Tyrosine-1 of RNA Polymerase II CTD controls global termination of gene transcription in mammals

Nilay Shah[1#], Muhammad Ahmad Maqbool[2#], Yousra Yahia[2], Amal Zine El Aabidine[2], Cyril Esnault[2], Ignasi Forné[3], Tim-Michael Decker[1], David Martin[2], Roland Schüller[1], Stefan Krebs[4], Helmut Blum[4], Axel Imhof[3], Dirk Eick[1*] and Jean-Christophe Andrau[2,5*]

Affiliations

[1]Department of Molecular Epigenetics, Helmholtz Center Munich, Center of Integrated Protein Science Munich, Marchioninistrasse 25, 81377 Munich, Germany;

[2]Institut de Génétique Moléculaire de Montpellier (IGMM), CNRS-UMR5535, Montpellier, France;

[3]Biomedical Center Munich, ZFP, Großhaderner Strasse 9, 82152 Planegg-Martinsried, Germany;

[4]Laboratory for functional genome analysis, Gene Center, Ludwig-Maximilians-Universität, Munich, Germany.

# These authors contributed equally

[5]Lead Contact

* Corresponding authors: jean-christophe.andrau@igmm.cnrs.fr and eick@helmholtz-muenchen.de

Summary

The carboxy-terminal domain (CTD) of RNA polymerase (Pol) II is composed of a repetition of YSPTSPS heptads and functions as a loading platform for protein complexes that regulate transcription, splicing and maturation of RNAs. Here, we studied mammalian CTD mutants to analyze the function of tyrosine1 residues in the transcription cycle. Mutation of 3/4 of the tyrosine residues (YFFF mutant) resulted in

a massive read-through transcription phenotype in antisense direction of promoters as well as in 3' direction several hundred kb downstream of genes. The YFFF mutant shows reduced Pol II at promoter-proximal pause sites, a loss of interaction with the Mediator and Integrator complexes and impaired recruitment of these complexes to chromatin. Consistent with these observations, Pol II loading at enhancers and maturation of snRNAs are altered in the YFFF context genome wide. We conclude that tyrosine1 residues of the CTD control termination of transcription by Pol II.

Keywords

RNA Polymerase II, CTD, Tyrosine1, Transcription termination, Read-through transcription, Promoter-proximal pausing, Divergent transcription.

Introduction

The control of transcription requires RNA polymerase (Pol) II recruitment at promoter, transcription initiation and transition to processive elongation. It also requires a proper control of transcription termination (Proudfoot, 2016). Despite many efforts during the last years to understand this process in vivo, it remains poorly understood. Transcription termination by Pol II generally succeeds polyadenylation at 3' ends of genes and can occur up to several kb after the annotated 3' ends. Recent works have involved proteins or protein complexes in this process such as the cleavage and polyadenylation complex, and the histone methyl-transferase SetD2 (Grosso et al., 2015; Nojima et al., 2015). Termination also occurs at 5' ends of genes. This process concerns a large fraction of mammalian promoters in which pausing of Pol II and divergent transcription is observed (Core et al., 2008; Seila et al., 2008). Longer upstream antisense (AS) non-coding transcripts can also be observed at many

promoters in normal cells and accumulate to high levels after exosome inhibition (Lepoivre et al., 2013; Preker et al., 2008; Schlackow et al., 2017). Current models propose that termination around promoters also requires polyadenylation, a process that would be partially repressed in the sense but not in the AS orientation of the genes by the presence of U1 snRNP recognition sites (Almada et al., 2013; Ntini et al., 2013).

Among the questions left open is that of the determinants targeted within the Pol II enzyme allowing termination in living cells. Although mutations in catalytic subunits were characterized in Pol III active sites that impair transcription termination (Iben et al., 2011; Shaaban et al., 1995), little is known about Pol II despite reports of slow/fast Pol II mutants displaying impaired termination at a subset of genes (Fong et al., 2015; Hazelbaker et al., 2013). In contrast to Pol I and Pol III, Pol II produces transcripts of widely varying sizes and types including polyadenylated, non-polyadenylated, coding and non-coding transcripts with various functions and thus different modes of regulation. As a consequence, Pol II activity is tightly controlled through the action of many proteins or protein complexes that can act at all steps of transcription including recruitment, initiation, pausing, pause release, processive elongation and termination.

The carboxy-terminal domain (CTD) of Pol II's largest subunit, Rpb1, is an essential platform for recruitment of factors controlling transcriptional and post-transcriptional events (Eick and Geyer, 2013; McCracken et al., 1997). The CTD is evolutionarily conserved and consists of repetitions of heptads ($Y_1S_2P_3T_4S_5P_6S_7$) that are phosphorylated in the transcription cycle. Phosphorylations of serine 2 and 5 (Ser2P and Ser5P) residues are the most studied and represent strong hallmarks of early transcription and processive elongation, respectively. The more recently

characterized Ser7P and Thr4P were proposed to be associated with snRNA or histone gene transcription and transcription termination (Chapman et al., 2007; Egloff et al., 2007; Harlen et al., 2016; Hintermair et al., 2012b; Hsin et al., 2011).

We and others have recently described that phosphorylation of Tyr1 in metazoans occurs at promoters (Descostes et al., 2014; Hsin et al., 2014) and in mammals Tyr1P is also found at enhancer locations. ChIP-seq signals for Tyr1P were also observed to a lesser extent at 3'ends. Overall the mammalian Tyr1P genomic locations were quite distinct from the ones described in yeast, where enrichments were essentially found over gene bodies and proposed to prevent early termination (Mayer et al., 2012). However, we were previously unable to describe the functional significance of Tyr1 residues due to the lack of stable mutants, as mutations of all Tyr1 residues of the CTD resulted in degradation of Rpb1 (Descostes et al., 2014). To circumvent this problem, we have generated novel mutations in the CTD and focused our analyses on a mutant, YFFF, in which Tyr1 residues are replaced by Phe in the last ¾ of the CTD repeats. This mutant reveals a role of Tyr1 residues in the control of termination of 5' anti-sense (AS) and 3' sense transcripts. In the YFFF mutant, a massive transcription read-through (RT) is observed, accompanied by reduced Pol II at the promoter-proximal pause, apparent transcriptional interference, snRNA maturation defect and decrease of Pol II accumulation at active enhancers. Further proteomic characterization of the YFFF mutant showed that tyrosine mutations resulted in loss of Pol II interaction with Mediator (Med) and Integrator (Int) complexes, suggesting that they might be involved in the pause/termination processes. Finally, both Med and Int also show impaired DNA recruitment as revealed by ChIP experiments.

Results

Phenotypes of CTD tyrosine mutants

We previously investigated the function of Tyr1P residues in the mammalian CTD by genome-wide location analysis (ChIP-seq) and by generating mutations in the CTD, replacing all Tyr1 residues of CTD heptads into phenylalanine (Descostes et al., 2014). These mutations resulted in a lethal phenotype and CTD degradation in Rpb1, restricting further functional investigation. To circumvent this problem, we designed four new CTD mutants (Figure 1A) in which only Tyr1 residues of 2 or 3 quarters of the heptads were mutated to Phe residues. The control used in our experiments contains the wt CTD sequence, including the non-canonical repeats and is designated as rWT. All mutants as well as rWT contain an α-amanitin resistance mutation that allows to express a recombinant Rpb1 while the endogenous Rpb1 is suppressed as described (Bartolomei et al., 1988; Meininghaus et al., 2000).

After induction of the mutants and rWT control cells, endogenous Rpb1 was shut down by α-amanitin treatment. We then analyzed the growth phenotype and the stability of the mutants. Mutants with half of the repeats mutated were found either lethal (YFFY) or viable and proliferated (YYFF and FYYF) for five to ten days after addition of α-amanitin to the medium, suggesting that the position of the heptads within the CTD is important for tyrosine function (Figure S1A and S1B). Mutation of last three quarters of the repeats (YFFF) also resulted in a lethal phenotype. Despite their variable phenotypes, all Rpb1 mutants were stable at the protein level with comparable amount of hyper- (IIO) versus hypo-phosphorylated (IIA) form of Rpb1 (Figure 1B) as well as a comparable level of Ser2P, suggesting that they are competent for elongation and allowing us to pursue functional study on the mutants. We also monitored expression of the various phospho-isoforms of the CTD and

found comparable levels of Ser2P, Ser5P, Thr4P and Ser7P (Figure S1C). At the time of sample collection for further experiments, all mutant cells displayed around 80% viability.

We next assessed how transcriptomes of mutants were affected by performing RNA-seq experiments after induction of recombinant Pol II and inhibition of endogenous Pol II with α-amanitin treatment (Figure S1A). In global differential expression (DEseq) analysis, we found a large dysregulation essentially in the YFFF mutant with many genes down (48) and up (810) regulated (Figure S1D). However, gene ontology analyses did not reveal specific functional categories lost or enriched in the mutant (data not show). Rather than an effect at specific categories of genes, our observations pointed to a global effect characterized by a 3' read-through (RT) phenotype visible weakly in mutants YYFF, FYYF, and YFFY, but strongly pronounced in YFFF mutant (Figures 1C-1D and S1E-S1F). The extent of the observed RT in the YFFF mutant appeared extreme, spanning from several kb up to hundreds of kb from the annotated 3' ends, suggesting a global pervasive phenotype. The phenomenon of 3' RT has been reported for WT Pol II before (Proudfoot, 2016) and has been described with a more amplified phenotype after knock-down of Setd2, Xrn2, CPSF or WDR82 proteins (Austenaa et al., 2015; Fong et al., 2015; Grosso et al., 2015; Nojima et al., 2015). All these factors are known to interact with CTD and function in the control of RNA elongation/termination. Interestingly, YFFF mutation has little effect on the binding of these factors to CTD, while the interaction with other factors and cellular complexes is fully abolished (mass spec data below). In sum, the phenotype of the YFFF mutant suggests a strong functional link between Tyr1 in the CTD and the control of termination.

Tyrosine mutations cause a massive read-through at 5’ and 3’ends of genes

To examine the consequences of the YFFF mutations in more detail and to strengthen our initial observations, we undertook further total RNA-seq experiments in which we improved the signal to noise ratio in intergenic regions (see methods) as exemplified in Figure S2A and quantified genome-wide in Figure S2B. Using this procedure, we confirmed a massive 3' RT phenotype in the YFFF mutant and also observed a RT for 5’ antisense (AS) transcription. An example for both phenotypes is shown for the PDCD6IP gene in Figure 2A. 5’ AS transcription is a hallmark of mammalian genes (Core et al., 2008; Seila et al., 2008) that occurs roughly at half of the promoters (Fenouil et al., 2012b). To consolidate this observation at the genome-wide scale, we performed RNA-seq composite average metagene profiles for protein coding genes by rescaling rWT and YFFF RNA signals at the same levels over the gene bodies (Figure 2B and Figure S2C) to better visualize the RT phenoytype. This demonstrated a clear RT effect at 3’ ends of genes in sense direction and 5’ ends of genes in antisense direction. This effect was also clearly visible and significant without normalization of signals at gene body (Figure S2D). We confirmed this independently by plotting the transcript densities over the gene bodies and 20 kb upstream of 5’ and downstream of 3’ ends (Figure 2C). The transcriptome (gene bodies, middle panel) shows a typical bimodal distribution representing lowly and moderately/highly expressed genes. The YFFF mutant displays more low expression values (first Gaussian) and less moderate/high values (second Gaussian) as compared to the rWT. The distributions of the 5’ AS and 3’ sense signals of the 20 kb regions surrounding the gene bodies indicate an inverse trend with more signal for the YFFF mutant. The quantification of upstream AS and downstream sense RT indices in rWT and YFFF mutant is shown in Figure 2D. Finally, a larger

chromosomal view (Figure 2E) further supports the genome-wide effect of the RT phenotype. Altogether, our data suggest a strong termination defect in mutant YFFF occurring at both ends of genes in sense (downstream) and antisense (upstream) orientations.

We next asked whether the termination defect was specific to the tyrosine mutations of the CTD. To this end, we built a serine2-to-alanine mutant in which ¾ of the distal CTD repeats are mutated (S2AAA) in a manner similar to the YFFF mutant. After shutdown of the endogenous Pol II, the S2AAA mutant also showed a lethal phenotype but no significant 3’ RT and only a slight increase in 5’ AS transcription (Figure S2E and data not shown) at few genomic locations. Thus the observed RT phenotype is specific to the mutation of tyrosine residues in the CTD.

Previous works proposed that Pol II loading at 5’ end of genes could influence termination at 3’ ends (Nagaike et al., 2011; Pinto et al., 2011). We sought to address this question in the context of the YFFF mutations and more specifically, if a marked 3' RT is linked to increase in 5’ AS transcription and vice versa. We ranked the genes for decreasing ratio of RNA-seq signal downstream of 3' ends in YFFF mutant versus rWT, split them into 4 groups (A to D) from the highest to the lowest RT effect at 3’ end and plotted the 5' AS RNA signal correspondingly (Figure S2F and table S4). Our analysis reveals that a high RT transcription index at 3' end of genes in groups A and B correlated with a high RT transcription index for divergent transcription at 5' ends of genes. Lower levels of 3' end RT transcription in groups C and D were paralleled by lower levels of RT AS transcription at 5' ends of genes. Similar analyses were performed for larger intervals (20 and 50 kb downstream of 3’ ends) and also by ranking the genes for decreasing ratio of 5’ AS RT transcription (data not shown) and indicated a link between 5’ AS transcription and 3’ RT phenotype in the YFFF mutant.

Further investigations however, did not allow us to directly correlate the 5' to 3' RT levels in rWT or YFFF cells.

Altogether our investigations support a model in which the AS RT transcription at 5' ends of genes and the 3' end RT is linked in the context of the YFFF mutant. They also indicate that at least half of the genes do present a significant read-through phenotype. In summary, our RNA-seq experiments strongly suggest that Tyr1 mutations of the CTD result in a massive and specific termination defect that occurs both for 5' antisense transcripts and sense transcripts at 3' ends of the genes.

The YFFF mutations result in transcriptional activation of downstream genes and transcriptional interference

Because the YFFF mutant displayed an apparent pervasive transcription phenotype, we asked if this could result in transcriptional interference or transcription of previously silent genes due to RT. Visual inspection of our data reveals many examples in which RT transcription of one gene resulted in increased transcription of the downstream gene (see PPFIA4 gene, Figure S2G). In this case, we could exclude that signals originated from new initiation as no H3K4me3 (or H3K27ac) signal was observed in the intergenic regions or at the promoter of the downstream gene. Conversely, when two genes were oriented head to head, we found many examples of apparent interference of the RT with transcription of adjacent genes (see ST14 gene in Figure S2H). However, we did not observe the loss of H3K4me3 marks at adjacent promoters, possibly because erasing of this histone modification is not very dynamic. In attempt to quantify global interference, we overlapped genes that were both down-regulated and with an increased antisense RNA-seq signal over the gene bodies. Our analysis revealed that 14% of the down-regulated genes also

display increased AS RNAs, suggesting a relatively spread interfering effect (Figure S2I). Overall, we conclude that activation of silent genes and interference is a very likely consequence of the YFFF mutations but this point will require further investigation.

YFFF transcription yields polyadenylated RNAs

An elevation of intergenic RNA levels at both gene ends could not only result from a transcriptional RT but also from an increased stabilization of the transcripts produced by natural RT (or both). To address this question, we purified chromatin associated RNAs to perform chrRNA-seq (Bhatt et al., 2012). This method allows scoring for nascent RNAs, associated to chromatin and gives similar read-out compared to other nascent RNA-seq methods (Mayer et al., 2015; Nojima et al., 2015). Both individual and meta-gene profiling of this data, using the same gene body normalization approach as before, confirmed that the RT observed in the YFFF mutant originates at least from a transcriptional effect (Figure 3A-C) both at 3' and 5' ends. We also note that in both rWT and YFFF, chrRNAs tend to accumulate in 5' AS but not sense orientation.

Pol II ChIP-seq allowed further confirmation of the RT at 3' ends with a delayed 3' pause around 2.6 kb after the annotated 3' ends (Figure 3D-F). In these analyses and as for RNA-seq, we rescaled the signals so that Pol II has comparable levels on the gene bodies (Figure 3D and S3A). At and after 3' ends we observed both an increased signal density (for at least 20 kb) and a delayed Pol II accumulation/pause occurring approximately 2.6 kb downstream of annotated 3' ends. This delay is more pronounced and extends further than the one recently described for an Xrn2 D235A dominant mutation, also showing 3' RT (Figure S3C) (Fong et al., 2015). We also

note that the result remains clearly visible even without performing the mentioned normalization, by just scaling the data to the same amount of sequenced tags (Figure S3B). We further calculated Pol II downstream (10 kb after annotated 3' ends) RT indices and find significantly higher values in the mutant (Figure 3E, right panel) in contrast to little difference observed upstream of 5' ends (Figure 3E, left panel). This latter result could be due to the fact that Pol II ChIP-seq is less sensitive over such large intervals to detect significant differences as compared to RNA-seq or chrRNA-seq.

Next, we wondered if RT transcripts were polyadenylated as the observed RT could arise from Pol II proceeding transcription following cleavage of the poly(A) transcripts, with no subsequent polyadenylation, as proposed in the torpedo termination model. To address this question, we performed polyA-RNA-seq on rWT and YFFF cells and analyzed RT poly(A) transcription. As shown in Figure 3G and H and exemplified at the CCR7 locus (Figure 3I), 3' RT is clearly accompanied by apparent polyadenylation. Thus the polyadenylation complex (CPA) might be associated to Pol II following the first poly(A) signal generally located a bit before the end of 3' UTRs. Intriguingly, increased poly(A) signal was also detected at 5' ends of the genes in the AS transcripts indicating that CPA could load at these locations. However, we cannot rule out that despite two rounds of poly(A) RNA enrichment, the sequenced libraries may contain residual non-polydenylated RNA. The observation that transcripts from non-polyadenylated histone genes show around 100-fold lower enrichment in our polyA-RNA-seq data as compared to total RNA-seq (Figure S3D), pleads overall against non-specific signal explaining our apparent RT phenotype. Currently we cannot discriminate whether the high level of intergenic and antisense RT RNA in the poly(A) fraction of mutant YFFF originates from constant polyadenylation at cryptic

poly(A) sites or from A-rich transcribed intergenic sequences. If polyadenylation of intergenic RNAs should occur, as our data suggest, this event obviously does not support termination of YFFF Pol II mutant. Together, our data indicate both 5' and 3' RT in the YFFF are linked to a transcriptional effect and that these RNAs may be subject of polyadenylation long after the normal poly(A) signal. We note however that polyadenylation can occur normally in the mutant despite the RT effect as highlighted by the high density of reads at 3' ends of genes, suggesting that polyadenylation and termination are uncoupled processes.

Tyrosine mutations are associated with reduced Pol II at the promoter-proximal pause sites and reduced nucleosome depletion around TSSs

Our Pol II ChIP-seq experiments showed a clear loss of Pol II accumulation in YFFF at promoters as exemplified at the MYCBP locus (Figure 4A). This was also evidenced in metagene profile analyses by applying normalization to gene bodies (Figure 4B) as before. When indexing genes according to pausing score classes from low to high, we also found that reduced Pol II levels were more pronounced at highly paused genes in mutant YFFF (Figure 4C, D and S4A). Such an effect was recently described following knock-down of the PAF1 complex, which also resulted in a global reduction of Pol II at pause sites in HCT116 cells (Chen et al., 2015). We then assessed if this could be accompanied by a change in nucleosome occupancy at promoters and performed MNase-seq in both rWT and YFFF. Interestingly, we found that nucleosome densities in proximity of the nucleosome-depleted regions (NDRs), upstream of the TSSs, were increased in the mutant (Figure 4E). This suggests that reduced Pol II levels at the pause site shortens the extent of NDRs and results in increased nucleosome occupancy, probably through reduced average Pol II

occupancy. Analyzing our ChIP-seq and RNA-seq data, we asked if reduced Pol II levels at promoters does correlate the 3' RT in the YFFF mutant (Figure S4B-D). This was not the case as the 3 groups with low medium and high Pol II at the pause site showed similar effects. Our data support a global reduction of Pol II at promoter-proximal pause sites in the YFFF mutant. This reduced Pol II accumulation could equally result from a defect in initiation, pausing or early elongation of the enzyme.

Transcribed enhancers and their epigenetic profiles are affected in the mutant

We previously showed that Tyr1P of Pol II is enriched at transcribed enhancers (TEs) (Descostes et al., 2014). Importantly, TEs are more active and more tissue-specific but the act of transcription itself at these regions does not necessarily yield stable-elongated RNAs (Lubas et al., 2015; Natoli and Andrau, 2012). Pausing of Pol II is also a hallmark of TEs (Core et al., 2014). The question therefore arises whether TEs also show reduced Pol II levels in the YFFF mutant as is observed at promoters. To investigate this, we compared Pol II and epigenetic marks characteristic of regulatory regions at both promoters and TEs. We first isolated 1316 intergenic TEs based on H3K27ac/H3K4me1/Pol II selection as described before (Descostes et al., 2014) that we compared to a selection of active control promoters. Interestingly, enhancers showed a strong reduction of Pol II and a more modest but significant loss of H3K27ac and H3K4me1 when compared to promoters (see Figure 5A and S5A for examples). This effect was confirmed genome-wide at most isolated enhancers (Figure 5B, C) but did not hold true for H3K4me3 that remained comparable in rWT and YFFF. We did not observe significant alteration of nucleosome positioning or NDRs at enhancers. We also analyzed chrRNA-seq in this context and found little difference in rWT or YFFF, suggesting that the defect in Pol II density, most likely

reflecting initiating or reduced Pol II pausing (Core et al., 2014), does not impair transcription at TEs (Figure S5B). This was in contrast with the situation observed at promoters in which reduced pausing was detectable after RNA-seq and ChIP-seq profiling at the 5' ends of the genes, both, in sense and AS direction (Figures 3 and 4).

Together, the data indicates that promoters and enhancers display similar molecular phenotypes due to YFFF mutations: both show reduced accumulation of Pol II at the proximity of pause site.

Mutant YFFF is impaired in its interaction with the Mediator and Integrator complexes

Our experiments shed light on a strong transcription termination defect phenotype at both 5' and 3' ends. To get further insight into what could be the mechanism of tyrosine involvement in termination, we immunoprecipitated Pol II and analyzed its associated proteins in rWT and YFFF cells by mass spectrometry (MS) experiments. To improve the signal to noise ratio, we performed 5 biological replicates for each of the two pull-downs following induction of recombinant Rpb1 and inhibition of endogenous Pol II by α-amanitin. The results highlight a marked loss of the Mediator (Med) and Integrator (Int) complexes (Figure 6A), two major interactors of the Pol II CTD (Baillat et al., 2005; Conaway and Conaway, 2015). Most of the 31 and 12 subunits of the Med and Int complexes, respectively, were lost in the YFFF mutant and in all biological replicates (Tables S1 and S2). We note that subunits of the kinase module of Mediator were not associated with Pol II in rWT and YFFF cells. This is consistent with the observation that binding of CTD and kinase module to Mediator is mutually exclusive(Allen and Taatjes, 2015; Tsai et al., 2013).

Because the cleavage/polyadenylation (CPA) complexes were previously linked to impaired termination phenotypes (Nojima et al., 2015), we searched for proteins associated to these complexes. We found that all of the CPSF subunits associated with Pol II in rWT in MS were also associated with Pol II in the YFFF mutant. Furthermore, XRN2 as well as most splicing factors peptides were found in comparable amounts in both fractions (Table S3).

We then asked if the loss of interaction with Med and Int complexes were specific to mutation of the tyrosine residues, and again made use of our S2AAA in which the same ¾ of repeats are mutated as compared to YFFF. We analyzed S2AAA mutant in MS experiments using the same induction/expression set-up. Interestingly, our results indicate that while Int complex subunits remain associated with Pol II, many Med subunits are lost or show decreased interaction in the S2AAA mutant (Figure S6A and table S5 and S6), suggesting that the loss of interaction with Integrator might be more critical for the observed RT phenotype when tyrosine residues are mutated in the CTD.

To exclude the possibility that reduced Med and Int levels in MS were due to reduced protein expression of these complexes, we determined expression of the Med15 and Int11 subunits by western blot and found no significant difference (Figure S6B, C).

Among other interesting proteins that lost interaction with the YFFF Pol II mutant, we found two CTD phosphatase RPAP2-associated proteins, RPRD1a and RPRD2, and one subunit of the PAF1c (WDR61) complex (Table S2).

We next wondered whether the loss of association of Med and Int with Pol II in YFFF would result in their impaired recruitment on DNA. To address this question, we performed ChIP experiments in rWT and YFFF cells at several target characteristic locations (including promoters, enhancer and snRNA gene). Following ChIP with

Med1 and Ints11 Abs, we found a decreased signal to background levels in YFFF as compared to rWT cells, showing that lost Pol II contacts resulted in reduced Med/Int occupancy on DNA as well (Figure 6B).

Altogether, these data indicate that loss of interaction with Med and Int complexes is a major consequence of tyrosine mutations and strongly suggest that loss of one, the other or both complexes might relate to the phenotypes linked to termination failure, and promoter/enhancer defects. However, they might also relate to yet uncharacterized independent function of Tyr1.

The YFFF mutations impair maturation of snRNAs and histone non-polyadenylated transcripts

The Integrator complex was previously described to function in synthesis and/or maturation of the snRNAs (Baillat et al., 2005; Egloff et al., 2007). Given, the major interaction defect observed in our MS experiments, we wondered if the YFFF mutant displayed impaired transcription/maturation at U1-5 snRNA genes. We thus analyzed transcription at these genes. We found little effect for the YFFF mutant in chrRNA-seq experiments for U1, U2 and U5 genes and a slight RNA signal increase at the U4 genes (Figure 7A), suggesting that nascent transcription was essentially not affected for these genes. A rather important increase, varying from 3-6x, was detected for total RNAs (Figure 7A), indicating that YFFF mutations may result in processing defects and stabilization of snRNAs. The opposite effect was observed for RNAs of non-poly(A) histone genes (Figure 7C). As for snRNA genes, nascent transcription (chrRNA-seq) was unaffected but the total RNA-seq signal was reduced several fold for histone RNAs (Figure 7C and 7D), indicating that non-poly(A) histone genes

undergo massive destabilization and that proper processing of transcripts from histone genes is also affected in YFFF.

Overall, our results show strong and opposite stabilization/maturation defects for histone and snRNA genes in the YFFF mutant that might be associated with the loss of Integrator interaction. We conclude that the lack of tyrosine residues in the CTD can lead to the failure of specific CTD-coupled processes as proper termination and processing of non-polyadenylated RNAs, while other processes as transcript elongation or 3' processing and polyadenylation of mRNAs remain unaffected.

Discussion

In this article, we report a novel function of mammalian CTD for transcription termination at 5' and 3' ends of genes. We show that tyrosine 1 residues of the CTD are required for termination thereby strongly limiting the extent of pervasive transcription. Among the phenotypes of Tyr1 mutants analyzed in this study, the read-through defect of the YFFF mutant was most striking. Transcription in this mutant remains high up to several hundreds of kb downstream of poly(A) sites, thus representing an exceptional case in which Pol II has lost the ability to terminate transcription. Although we can’t completely rule out that Pol II association to a termination factor could be impaired due to its altered expression, we do not favor this possibility based on our MS data where the number of CPSF peptides is comparable in both rWT and YFFF cells. Furthermore, in comparison to YFFF, the mutant S2AAA in which Ser2 residues are mutated in the same distal CTD repeats did not show significant read-through, indicating the high specificity of our Tyr1 mutant phenotype. The observed read-through phenotype of 5' antisense transcripts in YFFF mutant is consistent with our previous analysis showing association of Tyr1P

with antisense divergent transcription at the TSS of the genes(Descostes et al., 2014).

Pervasive transcription occurring after 3'ends of genes has been reported in specific WT cells or specific cellular context for Pol II mutants before. Recent works show that 3' read-through can be induced by osmotic stress or following Herpes-simplex 1 virus infection (Rutkowski et al., 2015; Vilborg et al., 2015). Given the overlap with the YFFF phenotype, it seems plausible that the CTD is involved in these processes by triggering transcriptional response to stress or viral infection. A mutation in the largest Pol II subunit was also described resulting in a faster and less processive enzyme (Kaplan et al., 2008; Kireeva et al., 2008). This mutation provokes distal termination at many genes, which was correlated with the fast elongation rate of the mutant enzyme (Fong et al., 2015). Although this phenotype was less pronounced than the one described here, we do not exclude that Tyr mutations might also alter the velocity of Pol II. Our study also points out little if any termination defect in the S2AAA mutant. This is surprising given a previous report that highlights the role of Ser2 for termination (Gu et al., 2013). However, we cannot rule out that the intact heptads (1-13) that were not mutated in our study still allow for the Ser2 to display a possible termination-related function.

How Pol II terminates transcription at 3' ends of genes remains a completely open question. Two main models were advanced in the past, the allosteric and the torpedo, the latter being prevalent in recent literature (Proudfoot, 2016). In this model, the exonuclease XRN2 attacks the uncapped 5' end of the nascent RNA after 3' cleavage and causes termination of transcription. Inactivation of XRN2 can result in

termination defects downstream of poly(A) site and shift termination to further downstream sequences (Fong et al., 2015; West et al., 2004). However, XRN2 knock-down does not result in massive, genome-wide, pervasive transcription of intergenic sequences (Nojima et al., 2015), suggesting that XRN2 contributes to tuning of termination but not to removal of Pol II from the template. We also found that XRN2 recruitment to Pol II is not altered in the YFFF mutant (Figure 6), further supporting the notion that its association with Pol II cannot prevent pervasive transcription. More expectedly, knock-down of the CPA subunits CPSF73 (CPSF3) and CstF64 was shown to lead to reduced termination (Nojima et al., 2015) but did not result in a massive pervasive transcription phenotype. Since the major CPSF subunits, including CPSF3, are recruited to Pol II and since polyadenylation of RNA occurs at least to the same extent in YFFF mutant as compared to WT at 3’ ends, we assume that the failure of termination occurs downstream of a functional 3' processing machinery. Overall, we conclude that one of the main functions of the missing tyrosine residues in the CTD of YFFF mutant is the control of transcription termination.

Another striking characteristic of the YFFF mutant is the reduced Pol II accumulation at 5’ ends of genes. This could result from a loading defect of essential CTD-associated factors such as the ones we identified in our MS analyses and we propose that an impaired promoter proximal pausing could be the cause of the termination defect 5’ of the genes. We also found a delayed, but not decreased in amplitude, Pol II accumulation at 3’ ends of genes, indicating that the lack of tyrosine residues generally affected pausing. At 3’ ends, our result suggests that impaired complex(es) association with Pol II would not allow proper pausing of the enzyme at

the first encountered poly(A) sites but instead at regions located on average 2.6 kb downstream. The consequence of this late pause could result in inefficient Pol II release from the template, possibly because of an impaired conformational transition in the enzyme or the lack of required signal, such as Tyr1P, required for efficient Pol II release from DNA. In either way, further works will be required to address these possibilities.

Our MS experiments indicate that Integrator and Mediator complexes do no longer associate with the CTD in the YFFF mutant. Both complexes were previously described as major CTD interactors, based on affinity purification (Baillat et al., 2005; Kim et al., 1994). The Mediator can act positively and negatively in the regulation of gene expression. It first supports the recruitment of Pol II to the promoter and later controls promoter release of Pol II in a CTD dependent manner (Allen and Taatjes, 2015). This negative regulation of gene expression by Mediator was first described for mutants with truncated versions of CTD in yeast, which were able to maintain cell growth if specific subunits of the Mediator were mutated(Kim et al., 1994; Koleske and Young, 1994). Therefore, it appears likely that a potential promoter release phenotype observed in YFFF mutant may be the consequence of the lack of interaction of CTD with the Mediator. This reduced pause could explain at least in part the 5’ AS pervasive effect in the mutant. Studies in yeast and plants also proposed involvement of the MED18 subunit of the Mediator head module in transcription termination (Lai et al., 2014; Mukundan and Ansari, 2011). Finally, reports have described a possible role for Mediator and Integrator in Pol II release through recruitment of the Super Elongation Complex (Donner et al., 2010; Gardini et al., 2014; Takahashi et al., 2011) and for Integrator in transcription termination (Skaar

et al., 2015). Therefore, both Int and Med complexes were previously connected to transcription pausing and termination-associated functions, making a direct link with the phenotypes described in this article.

At enhancers, we observed no apparent read-through phenotype, unlike at 5' or 3' ends of genes. However, Pol II occupancy and acetylation of histone H3K27 were impaired, suggesting that pausing is affected at enhancers in mutant YFFF. Given the known similarities and differences of promoters and enhancers (Core et al., 2014; Koch et al., 2011), Tyr1 residues of CTD might provide regulatory information that has different consequences at promoters and enhancers. A recent report indicated that WDR82 knockdown in macrophages results in pervasive transcription at enhancers(Austenaa et al., 2015) that we do not observe at enhancers in YFFF mutant. Conversely, when analyzing the transcriptome and Pol II data after WDR82 knockdown, we did not find strong read-through at gene units (unpublished observation). Altogether, this suggests that the control of termination might differ mechanistically for Pol II transcription initiated at promoters and enhancers.

Our work provides novel insights in the process of transcription termination and directly supports the involvement of Pol II CTD in this process. Future experiments should help to further dissect the mechanism of termination and establish possible roles of Mediator and Integrator complexes in termination and pause release. The YFFF mutant described here should also provide a great resource material to investigate the influence of extensive pervasive transcription on the frequency of DNA breaks in the genome, including the occurrence of DNA vs RNA polymerases collisions.

Acknowledgments

In the JCA lab, this work was supported by institutional grants from the CNRS, and by specific grants from “Agence Nationale de la Recherche” (ANR), ‘amorcage jeunes équipes’ Fondation pour la Recherche Medicale FRM AJE20130728183. In both JCA and DE labs, the work was also supported by a German-French BMBF-ANR grant ‘EpiGlyco’. This work was also supported by the Deutsche Forschungsgemeinschaft, SFB 1064 collaborative research center (CRC) – Chromatin Dynamics (DE). We thank Eric Soler and Mounia Lagha for critical reading of this manuscript.

Author Contributions

JCA and DE conceived the study and most of the experimental frame. NS made all constructs and performed all phenotypic characterization of the CTD mutants. MAM and YY prepared chromatin extracts for ChIP and RNA (chrRNA) or total RNAs. MAM and YY performed ChIP-seq and RNA-seq experiments, including QCs and library preparations. MAM, DM, and AZA performed bioinformatics analyses. NS, IF and AI performed MS experiments and data analysis. CE performed Mediator and Integrator ChIP experiments. TMD and RS contributed in the constructions and phenotypic characterization of the mutants. SK and HB performed the sequencing of the libraries. MAM, NS, JCA and DE prepared and finalized the figures. JCA and DE wrote the manuscript, which was reviewed by all authors.

Declaration of Interests

The authors declare no competing financial interests.

References

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

553 Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter
554 directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-
555 363.
556 Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count
557 data. Genome Biol *11*, R106.
558 Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with
559 high-throughput sequencing data. Bioinformatics *31*, 166-169.
560 Austenaa, L.M., Barozzi, I., Simonatto, M., Masella, S., Della Chiara, G., Ghisletti, S., Curina,
561 A., de Wit, E., Bouwman, B.A., de Pretis, S.*, et al.* (2015). Transcription of Mammalian cis-
562 Regulatory Elements Is Restrained by Actively Enforced Early Termination. Molecular
563 cell *60*, 460-474.
564 Baillat, D., Hakimi, M.A., Naar, A.M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005).
565 Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the
566 C-terminal repeat of RNA polymerase II. Cell *123*, 265-276.
567 Bartolomei, M.S., Halden, N.F., Cullen, C.R., and Corden, J.L. (1988). Genetic analysis of
568 the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA
569 polymerase II. Molecular and cellular biology *8*, 330-339.
570 Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L.,
571 and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by
572 sequence analysis of subcellular RNA fractions. Cell *150*, 279-290.
573 Chapman, R.D., Heidemann, M., Albert, T.K., Mailhammer, R., Flatley, A., Meisterernst, M.,
574 Kremmer, E., and Eick, D. (2007). Transcribing RNA polymerase II is phosphorylated at
575 CTD residue serine-7. Science *318*, 1780-1782.
576 Chen, F.X., Woodfin, A.R., Gardini, A., Rickels, R.A., Marshall, S.A., Smith, E.R., Shiekhattar,
577 R., and Shilatifard, A. (2015). PAF1, a Molecular Regulator of Promoter-Proximal Pausing
578 by RNA Polymerase II. Cell *162*, 1003-1015.
579 Conaway, R.C., and Conaway, J.W. (2015). Orchestrating transcription with the Pol II CTD.
580 Nature reviews Molecular cell biology *16*, 128.
581 Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis
582 of nascent RNA identifies a unified architecture of initiation regions at mammalian
583 promoters and enhancers. Nature genetics *46*, 1311-1320.
584 Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread
585 pausing and divergent initiation at human promoters. Science *322*, 1845-1848.
586 Descostes, N., Heidemann, M., Spinelli, L., Schuller, R., Maqbool, M.A., Fenouil, R., Koch, F.,
587 Innocenti, C., Gut, M., Gut, I.*, et al.* (2014). Tyrosine phosphorylation of RNA polymerase
588 II CTD is associated with antisense promoter transcription and active enhancers in
589 mammalian cells. eLife *3*, e02105.
590 Donner, A.J., Ebmeier, C.C., Taatjes, D.J., and Espinosa, J.M. (2010). CDK8 is a positive
591 regulator of transcriptional elongation within the serum response network. Nature
592 structural & molecular biology *17*, 194-201.
593 Egloff, S., O'Reilly, D., Chapman, R.D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D., and
594 Murphy, S. (2007). Serine-7 of the RNA polymerase II CTD is specifically required for
595 snRNA gene expression. Science *318*, 1777-1779.
596 Eick, D., and Geyer, M. (2013). The RNA polymerase II carboxy-terminal domain (CTD)
597 code. Chemical reviews *113*, 8456-8490.
598 Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P.,
599 Spicuglia, S., Gut, M., Gut, I.*, et al.* (2012a). CpG islands and GC content dictate
600 nucleosome depletion in a transcription-independent manner at mammalian promoters.
601 Genome Research.

602 Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P.,
603 Spicuglia, S., Gut, M., Gut, I.*, et al.* (2012b). CpG islands and GC content dictate
604 nucleosome depletion in a transcription-independent manner at mammalian promoters.
605 Genome Res *22*, 2399-2408.
606 Fenouil, R., Descostes, N., Spinelli, L., Koch, F., Maqbool, M.A., Benoukraf, T., Cauchy, P.,
607 Innocenti, C., Ferrier, P., and Andrau, J.-C. (2016). Pasha : a versatile R package for piling
608 chromatin HTS data. Bioinformatics.
609 Fong, N., Brannan, K., Erickson, B., Kim, H., Cortazar, M.A., Sheridan, R.M., Nguyen, T.,
610 Karp, S., and Bentley, D.L. (2015). Effects of Transcription Elongation Rate and Xrn2
611 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic
612 Competition. Mol Cell *60*, 256-267.
613 Gardini, A., Baillat, D., Cesaroni, M., Hu, D., Marinis, J.M., Wagner, E.J., Lazar, M.A.,
614 Shilatifard, A., and Shiekhattar, R. (2014). Integrator regulates transcriptional initiation
615 and pause release following activation. Molecular cell *56*, 128-139.
616 Grosso, A.R., Leite, A.P., Carvalho, S., Matos, M.R., Martins, F.B., Vitor, A.C., Desterro, J.M.,
617 Carmo-Fonseca, M., and de Almeida, S.F. (2015). Pervasive transcription read-through
618 promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. eLife
619 *4*.
620 Gu, B., Eick, D., and Bensaude, O. (2013). CTD serine-2 plays a critical role in splicing and
621 termination factor recruitment to RNA polymerase II in vivo. Nucleic acids research *41*,
622 1591-1603.
623 Harlen, K.M., Trotta, K.L., Smith, E.E., Mosaheb, M.M., Fuchs, S.M., and Churchman, L.S.
624 (2016). Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied
625 Roles for Each Phospho-CTD Residue. Cell reports *15*, 2147-2158.
626 Hazelbaker, D.Z., Marquardt, S., Wlotzka, W., and Buratowski, S. (2013). Kinetic
627 competition between RNA Polymerase II and Sen1-dependent transcription termination.
628 Molecular cell *49*, 55-66.
629 Hintermair, C., Heidemann, M., Koch, F., Descostes, N., Gut, M., Gut, I., Fenouil, R., Ferrier,
630 P., Flatley, A., Kremmer, E.*, et al.* (2012a). Threonine-4 of mammalian RNA polymerase II
631 CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. Embo J
632 *31*, 2784-2797.
633 Hintermair, C., Heidemann, M., Koch, F., Descostes, N., Gut, M., Gut, I., Fenouil, R., Ferrier,
634 P., Flatley, A., Kremmer, E.*, et al.* (2012b). Threonine-4 of mammalian RNA polymerase II
635 CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. EMBO
636 J *31*, 2784-2797.
637 Hsin, J.P., Li, W., Hoque, M., Tian, B., and Manley, J.L. (2014). RNAP II CTD tyrosine 1
638 performs diverse functions in vertebrate cells. eLife *3*, e02112.
639 Hsin, J.P., Sheth, A., and Manley, J.L. (2011). RNAP II CTD phosphorylated on threonine-4
640 is required for histone mRNA 3' end processing. Science *334*, 683-686.
641 Iben, J.R., Mazeika, J.K., Hasson, S., Rijal, K., Arimbasseri, A.G., Russo, A.N., and Maraia, R.J.
642 (2011). Point mutations in the Rpb9-homologous domain of Rpc11 that impair
643 transcription termination by RNA polymerase III. Nucleic acids research *39*, 6100-6113.
644 Ishihama, Y., Rappsilber, J., and Mann, M. (2006). Modular stop and go extraction tips
645 with stacked disks for parallel and multidimensional Peptide fractionation in proteomics.
646 J Proteome Res *5*, 988-994.
647 Kaplan, C.D., Larsson, K.M., and Kornberg, R.D. (2008). The RNA polymerase II trigger
648 loop functions in substrate selection and is directly targeted by alpha-amanitin.
649 Molecular cell *30*, 547-556.

650 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013).
651 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions
652 and gene fusions. Genome Biol *14*, R36.
653 Kim, Y.J., Bjorklund, S., Li, Y., Sayre, M.H., and Kornberg, R.D. (1994). A multiprotein
654 mediator of transcriptional activation and its interaction with the C-terminal repeat
655 domain of RNA polymerase II. Cell *77*, 599-608.
656 Kireeva, M.L., Nedialkov, Y.A., Cremona, G.H., Purtov, Y.A., Lubkowska, L., Malagon, F.,
657 Burton, Z.F., Strathern, J.N., and Kashlev, M. (2008). Transient reversal of RNA
658 polymerase II active site closing controls fidelity of transcription elongation. Molecular
659 cell *30*, 557-566.
660 Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T.K., Zacarias-Cabeza, J., Spicuglia, S., de la
661 Chapelle, A.L., Heidemann, M., Hintermair, C.*, et al.* (2011). Transcription initiation
662 platforms and GTF recruitment at tissue-specific enhancers and promoters. Nat Struct
663 Mol Biol *18*, 956-963.
664 Koleske, A.J., and Young, R.A. (1994). An RNA polymerase II holoenzyme responsive to
665 activators. Nature *368*, 466-469.
666 Lai, Z., Schluttenhofer, C.M., Bhide, K., Shreve, J., Thimmapuram, J., Lee, S.Y., Yun, D.J., and
667 Mengiste, T. (2014). MED18 interaction with distinct transcription factors regulates
668 multiple plant functions. Nature communications *5*, 3064.
669 Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat
670 Methods *9*, 357-359.
671 Lepoivre, C., Belhocine, M., Bergon, A., Griffon, A., Yammine, M., Vanhille, L., Zacarias-
672 Cabeza, J., Garibal, M.A., Koch, F., Maqbool, M.A.*, et al.* (2013). Divergent transcription is
673 associated with promoters of transcriptional regulators. BMC genomics *14*, 914.
674 Lubas, M., Andersen, P.R., Schein, A., Dziembowski, A., Kudla, G., and Jensen, T.H. (2015).
675 The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to
676 direct early ribonucleolysis. Cell reports *10*, 178-192.
677 Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R.,
678 Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript
679 sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*,
680 541-554.
681 Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C.,
682 Kremmer, E., Eick, D., and Cramer, P. (2012). CTD tyrosine phosphorylation impairs
683 termination factor recruitment to RNA polymerase II. Science *336*, 1723-1725.
684 McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S.D.,
685 Wickens, M., and Bentley, D.L. (1997). The C-terminal domain of RNA polymerase II
686 couples mRNA processing to transcription. Nature *385*, 357-361.
687 Meininghaus, M., Chapman, R.D., Horndasch, M., and Eick, D. (2000). Conditional
688 expression of RNA polymerase II in mammalian cells. Deletion of the carboxyl-terminal
689 domain of the large subunit affects early steps in transcription. The Journal of biological
690 chemistry *275*, 24375-24382.
691 Mukundan, B., and Ansari, A. (2011). Novel role for mediator complex subunit
692 Srb5/Med18 in termination of transcription. The Journal of biological chemistry *286*,
693 37053-37057.
694 Nagaike, T., Logan, C., Hotta, I., Rozenblatt-Rosen, O., Meyerson, M., and Manley, J.L.
695 (2011). Transcriptional activators enhance polyadenylation of mRNA precursors.
696 Molecular cell *41*, 409-418.
697 Natoli, G., and Andrau, J.C. (2012). Noncoding transcription at enhancers: general
698 principles and functional models. Annual review of genetics *46*, 1-19.

699 Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and
700 Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent
701 Transcription Coupled to RNA Processing. Cell *161*, 526-540.
702 Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof,
703 I., Schein, A., Andersen, P.R.*, et al.* (2013). Polyadenylation site-induced decay of
704 upstream transcripts enforces promoter directionality. Nature structural & molecular
705 biology *20*, 923-928.
706 Pinto, P.A., Henriques, T., Freitas, M.O., Martins, T., Domingues, R.G., Wyrzykowska, P.S.,
707 Coelho, P.A., Carmo, A.M., Sunkel, C.E., Proudfoot, N.J.*, et al.* (2011). RNA polymerase II
708 kinetics in polo polyadenylation signal selection. EMBO J *30*, 2431-2444.
709 Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano,
710 C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals
711 transcription upstream of active human promoters. Science *322*, 1851-1854.
712 Proudfoot, N.J. (2016). Transcriptional termination in mammals: Stopping the RNA
713 polymerase II juggernaut. Science *352*, aad9926.
714 Rutkowski, A.J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel,
715 P., Efstathiou, S., Zimmer, R., Friedel, C.C.*, et al.* (2015). Widespread disruption of host
716 transcription termination in HSV-1 infection. Nature communications *6*, 7126.
717 Schlackow, M., Nojima, T., Gomes, T., Dhir, A., Carmo-Fonseca, M., and Proudfoot, N.J.
718 (2017). Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs.
719 Molecular cell *65*, 25-38.
720 Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and
721 Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-
722 1851.
723 Shaaban, S.A., Krupp, B.M., and Hall, B.D. (1995). Termination-altering mutations in the
724 second-largest subunit of yeast RNA polymerase III. Molecular and cellular biology *15*,
725 1467-1478.
726 Shevchenko, A., Chernushevich, I., Wilm, M., and Mann, M. (2000). De Novo peptide
727 sequencing by nanoelectrospray tandem mass spectrometry using triple quadrupole
728 and quadrupole/time-of-flight instruments. Methods Mol Biol *146*, 1-16.
729 Skaar, J.R., Ferris, A.L., Wu, X., Saraf, A., Khanna, K.K., Florens, L., Washburn, M.P., Hughes,
730 S.H., and Pagano, M. (2015). The Integrator complex controls the termination of
731 transcription at diverse classes of gene targets. Cell research *25*, 288-305.
732 Takahashi, H., Parmely, T.J., Sato, S., Tomomori-Sato, C., Banks, C.A., Kong, S.E., Szutorisz,
733 H., Swanson, S.K., Martin-Brown, S., Washburn, M.P.*, et al.* (2011). Human mediator
734 subunit MED26 functions as a docking site for transcription elongation factors. Cell *146*,
735 92-104.
736 Tsai, K.L., Sato, S., Tomomori-Sato, C., Conaway, R.C., Conaway, J.W., and Asturias, F.J.
737 (2013). A conserved Mediator-CDK8 kinase module association regulates Mediator-RNA
738 polymerase II interaction. Nature structural & molecular biology *20*, 611-619.
739 Vilborg, A., Passarelli, M.C., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2015). Widespread
740 Inducible Transcription Downstream of Human Genes. Molecular cell *59*, 449-461.
741 West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2
742 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*,
743 522-525.
744 Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M.
745 (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-
746 electrospray mass spectrometry. Nature *379*, 466-469.
747

Figure legends

Figure 1: Screening of human CTD tyrosine mutants

(A) Schematic representation of Pol II CTD tyrosine mutants. WT and mutant heptads are represented in light and dark blue, respectively. **(B)** Western blot of rWT and CTD mutants following 24h induction and 48h α-amanitin treatment of the cells (72h induction). **(C)** Example of read-through phenotype at 3' end of the Znf621 gene. **(D)** Average metagene profile of total sense RNA-seq signal (asinh transformed) over the gene bodies and 20kb upstream and downstream regions. All profiles were normalized so that signals are equivalent on gene bodies (see methods). The 3 stars indicate a p-value < $2x10^{-16}$ (2 sided Wilcoxon test) between rWT and YFFF. See also Figure S1

Figure 2: YFFF mutations cause a massive read-through (RT) both at 3' and 5' (antisense) ends of genes

A) Example of RNA-seq signal (y axis) for a coding gene showing both 5'(AS) and 3'(S) RT that extends to at least 100 kb upstream of 5' end and 300 kb downstream of 3' end. **B)** Average metagene profile of total RNA-seq signal (asinh) in sense (blue) and AS (red) orientation of the gene bodies and 20kb upstream and downstream regions. **C)** Density plots of antisense RNA-seq signal in 20kb region upstream (a) or downstream (c) of the genes (FPM) or sense signal on gene body (FPKM in (b)) in rWT and YFFF cells. Selected regions were excluding genes <2kb and/or having other genes within 20 kb. Regions concern 1160 upstream AS, 3999 gene bodies, and 1263 downstream areas of the genome. All pairs of distribution are significantly different with a p-value < $2x10^{-16}$ (2 sided Wilcoxon test). **D)** Boxplot of upstream AS indices (left panel) and downstream RT indices (right panel). Units are

asinh transformed. FPM: fragments per million nucleotides. FPKM: fragments per kilobase of transcript per million mapped reads. See also Figure S2 **E)** Chromosome 2 snapshot of total RNA-seq data illustrating the generality of the YFFF read-through phenotype.

Figure 3: RT phenotype of the YFFF mutant is due to reduced transcription termination and gives rise to polyadenylated transcripts

A) Average metagene profile of chrRNA-seq (asinh) in sense (blue) and AS (red) orientation of the gene bodies and 20kb upstream and downstream regions. Profiles were normalized so that sense RNA signals are equivalent on gene bodies. **B)** Boxplot of upstream AS transcription index (left panel) and downstream RT transcription index (right panel) calculated with chrRNA signal. **C)** CCR7 example of chrRNA-seq signal (y axis) RT in YFFF. **D)** Average Pol II ChIP-seq profiles of significantly bound genes in rWT (top 30% protein coding genes) in rWT and YFFF around 3' ends. Data are normalized to the same gene body level. **E)** Boxplot of upstream AS transcription index (left panel) and downstream RT transcription index (right panel) calculated with Pol II ChIP signal. **F)** CCR7 locus showing Pol II RT activity (ChIP-seq signal is shown in y axis). **G)** Average profile of sense and AS poly(A) RNA signal in rWT and YFFF cells. **H)** Boxplot of upstream AS transcription index (left panel) and downstream RT transcription index (right panel) calculated with poly(A) RNA. **I)** CCR7 locus showing that RT RNA is polyadenylated (RNA-seq signal in y axis). In A) to I), p-values are < 2.2e-16. See also Figure S3. Figure S3E shows the non-normalized metagene chrRNA-seq and poly-RNA-seq show in 3A and 3G.

Figure 4: YFFF mutations result in massive loss of Pol II accumulation at promoter proximal pause sites

A) Examples of Pol II ChIP-seq showing promoter-proximal pausing loss in YFFF cells occurring at multiple genes (ChIP-seq signal is shown in y axis). **B)** Average Pol II profiles at TSS of significantly bound genes (top 30% coding genes) in rWT and corresponding profile in YFFF. Data are normalized to bring the signals to the same level in gene body. **C)** Pol II density heatmaps at TSS of genes ranked by increasing pausing score in rWT and shown at rWT and YFFF TSSs. The boundaries of the 3 pausing groups 1-3 are shown on the left of the heatmaps. **D)** Box plots of pausing scores for the 3 groups in rWT and YFFF. Only the Groups 2 and 3 show significant differences. **E)** Nucleosome densities at promoters in rWT and YFFF mutant. Data are normalized so that MNase-seq counts are equivalent in both experiments (scaling). In B) and E) the light blue rectangles indicate the areas that were taken in account for calculation of the indicated p-values. See also Figure S4

Figure 5: YFFF mutations result in impaired Pol II recruitment and epigenetic marking at active enhancers

A) Putative enhancers or enhancer stretches (in pink rectangles) around the DNAJC12 locus show altered Pol II loading in YFFF. **B)** Heatmap of Pol II densities at enhancers ranked by increasing Pol II signal in rWT and corresponding heatmap in YFFF. **C)** Average Pol II profiles, histone marks and nucleosome density at intergenic enhancers (upper row) and control promoters (lower row, top 30% Pol II promoters in rWT without any other genes in surrounding 5kb interval). P-values are indicated on the top right. Light blue rectangles indicate the areas that were taken in account for their calculation. See also Figure S5

Figure 6: Mass spectrometry differential analysis of rWT and YFFF Pol II interactome

A) Volcano plot comparing the Pol II interactome in rWT and the YFFF mutant. The table on the right lists selected proteins and complexes that interact with the Pol II of both, rWT and the YFFF mutant. All 12 subunits of the Pol II, several splicing factors and 3' end processing factors are listed in the table. Represented on the left are the proteins and complexes that do not interact with the YFFF Pol II. Highlighted are the 25 subunits of the Mediator complex (green); 11 subunits of the Integrator (red), CTD phosphatase associated proteins (magenta), E3-ubiquitin ligase; components of SOSS complex and few other proteins (blue). Threshold: log2 fold change ≥ 5; p-value < 0.05. Data is based on five independent biological replicates. See also Figure S6, Table S1, S2, S3. For these experiments, cells were collected after 24h of induction and 48h of amanitin treatment as for the other assays. **B)** qPCR ChIP of Mediator (Med1) and Integrator (Inst11) at the *Ets1* enhancer (~24kb upsteam) and at the *Rnu11*, *Myc*, *Snhg3-Rcc1*, *Kxd1* and *Taf12* promoters in rWT and YFFF cells. Dashed lines highlight signals observed at the negative control region. Data are means ± SEM, n=2.

Figure 7: YFFF mutations affect maturation of transcripts from snRNA and histone genes

A) Average metagene profile of chrRNA, total RNA (sense orientation) at 50% most highly transcribed U1, U2, U4 and U5 snRNA genes. Grey rectangles indicate the corresponding gene size. **B)** Examples of total RNA and chrRNA signal at representative U1 and U2 snRNA loci. **C)** Average metagene profile of chrRNA and

total RNA (asinh) in sense orientation at 50% most highly transcribed non-polyadenylated histone genes. **D)** Example of total RNA and chrRNA signal at histone genes cluster.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents should be directed to and will be fulfilled by Lead Contact Jean-Christophe Andrau (jean-christophe.andrau@igmm.cnrs.fr).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Establishing stable Cell lines:

Raji is an Epstein-Barr-virus-positive Burkitt's lymphoma cell line of Male origin. Full-length Rpb1 expression vector (rWT, YFFF, YYFF, FYYF and YFFY) were transfected into Raji cells using 1 X 10^7 cells (10 µg plasmid, 960 µF, 250V). Polyclonal cell lines were established after selection with G418 (1 mg/ml) for 2-3 weeks. Tetracycline was removed to induce the expression of recombinant Rpb1 by washing the cells three times with 50 ml of phosphate-buffered saline (PBS) supplemented with 1% fetal calf serum (FCS) (Gibco, Invitrogen). 24 h after induction, cells were cultured in the presence of 2 µg/ml of α-amanitin (Sigma) to inhibit endogenous Pol II. Cell lines were maintained in RPMI 1640 medium supplemented with 10% fetal calf serum, 100 U/ml penicillin, 100 µg/ml streptomycin, 2mM L-glutamine (Gibco, Invitrogen) at 37°C and 5% CO2.

METHOD DETAILS

I- Experimental Procedures

Antibodies:

Monoclonal antibodies specific for haemagglutinin (HA)-tag (3F10, Roche), (12CA5, Sigma) and polyclonal antibodies against MED15 (11566-1-AP, Proteintech), MED1 (A301-793A, Bethyl laboratories) and INT11 (A301-274A, Bethyl laboratories) are commercially available. Monoclonal antibodies against Rpb1 (Pol 3.3), Ser2P (3E10), Ser5P (3E8), Ser7P (4E12) and Thr4P (6D7) were described previously (Chapman et al., 2007; Hintermair et al., 2012a) and monoclonal antibody against GAPDH (5C4) was received from Elisabeth Kremmer, Helmholtz Zentrum Munich.

Construction of the CTD mutants:

Mouse CTD sequences of rWT and tyrosine mutants (YYYF, YYFF, FYYF and YFFY) with an optimized human codon usage were synthesized by Gene Art (Regensburg) and cloned into LS*mock vector (Meininghaus et al., 2000). All final constructs were sequenced before usage.

Western blot analysis:

Cells were washed twice with PBS and directly lysed with 2X Laemmli buffer. Whole cell lysates were separated on SDS-PAGE (6.5% gel) and blotted on a nitrocellulose membrane (GE healthcare). The membranes were blocked with 5% milk/TBS-T solution for 1 h and incubated overnight with the primary antibody at 4°C. Afterwards, the membranes were incubated, either with IRDye-labelled secondary antibodies

against rat (680 nm; Alexa, Invitrogen) and/or mouse (800 nm; Rockford, Biomol) and analyzed using an Odyssey Imaging System (Li-Cor) or they were stained with HRP-conjugated secondary antibodies against rat (Sigma) or mouse (Promega) to be detected by chemiluminescence.

Growth kinetics:

Growth kinetics of rWT, CTD mutants and wild-type Raji cells were monitored over a period of 10 days. For each cell line, 20 X 10^6 cells were induced and the number of living cells (Nl) and the number of dead cells (Nd) were calculated every day using trypan-blue staining. Cumulative living cell number was calculated by multiplying the total number of living cells (Nl) with the factor by which the culture was split over the course of the experiment. These kinetics recapitulate growth features of samples collection for ChIPseq and RNAseq experiments.

Purification of Pol II interacting proteins for mass spectrometric analysis:

For purification of recombinant Rpb1, α-HA antibody (12CA5) was coupled to sepharose A/G beads for 4 h at 4°C. Simultaneously, cells (7.5 X 10^7) were washed twice with ice cold PBS and lysed in lysis buffer [50mM Tris-HCl pH 8.0, 150mM NaCl, 1% NP-40 (Roche), 1 X PhosStop (Roche), 1 X protease cocktail (Roche)] for 30 min on ice. Samples were sonified (Sonifier 250 BRANSON, 3 × 20 cycles, output 5, duty cycle 50) and incubated on a shaker for 1 h at 4°C. Samples were then centrifuged at 10,000g for 15 minutes and the supernatants were incubated with antibody-coupled sepharose A/G beads for overnight at 4°C. Next day, beads were washed three times with lysis buffer and continued with either on-beads trypsin

digest or boiled with 2X Laemmli buffer (95°C, 8 min) to load proteins on SDS-PAGE for the subsequent in-gel trypsin digest.

On-beads trypsin digest:

Following the standard immunoprecipitation procedure, beads were first washed with lysis buffer (three times) and then with 50mM NH_4HCO_3 (ammonium bicarbonate). For trypsin digest, beads were incubated with 100 µl of 10 ng/µl of trypsin solution in 1M Urea and 50mM NH_4HCO_3 for 30 minutes at 25°C. The supernatant was collected, beads washed twice with 50mM NH_4HCO_3 and all three supernatants collected together and incubated overnight at 25°C after addition of 1mM DTT. 27mM of iodoacetamide (IAA) was then added to the samples and incubated at 25°C for 30 minutes in dark. Next, 1 µl of 1M DTT was added to the samples and incubated for 10 minutes to quench the IAA. Finally, 2.5 µl of trifluoroacetic acid (TFA) were added to the samples and desalted using C18 stage tips (Ishihama et al., 2006). Samples were evaporated to dryness, re-suspended in 30 µl of 0.1% formic acid solution and stored at -20°C until LC-MS analysis.

In-gel trypsin digest:

A standardized protocol was used for in-gel digestion with minor modifications (Shevchenko et al., 2000; Wilm et al., 1996). The digested peptides were evaporated to 5 µl and re-suspended in 30 µl of 0.1% TFA solution prior to desalting by C18 stage tips. Samples were evaporated to dryness and re-suspended in 30 µl of 0.1% formic acid solution and stored at -20°C until LC-MS analysis.

Liquid Chromatography Coupled to Tandem Mass Spectrometry:

For LC-MS/MS purposes, desalted peptides were injected in an Ultimate 3000 RSLCnano system (Thermo), separated in a 15-cm analytical column (75µm ID with ReproSil-Pur C18-AQ 2.4 µm from Dr. Maisch) with a 50 min gradient from 5 to 60% acetonitrile in 0.1% formic acid. The effluent from the HPLC was directly electrosprayed into a QexactiveHF (Thermo) operated in data dependent mode to automatically switch between full scan MS and MS/MS acquisition. Survey full scan MS spectra (from m/z 375–1600) were acquired with resolution R=60,000 at m/z 400 (AGC target of 3x106). The 10 most intense peptide ions with charge states between 2 and 5 were sequentially isolated to a target value of $1x10^5$, and fragmented at 27% normalized collision energy. Typical mass spectrometric conditions were: spray voltage, 1.5 kV; no sheath and auxiliary gas flow; heated capillary temperature, 250ºC; ion selection threshold, 33.000 counts. MaxQuant 1.5.2.8 was used to identify proteins and quantify by iBAQ with the following parameters: Database, Uniprot_Hsapiens_3AUP000005640_151111; MS tol, 10ppm; MS/MS tol, 0.5 Da; Peptide FDR, 0.1; Protein FDR, 0.01 Min. peptide Length, 5; Variable modifications, Oxidation (M); Fixed modifications, Carbamidomethyl (C); Peptides for protein quantitation, razor and unique; Min. peptides, 1; Min. ratio count, 2. Identified proteins were considered as interaction partners if their MaxQuant iBAQ values displayed a greater than log2 5-fold enrichment and p-value 0.05 (ANOVA) when compared to the rWT control. The data was processed for visualization using R (https://www.r-project.org/).

ChIP-seq and ChIP-qPCR:

To cross-link the cells for ChIP, 1/10th volume of 10X crosslinking solution (100mM NaCl, 1mM EDTA pH 8, 0.5mM EGTA pH 8, 50mM HEPES pH 7.8 and 11%

formaldehyde) was added to the raji cells in culture medium. After 10 minutes' incubation at room temperature, glycine was added to a final concentration of 250mM to quench the remaining formaldehyde and stop cross-linking. After five minutes of quenching, cells were washed twice with cold PBS. Cells were then sonicated as described in next paragraph or snap frozen in liquid nitrogen and stored at -80°C for sonication at a later stage.

For sonication, 50 x10^6 cross-linked raji cells were lysed by resuspending in cold 2.5mL LB1 (50mM Hepes pH 7.5, 140mM NaCl, 1mM EDTA pH 8, 10% glycerol, 0.75% NP-40, 0.25% Triton X-100) at 4°C for 20 minutes on a rotating wheel. Nuclei were pelleted down by spinning at 1350 rcf in a refrigerated centrifuge and washed in 2.5mL LB2 (200mM NaCl, 1mM EDTA pH 8, 0.5mM EGTA pH 8, 10mM Tris pH 8) for 10 minutes at 4°C on a rotating wheel followed by centrifugation to collect nuclei. Nuclei were then resuspended in 1mL LB3 (1mM EDTA pH 8, 0.5mM EGTA pH 8, 10mM Tris pH 8, 100mM NaCl, 0.1% Na-Deoxycholate, 0.5% N-lauroylsarcosine) and sonicated using Bioruptor Pico (Diagenode) in 15mL tubes for 25 cycles of 30 sec ON and 30 sec OFF pulses in 4°C water bath. All buffers (LB1, LB2 and LB3) were complemented with EDTA free Protease inhibitor cocktail (Roche), 0.2mM PMSF and 1µg/mL Pepstatin just before use. After sonication, Triton X-100 was added to a final concentration of 1% followed by centrifugation at 20000 rcf and 4°C for 10 minutes to remove particulate matter. After taking a 50µl aliquot to serve as input and to analyze fragmentation, chromatin was aliquoted and snap-frozen in liquid nitrogen and stored at -80°C until use in ChIP assays.

Input aliquots were mixed with equal volume of 2X elution buffer (100mM Tris pH 8.0, 20mM EDTA, 2% SDS) and incubated at 65°C for 12 hours for reverse-crosslinking. An equal volume of TE buffer (10mM Tris pH 8 and 1mM EDTA pH 8) was added to

dilute the SDS to 0.5% followed by treatment with RNase A (0.2µg/mL) at 37°C for one hour and Proteinase K (0.2µg/mL) for two hours at 55°C. DNA was isolated by phenol:chloroform: isoamylalcohol (25:24:1 pH 8) extraction followed by Qiaquick PCR Purification (Qiagen, Germany). Purified DNA was then analyzed on a 2% agarose gel or on Bioanalyzer (Agilent, USA) using a High Sensitivity DNA Assay.

Protein-G coated Dynabeads were incubated at 4°C in blocking solution (0.5% BSA in PBS) carrying specific antibodies to prepare beads pre-coated with specific antibody which were then used for ChIP. Sonicated chromatin was added to pre-coated beads and the mix was incubated overnight at 4°C on a rotating wheel (please refer to the Table S4 for information on specific antibodies and number of cells used for each ChIP). After incubation with chromatin, beads were washed 7 times with Wash buffer (50mM Hepes pH 7.6, 500mM LiCl, 1mM EDTA pH 8, 1% NP-40, 0.7% Na-Deoxycholate, 1X protease inhibitor cocktail) followed by one wash with TE-NaCl buffer (10mM Tris pH 8 and 1mM EDTA pH 8, 50mM NaCl) and a final wash with TE buffer (10mM Tris pH 8 and 1mM EDTA pH 8). Immunoprecipitated chromatin was eluted by two sequential incubations with 50µl Elution buffer (50mM Tris pH 8, 10mM EDTA pH 8, 1% SDS) at 65°C for 15 minutes. The two eluates were pooled and incubated at 65°C for 12 hours to reverse-crosslink the chromatin followed by treatment with RNase A and Proteinase K and purification of DNA as described above for Input samples. Med1 and Ints11 IPs were analyzed by qPCR (Stratagene) following manufacturer recommendations. Purified DNA was quantified with Qubit DS DNA HS Assay (ThermoFisher Scientific, USA).

At lease 1ng of ChIP DNA was used to prepare sequencing library with Illumina ChIP Sample Library Prep Kit (Illumina, USA). After end-repair and adapter ligation, library fragments were size-selected using E-Gel SizeSelect 2% Agarose Gel

(ThermoFisher Scientific, USA) followed by 12 cycles of PCR amplification. Barcoded libraries from different samples were pooled together and sequenced on Illumina HiSeq2000 platform in paired-end sequencing runs.

Total RNA-seq:

RNA was extracted from cells using TRIzol Reagent (ThermoFisher Scientific, USA) according to manufacturer's instructions. Any contaminating DNA was digested with rigorous Turbo DNase (ThermoFisher Scientific, USA) treatment according to manufacturer's instruction followed by a second extraction with TRIzol reagent to eliminate traces of contaminants. Purified RNA was quantified with Nanodrop 1000 instrument and quality was assessed using RNA Nano or Pico Assay kit with Bioanalyzer (Agilent Technologies, USA). Only the RNA samples with RIN above 8 were used for sequencing.

For strand-specific sequencing, ribosomal RNA was removed from total RNA with Ribo-Zero rRNA Removal Kit (EpiCenter, USA) according to manufacturer's instructions and depletion of rRNA was confirmed by analyzing the samples on RNA Pico Assay on Bioanalyzer. Libraries were prepared either with ScriptSeq Total RNA Library prep kit (EpiCenter, USA) according to manufacturer's instructions for the comparison of rWT and the 4 mutants shown in Figure 1 or with Small RNA Library Prep Kit (Illumina, USA) using a modified protocol for the data showed in Figure 2 and later as follows: 50ng rRNA depleted total RNA was fragmented to ~150bp by digesting with 1U of RNaseIII (ThermoFisher Scientific, USA) for 10 minutes at 37°C in a 10µl reaction. Fragmentation reaction was stopped by adding 90µl nuclease-free water and quickly adding 350µl RLT buffer from RNeasy Mini Kit (Qiagen, Germany) followed by purification of fragmented RNA using RNA Cleanup Protocol from this kit

however to enhance the recovery of smaller fragments, we added 500µl ethanol instead of recommended 250µl. 20ng RNaseIII fragmented RNA was used as input for ligation of 3' and 5' adapters according to Small RNA Library Prep Protocol followed by cDNA synthesis from adapter ligated RNA and 10 cycles of PCR amplification. However instead of performing a size-selection of agarose gel (as recommended by manufacturer for sequencing of small RNAs e.g., miRNAs), we used 1 volume of Ampure XP Beads (Beckman Coulter, USA) to clean up the amplified library and remove adapter dimers according to manufacturer's instructions. Purified libraries were then analyzed with HS DNA Assay Kit on Bioanalyzer (Agilent Technologoes, USA) and sequenced on Illumina HiSeq2000 platform.

PolyA RNA-seq:

Polyadenylated RNA was isolated from 5µg total RNA sample by two sequential purifications using Dynabeads mRNA Purification Kit (ThermoFisher Scientific, USA) according to manufacturer's instruction. Purified Poly(A) RNA was analyzed on Bioanalyzer using an RNA Pico Assay chip. Sequencing libraries were then prepared using Small RNA Library Prep Kit (Illumina, USA) using the modified protocol as described above for total RNA-seq.

chrRNA-seq:

Chromatin associated RNA was isolated from $20x10^6$ cells essentially as described previously by (Nojima et al, 2015) followed by rigorous treatment with TurboDNase. Before library preparation, any contaminating rRNA was removed with Ribo-Zero rRNA Removal Kit and libraries were prepared using Small RNA Library Prep Kit as described above for total RNA-seq.

MNase-Seq:

Nucleosomal DNA was obtained by digesting the chromatin with micrococcal nuclease (MNase). For this purpose, 5 x10^6 Raji cells were resuspended in 50µl Buffer I (150mM sucrose, 80mM KCl, 5mM K2HPO4, 5mM MgCl2, 0.5mM CaCl2, 35mM HEPES pH 7.4) and then permeabilized by adding NP40 to a final concentration of 0.2% while incubating at 37°C for one minute. Then 500µl of Buffer II (150mM sucrose, 50mM Tris pH 8, 50mM NaCl, 2mM CaCl2) was added along with 25 units of MNase enzyme (Roche Diagnostics, USA) and incubated at 37°C for 10 minutes. Reaction was stopped by adding EDTA to a final concentration of 10mM quickly followed by addition of 1.45 mL of SDS Lysis Buffer (1% SDS, 10mM EDTA pH 8, 50mM Tris pH 8). After 10 minutes of incubation at 4°C, 200µl aliquot was processed for extraction of DNA after treatment with RNase A and Proteinase K followed by an extraction with phenol:chloroform:isoamylalcohol (25:24:1). Only those nucleosomal DNA preps were used subsequently where DNA fragments corresponding to mononucleosomal fraction (~147bp) formed at least 70% of all DNA fragments. Sequencing libraries were then prepared with Illumina ChIP Sample Library Prep Kit (Illumina, USA) as described above for ChIP-seq libraries.

II- Bioinformatic Procedures

ChIP-seq Data Processing:

For ChIP-seq, raw sequencing reads were aligned to human genome (hg19) using Bowtie2 (Langmead and Salzberg, 2012). Sequence reads that aligned multiple times in genome with equal alignment score, were discarded as well as the duplicate reads with identical coordinates (sequencing depth taken into account) were

discarded to remove potential sequencing and alignment artefacts. Aligned reads were elongated *in silico* using the DNA fragment size inferred from paired-reads or an estimated optimal fragment size for orphan reads using an in-house developed R pipeline named PASHA (Fenouil et al., 2016). These elongated reads were then used to calculate the number of fragments that overlapped at a given nucleotide thus representing an enrichment score for each nucleotide in the genome. Wiggle files representing average enrichment score every 50bp were generated. Sequencing data from Input samples were treated in the same way to generate Input wiggle files. All wiggle files were then rescaled to normalize the enrichment scores to reads per million. Enrichment scores from Input sample wiggle files were then subtracted from ChIP sample wiggle files. This allowed us to remove/reduce the over-representation of certain genomic regions due to biased sonication and DNA sequencing. Besides this, input subtraction also improves the signal/noise ratio especially for ChIPs with low enrichment. Rescaled and Input subtracted wiggle files from biological replicate experiments were then used to generate a wiggle file that represents the average signal from several biological replicates.

RNA-seq Data Processing:

Raw sequencing reads were aligned to human genome (hg19) using TopHat2 (Kim et al., 2013). Sequence reads that aligned multiple times in genome with equal alignment score, were discarded. Thanks to strand-specific library prep of RNA samples, we could infer the strand from which the RNA was originally transcribed hence we separated the reads that align to Watson and Crick strands and processed them separately using PASHA (Fenouil et al., 2016) pipeline to generate strand-specific wiggle files. All wiggle files were then rescaled to normalize the enrichment

scores to reads per million. Rescaled wiggle files from biological replicate experiments were then used to generate a wiggle file that represents the average strand-specific RNA signal from several biological replicates.

Gene Expression Analysis:

Differential Gene Expression (DGE) analysis was performed by using the DESeq package (Anders and Huber, 2010) from Bioconductor. First, HTseq-count program from the HTSeq framework (Anders et al., 2015) was used to count the sequence reads mapping to gene annotations and then these counts were processed using the DESeq package to identify genes that are at least 3 fold (log2) differentially expressed relative to the reference sample.

Identification of genes down-regulated due to interference of antisense transcription:

We identified all the genes that were down regulated (log2 fold change > 1) in sense transcription as well as the genes that showed up regulation (log2 fold change > 1) in YFFF mutant as compared to rWT with FDR 0.05 and pval < 0.05. Intersection of the two lists of genes gave us the genes that are potentially down regulated due to interference from antisense transcription.

Peak calling:

We used wiggle files to detect the genomic regions with enrichment signals beyond background signal. For this purpose, we used *Thresholding* function of the Integrated Genome Browser (IGB) to determine the enrichment score above which we considered a genomic region to be enriched relative to background noise (*Threshold*)

as well as minimum number of consecutive enriched bins to be considered an enriched region (*Min.Run*) and finally the minimum gap beyond which two enriched regions were considered to be distinct (*Max.Gap*) (see Table S7 for parameters used). These parameters were then fed to an in-house script that performs peak-calling by using algorithm employed by *Thresholding* function of IGB.

Identification of Active Enhancers:

Genomic regions that show simultaneous enrichment with H3K4me1, H3K27ac and Pol II and are at least +/- 1500bp away from any annotated gene were considered to be putative active enhancers. To remove any unannotated promoters from our enhancer selection, we filtered out any regions that were more enriched with H3K4me3 as compared to H3K4me1. Method used for this filter has been described previously in Descostes et al, 2014. In identified enhancers, position of the minimum signal of H3K27ac (nucleosome depleted region – NDR) which was closest to location of maximum signal of Pol II was defined as center of the region.

Average Metagene Profiles:

To generate average signal profiles, we selected the hg19 genes or identified enhancer regions that do not have any other annotation within 20Kb (Figure 1-3), 10kb (Figure 4), 2kb (Figure 5) around boundaries. Removal of the annotations too close to each other is necessary to avoid mixing signals from close-by annotations which can cause misinterpretation of the results. ChIP-seq, MNase-seq and strand-specific RNA-seq values from wiggle files were retrieved with in-house R and Perl scripts for selected genes and enhancer regions. Then we used an algorithm as described previously (Koch et al., 2011) to rescale the genes to same length by

interpolating the values on 1000 points and build a matrix on which each column is averaged and resulting values are used to plot average metagene profiles.

RNA read-through index:

Upstream and downstream read-through transcription indices (Figure 2D, 3B, 3H, S4D) were calculated by dividing average sense (for downstream RT) and antisense (for upstream RT) signal in 20kb region upstream or downstream of the gene with average signal in first half of the corresponding gene body. Asinh transformation was applied to the values for graphical representation.

Pol II read through index:

Upstream and Downstream Pol II read-through indices (Figure 3E, S4B, S4C) were calculated by dividing average signal in 10kb region upstream or downstream of the gene respectively with average signal in second half of the corresponding gene body. Asinh transformation was applied to the values for graphical representation.

Pol II pausing score:

Pol II pausing score (Figure 4D) was calculated as described earlier (Fenouil et al., 2012a). Briefly, the average Pol II ChIP-seq signal in -300bp / +100bp region around TSS was divided by average signal in second half of the corresponding gene body. Asinh transformation was applied to the values for graphical representation.

QUANTIFICATION AND STATISTICAL ANALYSIS

All ChIP-seq, RNA-seq and MNase-seq experiments were performed in at least two biological replicates. Statistical significance of differential metagene profiles was

calculated by two sided Wilcoxon test. p-values associated to the number of asterisks in figures are described in figure legends. Significance of differential gene expressions were calculated by non-parametric Mann-Whitney test and p-values were adjusted for FDR < 0.05. Genes with at least 3-fold change in expression level relative to rWT were considered to differentially regulated.

DATA AND SOFTWARE AVAILABILITY

All high throughput sequencing data used in this study have been deposited at GEO under accession number GSE94330.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [1] partner repository with the dataset identifier PXD008270.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-H3K4me1	Abcam	ab8895
Rabbit polyclonal anti-H3K4me3	Abcam	ab8580
Rabbit polyclonal anti-H3K27ac	Abcam	ab4729
Rabbit polyclonal anti-HA	Abcam	ab9110
Rat monoclonal anti-HA	Roche	3F10
Rabbit polyclonal anti-MED15	Proteintech	11566-1-AP
Rabbit polyclonal anti-Med1	Bethyly Labs	A301-793A
Rabbit polyclonal anti-INT11	Bethyl Labs	A301-274A

Mouse Monoclonal anti-RPB1	Elisabeth Kremmer, Helmholtz Zentrum, Munich	Pol3.3
GAPDH	Elisabeth Kremmer, Helmholtz Zentrum, Munich	5C4
Rat monoclonal anti-Ser2P	Helmholtz Zentrum Munich	3E10
Rat monoclonal anti-Ser5P	Helmholtz Zentrum Munich	3E8
Rat monoclonal anti-Ser7P	Helmholtz Zentrum Munich	4E12
Rat monoclonal anti-Thr4P	Helmholtz Zentrum Munich	6D7
Chemicals, Peptides, and Recombinant Proteins		
Micrococcal nuclease	Roche	10107921001
RNaseIII	Thermo Fisher	AM2290
Turbo DNA-Free	Thermo Fisher	AM1907
Alpha-amanitin	Sigma Aldrich	A2263
G-418 Solution	Sigma Aldrich	000000472787800 1
Deposited Data		
ChIP-seq, RNA-seq and MNase-seq data	This study	GSE94330
Experimental Models: Cell Lines		
Raji cells	ATCC	CCL-86

Recombinant DNA		
Mouse Rpb1 gene cloned into LS*mock vector	Meininghaus et al.,2000	rWT
Mouse Rpb1 gene cloned into LS*mock vector	Meininghaus et al.,2000	YFFY
Mouse Rpb1 gene cloned into LS*mock vector	Meininghaus et al.,2000	YYFF
Mouse Rpb1 gene cloned into LS*mock vector	Meininghaus et al.,2000	FYYF
Mouse Rpb1 gene cloned into LS*mock vector	Meininghaus et al.,2000	YFFF
Mouse Rpb1 gene cloned into LS*mock vector	Meininghaus et al.,2000	S2AAA
Oligonucleotides		
SNHG3-prom-F : GTGGTCGCTTCTTCTCCTTG	This study	
SNHG3-prom-R : TAGGGAAGCTCGGCTACTGA	This study	
ETS1-Enh-UPS-1-F : GGCTGTTCGTCTCCCAAGTA	This study	
ETS1-Enh-UPS-1-R : CACTGCAGGTGGTAATTTGC	This study	
Myc-prom-F : AGGGATCGCGCTGAGTATAA	This study	
Myc-prom-R : TGCCTCTCGCTGGAATTACT	This study	
TAF12-prom-F : ACCTGGTCCTTCGAACACTG	This study	
TAF12-prom-R : GGCAGTTGAGGAACAAGAGC	This study	
Rnu11-prom-F : ACCCTGCTTTGGTGACAGAG	This study	
Rnu11-prom-R : ATCACCAGCTGCCCAAATAC	This study	
Kxd1-prom-F : CAAAAGTGGAGCAGGGATGT	This study	
Kxd1-prom-R : CCCCAAGGTCGTAAATGCTA	This study	
Software and Algorithms		
PASHA	Fenouil et al, 2016	https://cran.r-project.org/web/packages/Pasha/

Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
TopHat2	Kim et al, 2013	http://ccb.jhu.edu/software

Excel Table Legend:

Table S8: List of genes with downstream and upstream read-through transcription. Related to Figure S2F.

This table describes the values of the ratio of the genes shown in Fig. S2F ranked from higher to lower ratio downstream of 3'ends. Corresponding ratios for AS signal uptream of 5' ends is also indicated. ND ratio values could not be determined due to 0 values in wt cells. NA could not be determined in upstream ratio due to overlapping genes within 5kb upstream.

Figure 1

Shah, Maqbool et al, Figure 1

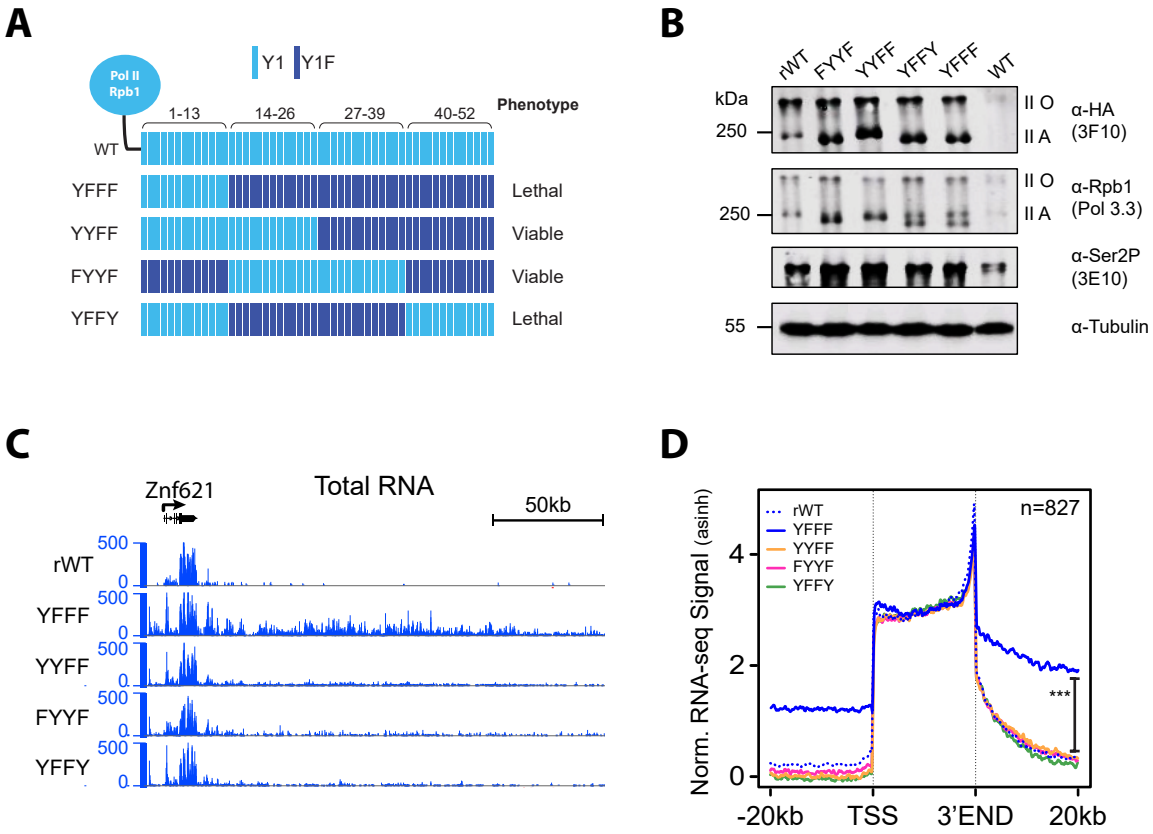


Figure 2

Shah, Maqbool et al, Figure 2

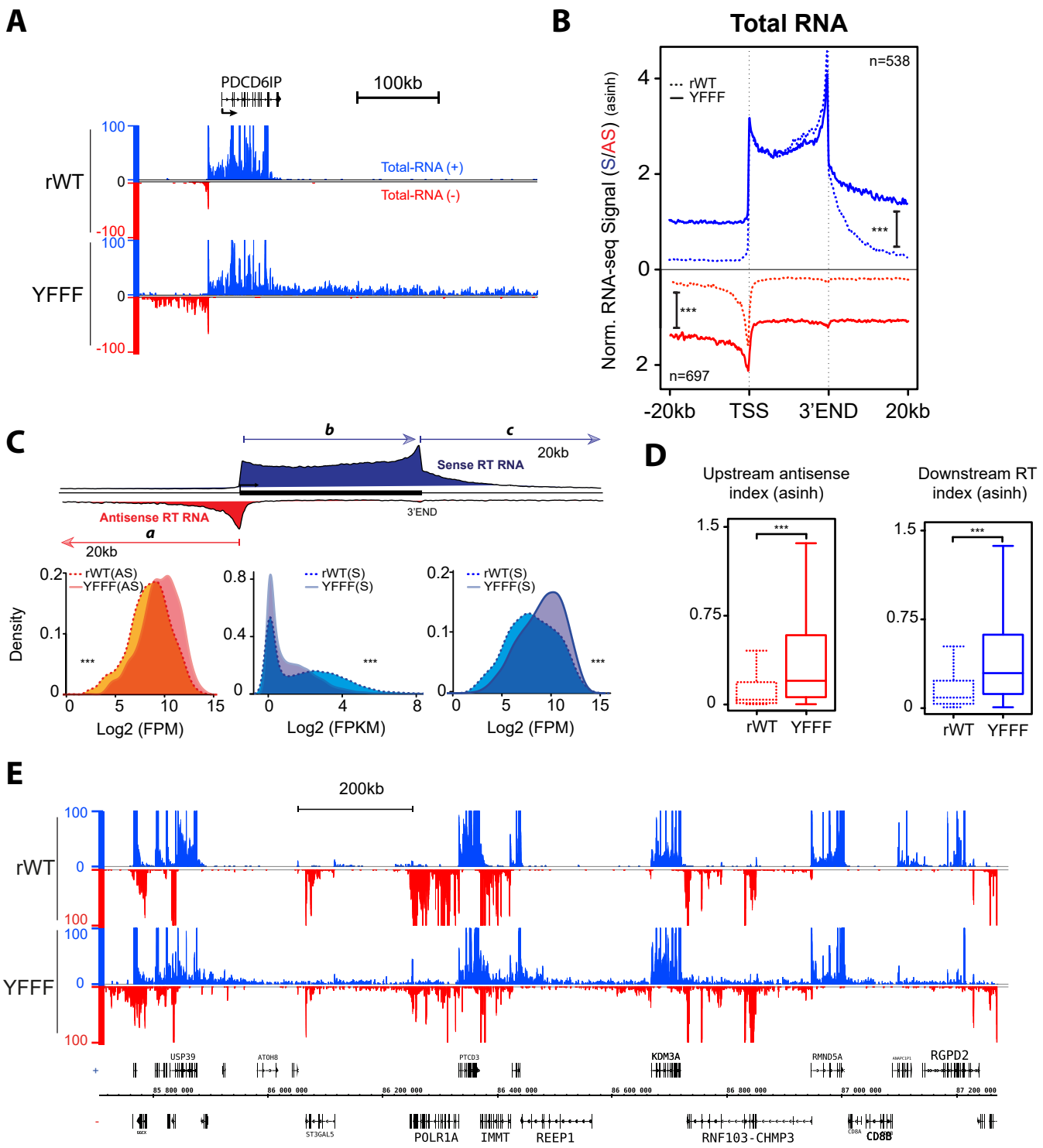


Figure 3

Shah, Maqbool et al, Figure 3

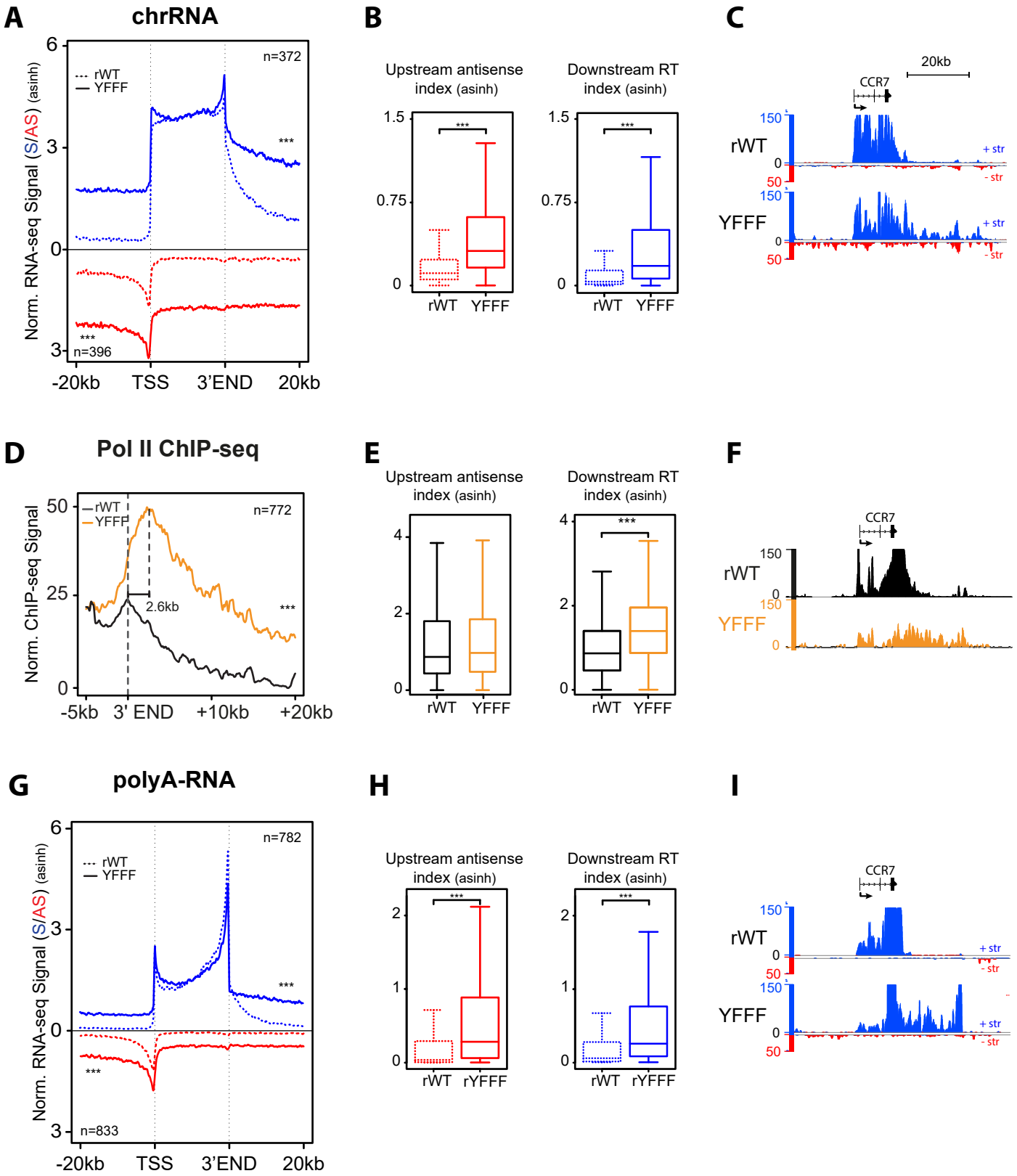


Figure 4

Shah, Maqbool et al, Figure 4

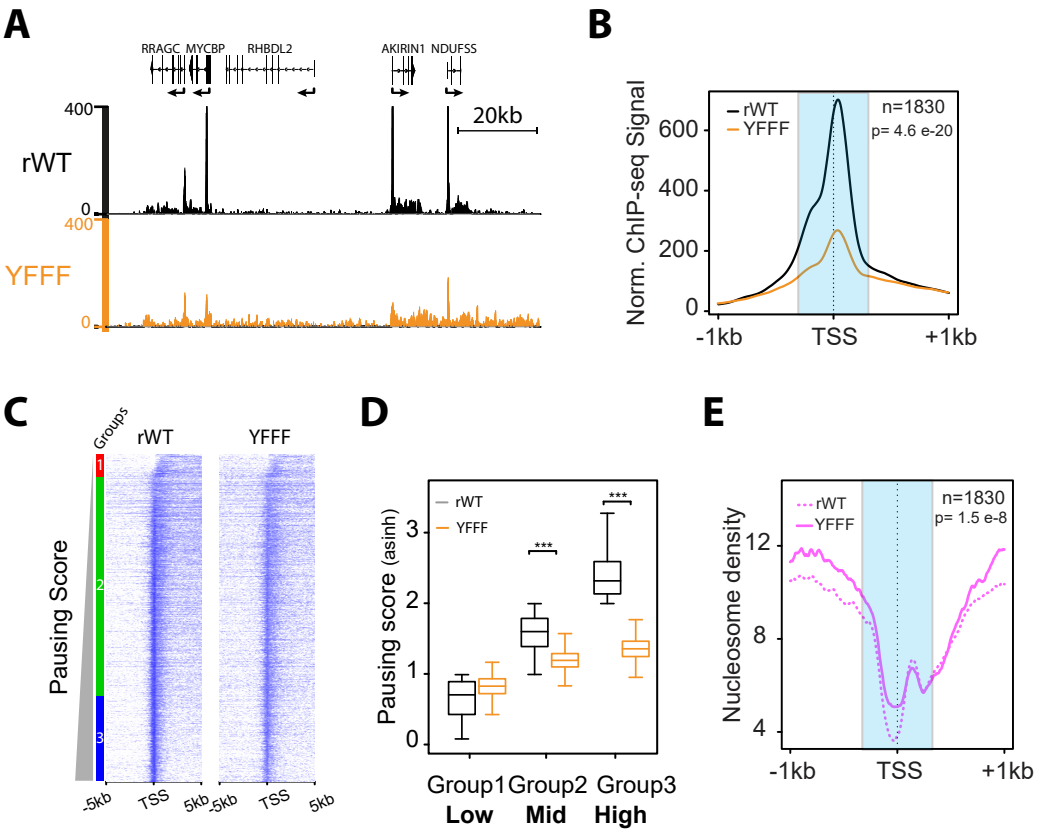


Figure 5

Shah, Maqbool et al, Figure 5

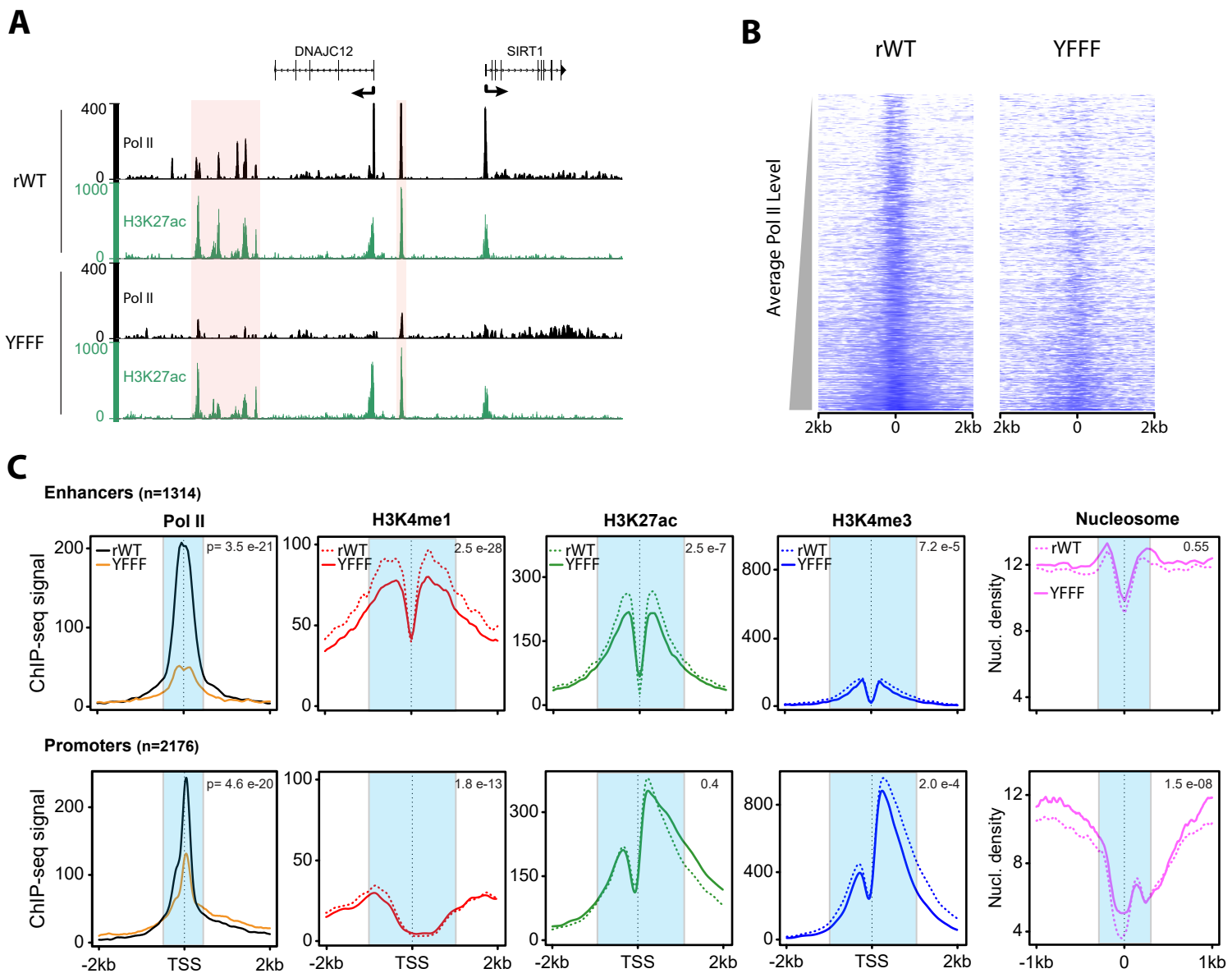
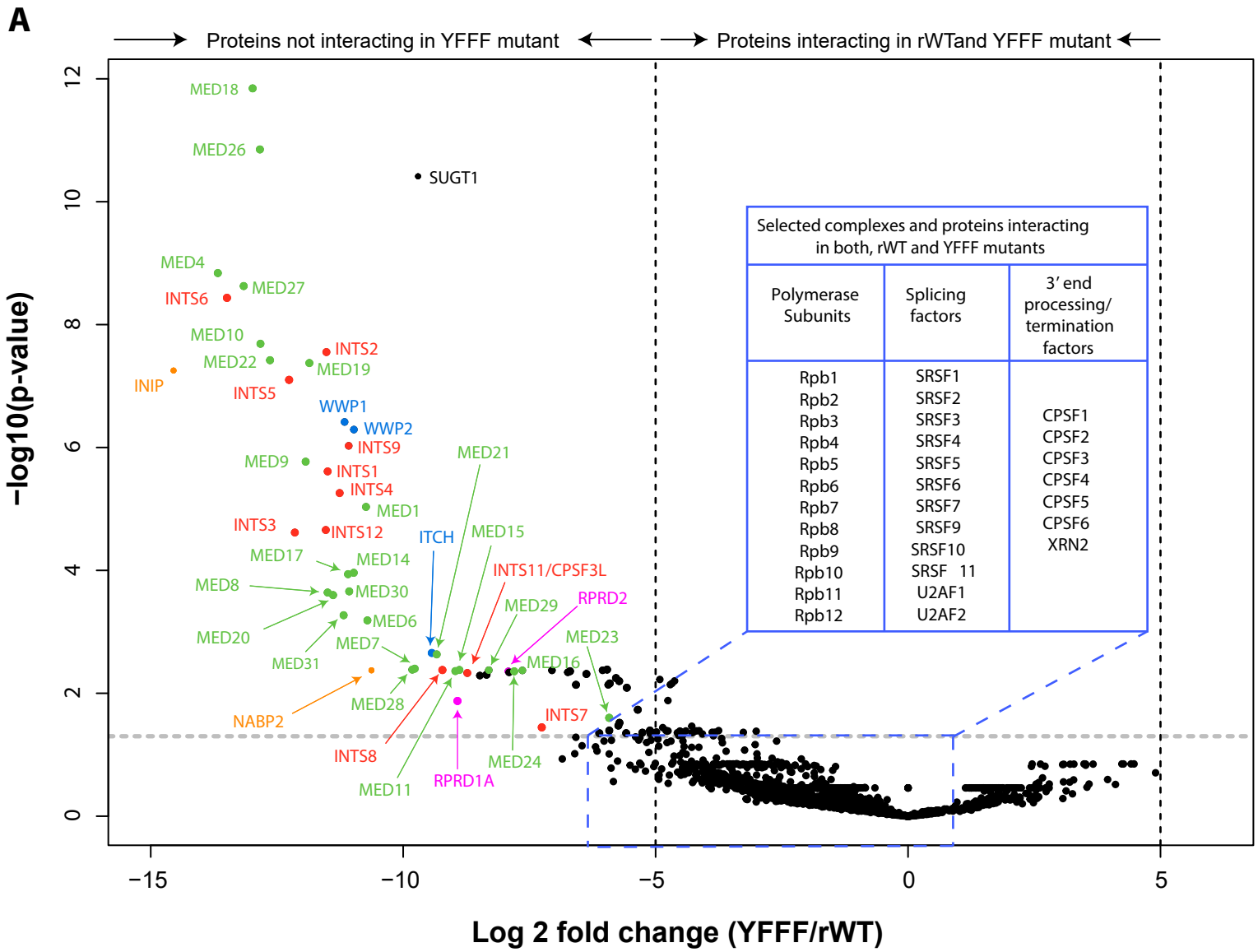


Figure 6

Shah, Maqbool et al, Figure 6



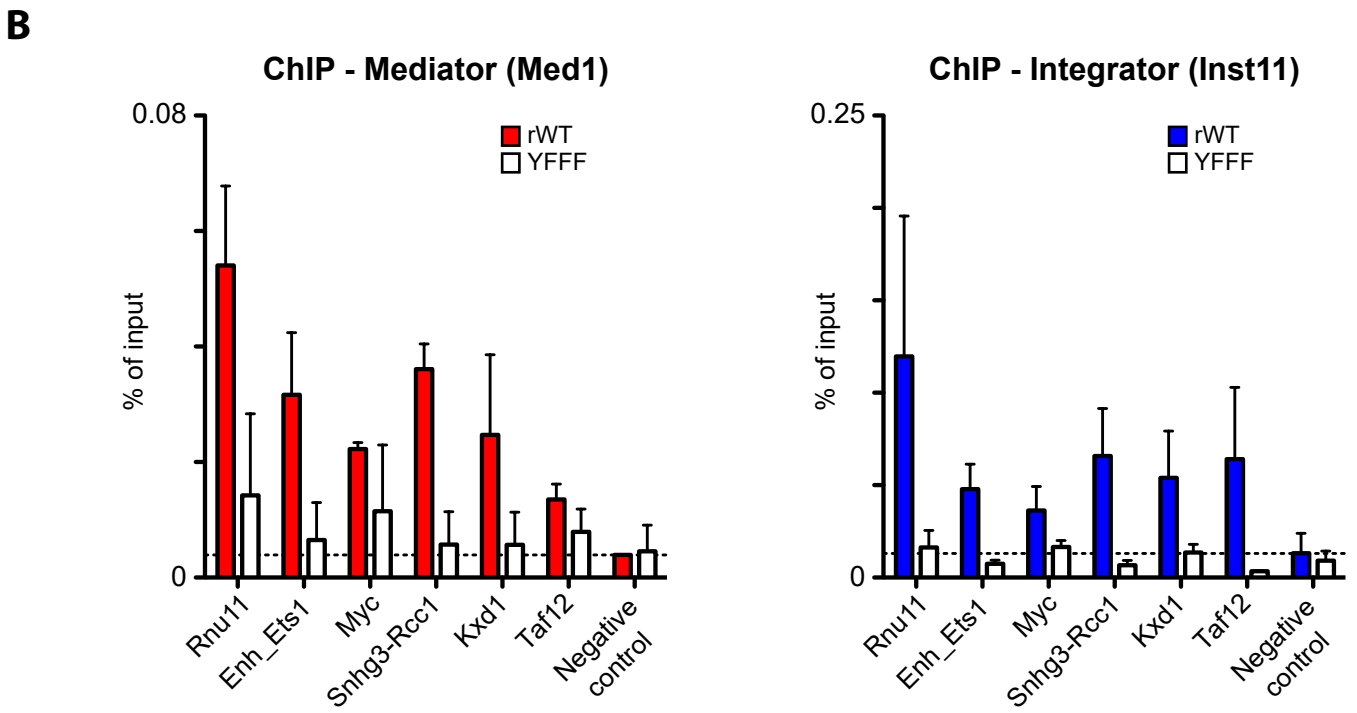
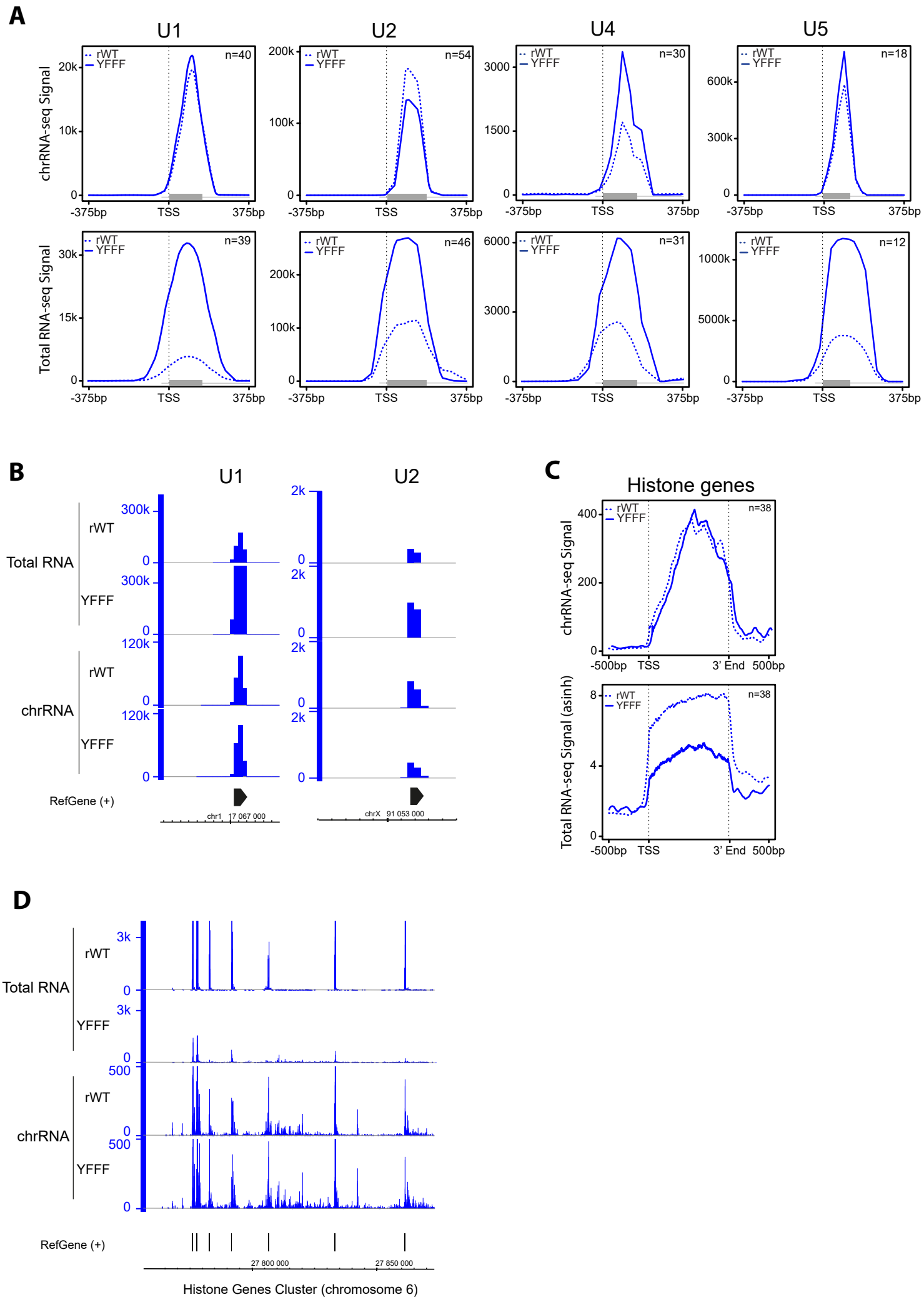


Figure 7

Shah, Maqbool et al, Figure 7



Supplemental Material

Tyrosine-1 of RNA Polymerase II CTD controls global termination of gene transcription in mammals

Nilay Shah[1#], Muhammad Ahmad Maqbool[2#], Yousra Yahia[2], Amal Zine El Aabidine[2], Cyril Esnault[2], Ignasi Forné[3], Tim-Michael Decker[1], David Martin[2], Roland Schüller[1], Stefan Krebs[4], Helmut Blum[4], Axel Imhof[3], Dirk Eick[1*] and Jean-Christophe Andrau[2,5*]

Affiliations

[1]Department of Molecular Epigenetics, Helmholtz Center Munich, Center of Integrated Protein Science Munich, Marchioninistrasse 25, 81377 Munich, Germany;

[2]Institut de Génétique Moléculaire de Montpellier (IGMM), CNRS-UMR5535, Montpellier, France;

[3]Biomedical Center Munich, ZFP, Großhaderner Strasse 9, 82152 Planegg-Martinsried, Germany;

[4]Laboratory for functional genome analysis, Gene Center, Ludwig-Maximilians-Universität, Munich, Germany.

# These authors contributed equally

[5]Lead Contact

* Corresponding authors: jean-christophe.andrau@igmm.cnrs.fr and eick@helmholtz-muenchen.de

Shah, Maqbool et al, Figure S1

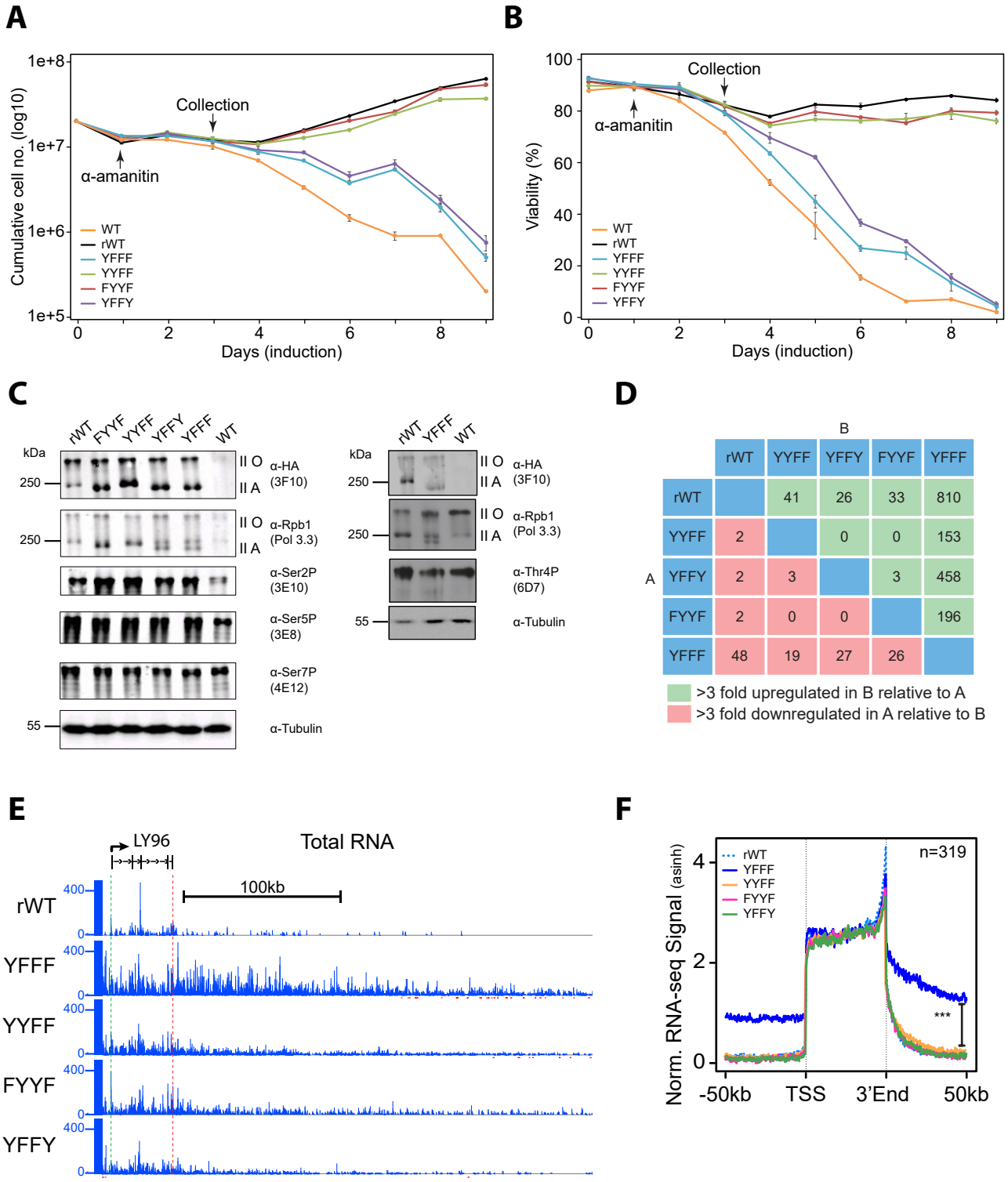


Figure S1: Phenotype and differential transcriptome of CTD Tyrosine mutants, Related to Figure 1

A, B) Proliferation kinetics and viability curve following induction of rWT and tyrosine mutants by removal of tetracycline (tet-off system) and treatment with α-amanitin. The time of sample collection for RNA-seq and ChIP-seq experiments presented further is indicated by an arrow. **C)** Western blots probing for Ser2P, Ser5P and Ser7P Abs in all mutants (left panel) and Thr4P in rWT, YFFF and WT cells (right panel), indicate no major alteration on the phosphorylation pattern of the CTD. **D)** Differential gene expression analysis of genes up and down regulated in the tyrosine mutants relative to rWT (3-fold change, FDR<0.05). **E)** Example of read-through phenotype at 3' end of the LY96 gene. **F)** Average metagene profile of total sense RNA-seq signal (asinh) over the gene bodies and 50kb upstream and downstream regions. The 3 stars indicate a p-value < $2x10^{-16}$ (2 sided Wilcoxon test) between rWT and YFFF.

Shah, Maqbool et al, Figure S2

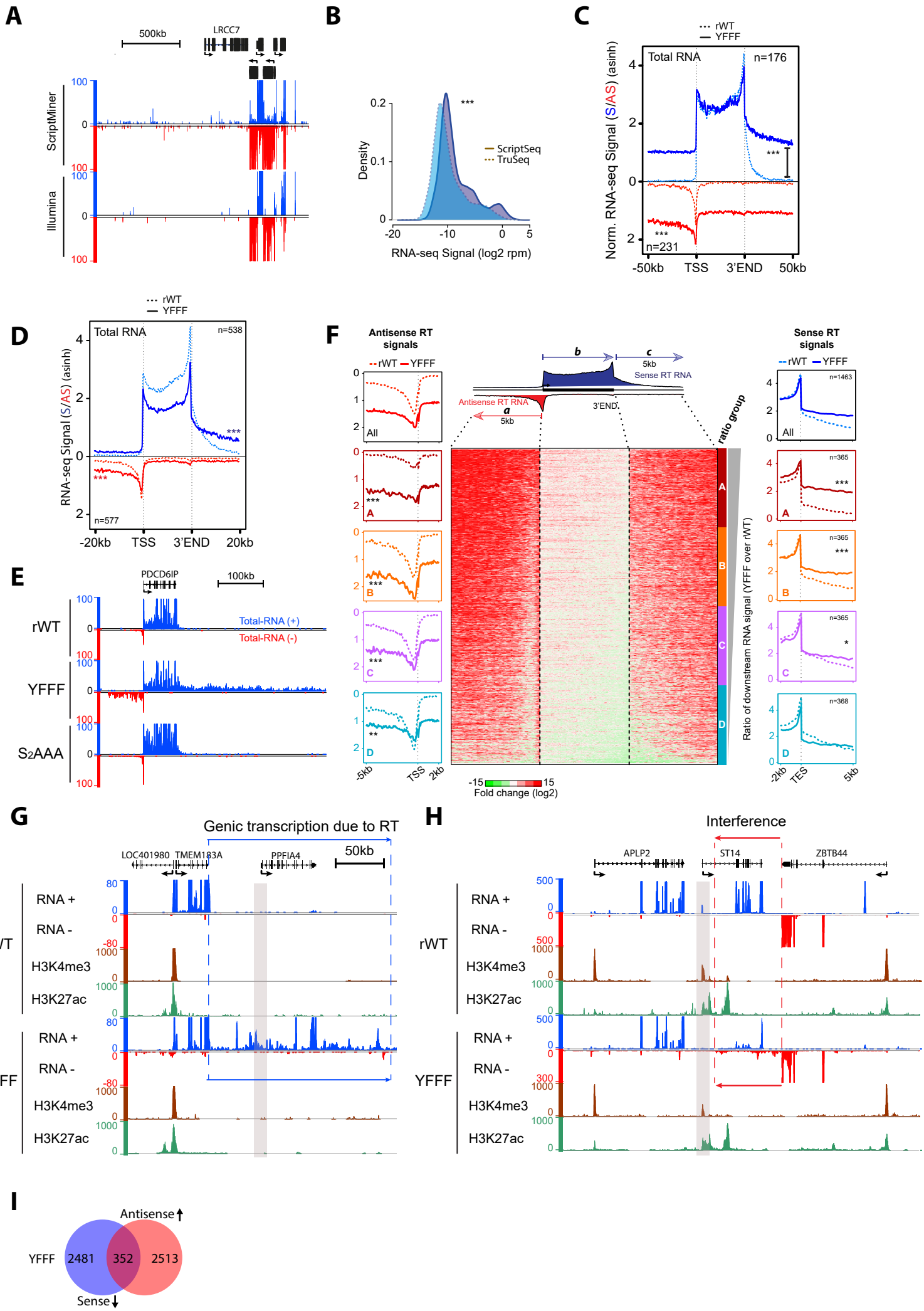


Figure S2: YFFF RT phenotype at 3' sense and 5' antisense transcription is specific and likely causes transcriptional interference, Related to Figure 2

A) Comparison of RNA-seq in WT Raji cells performed by using ScriptMiner RNA Library Prep Kit vs Illumina TruSeq small RNA Library Prep Kit. As compared to ScriptMiner Kit, TruSeq kit reduces the background noise in intergenic regions as well as increases strand-specificity. **B)** Distribution of intergenic RNA-seq signals obtained from TruSeq and ScriptMiner libraries in a selection of 21792 intergenic regions, excluding genes within 100kb for an assessment of the intrinsic experimental noise. The 3 stars indicate a p-value < $2x10^{-16}$ (2 sided Wilcoxon test). **C)** Average metagene profile of total RNA-seq signal (asinh) in sense (blue) and AS (red) orientation of the gene bodies and 50kb upstream and downstream regions. The 3 stars indicate a p-value < $2x10^{-16}$ (2 sided Wilcoxon test). **D)** Average metagene profile of total RNA-seq signal (asinh) without normalization on gene bodies, in sense (blue) and AS (red) orientation of the gene bodies and 20kb upstream and downstream regions. **E)** Tyrosine mutations of the YFFF induce a specific 5' AS and 3' sense RT phenotype as exemplified at the PDCD6IP locus and as compared to the S2AAA control mutant (lower panel). The S2AAA mutant has Ser2 positions of the last 3/4 of the CTD heptads replaced by Ala (lower panel). The PDCD6IP gene is representative for the RT phenotype observed genome-wide. **F)** Comparison of the 3' sense and the 5' AS RT phenotypes in the YFFF mutant using total RNA-seq. Genes were ranked according to 3' RT decreasing ratio in the YFFF vs rWT within 5kb after the annotated 3' ends and further divided in 4 equal sized groups A-D (colored profiles on the right). The corresponding ratio profiles in 5' AS RNAs are shown on the left of the density heat maps and in the middle for the gene bodies rescaled (0-100%). The upper plots represent the global average profiles whereas the 4 below, represent the groups A to D (more to less affected in 3' RT from top to bottom). The p-values of the YFFF vs rWT comparison are for (1) 3' S: A < 2.2e-16; B< 2.2e-16; C = 9.715e-11; D= 0.037 and (2) for 5' AS: A < 2.2e-16; B< 2.2e-16; C< 2.2e-16; D= 2.35e-15. **G)** Augmentation of adjacent transcript densities, PPFIA4 as example. The absence of the epigenetic marks H3K4me3 and H3K27ac at the PPFIA4 promoter pleads for a RT effect of TMEM183A rather than a neo-initiation event. **H)** Example of apparent RNA interference resulting in transcription inhibition (ST14 example). **I)** Venn Diagram showing overlap of protein coding genes down regulated with increased AS signal in YFFF mutant. These selections were isolated using DESeq package with log2 FC >1, FDR 0.05, oval 0.05.

Shah, Maqbool et al, Figure S3

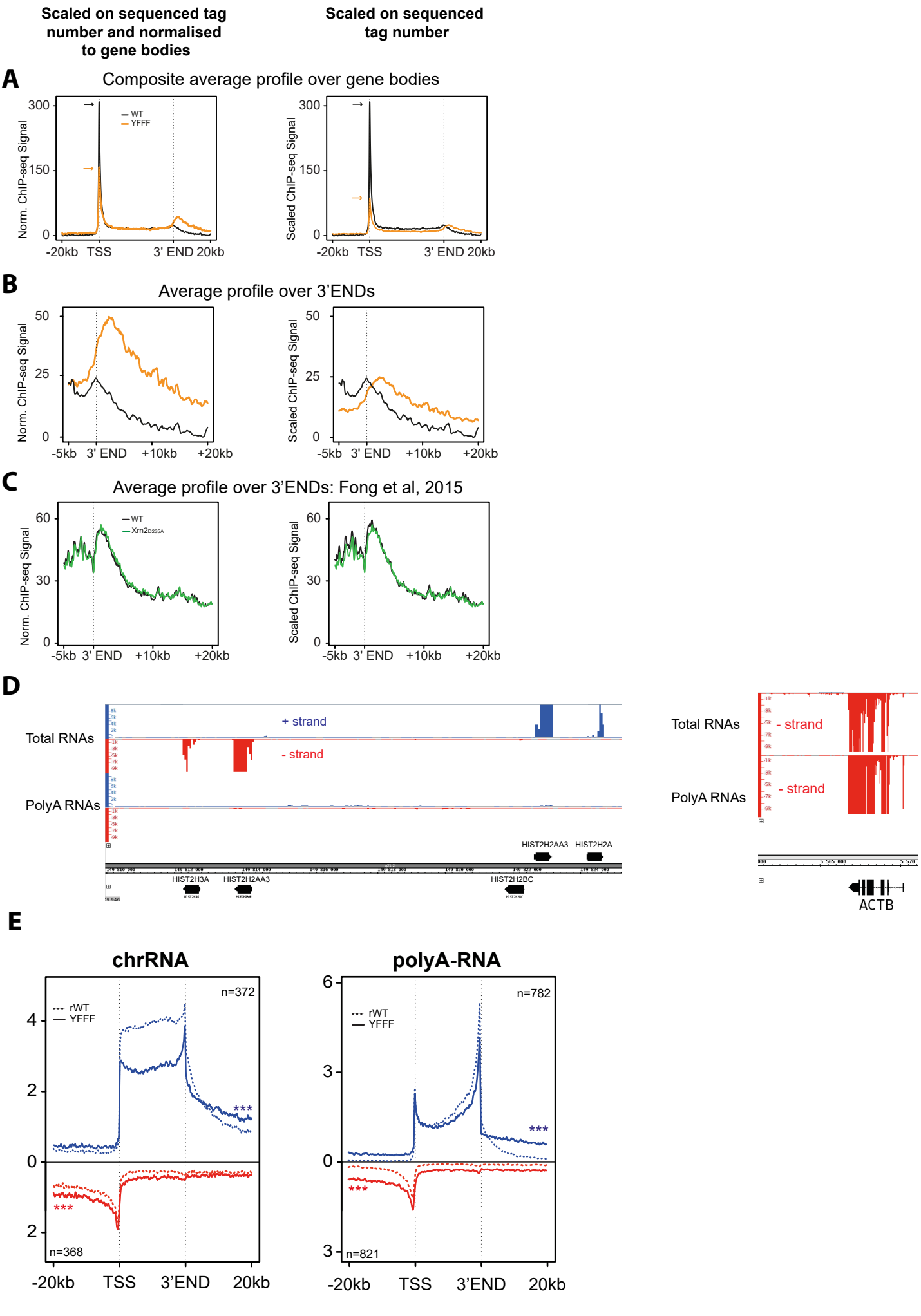


Figure S3: Pol II ChIP-seq and Poly(A) vs total RNA-seq analyses of YFFF mutant phenotype, Related to Figure 3

A) Composite Pol II ChIP-seq average profile (top 30% coding genes) over gene bodies normalized as described in the methods section for rWT and YFFF mutant (left) or just based on the sequence tag counts (right). The orange and black arrows represent the positions of the max peak at TSSs for WT and YFFF Pol II. **B)** Pol II average profiles for rWT and YFFF around 3' end of genes, normalized/scaled as in A). **C)** Pol II average profiles for WT and Xrn2 dominant mutation (Fong et al, 2015) around 3' end of genes, normalized/scaled as in A) in a selection of the top30% of coding genes. **D)** RNA-seq signals in rWT cells over 4 genes (non-polyadenylated histones) of the histone cluster located on chromosome 1. Around 100 times less signal is observed in poly(A) RNA-seq as compared to total RNA-seq suggesting that the protocols used allow discrimination of both populations. The highly transcribed ACTB poly(A) coding gene is shown as control, enriched in both poly(A) and total RNA-seq experiments. E) Average metagene profiles of chr- and PolyA- RNA-seq signal (asinh) without normalization on gene bodies, in sense (blue) and AS (red) orientation of the gene bodies and 20kb upstream and downstream regions.

Shah, Maqbool et al, Figure S4

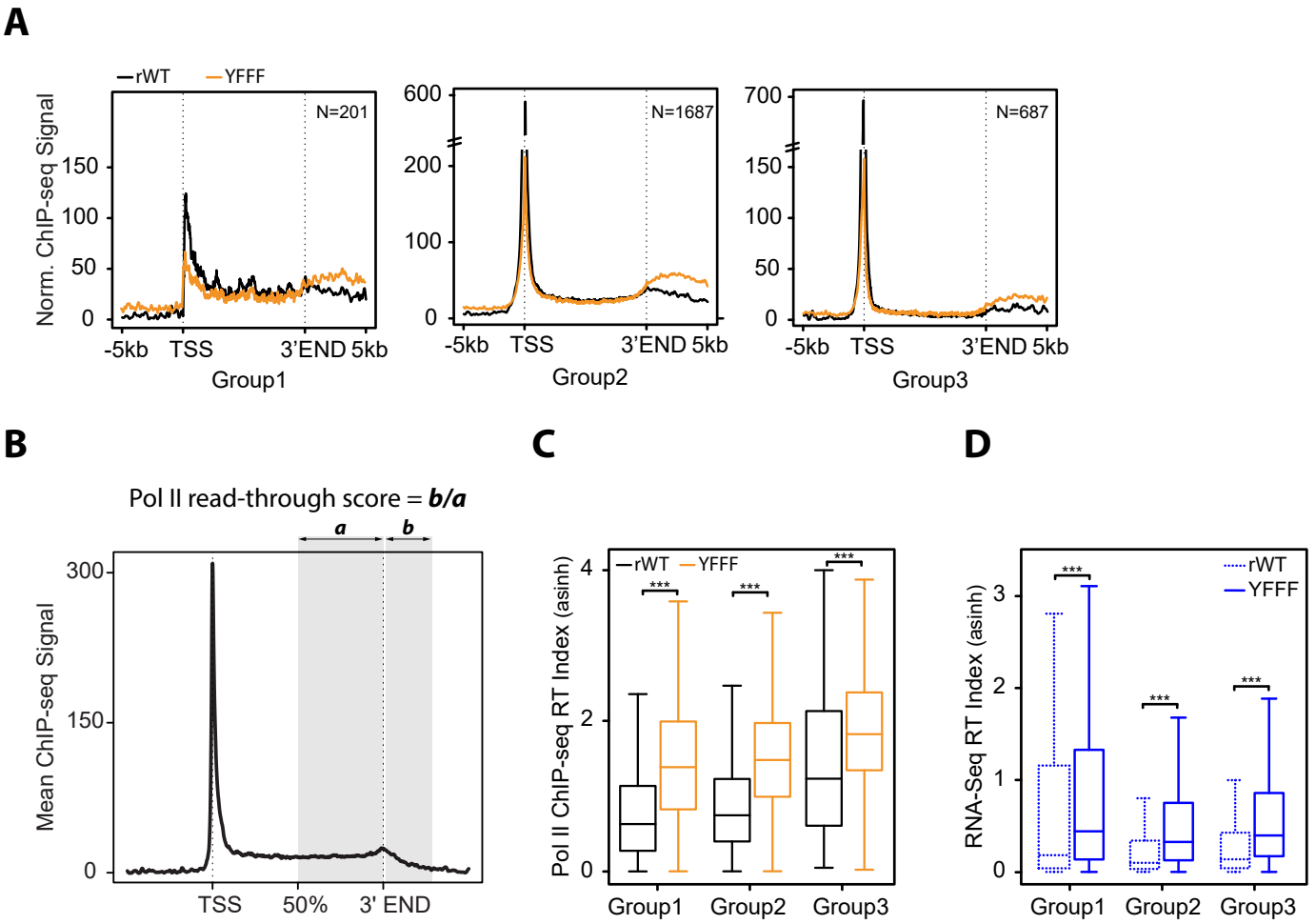


Figure S4: Pol II Pausing and 3' end RT of RNA and Pol II, Related to Figure 4
A) Pol II average profile on the 3 groups of pausing scores defined in Figure 4. **B)** Pol II RT score shown in C) was calculated by dividing the average Pol II signal downstream of 3' end (10kb) with average signal in the second half of gene body (50-100%). **C)** RT indices Box plot in rWT and YFFF calculated with Pol II signal from three groups of genes. **D)** Box plot of read-through score calculated with total RNA signal from three groups of genes.

Shah, Maqbool et al, Figure S5

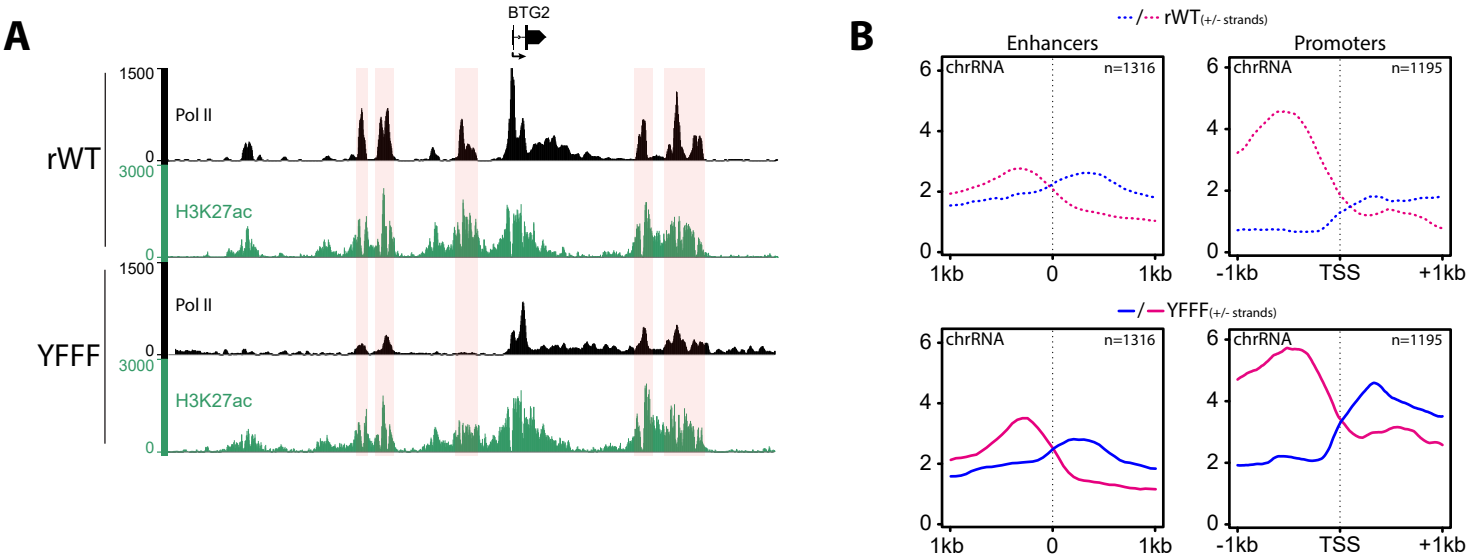
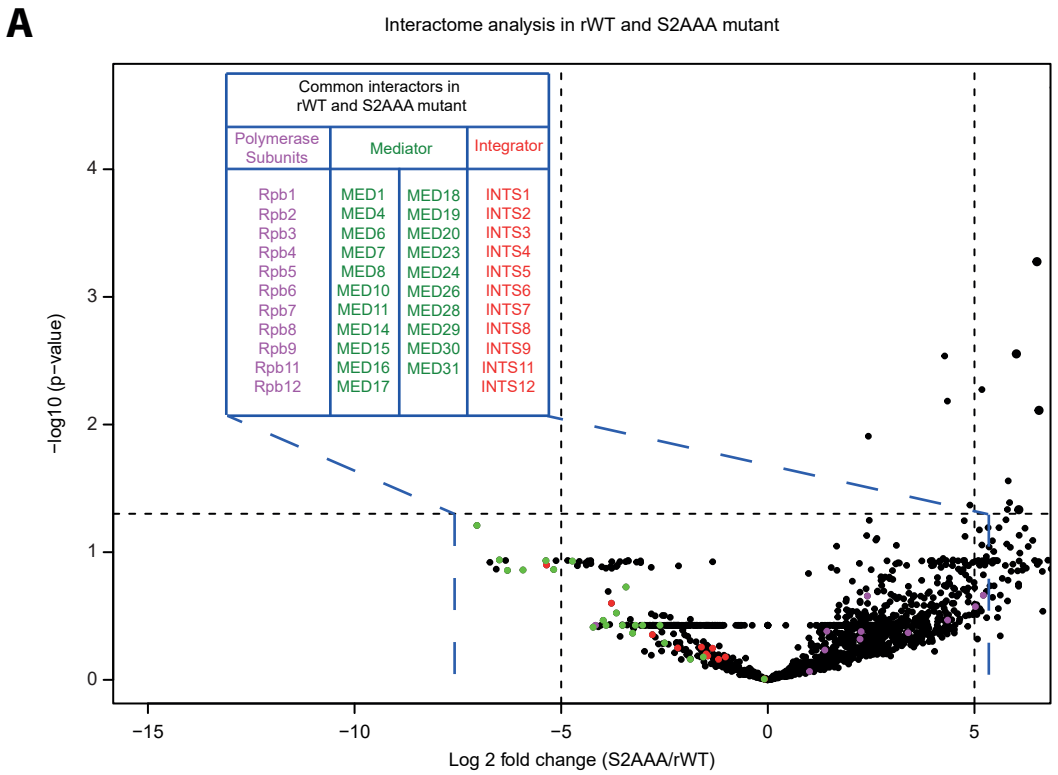
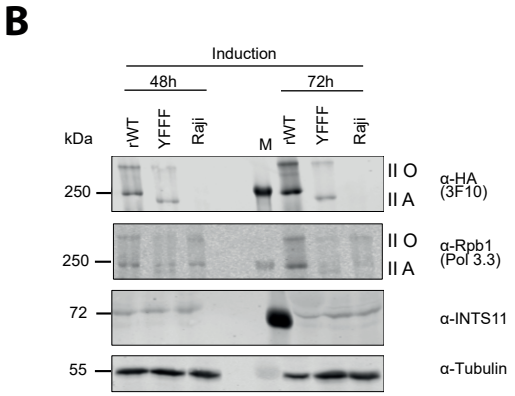
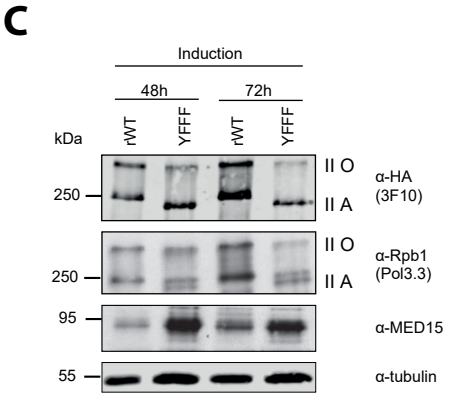


Figure S5: YFFF mutations do not impair transcription at enhancers, Related to Figure 5
A) Additional example of impaired Pol II loading at enhancers around the BTG2 gene indicated by light pink rectangles. **B)** ChrRNA-seq average profiles at enhancers reveal that nascent transcription is not affected when compared to promoters (right).

Shah, Maqbool et al, Figure S6







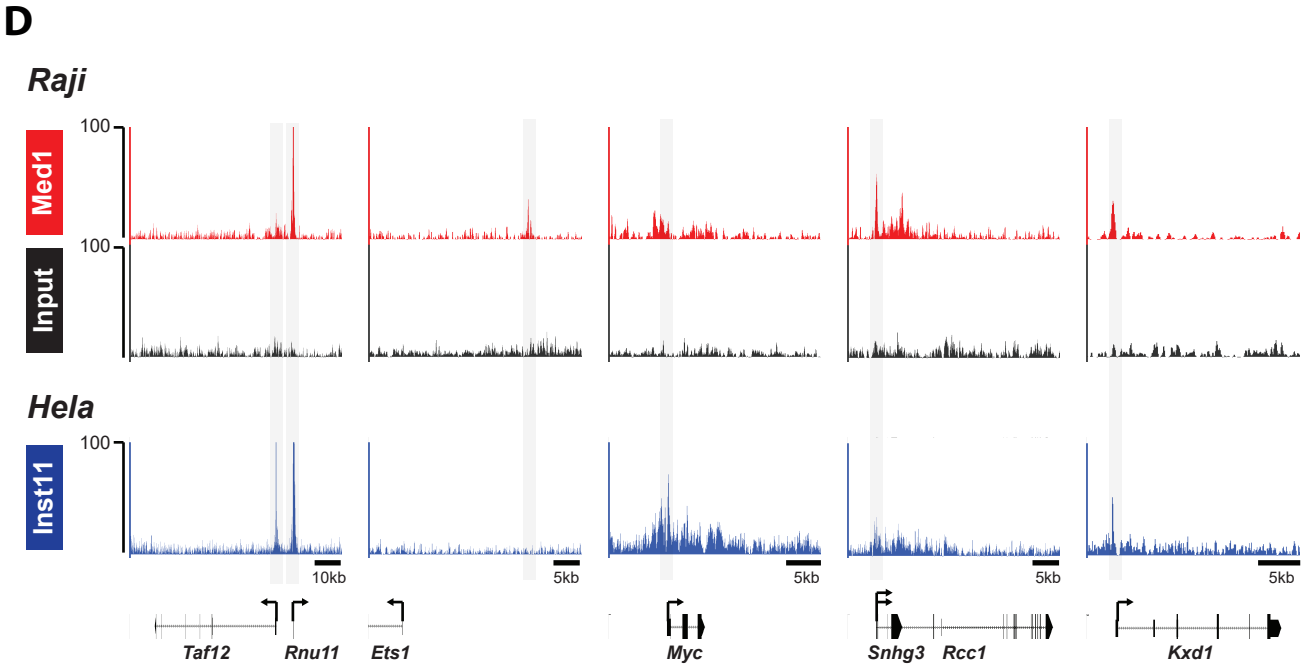


Figure S6: Control experiments for MS specificity and Mediator/Integrator complexes integrity in the YFFF mutant, Related to Figure 6

A) Volcano plot showing MS differential analysis of Mediator and Integrator interactions in a S2AAA mutant. No major interaction loss with Int subunits is observed. **B)** Western blot of HA-Rpb1, Rpb1, Ints11 in rWT or YFFF cells following 48h and 72h of induction (24h and 48h of α-amanitin treatment). **C)** Western blot of HA-Rpb1, Rpb1, Med15 in rWT or YFFF cells following 48h and 72h of induction (24h and 48h of α-amanitin treatment). **C)** ChIP-seq signals at selected loci for ChIP-qPCR analysis. The Med1 (red) and Input (Black) tracks from Raji cells are shown on top, the Ints11 track from HeLa cells is shown at bottom (Stadelmayer et al., 2014). Grey rectangles highlight areas that were used for qPCR analysis (see Figure 6).

Supplementary Tables

Table S1: Peptide counts of proteins and complexes not interacting with the YFFF mutant. Related to Figure 6

Peptide counts of 69 proteins that do not interact with the YFFF mutant for all five biological replicates. Samples in the experiments 1 and 2 were subjected to on-beads trypsin digest, while samples in the experiments 3, 4 and 5 were subjected to in-gel trypsin digest.

			rWT						YFFF				
	Experiment Number		1	2	3	4	5		1	2	3	4	5*
	Uniprot ID	Gene Name	Peptide counts						Peptide counts				
1	Q9NRY2	INIP	2	5	3	5	8		0	0	0	0	0
2	Q9NPJ6	MED4	6	10	13	13	13		0	0	0	0	0
3	Q9UL03	INTS6	10	27	32	41	45		0	0	0	0	0
4	Q6P2C8	MED27	3	9	11	12	10		0	0	0	0	0
5	Q9BUE0	MED18	3	4	3	3	6		0	0	0	0	0
6	O95402	MED26	5	14	14	18	11		0	0	0	0	0
7	Q9BTT4	MED10	1	5	6	6	7		0	0	0	0	0
8	Q5T8T7	MED22	1	2	6	7	6		0	0	0	0	0
9	Q6P9B9	INTS5	3	15	13	17	26		0	0	0	0	0
10	Q68E01	INTS3	12	31	35	34	50		0	1	1	0	0
11	Q9NWA0	MED9	1	1	6	3	6		0	0	0	0	0
12	A0JLT2	MED19	1	3	3	3	5		0	0	0	0	0
13	Q96CB8	INTS12	1	6	9	15	13		0	0	0	0	0
14	Q9H0H0	INTS2	4	11	17	18	30		0	0	0	0	0
15	Q96G25	MED8	4	8	8	9	11		0	0	1	0	0
16	Q8N201	INTS1	16	43	48	60	74		0	0	1	0	0
17	Q9H944	MED20	1	2	4	7	6		0	0	1	0	0
18	Q96HW7	INTS4	2	12	15	31	41		0	0	0	0	0
19	Q9Y3C7	MED31	2	4	5	5	5		0	1	0	0	0
20	Q9H0M0	WWP1	3	12	18	20	27		0	0	0	0	0
21	Q9NVC6	MED17	3	6	19	17	21		0	0	1	0	0
22	Q9NV88	INTS9	2	5	7	13	21		0	0	0	0	0
23	Q96HR3	MED30	1	4	6	6	5		0	0	1	0	0
24	O60244	MED14	6	29	33	30	34		0	0	2	0	0
25	O00308	WWP2	5	12	13	27	24		0	0	0	0	0
26	Q15648	MED1	6	10	28	31	31		0	0	1	0	0
27	O75586	MED6	1	8	9	6	6		0	0	1	0	0
28	Q9BQ15	NABP2	0	5	1	4	7		0	0	0	0	0
29	Q9H204	MED28	0	1	3	3	4		0	0	0	0	0
30	O43513	MED7	0	1	3	7	8		0	0	0	0	0
31	Q9Y2Z0	SUGT1	1	1	2	3	4		1	0	0	0	0
32	Q96J02	ITCH	13	25	26	29	34		0	2	4	1	0
33	Q13503	MED21	1	2	1	4	3		0	0	1	0	0
34	Q75QN2	INTS8	0	6	6	18	23		0	0	0	0	0
35	Q9P086	MED11	0	1	2	4	3		0	0	0	0	0
36	Q96P16	RPRD1A	1	15	23	12	15		0	8	8	0	0
37	Q96RN5	MED15	0	1	5	6	7		0	0	0	0	0
38	Q5TA45	CPSF3L	0	2	8	8	13		0	0	0	0	0
39	Q15369	TCEB1	1	1	2	5	3		0	0	0	3	0
40	O95104	SCAF4	0	2	11	7	5		0	0	0	0	0
41	Q9NX70	MED29	0	3	2	3	3		0	0	0	0	0
42	O75448	MED24	0	4	6	17	20		0	0	0	0	0
43	Q5VT52	RPRD2	0	8	18	15	7		0	0	0	0	0
44	Q6DN90	IQSEC1	0	6	12	6	11		0	0	0	0	0
45	Q9Y2X0	MED16	0	3	2	8	12		0	0	0	0	0

46	Q9NVH2	INTS7	0	6	6	23	30		0	0	0	1	0
47	A8MU58	AIMP2	0	1	2	2	2		0	0	0	0	0
48	Q5JSJ4	INTS6L	2	8	8	10	8		0	0	0	0	0
49	Q5TEJ8	THEMIS2	0	1	3	3	6		0	0	0	0	0
50	P30153	PPP2R1A	0	2	1	4	6		0	0	1	0	0
51	Q99590	SCAF11	0	3	5	3	3		0	0	0	0	0
52	Q13418	ILK	0	3	5	3	3		0	0	2	0	0
53	Q53G59	KLHL12	0	1	2	1	1		0	0	0	0	0
54	O00329	PIK3CD	0	2	3	4	2		0	0	0	0	0
55	Q13049	TRIM32	0	1	1	2	3		0	0	0	0	0
56	Q14145	KEAP1	0	2	8	2	3		0	0	0	0	0
57	Q13501	SQSTM1	0	1	3	0	3		0	0	0	1	0
58	Q14344	GNA13	0	1	2	1	1		0	0	0	0	0
59	H3BQA8	WDR61	1	1	0	1	1		0	0	0	0	0
60	O00505	KPNA3	1	3	4	4	4		0	1	1	0	0
61	Q14157	UBAP2L	0	2	2	1	1		0	0	0	0	0
62	Q15418	RPS6KA1	2	6	6	5	7		0	2	2	0	0
63	Q8ND56	LSM14A	0	2	4	0	1		0	0	0	0	0
64	Q9ULK4	MED23	0	1	1	9	12		0	0	0	0	0
65	Q13451	FKBP5	2	3	2	4	8		1	1	2	0	0
66	P04637	TP53	0	2	1	0	1		0	0	1	0	0
67	Q71RC2	LARP4	0	0	2	3	3		0	0	0	0	0
68	Q16576	RBBP7	0	3	4	4	2		0	1	3	2	0
69	P13807	GYS1	0	3	2	0	1		0	0	0	0	0

Table S2: List of proteins and complexes not interacting with the YFFF mutant. Related to Figure 6

A total of 69 proteins were found that shows loss of interaction with the YFFF Pol II mutant compared to rWT. Listed in the table are 25 subunits of the Mediator complex (green); 11 subunits of the Integrator complex (red); CTD phosphatase (magenta); E3-ubiquitin ligase, components of SOSS complex (blue) and few others. Log2 fold change (YFFF/rWT) and p-values for each protein is shown in the table. Data is based on five independent biological replicates.

	Uniprot ID	Gene Name	Description	Log2Fold Change (YFFF/rWT)	p-value
1	Q9NRY2	INIP	INTS3 and NABP interacting protein	-14.548	5.588E-08
2	Q9NPJ6	MED4	Mediator Complex Subunit 4	-13.671	1.447E-09
3	Q9UL03	INTS6	Integrator Complex Subunit 6	-13.490	3.673E-09
4	Q6P2C8	MED27	Mediator Complex Subunit 27	-13.159	2.356E-09
5	Q9BUE0	MED18	Mediator Complex Subunit 18	-12.981	1.433E-12
6	O95402	MED26	Mediator Complex Subunit 26	-12.838	1.41E-11
7	Q9BTT4	MED10	Mediator Complex Subunit 10	-12.827	2.053E-08
8	Q5T8T7	MED22	Mediator Complex Subunit 22	-12.637	3.805E-08
9	Q6P9B9	INTS5	Integrator Complex Subunit 5	-12.259	7.930E-08
10	Q68E01	INTS3	Integrator Complex Subunit 3	-12.146	2.420E-05
11	Q9NWA0	MED9	Mediator Complex Subunit 9	-11.932	1.696E-06
12	A0JLT2	MED19	Mediator Complex Subunit 19	-11.858	4.219E-08
13	Q96CB8	INTS12	Integrator Complex Subunit 12	-11.533	2.203E-05
14	Q9H0H0	INTS2	Integrator Complex Subunit 2	-11.520	2.799E-08
15	Q96G25	MED8	Mediator Complex Subunit 8	-11.496	2.292E-04
16	Q8N201	INTS1	Integrator Complex Subunit 1	-11.496	2.449E-06
17	Q9H944	MED20	Mediator Complex Subunit 20	-11.388	2.521E-04
18	Q96HW7	INTS4	Integrator Complex Subunit 4	-11.258	5.507E-06
19	Q9Y3C7	MED31	Mediator Complex Subunit 31	-11.180	5.368E-04
20	Q9H0M0	WWP1	WW Domain containing E3 Ubiquitin Protein Ligase 1	-11.161	3.837E-07
21	Q9NVC6	MED17	Mediator Complex Subunit 17	-11.092	1.153E-04
22	Q9NV88	INTS9	Integrator Complex Subunit 9	-11.079	9.394E-07
23	Q96HR3	MED30	Mediator Complex Subunit 30	-11.069	2.195E-04
24	O60244	MED14	Mediator Complex Subunit 14	-10.979	1.094E-04
25	O00308	WWP2	WW Domain containing E3 Ubiquitin Protein Ligase 2	-10.976	5.113E-07
26	Q15648	MED1	Mediator Complex Subunit 1	-10.736	9.263E-06
27	O75586	MED6	Mediator Complex Subunit 6	-10.708	6.511E-04
28	Q9BQ15	NABP2	Nucleic acid binding protein 2	-10.628	4.230E-03
29	Q9H204	MED28	Mediator Complex Subunit 28	-9.820	4.124E-03
30	O43513	MED7	Mediator Complex Subunit 7	-9.771	3.981E-03
31	Q9Y2Z0	SUGT1	SGT1 Homolog, MIS12 Kinetochore Complex Assembly Cochaperone	-9.705	3.864E-11
32	Q96J02	ITCH	Itchy E3 Ubiquitin Protein Ligase	-9.434	2.206E-03
33	Q13503	MED21	Mediator Complex Subunit 21	-9.335	2.326E-03
34	Q75QN2	INTS8	Integrator Complex Subunit 8	-9.223	4.178E-03
35	Q9P086	MED11	Mediator Complex Subunit 11	-8.966	4.348E-03
36	Q96P16	RPRD1A	Regulation of nuclear pre-mRNA domain containing 1A (CTD phosphatase)	-8.924	1.335E-02
37	Q96RN5	MED15	Mediator Complex Subunit 15	-8.887	4.168E-03
38	Q5TA45	CPSF3L	Cleavage and polyadenylation specificity factor 3-like (Integrator Complex Subunit 11)	-8.729	4.677E-03
39	Q15369	TCEB1	Transcription elongation factor B subunit 1	-8.481	5.145E-03
40	O95104	SCAF4	SR-related CTD associated factor 4	-8.349	4.985E-03
41	Q9NX70	MED29	Mediator Complex Subunit 29	-8.304	4.207E-03
42	O75448	MED24	Mediator Complex Subunit 24	-7.916	4.396E-03
43	Q5VT52	RPRD2	Regulation of nuclear pre-mRNA domain containing 2 (CTD phosphatase)	-7.903	4.556E-03
44	Q6DN90	IQSEC1	IQ motif and Sec7 Domain 1	-7.801	4.395E-03
45	Q9Y2X0	MED16	Mediator Complex Subunit 16	-7.639	4.243E-03
46	Q9NVH2	INTS7	Integrator Complex Subunit 7	-7.255	3.588E-02
47	A8MU58	AIMP2	Aminoacyl tRNA Synthetase Complex-interacting Multifuntional protein 2	-7.053	4.231E-03
48	Q5JSJ4	INTS6L	Integrator Complex Subunit 6 Like	-6.739	4.555E-03
49	Q5TEJ8	THEMIS2	Thymocyte selection associated family member 2	-6.695	4.375E-03

50	P30153	PPP2R1A	Protein phosphatase 2 regulatory subunit A, alpha	-6.577	7.267E-03
51	Q99590	SCAF11	SR-related CTD associated factor 11	-6.371	4.849E-03
52	Q13418	ILK	Integrin linked kinase	-6.116	4.462E-02
53	Q53G59	KLHL12	Kelch like family member 12	-6.051	4.223E-03
54	O00329	PIK3CD	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Delta	-5.961	4.118E-03
55	Q13049	TRIM32	Tripartite Motif Containing 32	-5.938	7.280E-03
56	Q14145	KEAP1	Kelch Like ECH Associated Protein 1	-5.919	2.511E-02
57	Q13501	SQSTM1	Sequestosome 1	-5.918	4.041E-02
58	Q14344	GNA13	G Protein Subunit Alpha 13	-5.908	6.915E-03
59	H3BQA8	WDR61	WD Repeat Domain 61	-5.784	5.656E-03
60	O00505	KPNA3	Karyopherin Subunit Alpha 3	-5.738	3.355E-02
61	Q14157	UBAP2L	Ubiquitin associated protein 2 like	-5.721	6.337E-03
62	Q15418	RPS6KA1	Ribosomal Protein S6 kinase A1	-5.715	3.025E-02
63	Q8ND56	LSM14A	LSM14A mRNA processing body assembly factor	-5.583	4.459E-02
64	Q9ULK4	MED23	Mediator Complex Subunit 23	-5.570	8.126E-03
65	Q13451	FKBP5	FK506 Binding protein 5	-5.353	1.853E-02
66	P04637	TP53	Tumor protein p53	-5.263	4.374E-02
67	Q71RC2	LARP4	La Ribonucleoprotein Domain Family Member 4	-5.112	4.040E-02
68	Q16576	RBBP7	Retinoblastoma Binding Protein 7	-5.100	4.517E-02
69	P13807	GYS1	Glycogen Synthase 1	-5.082	4.381E-02

Table S3: Peptide counts of selected proteins interacting with the Pol II of both, the rWT and the YFFF mutant. Related to Figure 6

Peptide counts of selected proteins and complexes that interact with Pol II of both, the rWT and the YFFF mutant for all five biological replicates. Samples in the experiments 1 and 2 were subjected to on-beads trypsin digest, while samples in the experiments 3, 4 and 5 were subjected to in-gel trypsin digest.

			rWT						YFFF				
	Experimental Number		1	2	3	4	5		1	2	3	4	5*
	Uniprot ID	Gene Name	Peptide counts						Peptide counts				
			Polymerase Subunits										
1	P24928	RPB1	119	132	156	153	153		98	126	139	102	11
2	P30876	RPB2	33	53	73	68	70		26	49	61	35	1
3	P19387	RPB3	7	15	17	19	16		5	14	13	9	0
4	O15514	RPB4	0	1	3	4	11		0	0	1	2	0
5	P19388	RPB5	4	11	14	11	10		6	11	9	5	0
6	U3KPY1	RPB6	0	0	1	1	2		0	0	1	0	0
7	P62487	RPB7	1	0	3	4	4		0	1	1	0	0
8	P52434	RPB8	8	12	12	11	12		7	11	11	9	2
9	P36954	RPB9	2	5	7	9	8		2	5	5	3	0
10	P62875	RPB10	3	3	1	1	2		3	3	1	1	0
11	P52435	RPB11	5	7	6	4	9		5	6	6	4	0
12	P53803	RPB12	2	1	2	1	3		0	2	2	1	0
			Splicing factors										
13	Q07955	SRSF1	5	14	20	11	15		5	16	20	13	2
14	J3KP15	SRSF2	0	0	7	0	2		0	2	6	1	0
15	P84103	SRSF3	4	11	10	7	6		3	11	10	10	1
16	Q08170	SRSF4	3	8	7	5	4		2	8	8	4	1
17	Q13243	SRSF5	0	4	5	4	3		1	5	7	3	1
18	Q13247	SRSF6	3	9	9	8	8		3	8	9	9	1
19	Q16629	SRSF7	3	10	11	8	11		7	11	10	10	2
20	Q13242	SRSF9	1	9	16	11	12		4	11	20	17	2
21	O75494	SRSF10	1	9	11	8	10		1	9	12	9	0
22	Q5T760	SRSF11	0	1	2	0	0		0	2	4	0	0
23	Q01081	U2AF1	1	7	7	4	5		1	6	7	3	1
24	P26368	U2AF2	2	2	14	2	4		0	4	8	0	0
			3' end processing and termination factors										
25	Q10570	CPSF1	0	3	13	4	3		0	2	8	1	0
26	Q9P2I0	CPSF2	0	1	2	1	1		0	1	3	0	0
27	G5E9W3	CPSF3	0	0	2	1	2		0	0	1	0	0
28	B7Z7B0	CPSF4	0	1	0	0	1		0	0	1	1	0
29	O43809	CPSF5	0	3	3	0	0		0	3	5	0	0
30	F8WJN3	CPSF6	0	1	2	0	1		0	2	1	0	0
31	Q9H0D6	XRN2	0	8	23	12	14		1	8	17	2	0

Table S4: Conditions for chromatin immunoprecipitation experiments. Related to STAR Methods section "ChIP-seq and ChIP-qPCR"

ChIP-ed Protein	Antibody Ref.	Antibody Quantity	# cells/ChIP	Dynabeads Prot. G /ChIP	# RIPA Washes
Pol-II	ab9110	10 µg	25×10^6	100 µl	6
H3K4me1	ab8895	2 µg	5×10^6	20 µl	6
H3K4me3	ab8580	2 µg	5×10^6	20 µl	5
H3K27ac	ab4729	2 µg	5×10^6	20 µl	5

Table S5: List of proteins and complexes in MS with the rWT and Ser2AAA mutant Pol II. Related to Figure S6

Log2fold change (S2AAA/rWT) and p-values for subunits of Polymerase, Mediator and Integrator complexes. Data is based on three independent biological replicates.

	List of proteins in rWT and the mutant S2AAA				
	Uniprot ID	**Gene Name**	**Description**	**Log2Fold Change (S2AAA/rWT)**	**p-value**
	Polymerase Subunit				
1	P24928	POLR2A	RNA Polymerase II Subunit B1 (RPB1)	1.429	0.415
2	P30876	POLR2B	RNA Polymerase II Subunit B2 (RPB2)	1.380	0.584
3	P19387	POLR2C	RNA Polymerase II Subunit B3 (RPB3)	2.232	0.479
4	O15514	POLR2D	RNA Polymerase II Subunit B4 (RPB4)	1.009	0.859
5	P19388	POLR2E	RNA Polymerases I, II, And III Subunit ABC1 (RPB5)	2.257	0.418
6	U3KPY1	POLR2F	RNA Polymerases I, II, And III Subunit ABC2 (RPB6)	-4.169	0.374
7	P62487	POLR2G	RNA Polymerase II Subunit B7 (RPB7)	3.394	0.427
8	P52434	POLR2H	RNA Polymerases I, II, And III Subunit ABC3 (RPB8)	2.407	0.220
9	P36954	POLR2I	RNA Polymerase II Subunit B9 (RPB9)	4.350	0.340
10	P62875	POLR2L	RNA Polymerases I, II, And III Subunit ABC5 (RPB10)	7.651	0.111
11	P52435	POLR2J	RNA Polymerase II Subunit B11 (RPB11)	5.215	0.217
12	P53803	POLR2K	RNA Polymerases I, II, And III Subunit ABC4 (RPB12)	5.023	0.267
13	Q8N201	INTS1	Integrator Complex Subunit 1	-1.025	0.662
14	Q9H0H0	INTS2	Integrator Complex Subunit 2	-1.192	0.690
15	Q68E01	INTS3	Integrator Complex Subunit 3	-1.337	0.566
16	Q96HW7	INTS4	Integrator Complex Subunit 4	-1.453	0.650
17	Q6P9B9	INTS5	Integrator Complex Subunit 5	-2.179	0.562
18	Q9UL03	INTS6	Integrator Complex Subunit 6	-1.605	0.554
19	Q9NVH2	INTS7	Integrator Complex Subunit 7	-2.794	0.443
20	Q75QN2	INTS8	Integrator Complex Subunit 8	-5.354	0.126
21	Q9NV88	INTS9	Integrator Complex Subunit 9	-3.784	0.251
22	Q5TA45	CPSF3L	Integrator Complex Subunit 11	-1.492	0.626
23	Q96CB8	INTS12	Integrator Complex Subunit 12	-0.082	0.985
24	Q15648	MED1	Mediator Complex Subunit 1	-1.562	0.662
25	Q9NPJ6	MED4	Mediator Complex Subunit 4	-6.496	0.115
26	O75586	MED6	Mediator Complex Subunit 6	-2.498	0.514
27	O43513	MED7	Mediator Complex Subunit 7	-3.205	0.374
28	Q96G25	MED8	Mediator Complex Subunit 8	-4.227	0.387
29	Q9BTT4	MED10	Mediator Complex Subunit 10	-3.271	0.431
30	Q9P086	MED11	Mediator Complex Subunit 11	-5.924	0.138
31	O60244	MED14	Mediator Complex Subunit 14	-3.984	0.342
32	Q96RN5	MED15	Mediator Complex Subunit 15	-5.369	0.117
33	Q9Y2X0	MED16	Mediator Complex Subunit 16	-4.729	0.117
34	Q9NVC6	MED17	Mediator Complex Subunit 17	-5.181	0.137
35	Q9BUE0	MED18	Mediator Complex Subunit 18	-3.517	0.374
36	A0JLT2	MED19	Mediator Complex Subunit 19	-2.609	0.374
37	Q9H944	MED20	Mediator Complex Subunit 20	-1.875	0.689
38	Q9ULK4	MED23	Mediator Complex Subunit 23	-3.432	0.187
39	O75448	MED24	Mediator Complex Subunit 24	-3.665	0.299
40	Q9NX70	MED29	Mediator Complex Subunit 29	-3.028	0.374
41	Q96HR3	MED30	Mediator Complex Subunit 30	-0.086	0.986
42	Q9Y3C7	MED31	Mediator Complex Subunit 31	-3.914	0.374

Table S6: Peptide counts of proteins in MS with the rWT and Ser2AAA mutant Pol II. Related to Figure S6

Peptide counts of subunits for Polymerase, Mediator and the Integrator complexes in rWT and S2AAA mutant.

		Peptide counts					
Gene Name	**Description**	**rWT1**	**rWT2**	**rWT3**	**S2AAA_1**	**S2AAA_2**	**S2AAA_3**
POLR2A	RNA Polymerase II Subunit B1 (RPB1)	74	115	89	133	107	120
POLR2B	RNA Polymerase II Subunit B2 (RPB2)	14	61	38	53	56	57
POLR2C	RNA Polymerase II Subunit B3 (RPB3)	1	13	11	12	15	15
POLR2D	RNA Polymerase II Subunit B4 (RPB4)	0	4	3	0	6	4
POLR2E	RNA Polymerases I, II, And III Subunit ABC1 (RPB5)	2	5	6	9	8	8
POLR2F	RNA Polymerases I, II, And III Subunit ABC2 (RPB6)	0	0	1	0	0	0
POLR2G	RNA Polymerase II Subunit B7 (RPB7)	0	2	3	4	2	5
POLR2H	RNA Polymerases I, II, And III Subunit ABC3 (RPB8)	4	7	8	10	10	9
POLR2I	RNA Polymerase II Subunit B9 (RPB9)	0	3	5	6	4	5
POLR2L	RNA Polymerases I, II, And III Subunit ABC5 (RPB10)	0	1	0	1	2	2
POLR2J	RNA Polymerase II Subunit B11 (RPB11)	0	3	2	5	6	6
POLR2K	RNA Polymerases I, II, And III Subunit ABC4 (RPB12)	0	1	1	3	1	1
INTS1	Integrator Complex Subunit 1	3	48	23	21	6	22
INTS2	Integrator Complex Subunit 2	0	15	3	2	1	4
INTS3	Integrator Complex Subunit 3	4	29	14	11	10	9
INTS4	Integrator Complex Subunit 4	0	16	6	3	2	1
INTS5	Integrator Complex Subunit 5	0	14	1	1	0	2
INTS6	Integrator Complex Subunit 6	1	26	12	4	5	4
INTS7	Integrator Complex Subunit 7	0	16	5	1	0	1
INTS8	Integrator Complex Subunit 8	0	12	3	1	0	0
INTS9	Integrator Complex Subunit 9	0	3	2	1	0	0
CPSF3L	Integrator Complex Subunit 11	0	1	2	1	1	0
INTS12	Integrator Complex Subunit 12	0	1	7	6	0	3
MED1	Mediator Complex Subunit 1	0	26	12	3	1	4
MED4	Mediator Complex Subunit 4	2	6	6	0	0	1
MED6	Mediator Complex Subunit 6	1	0	1	3	0	0
MED7	Mediator Complex Subunit 7	0	0	2	0	0	0
MED8	Mediator Complex Subunit 8	3	6	0	0	3	0
MED10	Mediator Complex Subunit 10	0	2	2	0	0	1
MED11	Mediator Complex Subunit 11	0	1	2	0	0	0
MED14	Mediator Complex Subunit 14	0	29	10	2	0	2
MED15	Mediator Complex Subunit 15	0	2	2	0	0	0
MED16	Mediator Complex Subunit 16	0	5	4	0	0	0
MED17	Mediator Complex Subunit 17	3	14	14	6	0	1
MED18	Mediator Complex Subunit 18	0	0	2	0	1	0
MED19	Mediator Complex Subunit 19	0	0	2	0	0	0
MED20	Mediator Complex Subunit 20	0	0	4	1	0	0
MED23	Mediator Complex Subunit 23	0	6	3	0	0	1
MED24	Mediator Complex Subunit 24	0	8	4	0	0	1
MED29	Mediator Complex Subunit 29	0	1	2	1	0	1
MED30	Mediator Complex Subunit 30	0	0	3	1	0	1
MED31	Mediator Complex Subunit 31	0	0	3	0	0	0

Table S7: Parameters used for ChIP-seq peak calling with Integrated Genome Browser's *Thresholding* function. Related to STAR Methods section "ChIP-seq data analysis"

Sample	Threshold	Max.Gap	Min.Run
rWT_H3K4me1	110	3000	400
rWT_H3K27ac	130	2000	200
rWT_H3K4me3	100	1000	200
rWT_PolII	35	3000	300