complement genes C2 and factor B: common termination factor for C2 and c-myc? *EMBO J* 10: 4197 – 4207

Ashfield R, Enriquez-Harris P, Proudfoot NJ (1991b) Transcriptional termination between the closely linked human complement genes C2 and factor B: common termination factor for C2 and c-myc? *EMBO J* 10: 4197 – 4207

Ashfield R, Patel AJ, Bossone SA, Brown H, Campbell RD, Marcu KB, Proudfoot NJ (1994) MAZ-dependent termination between closely spaced human complement genes. *EMBO J* 13: 5656 – 5667

Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27: 1653 – 1659

Bailey TL (2021) STREME: Accurate and versatile sequence motif discovery. *Bioinformatics* 37: 2834 – 2840

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37: W202 – W208

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME suite. *Nucleic Acids Res* 43: W39 – W49

Bartosovic M, Kabbe M, Castelo-Branco G (2021) Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat Biotechnol* 39: 825 – 835

Bochman ML, Paeschke K, Zakian VA (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* 13: 770 – 780

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N *et al* (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 50: D165 – D173

Chen L, Chen J-Y, Huang Y-J, Gu Y, Qiu J, Qian H, Shao C, Zhang X, Hu J, Li H *et al* (2018) The augmented R-loop is a unifying mechanism for myelodysplastic syndromes induced by high-risk splicing factor mutations. *Mol Cell* 69: 412 – 425

Chen Y, Pai AA, Herudek J, Lubas M, Meola N, Järvelin AI, Andersson R, Pelechano V, Steinmetz LM, Jensen TH *et al* (2016) Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat Genet* 48: 984 – 994

Connelly S, Manley JL (1988) A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev* 2: 440 – 452

Dalziel M, Nunes NM, Furger A (2007) Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3' end processing. *Mol Cell Biol* 27: 1568 – 1580

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al* (2012) Landscape of transcription in human cells. *Nature* 489: 101 – 108

Drexler HL, Choquet K, Churchman LS (2020) Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Mol Cell* 77: 985 – 998.e8

Dye MJ, Proudfoot NJ (2001) Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II. *Cell* 105: 669 – 681

Egloff S, O'Reilly D, Chapman RD, Taylor A, Tanzhaus K, Pitts L, Eick D, Murphy S (2007) Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science* 318: 1777 – 1779

Erhard F, Baptista MAP, Krammer T, Hennig T, Lange M, Arampatzi P, Jürges CS, Theis FJ, Saliba A-E, Dölken L (2019) scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* 571: 419 – 423

Erickson B, Sheridan RM, Cortazar M, Bentley DL (2018) Dynamic turnover of paused Pol II complexes at human promoters. *Genes Dev* 32: 1215 – 1225

Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* 118: 555 – 566

Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27: 1017 – 1018

Greger IH, Proudfoot NJ (1998) Poly(A) signals control both transcriptional termination and initiation between the tandem GAL10 and GAL7 genes of Saccharomyces cerevisiae. *EMBO J* 17: 4771 – 4779

Gromak N, West S, Proudfoot NJ (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol* 26: 3986 – 3996

Gruber AJ, Zavolan M (2019) Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* 20: 599 – 614

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. *Genome Biol* 8: R24

Hagenbüchle O, Wellauer PK, Cribbs DL, Schibler U (1984) Termination of transcription in the mouse alpha-amylase gene Amy-2a occurs at multiple sites downstream of the polyadenylation site. *Cell* 38: 737 – 744

Hainer SJ, Pruneski JA, Mitchell RD, Monteverde RM, Martens JA (2011) Intergenic transcription causes repression by directing nucleosome assembly. *Genes Dev* 25: 29 – 40

Harlen KM, Churchman LS (2017) The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat Rev Mol Cell Biol* 18: 263 – 273

Heidemann M, Hintermair C, Voß K, Eick D (2013) Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochim Biophys Acta* 1829: 55 – 62

Ho CK, Shuman S (1999) Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. *Mol Cell* 3: 405 – 411

Hsin J-P, Manley JL (2012) The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev* 26: 2119 – 2137

Ivanov M, Sandelin A, Marquardt S (2021) TrancriptomeReconstructoR: data-driven annotation of complex transcriptomes. *BMC Bioinformatics* 22: 290

Jangi M, Fleet C, Cullen P, Gupta SV, Mekhoubad S, Chiao E, Allaire N, Bennett CF, Rigo F, Krainer AR *et al* (2017) SMN deficiency in severe models of spinal muscular atrophy causes widespread intron retention and DNA damage. *Proc Natl Acad Sci USA* 114: E2347 – E2356

Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt JF, Buratowski S (2004) The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* 432: 517 – 522

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A *et al* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011: bar030

Krzyszton M, Zakrzewska-Placzek M, Kwasnik A, Dojer N, Karlowski W, Kufel J (2018) Defective XRN3-mediated transcription termination in Arabidopsis affects the expression of protein-coding genes. *Plant J* 93: 1017 – 1031

Leng X, Ivanov M, Kindgren P, Malik I, Thieffry A, Brodersen P, Sandelin A, Kaplan CD, Marquardt S (2020) Organismal benefits of transcription speed control at gene boundaries. *EMBO Rep* 21: e49315

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094 – 3100

Lu X, Zhu X, Li Y, Liu M, Yu B, Wang YU, Rao M, Yang H, Zhou K, Wang Y *et al* (2016) Multiple P-TEFbs cooperatively regulate the release of promoter-proximally paused RNA polymerase II. *Nucleic Acids Res* 44: 6853 – 6867

Martens JA, Laprade L, Winston F (2004) Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature* 429: 571 – 574

Mayer A, Heidemann M, Lidschreiber M, Schreieck A, Sun M, Hintermair C, Kremmer E, Eick D, Cramer P (2012) CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* 336: 1723 – 1725

Mayfield JE, Irani S, Escobar EE, Zhang Z, Burkholder NT, Robinson MR, Mehaffey MR, Sipe SN, Yang W, Prescott NA *et al* (2019) Tyr1 phosphorylation promotes phosphorylation of Ser2 on the C-terminal domain of eukaryotic RNA polymerase II by P-TEFb. *Elife* 8: e48725

McCracken S, Fong N, Yankulov K, Ballantyne S, Pan G, Greenblatt J, Patterson SD, Wickens M, Bentley DL (1997) The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* 385: 357 – 361

Nagaoka M, Shiraishi Y, Sugiura Y (2001) Selected base sequence outside the target binding site of zinc finger protein Sp1. *Nucleic Acids Res* 29: 4920 – 4929

Nguyen HD, Leong WY, Li W, Reddy PNG, Sullivan JD, Walter MJ, Zou L, Graubert TA (2018) Spliceosome mutations induce R loop-associated sensitivity to ATR inhibition in myelodysplastic syndromes. *Cancer Res* 78: 5363 – 5374

Nguyen T, Fischl H, Howe FS, Woloszczuk R, Serra Barros A, Xu Z, Brown D, Murray SC, Haenni S, Halstead JM *et al* (2014) Transcription mediated insulation and interference direct gene cluster expression switches. *Elife* 3: e03635

Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ (2015) Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* 161: 526 – 540

Nojima T, Rebelo K, Gomes T, Grosso AR, Proudfoot NJ, Carmo-Fonseca M (2018) RNA polymerase II phosphorylated on CTD serine 5 interacts with the spliceosome during co-transcriptional splicing. *Mol Cell* 72: 369 – 379

Oberg D, Fay J, Lambkin H, Schwartz S (2005) A downstream polyadenylation element in human papillomavirus type 16 L2 encodes multiple GGG motifs and interacts with hnRNP H. *J Virol* 79: 9254 – 9269

Proudfoot NJ (2016) Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* 352: aad9926

Pruneski JA, Hainer SJ, Petrov KO, Martens JA (2011) The Paf1 complex represses SER3 transcription in Saccharomyces cerevisiae by facilitating intergenic transcription-dependent nucleosome occupancy of the SER3 promoter. *Eukaryot Cell* 10: 1283 – 1294

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 42: W187 – W191

Rom A, Melamed L, Gil N, Goldrich MJ, Kadir R, Golan M, Biton I, Perry RB-T, Ulitsky I (2019) Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nat Commun* 10: 5092

Sanz LA, Hartono SR, Lim YW, Steyaert S, Rajpurkar A, Ginno PA, Xu X, Chédin F (2016) Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Mol Cell* 63: 167 – 178

Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ (2017) Distinctive patterns of transcription and RNA processing for human lincRNAs. *Mol Cell* 65: 25 – 38

Shah N, Maqbool MA, Yahia Y, El Aabidine AZ, Esnault C, Forné I, Decker T-M, Martin D, Schüller R, Krebs S *et al* (2018) Tyrosine-1 of RNA polymerase II CTD controls global termination of gene transcription in mammals. *Mol Cell* 69: 48 – 61

Shearwin K, Callen B, Egan J (2005) Transcriptional interference – a crash course. *Trends Genet* 21: 339 – 345.

Shinkai Y, Tachibana M (2011) H3K9 methyltransferase G9a and the related molecule GLP. *Genes Dev* 25: 781 – 788

Shuman S (2020) Transcriptional interference at tandem lncRNA and protein-coding genes: an emerging theme in regulation of cellular nutrient homeostasis. *Nucleic Acids Res* 48: 8243 – 8254

Skourti-Stathaki K, Kamieniarz-Gdula K, Proudfoot NJ (2014) R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* 516: 436 – 439

Skourti-Stathaki K, Proudfoot NJ, Gromak N (2011) Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell* 42: 794 – 805

Stasevich TJ, Hayashi-Takanaka Y, Sato Y, Maehara K, Ohkawa Y, Sakata-Sogawa K, Tokunaga M, Nagase T, Nozaki N, McNally JG *et al* (2014) Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature* 516: 272 – 275

Tantravahi J, Alvira M, Falck-Pedersen E (1993) Characterization of the mouse beta maj globin transcription termination region: a spacing sequence is required between the poly(A) signal sequence and multiple downstream termination elements. *Mol Cell Biol* 13: 578 – 587

Thebault P, Boutin G, Bhat W, Rufiange A, Martens J, Nourani A (2011) Transcription regulation by the noncoding RNA SRG1 requires Spt2-dependent chromatin deposition in the wake of RNA polymerase II. *Mol Cell Biol* 31: 1288 – 1300

Veloso A, Kirkconnell KS, Magnuson B, Biewen B, Paulsen MT, Wilson TE, Ljungman M (2014) Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res* 24: 896 – 905

Vilborg A, Sabath N, Wiesel Y, Nathans J, Levy-Adam F, Yario TA, Steitz JA, Shalgi R (2017) Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. *Proc Natl Acad Sci USA* 114: E8362 – E8371

West S, Gromak N, Proudfoot NJ (2004) Human 5′→ 3′ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* 432: 522 – 525

Wu J, Xiao J, Wang L, Zhong J, Yin H, Wu S, Zhang Z, Yu J (2013) Systematic analysis of intron size and abundance parameters in diverse lineages. *Sci China Life Sci* 56: 968 – 974

Yanling Zhao D, Gish G, Braunschweig U, Li Y, Ni Z, Schmitges FW, Zhong G, Liu KE, Li W, Moffat J *et al* (2016) SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination. *Nature* 529: 48 – 53

Yonaha M, Proudfoot NJ (1999) Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol Cell* 3: 593 – 600

Yu X, Martin PGP, Michaels SD (2019) BORDER proteins protect expression of neighboring genes by promoting 3’ Pol II pausing in plants. *Nat Commun* 10: 4359

Zhang H, Rigo F, Martinson HG (2015) Poly(A) signal-dependent transcription termination occurs through a conformational change mechanism that does not require cleavage at the Poly(A) site. *Mol Cell* 59: 437 – 448

Zhu Z, Wang Y, Li X, Wang Y, Xu L, Wang X, Sun T, Dong X, Chen L, Mao H *et al* (2010) PHF8 is a histone H3K9me2 demethylase regulating rRNA synthesis. *Cell Res* 20: 794 – 801

Zuckerman B, Ulitsky I (2019) Predictive models of subcellular localization of long RNAs. *RNA* 25: 557 – 572

Žumer K, Maier KC, Farnung L, Jaeger MG, Rus P, Winter G, Cramer P (2021) Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo. *Mol Cell* 81: 3096 – 3109.e8