Illuminating the roles of transcription elongation factors TCEA1 and TCEA2 in human cells

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by Yelizaveta Mochalova

2023

The dissertation was prepared under the supervision of Dr. Andreas Mayer at the Max Planck Institute for Molecular Genetics in Berlin from November 2017 to May 2023.

1st reviewer: Dr. Andreas Mayer

2nd reviewer: Prof. Dr. Markus Wahl

Date of defense: 21.08.2023.

Acknowledgments

With all my heart, I would like to thank my supervisor Dr. Andreas Mayer for continuous support and guidance throughout my challenging project. I am grateful for the skills and the mindset that I developed under your mentorship, as you have always encouraged me to think independently and approach challenges more optimistically. Thank you for all the insightful discussions and the opportunity to be a part of such a motivating scientific environment.

I would like to express my gratitude to my thesis advisory committee, Dr. Tuğçe Aktaş and Prof. Dr. Markus Wahl for their supportive counseling and perceptive suggestions on experiments. Your ideas prompted me to selectively investigate the interactomes of the N-terminal domain of TFIIS and make interesting observations. I would also like to thank all the reviewers of my work for their time and feedback.

A big thank you goes to my collaborators, Mario Rubio and Dr. Martyna Gajos, for their extensive bioinformatic analyses and their patience while dealing with me in the state of eager anticipation of ground-breaking discoveries. I am very grateful to our excellent technicians, Ilka Langner, for her tremendous help with cell line generation, and Susanne Freier, for her stellar NET-seq library preparations. Thank you to Johanna Krüger who has masterfully confirmed our hypothesis about TFIIS evolution and Vanessa Treffner for her assistance with genotyping. I also highly appreciate the work done by Dr. David Meierhofer and Beata Lukaszewska-McGreal of the mass spectrometry facility and the excellent staff of our FACS, microscopy, and sequencing facilities as well as our IT team.

I would also like to thank all my other colleagues who have helped me along the way. I am grateful to my best friend and a star lab trooper Dr. Mirjam Arnold for sharing her knowledge and protocols and for hundreds of brainstorming sessions, Dr. Annkatrin Bressin for her insight in the computation analyses, and Dr. Elisabeth Altendorfer for sharing her knowledge about the cell cycle and being a delightful office neighbor. A big thank you to Jelena Ulicevic and Nicole Eischer of our group and the whole first floor of our tower, especially Kate Lübke, for such a friendly and lively working atmosphere. Many thanks go to the research groups of Drs. Tuğçe Aktaş and Sarah Kinkley who generously shared cells, antibodies, and experience with me throughout the years.

Every day I am thankful for my family and friends who have cheered me on, gave me strength to storm through the tough times, occasionally reminded me “to live a little,” and flew in on “international rescue missions.”

Selbstständigkeitserklärung

Hierdurch versichere ich, dass ich meine Dissertation selbstständig verfasst und keine anderen als die von mir angegeben Quellen und Hilfsmittel verwendet habe. Geistiges Eigentum anderer Autoren wurde als entsprechend gekennzeichnet. Ebenso versichere ich, dass ich an keiner anderen Stelle ein Prüfungsverfahren beantragt bzw. die Dissertation in dieser oder anderer Form an keiner anderen Fakultät als Dissertation vorgelegt habe.

Berlin, den 5. Juni 2023

Yelizaveta Mochalova

Contents

Summary

In eukaryotes, the transcription cycle, which includes initiation, elongation, and termination, is regulated at multiple steps by both *cis*-regulatory elements and *trans*-acting factors. At protein-coding genes, successful transcription initiation is often followed by a pivotal regulatory step, i.e. the promoter-proximal pause. During this pause, the elongation complex is supplemented by additional elongation factors enabling productive elongation, proper termination and co-transcriptional RNA processing. In addition to this scheduled regulated pause, Pol II encounters obstacles such as nucleosomes, pause predisposing *cis*-sequences or other physical barriers that cause Pol II to move backward - a phenomenon referred to as Pol II backtracking. In case of a transient pause, when Pol II backtracks by one nucleotide, it can cleave the backtracked nucleotide itself, as it has an intrinsic nuclease property. However, when more extensive backtracking is thermodynamically favorable, Pol II is trapped in an immobilized state known as transcriptional arrest and requires an auxiliary factor TFIIS (TCEA in human cells). TFIIS alleviates transcriptional arrest by enhancing the RNA cleavage activity of Pol II. Following the hydrolysis of backtracked RNA, a new 3' is generated and the active site of Pol II is free to resume RNA synthesis.

The molecular mechanism of TFIIS has been extensively characterized both structurally and biochemically; however, our understanding of its role in human cells remains incomplete because of the additional complexity incurred by four highly conserved TFIIS paralogs. To date, there have been no multiomics studies investigating paralog-specific roles. In our research project, we analyzed the function of two paralogs expressed in HEK293T cells: the ubiquitously expressed TCEA1 and lowly expressed cerebellum- and testis- specific TCEA2. We employed two cellular systems: an inducible epitope-tagged protein expression system to determine their individual interaction partners and genomic binding patterns and CRISPR-knockouts of each and both proteins to evaluate the impact on elongation at the level of nascent and messenger RNA.

Our results demonstrate that both paralogs bind Pol II shortly after initiation and, in their absence, Pol II occupancy is strikingly increased at the promoter-proximal region of the majority of active genes, as revealed by High-sensitive native elongating transcript sequencing (HiS-NET-seq). About a half of those genes have reduced Pol II occupancy in the gene body. This indicates that backtracking is a common event at the promoter-proximal

region and TCEA1 and TCEA2 are needed to progress into the gene body and, possibly, also for proper Pol II processivity. Unexpectedly, TCEA2 deletion was more detrimental to cellular growth than the TCEA1 knockout. The TCEA2 and the double knockout cells had a profound delay in DNA replication accompanied by activated DNA damage response. The deletion of TCEA2 had a more pronounced impact on the transcriptional output, indicating that, despite its higher expression level, TCEA1 alone is not sufficient to alleviate backtracking. This suggests that TCEA2 is a potent elongation factor and has its own target genes that regulate cell proliferation.

Zusammenfassung

In Eukaryoten wird der Transkriptionszyklus, bestehend aus Initiation, Elongation und Termination, durch mehrere Schritte sowohl von *cis*-regulatorischen Elementen als auch von *trans*-aktivierenden Faktoren reguliert. Bei Proteincodierenden Genen wird auf eine erfolgreiche Transkriptionsinitiation oft ein entscheidender regulatorischer Schritt folgen, nämlich der promoter-proximale Halt. Während dieses Halts wird der Elongationskomplex durch zusätzliche Elongationsfaktoren ergänzt, um eine produktive Elongation, eine korrekte Termination und die ko-transkriptionelle RNA-Verarbeitung zu ermöglichen. Neben diesem geplanten, regulierten Halt sieht sich die Pol II auch mit Hindernissen wie Nukleosomen, *cis*-Sequenzen, die Pausen begünstigen, oder anderen physikalischen Barrieren konfrontiert, die dazu führen, dass sich die Pol II zurückbewegt - ein Phänomen, das als Pol II-Backtracking bezeichnet wird. Bei einer vorübergehenden Pause, bei der die Pol II um ein Nukleotid zurückgeht, kann sie das zurückgegangene Nukleotid selbst spalten, da sie eine intrinsische Nuklease-Eigenschaft besitzt. Wenn jedoch ein umfangreicheres Backtracking thermodynamisch begünstigt wird, gerät die Pol II in einen immobilisierten Zustand, der als Transkriptionsarrest bekannt ist, und benötigt einen Hilfsfaktor namens TFIIS (TCEA in humanen Zellen). TFIIS löst den Transkriptionsarrest auf, indem es die RNA-Spaltaktivität der Pol II erhöht. Nach der Spaltung der zurückgegangenen RNA entsteht ein neues 3'-Ende, und die aktive Stelle der Pol II ist frei, die RNA-Synthese fortzusetzen.

Der molekulare Mechanismus von TFIIS wurde sowohl strukturell als auch biochemisch umfassend charakterisiert, jedoch ist unser Verständnis seiner Rolle in humanen Zellen aufgrund der zusätzlichen Komplexität durch vier stark konservierte TFIIS-Paraloge unvollständig. Bisher wurden keine Multiomik-Studien zur Untersuchung der paralogspezifischen Funktionen durchgeführt. In unserem Forschungsprojekt haben wir die Funktion von zwei in HEK293T-Zellen exprimierten Paralogen analysiert: dem ubiquitär exprimierten TCEA1 und dem in Gehirn und Hoden niedrig exprimierten TCEA2. Wir haben zwei zelluläre Systeme verwendet: ein induzierbares Epitop-markiertes Proteinausdruckssystem, um ihre individuellen Interaktionspartner und genomische Bindungsmuster zu bestimmen, sowie CRISPR-Knockouts für jedes einzelne Protein und beide zusammen, um die Auswirkungen auf die Elongation auf Ebene der neu synthetisierten und Boten-RNA zu bewerten.

Unsere Ergebnisse zeigen, dass beide Paraloge kurz nach der Initiation an die Pol II binden

und dass in ihrer Abwesenheit die Besetzung der Pol II im promoter-proximalen Bereich der Mehrzahl der aktiven Gene signifikant erhöht ist. Etwa die Hälfte dieser Gene weist eine reduzierte Besetzung der Pol II im Genkörper auf. Dies deutet darauf hin, dass Backtracking ein häufiges Ereignis im promoter-proximalen Bereich ist und dass TCEA1 und TCEA2 erforderlich sind, um in den Genkörper zu gelangen und möglicherweise auch für eine ordnungsgemäße Pol II-Prozessivität. Überraschenderweise hatte die Deletion von TCEA2 einen stärkeren negativen Einfluss auf das Zellwachstum als der Knockout von TCEA1. Die TCEA2- und die Doppel-Knockout-Zellen wiesen eine erhebliche Verzögerung bei der DNA-Replikation und eine aktivierte DNA-Schadensantwort auf. Die Deletion von TCEA2 hatte auch einen stärkeren Einfluss auf die transkriptionelle Ausgabe, was darauf hindeutet, dass TCEA1 allein trotz seiner höheren Expression nicht ausreicht, um das Backtracking zu beseitigen. Dies legt nahe, dass TCEA2 ein wirksamer Elongationsfaktor ist und möglicherweise eigene Zielgene besitzt, die die Zellproliferation regulieren.

Chapter 1

INTRODUCTION

1.1 The transcription cycle

In eukaryotes, RNA polymerase II (Pol II) is a large enzyme complex, comprising twelve subunits. It synthesizes all precursor messenger RNA (pre-mRNA) and many non-coding RNAs, including enhancer RNA [1, 2], long noncoding RNA [3, 4], and some primary microRNAs [5]. Transcription regulation is an active area of research as it is pivotal for proper gene expression. The progression of Pol II through a gene is discontinuous and regulated by *cis*-regulatory DNA elements and *trans*-acting factors that it encounters and interacts with throughout the transcription cycle. The transcription cycle is subdivided into three phases: initiation, elongation, and termination (reviewed in [6]).

1.1.1 Transcription initiation

Structural and biochemical studies demonstrated that transcription initiation proceeds in the following stages in metazoan organisms: the pre-initiation complex (PIC) assembly, unwinding of promoter DNA and activation of the PIC, and promoter escape [7-10]. This process is schematically summarized in Figure 1.

Promoter structure varies in terms of the core promoter elements and the number of transcription start sites (TSSs), depending on the function of the gene (reviewed in [11]). Promoters of genes carrying out cell-type specific roles in terminally differentiated adult cells have a single clearly defined TSS, containing a TATA box core promoter element and an Initiator element downstream [12], and imprecisely positioned nucleosomes [13]. Whereas ubiquitously expressed housekeeping genes harbor a CpG island and generally have a "dispersed" promoter with multiple closely spaced TSSs [12] and a defined nucleosome-depleted region [13]. Both types of promoters, when active, are epigenetically marked with H3K4me3 and H3K27ac [11]. Promoters of mammalian developmental genes also have CpG islands, which can be long or consist of multiple CpG islands, and in embryonic stem cells are bivalently marked by activating H3Kme3 and repressive H3K27me3 [14].

Prior to transcription initiation, a promoter region needs to become accessible for Pol II recruitment and binding. Promoter DNA, wrapped around histones, together forming a chromatin packing unit, called a nucleosome, is a barrier to transcription initiation [15-18]. Whole genome nucleosome-mapping studies revealed that promoters and TSSs of actively expressed genes are nucleosome-free [19-23] indicating that this promoter-proximal nucleosome, or the +1 nucleosome, needs to be slid or evicted [17, 24-26]. This pre-initiation regulatory step is mediated by nucleosomal DNA binding "pioneer" factors, free DNA binding transcription factors, histone acetyltransferases, and chromatin remodeling complexes [17, 27, 28]. Together, these regulators facilitate the binding of the general transcription factors (GTFs) and the Mediator complex, which subsequently recruit and position Pol II, forming the pre-initiation complex (PIC) at the correct genomic sites [29]. In humans, the PIC consists of the GTFs (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH), the Mediator complex, and Pol II, together amounting to about 4.5 MDa [30].

1.1.1.1 Assembly of the pre-initiation complex (PIC)

The PIC assembly begins with recognition and binding of the 1 MDa TFIID complex to the core promoter element, for example the TATA box [31]. The complex consists of the TATA-box binding protein (TBP) and 13 TBP-associated factors (TAFs) [32]. TBP can be stabilized by TFIIA, although its involvement appears to be not essential [33, 34]. Other GTFs and co-activators, including the Mediator complex, are sequentially recruited to the promoter [31, 35]. TFIID recruits TFIIB, which serves as a bridge for the binding of Pol II as it binds the "dock" and "wall" domains of Pol II [36-39]. TFIIB also positions the promoter DNA over the central cleft of Pol II [37]. TFIIB is involved in the early PIC stabilization and dissociates from Pol II during promoter escape [38]. Next, TFIIB is stabilized by TFIIF [40]. Finally, TFIIE and TFIIH are recruited to the assembling PIC. TFIIE interacts directly with Pol II and helps orient TFIIH [41-43] and stimulates its kinase and ATPase activities, required for PIC activation and transcription initiation [44].

A transcription co-activator, Mediator, is a large protein complex with variable multi-subunit combination and it serves as a global regulator of transcription pre-initiation, initiation, elongation, as well as chromatin architecture [45]. The two core modules of the Mediator complex interact with Pol II, TFIIB, and TFIIH [46-49]. Mediator was shown to regulate the recruitment and assembly of some GTFs, including TFIIB, TFIID, and TFIIH [50-52]. Mediator also influences the activity of TFIIH, by stimulating its kinase activity [49, 52, 53], thus inducing more activating signals.

The C-terminal domain (CTD) of RPB1, the largest subunit of Pol II, possesses numerous heptapeptide repeats, 52 and 26 for human and yeast respectively, of the consensus amino acid sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 [54-56]. As Pol II moves along a gene, the CTD gets post-translationally modified at various residues of the repeats, defining the appropriate interactions with initiation, elongation, termination, and RNA processing factors (reviewed in [57]). However, during initiation, the CTD is mostly unphosphorylated, which allows for a stabilizing interaction with the Mediator complex. Upon binding to the PIC, Mediator recruits TFIIH and stimulates phosphorylation of Ser5 (Ser5p) and Ser7 (Ser7p) of Pol II RPB1 CTD by the cyclin-dependent kinase CDK7 subunit of TFIIH [49, 52, 53]. Then Ser5p recruits the co-transcriptional 5' capping enzymes [58] and aids in promoter escape [59] (described later in this section). The exact function of Ser7p is not fully understood yet, but it is proposed to facilitate subsequent phosphorylation at the CTD of RPB1 [60]. CDK7 also acts as a CDK-activating kinase (CAK) and can phosphorylate CDK9 of P-TEFb and CDK12 [61], the kinases involved in elongation (described in detail in 1.1.2).

1.1.1.2 Promoter DNA unwinding

Importantly, the PIC is “closed” and inactive until a helicase subunit of TFIIH opens the DNA around the TSS. The ATP-dependent translocase subunit XBP of TFIIH binds downstream of Pol II and unwinds 11-15 base pairs around the TSS by moving along one DNA strand inducing torsion and conformation changes, which allow one strand to enter the active site of Pol II [43, 62-66]. This step leads to a formation of a transcription bubble and Pol II begins synthesizing nascent RNA, complementary to the inserted template DNA. The site of the initial synthesis is the TSS and it is usually found at about 25 - 30 bp and 40 - 150 bp downstream of core promoter elements, such as TATA box, in metazoans and yeast, respectively [12, 67-69].

1.1.1.3 Promoter escape

Having begun synthesizing, Pol II encounters a critical phase known as the promoter escape. Further movement of Pol II is impeded with gene-variable strengths by contacts with promoter DNA sequences and the PIC. Biophysical experiments demonstrated that when the first 8 - 15 nucleotides are being transcribed, Pol II is still contacting the promoter DNA, resulting in a formation of an extended transcription bubble with “scrunched” template DNA [70-72]. This structure is unstable and energetically unfavorable, leading to a collapse of the upstream part of the transcription bubble [73-75]. This results in a release of energy that either pushes Pol

II forward, allowing it to break off contacts with the promoter and escape it, or reverses Pol II to the active PIC state, releasing a short “abortive” transcript [76]. Promoter escape was shown to be facilitated by Ser5p at the CTD of Pol II RPB1 because it causes dissociation from Mediator [59]. Following a successful promoter escape, Pol II enters the next transcription cycle phase, elongation.

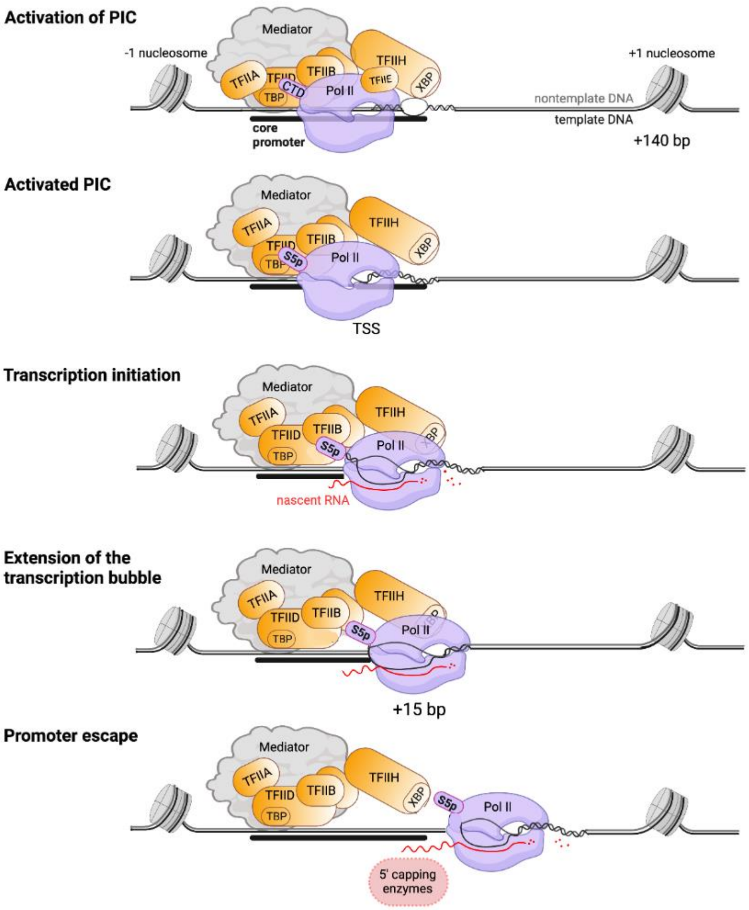


Figure 1: Schematic summary of transcription initiation. Prior to transcription initiation, the general transcription factors (GTFs) (in orange) and the Mediator complex find the core promoter (this step is omitted here). Subsequently, RNA polymerase II (Pol II) is recruited and positioned on the DNA, forming the pre-initiation complex (PIC). Pol II CTD is hypophosphorylated at that time. The PIC is activated after the TFIIH subunit XBP unwinds the DNA around the TSS, causing the insertion of one strand into the active site of Pol II. Pol II commences RNA synthesis, however, transcribing through the promoter is challenging because of interactions with the GTFs and the DNA sequence, resulting in a thermodynamically unstable extended transcription bubble. It collapses, releasing energy and pushing Pol II downstream. Ser5p of CTD, deposited by CDK7 of TFIIH, helps with promoter escape and recruits the 5’ capping enzymes.

1.1.1.4 The mechanism of RNA synthesis by Pol II

During RNA synthesis, Pol II catalyzes phosphodiester bond formation using a two-metal-ion mechanism [77]. One metal ion activates the attacking hydroxyl group, while the other stabilizes the oxyanion leaving group. Forward movement of Pol II occurs via the “Brownian ratchet” mechanism with the pre-translocation and post-translocation states [78, 79]. The pre-translocation state entails a paired template DNA NTP (nucleoside triphosphate) with a complementary RNA NTP, while in the post-translocated state, the DNA NTP is not yet paired, the latter state is more thermodynamically favorable. The incorporation of a new NTP causes forward translocation of Pol II because it is more thermodynamically favorable. In the active site of Pol II, 8 - 9 nucleotides of nascent RNA hybridize with the template DNA [80]. This RNA-DNA hybrid stabilizes Pol II on the DNA strand, thereby enhancing its commitment to elongation [75]. However, if the 3’ end of the nascent RNA disengages with the active site, Pol II can move backward, or backtrack, leading to a pause or arrest (described in detail in 1.2.3).

1.1.2 Transcription elongation

Soon after Pol II escapes the promoter and begins the elongation phase, it is challenged by regulatory checkpoints and physical barriers. The advancement of genome-wide sequencing methods led to the currently established view that Pol II generally pauses at the promoter-proximal regions, at exon-intron boundaries, and after passing the polyadenylation site, in the termination zone. The most studied pause enrichment is the one at the promoter-proximal region and it signifies a major regulatory step, during which Pol II is prepared for productive elongation. The key factors controlling the promoter-proximal regulation point are described in this section and summarized schematically in Figure 2. However, this field of research is currently advancing and the list of transcription factors is growing.

1.1.2.1 Promoter-proximal pause

In metazoan organisms, a pivotal regulatory process occurs during transcription elongation, and it is known as the promoter-proximal pausing, which occurs at about 50 base pairs downstream of the TSS. Over the last two decades, the view that Pol II recruitment is the rate-limiting step of transcription has been debated due to the discovery of global promoter-proximal pausing. Currently, promoter-proximal pausing is thought to be a regulatory checkpoint in gene expression, which involves general and gene target-specific *trans*-acting factors, *cis*-regulatory elements, and co-transcriptional RNA processing factors.

Promoter-proximal pausing was first observed at the heat-shock protein 70 gene, *Hsp70* in *Drosophila melanogaster* [81-84]. Pol II was already bound to the promoter-proximal region, prior to gene activation by heat shock, indicating that Pol II had been recruited and was remaining on standby for an activation signal to continue with elongation [81, 82]. Later, whole genome sequencing experiments revealed a striking Pol II density peak in the first 20 - 60 nucleotides downstream of TSS and that this phenomenon was not restricted to heat shock-responsive genes, but was a general event at protein-coding genes in *Drosophila* [83, 85-87]. Pol II pausing in the promoter-proximal region was also observed at the majority of genes in mammalian cells, using nascent RNA profiling methods [88-90]. A few studies proposed that the Pol II density peak should be interpreted as a dynamic result of initiation, pausing, and turnover in that region [91-94].

What exactly causes Pol II to stall has not been completely elucidated *in vivo* yet, however, the allosteric regulation of Pol II pausing in the promoter-proximal region was uncovered thanks to extensive structural studies. Cryogenic electron microscopy revealed that during a pause, the DNA-RNA hybrid is tilted in the active site of Pol II [95], precluding nucleotide addition [96]. The transcription elongation factors DSIF (5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) sensitivity inducing factor) and NELF (negative elongation factor) stabilize the conformation of the paused Pol II [97].

DSIF

DSIF is a heterodimer consisting of SPT4 and SPT5 [98]. The precise mode of binding of DSIF was revealed by cryo-EM and X-ray crystallography [99]. DSIF spans the surface of Pol II from the DNA cleft to the RNA exit tunnel. Two modules of DSIF form a DNA clamp, the other two form an RNA clamp, and one domain buttresses the RNA clamp. Based on these observations, DSIF functions in stabilizing the transcription bubble, navigating upstream DNA positioning and RNA exit. Besides binding Pol II near the RNA-exit channel [95, 99, 100], SPT5 also binds the upstream DNA and the emerging nascent RNA, longer than 18 nucleotides (nt) [101]. SPT5 also plays a role in promoting RNA-protecting 5’ capping [102]. Rapid degradation of SPT5 in human cells resulted in a global reduction of Pol II at promoters and enhancers [103] and ubiquitination and proteasomal degradation of RPB1 subunit of Pol

II, indicating that SPT5 plays a crucial role in stabilization of Pol II at promoter-proximal regions [104].

NELF

NELF senses SPT5-Pol II [105, 106] and binds Pol II on the opposite side from DSIF [95]. NELF is a complex consisting of four subunits, NELF-A, NELF-B, NELF-C or -D, and NELF-E [105, 107] and it folds into a three-lobed shape [95]. NELF associates with the funnel domain and the trigger loop domain of Pol II. It was demonstrated that in the paused Pol II complex, NELF restrains Pol II mobility, prevents nucleotide addition, and precludes binding of a positive elongation factor, transcription factor IIS (TFIIS) [95], which is required for the release of backtracked Pol II (described in detail in 1.2).

Studies relying on the loss of function of NELF coupled with global Pol II occupancy profiling uncovered more mechanistic insights into the pause regulation. NELF depletion led to a reduction of paused Pol II at the promoter-proximal region at the majority of *Drosophila* genes, but an increase in the gene body, which indicated that the pause release, was observed only at a subset of genes [83, 108, 109]. This finding confirms the function of NELF in establishing the promoter-proximal pause genome-wide, but, interestingly, the depletion of NELF is not sufficient to release Pol II into the gene body at all genes. Recent experiments in mammalian cells, using the rapid targeted protein degradation strategy, demonstrated that upon NELF degradation, Pol II is not released into the gene body because it is stalled at the proposed second pause site, around the +1 nucleosomal dyad-associated region [110]. The authors showed that the first pause, at the entrance of the +1 nucleosome, is regulated by NELF. However, following the loss of NELF, Pol II continues elongating regardless of P-TEFb activity (described in the next paragraph) until the +1 nucleosomal dyad, where it pauses, as evidenced by accumulated Pol II occupancy signal.

1.1.2.2 Promoter-proximal pause release

P-TEFb and BRD4

Release from the promoter-proximal pause requires the positive transcription elongation factor b (P-TEFb) [111], which is recruited to chromatin by TFs, Mediator, and coactivators [106]. In a conventional view, an elongation factor BRD4 and other BET family proteins bind acetylated histones and recruit P-TEFb [112]. However, this view was challenged by recent studies showing that, following a rapid BRD4 degradation, the P-TEFb occupancy was not impacted [112]. BRD4 was shown to interact with the elongation complex (EC) components and to be required for PAF binding (discussed further) [113]. Its loss led to a defect in the promoter-proximal pause release [113-115] and a global decrease in RNA synthesis [114] These findings indicate that BRD4 itself plays a major role in the release of Pol II into the gene body. Additionally, BRD4 was shown to recruit the 3' end RNA processing factors to Pol II in the promoter-proximal region [113].

P-TEFb consists of cyclin-dependent kinase CDK9 and cyclin T1/2/K [116]. CDK9 phosphorylates the CTD of Pol II subunit RPB1, SPT5, and NELF-E [106]. The majority of P-TEFb is recruited as a constituent of the superelongation complex (SEC), containing a wide assemblage of factors that facilitate the recruitment of P-TEFb to the specific target genes [106, 117]. SEC components include a stabilizing factor ALF Transcription Elongation Factor 4 (AFF4), eleven-nineteen lysine-rich leukemia (ELL) protein, and mixed lineage leukemia (MLL) translocation partners [118-120]. P-TEFb is inactive when bound by 7SK non-coding RNA and HEXIM1/2 [121], which can dissociate from P-TEFb upon cyclin T1 acetylation by P300 [122].

In our current understanding, the phosphorylation of SPT5 is what directly leads to promoter-proximal pause release. This phosphorylation causes NELF to dissociate and prevents its re-binding to Pol II (reviewed in [79, 106]). CDK9 phosphorylates SPT5 at the C-terminal repeat (CTR) domain and the Kyprides, Ouzounis, Woese (KOW) 4-KOW5 linker region [123], dephosphorylation of the latter was shown to lead to premature transcription termination [124]. Upon phosphorylation at its CTR, SPT5 stimulates elongation [125, 126] and stays associated with Pol II throughout the elongation phase [127, 128].

Phosphorylation at Ser2 of RPB1 CTD by P-TEFb was observed as a prominent peak immediately downstream of the TSS by ChIP-seq in mammalian cells and P-TEFb can phosphorylate RPB1 CTD at Ser5, Ser2 [54], and the linker to CTD, facilitating the re-activation of the paused elongation complex, by recruitment of additional elongation factors that help Pol II transcribe through obstacles downstream of the promoter-proximal pause, including nucleosomes and pause-predisposing *cis*-regulatory elements (reviewed in [18]).

1.1.2.3 Subsequent allosteric stimulation of elongation by additional elongation factors, PAF and SPT6

Soon after pause release, the EC encounters a +1 nucleosome, on average located around 140 bp from TSS, which is a known barrier to transcription because of the tight interaction of DNA with histones (reviewed in [106]). Recent studies proposed that the first nucleosome is a second pause site EC experiences, while the first promoter-proximal pause gives an opportunity for Pol II to assemble the necessary elongation factors to overcome nucleosomes [110, 129].

Polymerase-associated factor (PAF) complex consists of PAF1, LEO1, CTR9, WDR61, CDC73 subunits and a dissociable RTF1 subunit [130]. PAF and NELF binding appears to be mutually exclusive [95, 100]. Upon dissociation of NELF during pause release, PAF replaces NELF at the funnel domain of Pol II. The elongation stimulation by PAF binding is attributed to a conformational change in the DSIF DNA clamp, possibly, promoting the unwinding of upstream DNA [100]. Whole genome analysis revealed that PAF binds to Pol II when the EC encounters the +1 nucleosome and remains associated throughout the gene body (reviewed in [106]). In human cells, depletion of PAF1 resulted in reduced Pol II processivity [131] and depletion of RTF1 led to a reduced elongation velocity [129].

Additionally, EC is accessorized by SPT6, which binds the P-TEFb-phosphorylated linker to CTD of RPB1 and opens the RNA clamp formed by DSIF [99, 100, 132]. This conformational change is thought to facilitate the passage of RNA through the exit channel, thus stimulating elongation. SPT6 binds to EC in the promoter-proximal region and remains associated with it through elongation, and its depletion results in the accumulation of Pol II at the +1 nucleosome [129] and globally compromises processivity and elongation velocity [133]. SPT6 is also a histone chaperone [134] and, together with another histone chaperone FACT, maintains chromatin structure, when EC transcribes through nucleosomes [135]. Additionally, the transcription elongation factor IIS (TFIIS) also alleviates transcription through nucleosomes (reviewed in [79]), discussed in detail in section 1.2.

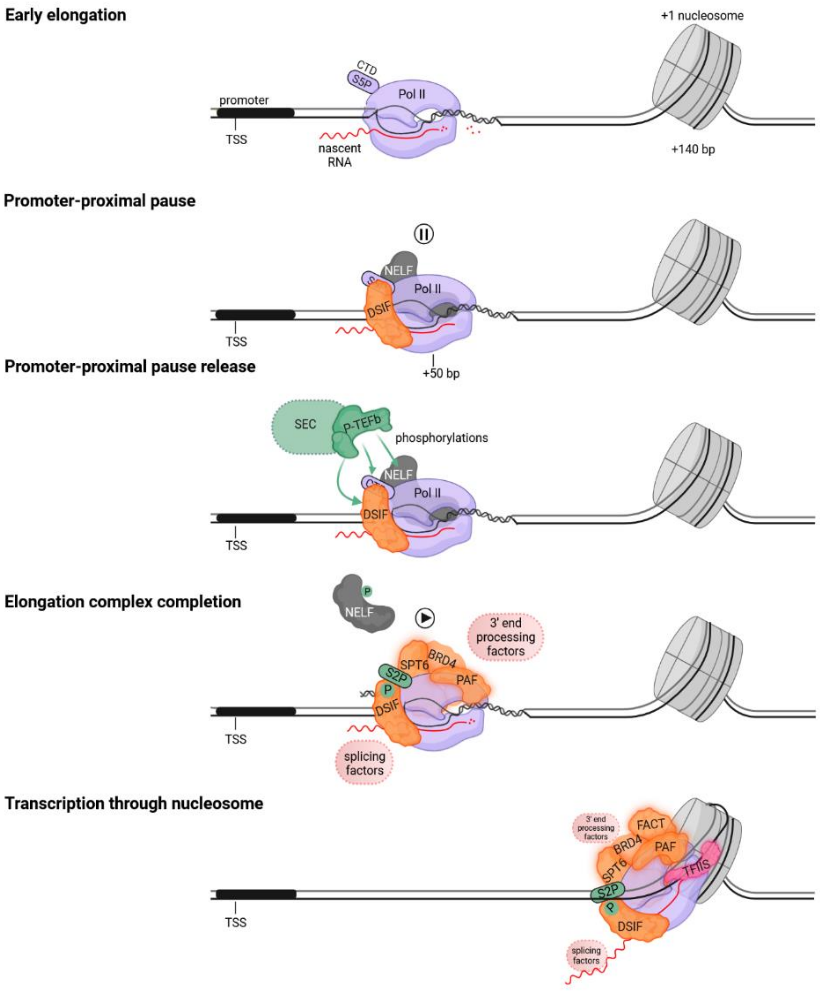


Figure 2: Schematic overview of elongation at the promoter-proximal region. Soon after the promoter escape, DSIF stabilizes Pol II on the DNA strand, however, next, the negative elongation NELF complex recognizes and binds DSIF-Pol II, causing conformational changes, which disable downstream translocation of Pol II. This paused Pol II is re-activated by phosphorylations of NELF, DSIF, and Ser2 of Pol II CTD carried out by P-TEFb. Phosphorylated NELF dissociates from Pol II and cannot rebind, supposedly, making room for TFIIS and PAF. Thanks to phosphorylated SPT5, the elongation is resumed, and the elongation complex is completed by the binding of other elongation factors, including SPT6, FACT, PAF (BRD4-dependent), and TFIIS to help Pol II to transcribe nucleosomal DNA. During this early elongation phase, pre-mRNA splicing factors are recruited by phosphorylated CTD, elongation factors, and the 3'- end RNA processing factors are recruited by BRD4.

1.1.3. Physical intrinsic barriers to Pol II transcription elongation

Transcription through nucleosomes

At highly transcribed genes, nucleosomes can be transiently removed [136, 137], however, the nucleosomal architecture is maintained at the majority of genes and Pol II needs to continue transcribing on them. It was shown *in vitro* that nucleosome-bound DNA is transcribed less efficiently than free DNA [138-140], suggesting that this may also be the case in vivo.

The nucleosome core consists of a central H3/H4 tetramer and flanking H2A/H2B dimers on each side. Nucleosomes undergo remodeling when Pol II transcribes on it: one H2A/H2B dimer dissociates from the rest of the structure [140-143]. Nucleosome dismantling is coupled with reassembly [144]. This is mediated by an elongation factor FACT complex (consisting of two highly conserved subunits SPT16 and SSRP1), which destabilizes the H2A/H2B dimer, but also acts as a histone chaperone and facilitates histone addition to DNA [141]. SPT6 (described above) also promotes transcription through nucleosomes and supports nucleosome reassembly [135]. Highly transcribed genes have an enrichment of the H2A.Z histone variant, which is thought to promote Pol II passage [145, 146].

Interestingly, biochemical analysis revealed that nucleosomes do not constitute a uniform, symmetrical barrier, but it is specifically the tetramer-associated DNA that poses the most challenge for transcription, this site is also known as nucleosome dyad [147]. Cryo-EM structure of Pol II transcribing through a nucleosome uncovered two major pause-inducing sites, coinciding with the tightest DNA-histone interactions: 1) the entry of nucleosome and 2) immediately prior to the dyad [148, 149]. Whole-genome studies profiling of Pol II pausing rendered additional insights. In the gene body, Pol II pausing was detected at the entry to the dyad, whereas in the promoter-proximal regions, Pol II pauses at the entry to the +1 nucleosomes [145]. A recent study demonstrated that upon NELF depletion, Pol II transcribed without stalling through the promoter-proximal pause site and the +1 nucleosome entry site, but accumulated at the dyad entry [110], which is a pause site occurring in the gene body. Supposedly depending on a gene-to-gene basis, the pause at the nucleosome entry can be at least partially a consequence of the *trans*-factor-regulated promoter-proximal pause or possibly be a separate pause enabling additional regulation.

During transcription of nucleosomal DNA, Pol II was observed to experience extensive backtracking *in vitro* [78, 150]. Backtracking is alleviated by the transcription elongation factor IIS (TFIIS). Backtracking occurs when Pol II moves backward to a more stabilizing DNA sequence in cases when forward translocation is thermodynamically unfavorable, described in more detail further (reviewed in [79]). A structural study found that TFIIS addition was necessary to alleviate backtracking at nucleosomes [148]. The global role of TFIIS in overcoming this barrier is yet to be determined *in vivo*.

Hard-to-transcribe DNA sequences

The nucleotide composition of the DNA-RNA hybrid in the active site of Pol II can affect elongation by altering the stability of the ternary complex [148]. In the current model, AT-rich sequences destabilize the ternary complex and Pol II moves back upstream to a thermodynamically more stable template DNA, this is known as Pol II backtracking [79]. Critically, during backtracking, Pol II can become locked in an inactive state, known as transcription arrest, and will require the auxiliary factor TFIIS. The mechanism of how TFIIS relieves Pol II from transcription arrest is described in section 1.2.3.

Some pause-predisposing motifs have been characterized *in vivo*. For example, in *Drosophila,* the GC-rich “pause button” is present at a quarter of promoters with stalled Pol II [151]. In human cells, predominantly, a motif, in which a GC-rich sequence followed by an AT-rich sequence, was shown to contribute to Pol II slowing and pausing [86, 152-155].

Collisions of Pol II with a replisome and other Pol II molecules

Despite transcription and replication being generally temporally separated, Pol II can collide with DNA replication machinery in some cases, for example at the few genes that are transcribed during the S phase of the cell cycle, at very long genes, and at common fragile sites causing replication stress (reviewed in [79]). The direction in which the collisions occur appears to have a different impact [156]. While co-directional collisions do not seem to have a detrimental effect, head-on collisions can compromise genome integrity via R-loop accumulation [157] and enable unplanned recombination [158, 159]. Proper regulation of transcription is crucial to avoid replication stress. For example, the loss of SPT6 caused aberrant transcription of long noncoding RNAs and increased R-loop formation leading to replication stress [159]. Similarly, R-loops caused by a termination and mRNA cleavage defect upon CPSF (cleavage and polyadenylation specificity factor) loss led to replication stress [160].

Usually, a few polymerases transcribe one gene simultaneously. Interestingly, the collisions of polymerases can have either a positive effect on elongation or can be deleterious to transcription, depending on their orientation: co-directional or heads-on (reviewed in [79]). This phenomenon has not been investigated *in vivo* yet, but a few *in vitro* studies demonstrated the outcomes of this event. Based on *in vitro* transcription reactions, co-directional collisions facilitate the overcoming of transient pauses [161]. However, when a downstream Pol II was in arrest, the upstream Pol II, upon bumping into the first one, backtracked extensively and required TFIIS to be reactivated. It was shown that a heads-on collision results in ubiquitylation of both polymerases, indicating a mechanism for resolving this type of elongation blockage at convergent genes. A transcription factor PCF11 is also involved in preventing or removal of collided polymerases [162].

Pausing at DNA damage sites

Elongation can be blocked when bulky lesions occur in the actively transcribed region and at DNA double strand breaks (DSBs) (reviewed in [79]). Ultraviolet (UV)-induced bulky DNA lesions, cyclobutane pyrimidine dimers, on the transcribed strand block Pol II and induce a special DNA damage repair pathway called transcription-coupled nucleotide excision repair (TC-NER) [79, 163]. A stepwise response to UV entails 1) a global release of polymerases into the gene body, supposedly, facilitating the recruitment of TC-NER factors [164], 2) Pol II is stopped and arrested at the DNA damage sites, 3) the polymerases are removed and degraded by the ubiquitin-proteasome system, while TC-NER is activated and the lesions are repaired. This depletion of polymerases during UV-damage response leads to a shutdown of RNA synthesis [79]. The Cockayne syndrome B-associated protein, a DNA translocase RAD26 (also known as ERCC6), is highly relevant to TC-NER, however its necessity is debatable, as the elongation complex itself facilitates TC-NER [79]. When RAD26 binds arrested Pol II, it tries to push it forward, however, forward movement is impossible at the lesion, this causes bending of upstream DNA [165, 166]. This structural conformation at the DNA damage sites has been proposed to lead to the activation of TC-NER and removal of Pol II.

The mechanism of how transcription is repressed at the DSBs is still not completely solved due to the complex coordination of numerous factors in cell cycle phase-temporal manner. In response to DSBs, ataxia telangiectasia mutated (ATM) recognizes the breaks and leads to transient repressive remodeling of chromatin across several kilobases [167-169], therefore, amounting to a barrier to transcription elongation at those sites. However, besides the epigenetic repression of transcription, the elongation is diminished via the direct regulation of the elongation complex. For example, NELF-E is recruited in a PARP1-dependent manner to Pol II at active genes with DSBs [170] and an RNA helicase senataxin (SETX), involved in transcription termination [171], localizes to DSBs at active genes, triggers DNA repair, resolves R-loops, and possibly promoters Pol II removal by termination [172]. Additionally, DNA-PK (DNA protein kinase) facilitates the removal of arrested Pol II via ubiquitylation and proteasomal degradation and ceases RNA synthesis at the genes with DNA damage [173, 174].

1.1.4 Transcription termination and RNA 3' end processing

Transcription termination is the final phase of a transcription cycle and is also carefully regulated. The regulatory mechanisms are not extensively understood yet and are an active area of research. Failure of termination can have a detrimental effect on gene expression of a downstream gene by transcriptional interference or cause aberrant intergenic transcription [175]. Additionally, Pol II needs to be released upon completion of the transcription cycle, so that it can initiate again. Termination is directly linked with RNA 3' end processing, which occurs co-transcriptionally [176, 177]. This process is crucial for the stability and targeting of mRNA to the cytoplasm for translation.

Sequence elements and 3' end processing machinery

In eukaryotes, termination at the end of protein-coding genes is the most characterized thus far. This process requires a polyadenylation signal (PAS), which, as it emerges in the nascent RNA transcript, is recognized by a macromolecular mRNA 3' end processing machinery [178]. Most protein-coding genes have a PAS, consisting of AAUAAA motif, at 69% of human pre-mRNAs, or AUUAAA, at 14% of pre-mRNAs [179]. The PAS-containing mRNAs also often have upstream U-rich and downstream U/GU-rich elements [180].

Following transcription through the PAS, Pol II slows down and pauses, partially because the cleavage and polyadenylation (CPA) complex is recruited (reviewed in [175]). The CPA complex consists of the following subcomplexes 1) the cleavage and polyadenylation specificity factor (CPSF), 2) cleavage stimulatory factor (CstF), 3) cleavage factor Im, and 4) cleavage factor IIm (reviewed in [180, 181]). *In vitro* transcription experiments revealed that CPSF and CstF, not only slow Pol II down, but also enable it to release the template DNA, indicating that the binding of these factors induces conformational changes in Pol II [182].

CPSF has two modules: a polyadenylation specificity module, consisting of CPSF160, CPSF30, FIP1, and WDR33 and a cleavage module, assembled by CPSF73, CPSF100, and symplekin [183]. Zinc-finger containing CPSF30 and WDR33 recognize the transcribed PAS [184-186]. FIP1 binds CPSF30, recruits and regulates PAP1, the poly(A) polymerase [187]. CPSF160 acts as a scaffold to the polyadenylation module. CPSF100 is a flexible tether connecting the two modules [183]. The PAS cleavage is carried out by the endonuclease CPSF73 [188]. CPSF73 belongs to the metallo-β-lactamase superfamily [189]; its β-CASP domain controls the access of two zinc ions to the active site [188]. CPSF73 cleaves pre-mRNA 15 - 30 nt downstream of the PAS after a short CA motif. Next, PAP1 adds about 250 adenosines to the generated free 3' hydroxyl end [190, 191]. Symplekin is not necessary for the assembly of CPSF and recognition of the PAS [192], but it associates with Pol II via PAF complex [193] and was shown to stimulate dephosphorylation of Ser5 of Pol II RPB1 CTD, in yeast [194].

CstF is recruited by CPSF and it is dispensable for polyadenylation, but necessary for cleavage (reviewed in [181]). CstF is a hexameric assembly of CstF50, CstF64, and CstF77, each present in two copies. CstF77 binds CPSF160 [195, 196], while CstF50 binds CTD of Pol II RPB1 [197]. CstF64 dimer recognizes and binds the downstream signal element [183] and interacts with the exonuclease XRN2 [198].

Another core component of the termination machinery is CFIm, which is also a regulator of alternative polyadenylation, a phenomenon occurring at the genes with more than one PAS, usually located in the 3'UTR (untranslated region). This results in an mRNA transcript with an extended 3'UTR, potentially with different regulatory elements that can alter its metabolism, localization, and translation efficiency (reviewed in [199]). CFIm is a heterotetramer and it consists of CFIm59, CFIm68, and CFIm25 homodimer [200]. CFIm25 specifically binds the UGUA element, usually found 20 nt upstream of PAS [201]. CFIm68 interacts with FIP1 [202] and SR splicing factors [203] via its arginine-serine repeat domain. CFIm25 and CFIm59 knock-down experiments showed increased usage of proximal PAS sites [200, 204]. This effect was also observed on a whole-genome level [205, 206].

CFIIm, consisting of CLP1 and PCF11, interacts with nascent RNA and Ser2-phosphorylated Pol II [207]. In *in vitro* experiments, PCF11 evicts Pol II from template DNA and is necessary for efficient termination and degradation of the 3' product of PAS cleavage [208]. In human cells, PCF11 is present at a sub-stoichiometric level to the CPA complex and binds to genes selectively, preferentially between closely-spaced genes, suggesting that PCF11 is a specialized accessory, rather than a core component of the CPA [162].

Although the majority of metazoan protein-coding genes have a PAS, the replication-dependent histone genes are an exception. Instead, they end with a stem-loop structure, which is cleaved by a specialized 3' end processing machinery, including the U7 small nuclear RNP [209], as well as CPSF73 endonuclease [210-212], and symplekin [213].

1.1.5 Co-transcriptional RNA splicing

Precursor messenger RNA (pre-mRNA) undergoes specific processing steps to become a functional mature mRNA and to be transported into the cytoplasm for translation. RNA processing is mediated by hundreds of proteins and includes 5' end capping, 3' end RNA processing (briefly described earlier) and pre-mRNA splicing. pre-mRNA is mainly spliced during transcription [214], as introns were observed to be spliced out as soon as the nascent transcript emerges from Pol II [215]. However, this is not always the case, as a recent study revealed that more downstream exons are frequently spliced first [216]. The coordination between transcription and splicing is further strengthened by the evidence of physical interactions between the transcription and splicing machineries [217]. Splicing is precisely and dynamically regulated by an enormous number of factors. How splicing of nuclear pre-mRNA is mediated is briefly summarized below.

The exon-intron boundaries are defined by conserved dinucleotide sequences GU at the 5' splice site (SS) and AG at the 3' SS, a branch point sequence (BPS), located in the intron, 18 - 40 nt upstream of the 3' SS, and a polypyrimidine tract downstream of the BPS (reviewed in [218]). During pre-mRNA splicing, an intron is removed and the two exons are joined via two transesterification reactions: 1) the 2'-hydroxyl group of an adenosine of the BPS in the intron attacks the phosphodiester bond at the 5' SS, resulting in a free 5' exon and an intron lariat-3' exon, 2) the 3'-hydroxyl group of the 5' exon attacks the phosphodiester bond at the 3' SS, resulting excision of intron lariat and ligation of the exons. The splicing of nuclear pre-mRNA is accomplished by the spliceosome and dynamically coordinated by a plethora of *trans*-acting factors, together comprising the splicing machinery and ensuring splicing precision. In higher eukaryotes, while the SS sequences are very short, simple and poorly conserved, it is the network of numerous splicing factors that enable frequent alternative splicing, rendering multiple protein isoforms (reviewed in [217, 218]).

The major spliceosome consists of the core U1, U2, U5, U4/U6 snRNPs and over two hundred accessory proteins [218, 219]. The spliceosome assembles in a stepwise manner (reviewed in [218, 220]). In a simplified summary, the assembly begins with the U1 snRNP binding to the 5' SS, by base-pairing of the U1 snRNA, while this interaction is stabilized by additional non-spliceosomal splicing factors, SF1 and U2AF, which cooperatively bind the BPS and the polypyrimidine tract, respectively. U2AF subunit U2AF35 also binds the 3' SS, forming spliceosomal E complex. Next, U2snRNP binds the BPS by base-pairing interaction, forming complex A. This interaction is stabilized by SF3a, SF3b, and a U2AF subunit U2AF65. Upon U2 snRNP binding, SF1 dissociates and is replaced by SF3b14. Next, the pre-assembled U4/U6.U5 tri-snRNP is recruited, forming a catalytically inactive complex B, which undergoes massive compositional and conformational changes. During activation, U1 and U4 snRNPs dissociate together with dozens of associated splicing factors, while some new ones, such as PRP19 complex, get recruited and catalytically activate the spliceosome. The activated B complex carries out the first transesterification reaction (described above), resulting in complex C, which undergoes additional catalytic rearrangements and catalyzes the second reaction. Then the spliceosome dissociates from RNA, disassembles and is reused in further splicing cycles.

Importantly, not only the *cis*-splicing elements and splicing factors regulate splicing, but Pol II itself is also functionally involved in splicing coordination: its phosphorylated CTD serves as an interaction platform for RNA processing factors [217]. For example, Ser5-phosphorylated CTD interacts with U1 snRNP [221, 222] and Ser2-phosphorylated CTD appears to recruit U2AF65 [223]. Additionally, the crosstalk between the transcription and splicing machineries is mediated by the physical interactions splicing factors with GTFs and elongation factors [217].

Additionally, the elongation rate can also influence splicing site choices. As Pol II transcribes a gene, only a portion of pre-mRNA is available for recognition by the spliceosome. Therefore, the rate of transcript synthesis affects the SS availability and recognition by splicing factors (reviewed in [217]. Elongation rate of Pol II depends on many factors, such as nucleosomes, the DNA sequences, elongation factors, and chromatin structure [79, 224]. Increased Pol II pausing was detected at the 5’ and 3’ SSs, indicative of a potential regulation point, and more pausing was detected at alternatively retained than skipped exons [89, 225], in agreement with the kinetic model. Slow elongation was shown to facilitate inclusion of alternative exons: weaker alternative SSs upstream of constitutive SSs were given a window of opportunity to be used [226, 227]. However, this is not always the case, as, during slow elongation, alternative exon skipping was preferred thanks to the binding of a splicing silencer [228]. Furthermore, slow and fast elongation resulted in both increased and decreased inclusion of the same exons and introns, indicating that the co-transcriptional splicing regulation is complex and an optimal elongation rate is necessary for a proper splicing pattern [229]. Additionally, in human cells, splicing does not always follow the order of transcribed exons, however it occurs in a defined order, such that splicing of two neighboring introns appears to be coordinated [216].

The splicing machinery components must be somehow retained close to the actively transcribed units to allow for co-transcriptional RNA processing. Many experiments have proposed that this proximity is mediated by liquid-liquid phase separation (LLPS) enabled by the intrinsically disordered regions (IDRs) of many splicing factors, which can congregate into membrane-less granules (reviewed in [217]). These structures are proposed to provide additional regulation of transcription and splicing. Recently, phosphorylation of CTD was demonstrated to cause Pol II to transition from the initiation mediator condensate to the condensate consisting of splicing factors at genes regulated by super-enhancers [230].

1.2 Transcription elongation factor IIS (TFIIS)

TFIIS is a positive elongation factor, which alleviates Pol II backtracking when it encounters physical barriers, described earlier. TFIIS itself is not an enzyme, but it stimulates the intrinsic nucleolytic cleavage of Pol II. The mechanism by which TFIIS relieves backtracked Pol II was vividly demonstrated by structural and *in vitro* transcription studies. This chapter describes the structure, molecular mechanism, and evolution of TFIIS.

1.2.1 Discovery of TFIIS and early insights from *in vitro* transcription experiments

Transcription factor IIS (TFIIS) was the first purified Pol II-associating transcription factor. It was extracted from Ehrlich ascites tumor cells in 1973 [231]. TFIIS was extensively studied *in vitro*, which elucidated its role as an elongation factor because its addition to the reconstituted transcription reactions stimulated RNA synthesis [232, 233]. Thorough biochemical analyses revealed that TFIIS potently enhances the ribonuclease activity of Pol II [234-236].

TFIIS was shown to help Pol II with “reading through” a transcription block, an intrinsic terminator/arrest site from histone H3.3 gene constructed on a template [237]. Further biochemical experiments were extremely insightful. Pol II was discovered to have a weak 3’ to 5’ exonuclease activity, in addition to its main function of phosphodiester bond formation catalyzation [234-236]. Pol II was shown to cleave the nascent RNA within the arrested ternary complex [234, 235]. The 3’ end of the transcript remained in the ternary complex and was extended. The catalytic site of Pol II was proposed to be repositioned in a way to enable addition of nucleotides (nt) to the transcript at the correct location on the template DNA [235]. This activity was shown to depend on the presence of TFIIS and a divalent metal ion [234, 235]. This nuclease activity was proposed to serve a proofreading function [238].

The length of the cleaved RNA was also investigated [239, 240]. Interestingly, the addition of TFIIS to *in vitro* reactions where Pol II was arrested at intrinsic arrest sites resulted in a release of RNA oligos that were 7 - 14 nt long. Whereas, in the reaction where one NTP was missing, thereby causing Pol II to stall, rather than arrest, the released fragments were mostly dinucleotides (5’-phosphodinucleotides), in some cases trinucleotides. Another study confirmed that 7 - 9 nucleotide-long RNA fragments are released from arrested ternary complexes following TFIIS addition, while dinucleotides were released from paused elongation complexes, however 6-nucleotide long oligos were also released from not-arrested complexes [241]. These observations indicated that Pol II backtracks at arrest sites, but it also facilitates cleavage at pause sites. Later, TFIIS was shown to increase Pol II velocity not generally, but by reducing the life-time of pauses [242, 243].

1.2.2 TFIIS structure

TFIIS is a highly conserved protein, its homologs can be found even in archaea and in some viral genomes. It is composed of three stable domains: the N-terminal (I), central (II), and C-terminal (III) domains [244]. The N-terminal domain (NTD) has a structured region and an unstructured part. The structured part forms a compact four-helix bundle [245]. More recent structural analysis specified that it consists of five alpha-helices [246]. The central domain is a three-helix bundle [244, 247], the end of the second domain is more flexible and comprises the linker region [248]. The C-terminal domain has a zinc-ribbon fold with a protruding β-hairpin [247, 249].

Expression of TFIIS mutants of various lengths in yeast elucidated which parts of the protein are required for binding to Pol II and RNA cleavage [250]. In yeast, TFIIS knockout is viable, without any obvious defects [251], except sensitivity to oxidative stress [252] and to 6-azauracil [250], a growth inhibitor which depletes GTP and dUTP nucleotide pools. Domains II and III were sufficient to reverse the 6-azauracil sensitivity and relieve backtracked Pol II. Structural analysis showed that the central domain with the linker are required for binding to Pol II [244, 245, 248, 249, 253], while the C-terminal domain is required for nucleic acid recognition and stimulation of nucleolytic cleavage [247, 253]. More specifically, the aspartic and glutamic acid residues in the zinc ribbon of the C-terminal domain are essential for RNA cleavage, which occurs via a two-metal ion mechanism, described further [254].

Although the NTD appears to be dispensable for the function of TFIIS in elongation, its role appears to be rather important. The NTD was shown to be required for interaction with Pol II holoenzyme, containing GTFs, in an affinity chromatography experiment [255]. The potential involvement of the NTD in transcription initiation was further suggested based on *in vitro* experiments showing that the NTD and central domain are needed during the PIC assembly [256]. Furthermore, the NTD-mediated interaction with the Mediator complex subunit Med13 and Spt8 of SAGA complex were reported in yeast [257]. Very recently, the NTD of TFIIS was highlighted as a shared structural feature in at least 15 transcription factors in the human proteome [246, 258], confirming some previous observations of structural similarity to: Elongin A [245, 246, 259], CRSP70/MED26 [260], PPP1R10, PIBP, IWS1 [246, 261], LEDGF and HRP2 [246]. The direct role of the NTD of TFIIS in human cells remains to be analyzed.

1.2.3 The mechanism of reactivation of backtracked Pol II by TFIIS

The precise mechanism of how TFIIS binds and stimulates Pol II to cleave backtracked RNA is based on a few landmark structural works that capture paused, arrested, and reactivated Pol II [262-265]. In one system, Pol II was backtracked by only one nucleotide by incorporating a DNA-RNA hybrid with a mismatch at the 3'-end of the RNA, therefore rendering mechanistic insight into mRNA proofreading [264]. Crystallization of Pol II in this short backtrack state revealed changes in the DNA-RNA hybrid helix positioning. The last paired base was slightly tilted out of the plane, while the phosphodiester backbone between the last matched and the mismatched bases was bent 120° from the hybrid helix, toward the bridge helix of Rpb1 of Pol II. This caused changes in contacts with the bridge helix, trigger loop and other residues of Pol II and formed a binding pocket for the backtracked base. This study indicated that backtracking by one nucleotide is a stabilizing state for Pol II, which generally stalls at this position.

During the one-nucleotide backtracking pause, the first backtracked RNA base stacks at Rpb2 tyrosine 769, known as the gating tyrosine [264, 265]. This gating tyrosine appears to define the end of the backtracking, however, supposedly, if the hybrid is weak, the RNA can backtrack beyond the gating tyrosine, resulting in irreversible backtracking and transcriptional arrest. When this happens, TFIIS is required, however TFIIS also assists during one-nucleotide backtracking inducing the release of dinucleotides [264].

Details about transcriptional arrest were solved in another system: extensive backtracking was induced with a specialized DNA template, and the crystallized Pol II was captured with a 6-bp DNA-RNA hybrid, 9 backtracked nucleotides, and 13 bp of downstream DNA [265]. In the elongation complex, the 3' end of nascent RNA matched with the template DNA is located at the position denoted as -1 relative to Pol II active site, while the site where a new NTP incorporates to the template DNA base is referred to as the +1. In the arrested complex, the hybrid is tilted towards the bridge helix, as the -1 base pair was tilted by 25° and its DNA base was moved to the +1 site. The base that is normally at the +1 site is displaced into the downstream cleft, thus +1 RNA base is left unpaired. The backtracked RNA associates with one side of the pore and the mobile trigger loop, the backtrack site [265]. The trigger loop controls the lateral oscillation of Pol II [266]. The backtracked RNA immobilizes the trigger loop in a conformation, distinct from the ones observed during active elongation [263, 267] and pause with a single nucleotide backtracking [264]. The trapped trigger loop constitutes the basis for transcription arrest: during arrest, backtracked RNA is long enough to bind the trigger loop. Both become immobile, thus preventing forward translocation, whereas when backtracking is not extensive, the interactions between the backtracked RNA and the trigger loop are weak enough for Pol II to overcome and continue transcription [265].

Arrested Pol II requires TFIIS to be released from the immobile state. TFIIS achieves this by loosening the grip of Pol II on backtracked RNA and recruiting a second metal ion for hydrolytic RNA cleavage. The TFIIS central domain directly binds at the rim of Pol II funnel domain, at the entrance of the pore and the TFIIS C-terminal domain is inserted into the pore, reaching the active site with the β-hairpin [262, 263]. The crystallization of the reactivation intermediate revealed the specific conformation changes in Pol II caused by TFIIS binding: 1) the trigger loop was physically moved causing it to be switched from the “trapped” to “locked” state [262, 265], 2) the gating tyrosine chain was rotated, and 3) the backtracked RNA was displaced from the backtrack site on the funnel domain to the pore [265, 268].

The two-metal-ion mechanism is thought to occur during both RNA polymerization and cleavage, however instead of the 3’ hydroxyl group, metal A binds the phosphate during the RNA cleavage [269, 270]. While metal A is at the active site of Pol II, the second metal ion (B) and a nucleophilic water molecule are proposed to be recruited and coordinated by TFIIS, specifically by the two essential acidic amino acid residues, aspartic (D290) and glutamic (E291) acids in the C-terminal hairpin [262, 265]. During cleavage, metal A is thought to bind the +1 RNA phosphate [236]. The TFIIS hairpin residues D290 and E291 are in close proximity to the RNA sugar-phosphate backbone [265]. The TFIIS D290 and E291 can bind metal B, while together with R287, they are, supposedly, required for two catalytic proton transfers: proton subtraction from a nucleophilic water molecule and a proton donation to the new 3’ RNA nucleotide [265].

In summary (Figure 3), when Pol II encounters a barrier, at first, it backtracks by one nucleotide to a stable pause position, as further backtracking is averted by the gating tyrosine. Pol II can either cleave the RNA by itself or more efficiently with the help of TFIIS, releasing a dinucleotide. In the case when the DNA-RNA hybrid is weak, further backtracking beyond the gating tyrosine can be thermodynamically favored. The longer backtracked RNA traps the trigger loop in the pore, thereby immobilizing Pol II. Upon binding, TFIIS pushes the trigger loop away and displaces the backtracked RNA from the stable position on the funnel domain into the pore. Besides inducing the conformational changes, TFIIS adds two acidic and one basic residues to the active site of Pol II that facilitate the binding of the second metal ion and recruit a water molecule to enhance the hydrolytic cleavage. Finally, a fresh 3’ end of RNA is generated and Pol II can resume polymerization.

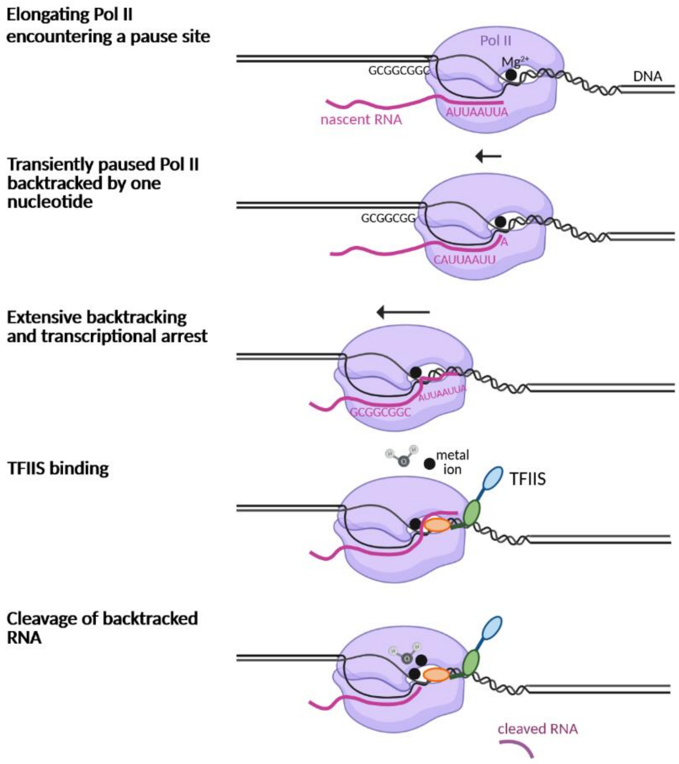

Elongating Pol II encountering a pause site
Pol II
Mg2+
GCGGCGGC
DNA
nascent RNA
AUUAAUUA
Transiently paused Pol II backtracked by one nucleotide
GCGGCGG
A
CAUUAAUU
Extensive backtracking and transcriptional arrest
AUUAAUUA
GCGGCGGC
TFIIS binding
metal ion
TFIIS
Cleavage of backtracked RNA
cleaved RNA

Figure 3: Simplified summary of Pol II arrest and reactivation by TFIIS. The details are described above. The scheme is based on [265]. TFIIS domains I, II, and III are denoted in blue, green, and orange, respectively.

1.2.4. TFIIS homologs

The notion of TFIIS as a highly conserved transcription factor is mentioned in almost all publications about TFIIS. TFIIS is an ancient protein and it is conserved in all eukaryotes, and its homolog, TFS, is found in archaea [271]. Bacteria have structurally different factors GreA/GreB that serve an analogous function [272]. “Homologs” is a general term signifying that the genes are related by common descent and can be subdivided into more specialized terms, orthologs and paralogs (reviewed in [273]). Orthologs are genes related via vertical descent, in contrast to paralogs that are related via gene duplication. The TFIIS orthologs, identified in various organisms, showed a striking structural similarity: the N-terminal ~80 amino acids and the C-terminal ~160 amino acids are highly conserved, while the region in between is much less conserved [251, 274-277].

Many early functional assays were performed in yeast and *in vitro*, at the same time, in the 1990's, TFIIS orthologous and paralogous genes were detected and characterized [251, 274-284]. These studies converged to form a view that in vertebrates, there is a family of three TFIIS proteins, encoded by *TCEA1*, *TCEA2*, and *TCEA3* with distinct tissue expression patterns. TCEA1 is ubiquitously expressed, TCEA2 expression is restricted to testis [280], and TCEA3 is expressed in intestine, heart, testis, kidney, and skeletal muscle [284]. Thanks to the advancement in genome editing, whole genome sequencing, and bioinformatics, we currently have a more complete overview of the human paralogs and compelling questions of their shared and unique roles in human cells.

Currently, in the Ensembl genome browser version 109 [285], an additional gene, *TCEANC*, is annotated as a TFIIS paralog. TCEANC is structurally similar to the rest of the human TFIIS family proteins: it contains three domains, including the C-terminal domain with the two acidic amino acids, responsible for RNA cleavage. Interestingly, the C-terminal domain also has an 11-amino acid stretch, unique to this paralog, which potentially could alter its interaction with Pol II. The aligned amino acid sequences of the four human TFIIS paralogs are shown in Figure 4.

1.3 What is currently known about human TFIIS paralogs?

Our notion of the organ-specific expression has also been updated. The Genotype-Tissue Expression (GTEx) database [286] confirms that *TCEA1* is generally expressed across all organs, while *TCEA2* is not only a testis-specific paralog, but it is even higher expressed in the cerebellum. *TCEA3* is predominantly expressed in skeletal muscle, but also in other organs at a considerably lower level, and *TCEANC* is ubiquitously, but lowly expressed. In mammalian cells, gene loss and depletion experiments were performed over the recent years and resulted in interesting observations sparking curiosity. However, studies aiming to elucidate the role of each TFIIS paralog are quite sparse. While *TCEA1* and *TCEA3* depletion was performed (described further in this chapter), no functional analyses were done to investigate TCEA2 and TCEANC. Based on high structural similarity between the paralogs, one can be inclined to think that the factors are functionally redundant. Nonetheless, tissue-restricted expression suggests that the paralogs are differentially regulated and may have different roles.

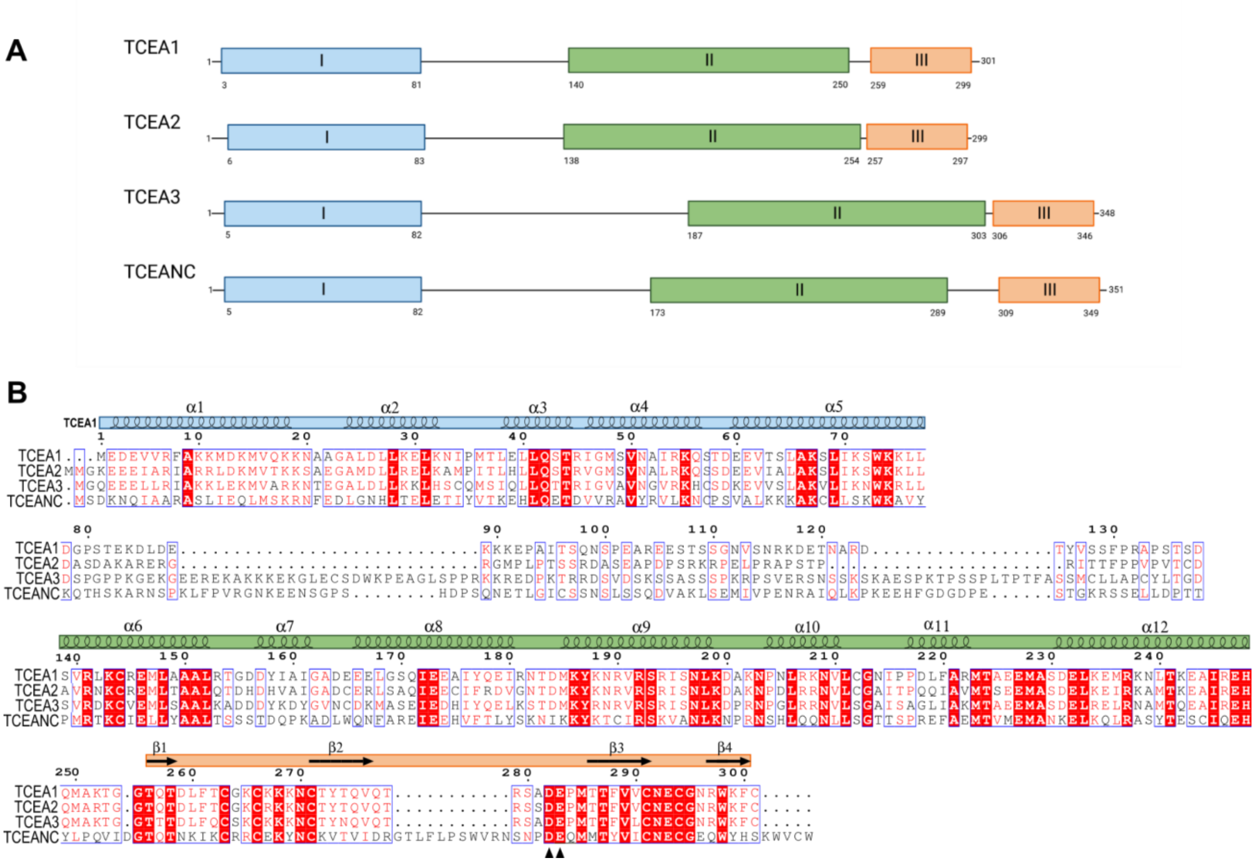


Figure 4. Human TFIIS (TCEA) paralogs.
A. Schematic of domain architecture of the TCEA paralogs. The domains are annotated based on the UniProt database [287], except TCEANC: the domain III borders were predicted based on the following aminoacid sequence alignment with the other paralogs; **B.** Amino acid sequence alignment of the paralogs. The amino acid sequences of human TCEA1, TCEA2, TCEA3 and TCEANC were aligned in Clustal W [288] and analyzed in ESPript 3.0 [289]. The predicted secondary structures are denoted as loops for α-helix and an arrow for a β-sheet. The domain architecture was added based on TCEA1 structures available in the Protein Data Bank. The N-terminal, central and C-terminal domains are denoted in blue, green, and orange bars, respectively. The region between the N-terminal and central domain is unstructured. The two acidic amino acids (D and E), responsible for RNA cleavage, are marked with black triangles. The identical residues are in red blocks. The residues in the red font are similar. The black residues are not conserved.

1.3.1 TCEA1

TCEA1 gene disruption revealed that this paralog is essential for definitive hematopoiesis [290]. Mouse embryos with one inactivated *TCEA1* allele were viable and did not have any obvious abnormalities, however homozygous null mutants died between E13.5 and E16.5. On E13.5, these embryos were overall smaller and had liver hypoplasia and pericardial edema. TCEA1 was shown to be not required for the generation of defined lineage-committed hematopoietic stem cells, but crucial for their growth in the fetal liver. Furthermore, differentiation into erythrocytes was impaired without TCEA1. Defected erythropoiesis was attributed to reduced expression of an antiapoptotic factor gene *Bcl-x*$_L$, the transcription of which was shown to be stimulated by TCEA1. The authors hypothesized that TCEA1 induces the transcription of *Bcl-x*$_L$, possibly with another transactivator. The same research group previously identified a direct interaction between the N-terminal domain of TCEA1 with two isoforms of FESTA using the yeast two-hybrid system [291]. FESTA, presently known as ELL associated factor 2 (EAF2), was shown to have a transcription activating ability and to be expressed in spleen and kidney. Together, these observations suggested a mechanism of how TCEA1 can stimulate elongation selectively.

A recent shRNA screen identified TCEA1 as a potential regulator of myeloid cell fate [292]. Interestingly, *TCEA1* knockdown in mouse myeloid cells (32Dcl3) increased the expression of myeloblast and promyelocyte markers, stimulated cellular proliferation, and inhibited apoptosis. However, the mRNA expression level of the differentiation markers was reduced and a differentiation block was also observed at the level of cellular morphology. These results indicate that TCEA1 regulates gene expression selectively via a mechanism that is presently not known.

Another study provided evidence that knockdown of *TCEA1* inhibits growth and proliferation of breast, lung, and pancreatic cancer cell lines [293]. Relevance of TCEA1 in breast cancer was further investigated because the reduced cell growth was more severe in cancerous MCF7 cells than in noncancerous MCF10A cells. Following the TCEA1 knockdown, c-myc/p53 and β-estradiol pathways were affected, but variably. Reduction of TCEA1 resulted in the elevation of the oncogenic c-myc and tumor suppressor p53 in MCF7, but not in MCF10A. Whereas, the mitogen-activated protein kinase (MAPK) signal transduction pathway was disrupted in MCF10A cells only. TCEA1 knockdown induced the expression of some estradiol-induced genes, which, supposedly, led to apoptosis in MCF7 cells. The differential impact on target genes of the pathways was proposed to be the cause of varying growth phenotype of cancer and noncancerous cells. Additionally, TCEA1 was also reported to be relevant to hepatocellular carcinoma as a target gene of YEATS4 oncogene [294]. Upregulated TCEA1 appeared to stabilize the RNA helicase DDX3 at the protein level, consequently leading to tumor proliferation.

1.3.2 TCEA2

To our knowledge, no gene loss experiments were performed to analyze the function of TCEA2 in transcription and cell growth. Although TCEA2 was described as a highly expressed TFIIS paralog in spermatocytes and ovaries more than 25 years ago [277, 281, 295], its biological significance is still unknown. In one report, in the yeast two-hybrid system, TCEA2 interacted with a testis-specific glutamate receptor-interacting protein isoform GRIP1τ [296]. TCEA2 was shown to stimulate the transactivating property of GRIP1τ in a reporter assay.

More recently, TCEA2 and TCEANC emerged in the context of DNA damage repair as interactors of a major tumor suppressor BRCA1 in a protein-protein interaction screening [297]. The study focused on BRCA1 involvement in transcription-associated DNA damage, as BRCA1-deficient cells had more DNA damage caused by transcription arrest induced by α-amanitin and DRB (5,6-dichlorobenzimidazole 1-β-D-ribofuranoside). Suppressing TCEA2 and TCEANC expression in cells with only one functional BRCA1 allele resulted in reduction of cell colonies, confirming that the interaction between BRCA1 and TCEA2/TCEANC has a physiological impact on cells. Further analysis of TCEA2 and TCEANC was not possible due to the lack of suitable antibodies, therefore only TCEA1 was investigated. Therefore, it is not clear if TCEA2 has an identical role to TCEA1 in this context, and, intriguingly, the study highlighted that the *TCEA2* locus is found among those that are amplified in breast and ovarian tumors, potentially contributing to tumorigenesis. Using UV to induce transcription-associated DNA damage and immunofluorescence as a readout, BRCA1 was found co-localizing with Pol II and TCEA1 at the sites of DNA damage. This observation suggested that BRCA1 is recruited to the DNA damage sites caused by transcription arrest and can act in a multifunctional way by directly or indirectly preventing R-loop accumulation, recruiting other DNA repair factors, and assisting in restart of transcription. As previously described, TCEA1 depletion dramatically reduced cell growth and induced apoptosis in a breast cancer cell line [293], further investigation of the role of TCEA2 in the context of breast and ovarian cancers will be a compelling direction with a therapeutic potential.

1.3.3 TCEA3

During the last ten years, interesting observations were collected about TCEA3. In contrast to *TCEA1* and *TCEA2*, *TCEA3* is prevalently expressed in mouse embryonic stem cells (mESCs) and oocytes [298]. Upon induced differentiation *in vitro*, overexpressed *TCEA3* maintained pluripotency, while *TCEA3* knockdown led to a more rapid differentiation than WT mESCs with an upregulation of mesodermal and endodermal markers [298]. Altered TCEA3 expression had a direct impact on the TGF-β family member *Lefty1* expression, which is a negative regulator of Nodal signaling. These results indicated that TCEA3 plays a crucial role in regulating cell fate commitment as an upstream regulator of the Lefty1-Nodal-Smad2 pathway. Furthermore, TCEA3 in mESCs knockdown led to an upregulation of vasculogenesis-promoting genes and highly vascularized skin in chimeric embryos [299].

TCEA3 was also proposed to play a role in myogenesis, however it has a differentiation-promoting role instead of a suppressive one. In one report, *TCEA3* was among the genes that were upregulated during human myocardial structure development [300]. In an *in vitro* differentiation system of bovine muscle-derived satellite cells, overexpression of TCEA3 drove the growth and elongation of myotubes and elevated the protein level of myogenin and MYH3, muscle cell differentiation-related factors [301]. Whereas, TCEA3 knockdown had the opposite effect. A recent study demonstrated a regulatory circuit of TCEA3 [302]. Myogenin directly binds the promoter of *TCEA3* and recruits Pol II. Upon TCEA3 knockdown and differentiation induction, C2C12 myoblast cells failed to differentiate and form myotubes. Interestingly, TCEA1, also expressed in the cell line, was not able to compensate for the depleted TCEA3, indicating that the paralogs have unique and differential roles. As a feed forward loop, TCEA3 was shown to stimulate the expression of myogenic regulatory factor genes, including myogenin itself and its activating precursor MYOD1, and to directly associate with both factors. The recruitment of TCEA3 to these genes was shown to be dependent on myogenin.

TCEA3 has also been brought forward in the context of cancer. TCEA3 is highly expressed in noncancerous ovarian epithelial cells, in contrast to ovarian cancer cells [303]. Interestingly, TCEA3 knockdown in noncancerous ovarian epithelial cells stimulated cellular growth, while ectopic TCEA3 expression in ovarian cancer cells triggered caspase-dependent apoptosis, suggesting that TCEA3 is involved in regulating apoptosis, specifically in cancer cells. The authors determined that apoptosis is induced by TCEA3 interaction with TGFβ receptor I, thereby activating TGFβ signaling. The role of TCEA3 as a tumor suppressor is further supported by other experiments. *TCEA3* was also reported as a target gene of TP53 in colon cancer HCT116 cells. TCEA3, in turn, stimulates the expression of apoptosis regulator *BAX* gene, thereby promoting cell death [304]. Overexpression of TCEA2 did not increase *BAX* expression, showing that this role is unique to TCEA3. Overexpression of TCEA3 inhibited the growth of HCT116, lung cancer H1299, and osteosarcoma U2OS cells, indicating that TCEA3 acts as a tumor suppressor. Accordingly, TCEA3 is downregulated in human gastric cancers, while another study showed that TCEA3 has an antiproliferative effect in gastric cancer cells [305].

1.3.4. Whole genome profiling of Pol II backtracking sites

The direct role of TFIIS in transcription *in vivo* was demonstrated in *Drosophila*, as an elongation factor acting in the promoter-proximal region alleviating Pol II pausing and backtracking [86, 306]. TFIIS is required at the heat shock protein gene *hsp70* for promoter release and recruitment of additional polymerases [306]. The global short RNA analysis upon TFIIS depletion by RNA interference revealed that the transcripts, in general, lengthened and 35 - 60 nt RNAs were particularly enriched, while 20 - 35 nt transcripts were reduced [86]. This finding rendered identification of backtracking sites that, under normal conditions, are resolved by TFIIS. This effect was characteristic to the genes with Pol II pausing in the promoter-proximal region, indicating that pausing usually co-occurs with backtracking. In summary, both studies found that Pol II is prone to backtracking during early elongation.

Backtracking was also profiled genome-wide in yeast by knocking out *Dst1* (TFIIS gene) and performing the native elongating transcript sequencing, NET-seq [307]. In the absence of TFIIS, the majority of pauses shifted downstream by 8 - 15 nt, with a strong preference to thymidine as the last nucleotide of the pause. The pauses with backtracking were detected throughout the whole genes in yeast, not only in the promoter-proximal region. The study also revealed that the pause peak intensity was the highest right before the nucleosomal dyad, directly showing that nucleosomes are barriers to transcription causing Pol II to backtrack.

In human cells, instead of a TCEA depletion and deletion, an induced overexpression of a mutant TFIIS, inducing transcription arrest, was utilized to get insights into the consequences of this perturbation genome-wide. Two recent reports used the same cellular system, derivative of HEK293 cells, to identify backtracking sites by overexpressing a TFIIS mutant, incapable of enhancing the nucleolytic cleavage, and quantitatively tracking RNA synthesis [154, 308]. In the TFIIS (TCEA1) mutant, the amino acid residues, aspartic and glutamic acids, essential for RNA cleavage in the C-terminal domain were replaced by alanine, therefore the protein was able to be recruited, bind stalled Pol II, and inhibit nucleolytic cleavage of Pol II. In this experimental setup, the mutant TCEA1 competed with the endogenous TCEA1 and TCEA2, therefore the increased backtracking phenotype is partial.

In the first study, Sheridan et al. determined that TFIIS (TCEA1) binds at the 5’ pause sites and downstream of polyadenylation sites (PAS) [154]. They performed the global run-on sequencing (GRO-seq) [88], native elongating transcript sequencing technology for mammalian chromatin (mNET-seq) [90] and Ser2p Pol II ChIP-seq to profile actively elongating and paused Pol II upon induction of the backtracking-stabilizing TFIIS mutant. GRO-seq, capturing only the actively elongating transcripts, revealed a decrease in the promoter-proximal region, specifically in the first 300 bp after the TSS. This observation indicated that Pol II experiences backtracking within that region and the functional TFIIS is, in general, needed to resume elongation. Counterintuitively, stabilization of backtracked Pol II by the mutant TFIIS did not lead to an overall accumulation of polymerases in that gene region as shown by Pol II ChIP-seq and mNET-seq. Possibly, the turnover is maintained by premature termination [92, 93]. Additionally, the RNA cleavage of Pol II was also demonstrated to be important for proper and timely elongation in the gene body and transcription of long genes. The elongation rate, determined by synchronizing transcription with DRB treatment and washout, followed by bromouridine (Bru)-seq and Pol II ChIP-seq, reduced by a half in the cells with the mutant TFIIS expression. The RNA cleavage inhibition by the mutant TFIIS also affected transcription at the 3’ ends of genes, observed as an upstream shift in mNET-seq and Pol II ChIP-seq signal, indicative of earlier (more proximal to PAS) termination. Earlier termination was proposed to be caused by the RNA exonuclease XRN2 catching on to Pol II faster because of its prolonged stalling due to the mutant TFIIS binding. Finally, the study also detected the backtracking sites with mNET-seq: upon expression of mutant TFIIS, the mNET-seq signal was extended up to 15 nucleotides downstream.

In the other study, Zatreanu et al. have also assessed transcriptional changes in TFIIS mutant-expressing cells and focused on the consequences of prolonged backtracking on genome stability [308]. TT-seq, a method based on 4-thiouridine (4sU) labeling and separately sequencing only the recently transcribed RNAs [309], revealed that Pol II activity was reduced at the 3’ ends of genes, pointing that elongation and RNA synthesis were impeded. The effect was especially strong at genes longer than 60 kb. The impact on the transcriptional output of genes was not severe, however, the downregulated genes were longer than average. Splicing was also affected in TFIIS-mutant expressing cells, possibly because of the increased pausing during elongation. Interestingly, alternatively spliced last exons were the major splicing change, this observation is in accordance with the TT-seq profile and could be a consequence of earlier termination described in the study above [154]. Zatreanu et al. investigated the transcriptional stress that mutant TFIIS causes. They found an increase in R-loops, which are DNA-RNA hybrids formed by annealing nascent RNA hybrid with template and nontemplate DNA strands. The R-loops were found to be formed in front of Pol II, rather than behind, in an *in vitro* transcription assay. These anterior R-loops were shown to cause DNA damage, observed as elevated P53BP1, γH2A.x, and single and double strand breaks in DNA, therefore, indicating that TFIIS plays a role in preventing genome instability by preventing the formation of such structures.

In summary, both studies profiled Pol II backtracking genome-wide in human cells and provided important insights into the consequences of induced backtracking, such as impaired completion of long transcripts, earlier termination, and genome instability. However, among limitations of the mutant TFIIS overexpression system are 1) the inhibition of Pol II release is partial because the endogenous TCEA1 and TCEA2 compete with the mutant and 2) the induced backtracking is possibly artificially prolonged by the mutant.

1.4 Motivation

Although high-resolution structural studies and *in vitro* transcription assays illustrate in detail the mechanism by which TFIIS alleviates backtracked Pol II, our understanding of the direct biological role of TFIIS *in vivo* is far from complete. Functional analysis of TFIIS in human cells is complicated because there are four paralogs. Their structural similarity led to a general assumption of their functional redundancy despite the intriguing tissue specificity that TCEA2 and TCEA3 have. TCEA2 has sparked our interest because almost nothing is known about its role and recent human organ expression data suggest its potential function in the cerebellum (Figure 5), challenging the established view that *TCEA2* expression is restricted to testis. We aimed to investigate this paralog and compare its function to the ubiquitously expressed TCEA1, which is generally regarded as the prevalent TFIIS protein.

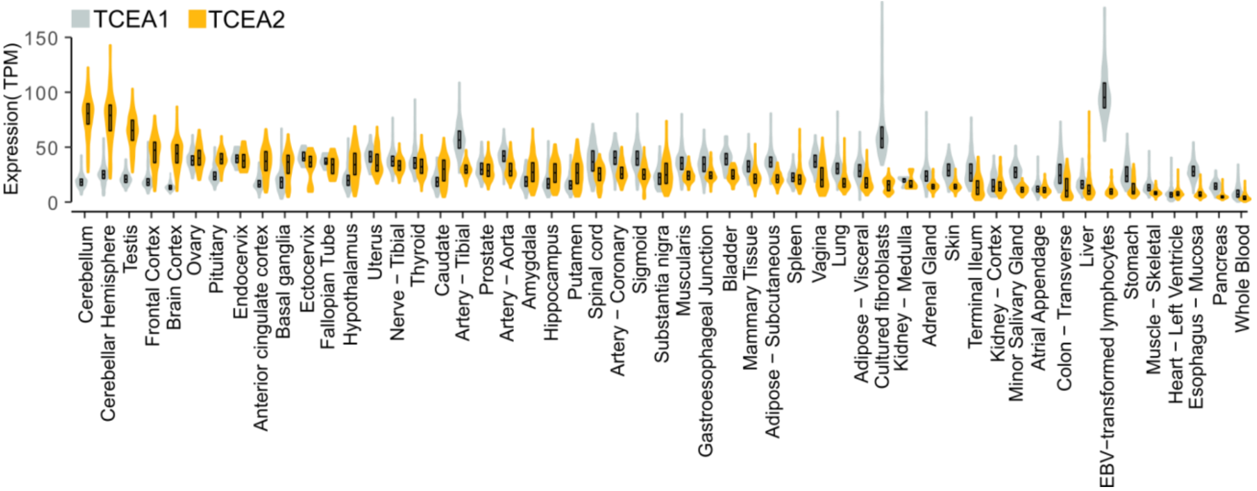


Figure 5. *TCEA2* is expressed higher than *TCEA1* in human brain and testis.
TCEA1 and *TCEA2* mRNA expression across human organs. The mRNA expression data were downloaded from the GTEx database [286], merged on one plot, and arranged in the order of *TCEA2* expression from high to low.

The following questions guided our investigation of TCEA1 and TCEA2:

1. What is the evolutionary history of TFIIS paralogs? Early TFIIS genes characterizing studies that were conducted more than 25 years ago converged to a general view that TCEA1, TCEA2, and TCEA3 exist in vertebrates [284]. Taking advantage of the whole genome databases, we attempted to elucidate when TFIIS diverged into paralogs in the vertebrate lineage.

2. Do TCEA1 and TCEA2 interact with the same interaction partners? TCEA1 is expected to directly interact with Pol II and we wondered whether TCEA2 also associates with Pol II. Additionally, we aimed to capture other interactions of both paralogs and to compare the affinities of the most dissimilar part of the proteins, the N-terminal domain with a flexible linker.

3. Do TCEA1 and TCEA2 relieve Pol II backtracking at the same genomic sites? TCEA1 was shown to predominantly bind at the 5’ and 3’ ends of genes [154], we set out to investigate whether TCEA2 behaves as an elongation factor as well.

4. What are the consequences of TCEA1 and TCEA2 loss on cells? In yeast, TFIIS is not essential under normal growth conditions [251], however TCEA1 depletion experiments in human cancer cells caused reduced cell growth [293] and inhibition of RNA cleavage by inactivated mutant TFIIS also impeded growth and caused DNA damage [308]. We aimed to characterize how the loss of each and both paralogs affects cell growth and transcription.

5. How is Pol II elongation affected by the loss of TCEA1, TCEA2 and both paralogs? Our goal was to determine how each paralog globally regulates elongation by profiling and comparing Pol II occupancy in the absence of each and both paralogs at single nucleotide resolution.

1.5 Experimental approaches to investigate the function of TCEA1 and TCEA2 in human cells

For the functional analysis, we chose to set up all experiments in the human embryonic kidney cell line HEK293T because both TCEA1 and TCEA2 are expressed (Figure 10A in the Results chapter 3.2.2) and these cells are easy to grow and transfect. First, we set out to determine the interactomes of TCEA1 and TCEA2, however we were limited by the lack of paralog-specific antibodies, suitable for immunoprecipitation, and a rather low *TCEA2* expression level. We circumvented these limitations by generating an inducible epitope-tagged TCEA1/2 overexpression system, Flp-In T-Rex system in HEK293 (the principle is described in Methods chapter 2.3.1). We have used the same approach to investigate the interactions that the most dissimilar parts of the proteins, their N-terminal domains, mediate. The interaction partners were determined by crosslink-assisted immunoprecipitation, followed by mass spectrometry (IP-MS). The inducible overexpression of epitope-tagged TCEA1 and TCEA2 also allowed us to resolve the genomic binding profiles with chromatin immunoprecipitation followed by sequencing (ChIP-seq), avoiding the likelihood of low efficiency in terms of signal enrichment because TFIIS associates with chromatin through Pol II, not directly with DNA.

For the gene loss analysis, our initial strategy was the state-of-the-art approach, entailing a CRISPR/Cas9-mediated endogenous insertion of a degron tag, allowing for rapid degradation [310], which would have enabled us to characterize the elongation defects with minimal indirect secondary effects. Due to the known interaction between Pol II and TFIIS, only the N-terminus could be modified. Unfortunately, tagging both genes turned out to be challenging.

The only homozygously tagged TCEA1 clone was not proper because the alleles were repaired slightly differently. *TCEA2* tagging was not successful because the tag was, seemingly, cleaved off and the rest of the protein was not expressed. After a few attempts, we changed our experimental approach to three monoclonal knockout cell lines with gene disruptions in *TCEA1*, *TCEA2*, and both genes. These cell lines experienced intriguing growth defects to different extents, which we characterized with cell cycle analysis and DNA damage response. The impact of the deletions on nascent transcription was profiled with a new version of the native elongating transcript sequencing, rendering higher sensitivity, HiS-NET-seq [311] (Figure 6).

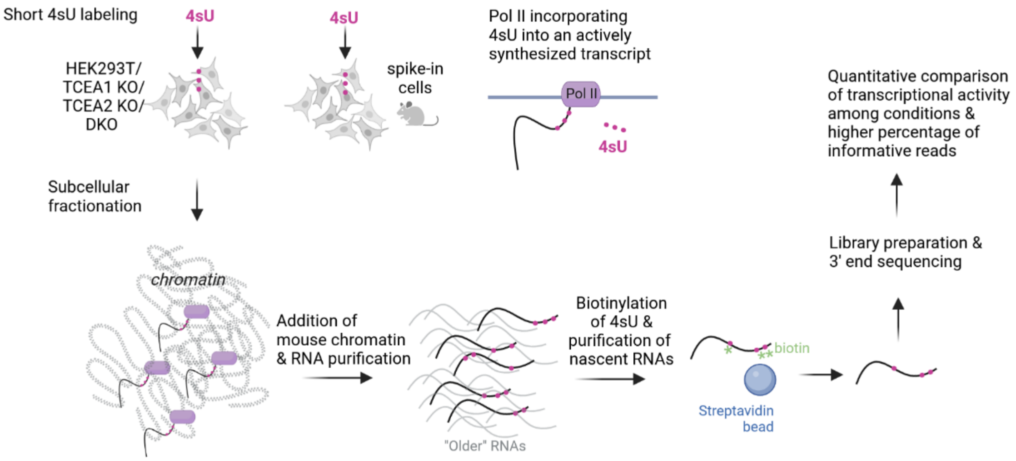


Figure 6. Schematic outline of High-sensitive nascent transcript sequencing (HiS-NET-seq)
The knockout (KO) cell lines, unmodified HEK293T, and mouse spike-in cells are exposed to 4-thiouridine (4sU) for 10 minutes, just enough time to label only the actively transcribed transcripts. Transcription is stopped by lysis in the presence of α-amanitin on ice, the chromatin fraction is isolated and solubilized, and all chromatin RNA is extracted. Next, nascent RNA is enriched by thiol-selective biotinylation, followed by binding to streptavidin microbeads and purified by magnetic separation. The library preparation is performed as in the SI-NET-seq protocol [113]. The scheme is based on [113, 311].

Chapter 2

MATERIALS & METHODS

2.1 *In silico* analyses of TFIIS homologs

2.1.1 Searching TFIIS homologs and TCEA paralogs

This analysis was performed by Johanna Krüger. The TFIIS paralogs were searched using the gene gain and loss function of Ensembl [312] and GenBank [313] databases. The following TFIIS names were searched: TCEA1, TCEA2, TCEA3, TCEANC, TCEA, TFIIS, TF2S, TFS, Transcription elongation factor S-II, Transcription elongation factor A. Upon examination of the list of the organisms, we inferred that the nomenclature changed from “TFS/TF2S” to “TCEA” based on invertebrate to vertebrate classification. Additionally, we found that the diversification into paralogs occurred as early in evolution of jawless fishes, which allowed us to narrow down our further search.

The number of the paralogs was determined using the HMMER software package with default parameters [314]. The input seed alignment was the cDNA alignment of TCEA1-3 of elephant shark calculated using MAFFT (Multiple Alignment using Fast Fourier Transform) (version 7), with default parameters [315]. We chose elephant shark because its genome is well assembled and it is in the evolution window of our interest: between jawless and jawed vertebrates. To assess the sensitivity of the hidden Markov model (HMM), we tested the recognition sensitivity by scanning Genome Reference Consortium Human Build 38 (GRCh38). The HMM successfully detected a distantly related paralog, *TCEANC*, and all four TCEA1 pseudogenes. With the seed input, we scanned the following organisms:

Table 1: Genome assemblies scanned for the presence of TCEA paralogs

Organism name		**Assembly**
Homo sapiens	human	GRCh38
Callorhhinchus milii	elephant shark	6.1.3
Eptatretus burgeri	inshore hagfish	3.2
Petromyzon marinus	sea lamprey	Somatic: 7.0; Germline: Pmar_germline_1.0
Lethenteron camtschaticum	arctic lamprey	LetJap1.0
Brachiostoma belcheri	Belcher’s lancelet	v18h27.r3
Brachiostoma floridae	Florida lancelet	Version 2
Branchiostoma lanceolatum	European lancelet	BraLan2
Ciona intestinalis	Vase tunicate	Hoya T-line assembly 2019

Every alignment was closely examined and, in cases when HMM aligned to an unannotated region in a genome, a region of 60 kbp, upstream and downstream of the aligned sequence, was extracted and tested for any amino acid sequence homologous to TFIIS with GENSCAN [316]. Furthermore, a more general HMM was tested with seed alignment based on a highly conserved exon 8 sequence of *TCEA1* among human and other vertebrates. This model detected *tcea2*, *tcea3*, and *tceanc* genes in two lamprey genomes (*P. marinus* and *L. camtschaticum*) and *tcea2* and *tceanc* in hagfish *E.burgeri.* To confirm that hagfish and lamprey lack *tcea1*, their genomes were scanned using HMM based on lamprey *tcea2* and *tcea3* cDNA alignment. We summarized our observations as a cladogram using Inkscape.

2.1.2 Paralog alignments

The human TCEA1, TCEA2, TCEA3, and TCEANC were aligned in Clustalw with the default settings [288]. The alignment was analyzed and visualized with ESpript 3.0 expert mode [289]. The secondary structure information was added based on the solved domain structures available in the RCSB Protein Data Bank: [249], 3NDQ and [246], which are C-terminal, central, and N-terminal domains, respectively.

2.2 Cell culture

All cell lines were grown at 37°C and 5% CO_2. All used cell lines are semi-adhesive and were maintained as follows: the cells were trypsinized, counted with an automated cell counter, and $2x10^6$ cells were seeded into a T75 flask every 3 days. The medium was changed 36 - 48 hours after plating. For experiments, the cells were seeded at least 36 hours prior. The cells were not kept in culture for longer than 4 weeks and were regularly tested for mycoplasma. All cell lines were gradually frozen in 10% DMSO and 20% fetal bovine serum (FBS) in DMEM, using a cell freezing container and stored in liquid nitrogen.

2.2.1 HEK293T and knockout cell lines

HEK293T, TCEA1KO, TCEA2KO, and the DKO were grown in DMEM, supplemented 1x GlutaMAX, 10% FBS, and 5% penicillin-streptomycin.

2.2.2 Flp-In T-Rex cell lines

Flp-In T-Rex 293 cells were kindly provided by the Aktas laboratory of MPIMG. They were grown in DMEM, 10% FBS, 1x GlutaMAX, 5% penicillin-streptomycin, 100 µg/mL Zeocin, and 15 µg/mL Blasticidin. Flp-In-293-FLAG-TCEA1, -TCEA2, -TCEA1-NTDL, and -TCEA2-NTDL were grown in DMEM, supplemented with 1x GlutaMAX, 5% penicillin-streptomycin, heat-inactivated FBS (incubated at 56°C for 30 min), 15 µg/mL Blasticidin, and 100 µg/mL Hygromycin B HCl. Expression of FLAG-tagged proteins was induced by adding 1 ug/mL Tetracycline (Gibco A39246), pre-mixed into fresh medium, for 24 hours.

2.2.3 NIH3T3 cells

Spike-in mouse NIH3T3 cells were grown in DMEM supplemented with 10% iron-fortified bovine calf serum, and 5% penicillin-streptomycin.

Table 2: Materials and equipment for cell culture

Materials	
DMEM without Glutamine	Gibco 21969-035
DMEM	Gibco 11995065
DMSO	Sigma-Aldrich D8418
DPBS	Gibco 14287080
FBS	Sigma FBS Superior S0615
FBS, iron-fortified	Sigma 12138C
GlutaMax	Gibco 35050-038
PBS	Life Technologies10010-023
TrypLE Express	ThermoFisher Scientific 12605028
Antibiotics	
Blasticidin	Gibco A11139-03
Hygromycin B HCl	Roth 1287.2
Penicillin-streptomycin	Sigma P4458
Zeocin	Gibco R250-01
Equipment	
Automated cell counter	Bio-Rad 1450102
Cell freezing container	Corning CLS432003
Incubator	Thermo Scientific 51030285

2.3 Inducible expression of epitope tagged TCEA1, TCEA2, TCEA1-NTDL, and TCEA2-NTDL

2.3.1 The principle of the Flp-In T-Rex system

The Flp-In T-Rex system entails the integration of the gene of interest (GOI) into a specific genomic location via Flp recombinase-mediated homologous recombination [317] and Tetracycline-regulated expression of the GOI [318]. We used the commercially available host Flp-In T-Rex 293 cell line, derived from HEK293, with the stably integrated 1) *lac*Z-Zeocin fusion gene with Flp recombinase binding and cleavage site (FRT) and 2) highly expressed tetracycline (Tet) repressor gene [319] under a human CMV promoter. The GOI is cloned into an expression vector after a CMV promoter, which also contains 2 copies of the tetracycline operator 2 (TetO2) sequences, to which the Tet repressor will bind [320, 321]. Additionally, the expression vector contains the second FRT and hygromycin resistance gene. The host cells are co-transfected with the GOI expression vector and Flp recombinase expressing vector. Flp recombinase catalyzes recombination between the two FRT sites: the one integrated into the genome of the host cells and the one in the expression vector. The hygromycin resistance gene of the expression vector does not have a promoter and an ATG codon, so only after successful recombination at the FRT, the previously integrated SV40 promoter and ATG codon are separated from *lac*Z by insertion of the hygromycin resistance gene and the GOI. This makes the cells Hygromycin-resistant and zeocin-sensitive, enabling the selection of an isogenic population with the successfully integrated GOI.

In this system, in the absence of tetracycline, the Tet repressor homodimer binds the two Tet operator (TetO2) sequences in the CMV promoter of the integrated GOI with high affinity, thereby repressing its transcription [320]. Once added, tetracycline binds the Tet repressor also with high affinity and induces conformational changes causing its dissociation from TetO2. This allows the GOI to be transcribed.

2.3.2 Cloning the gene of interest

RNA was extracted from HEK293T cells using the RNeasy Mini kit (Qiagen 74104). cDNA was synthesized using QuantiTect Reverse Transcription kit (Qiagen 205311). The TCEA1 and TCEA2 isoforms were PCR amplified with the primers, also containing the stop codon and overhangs to be cloned into the pcDNA5-FRT-TO-HBHT3xFLAG plasmid. The details of the PCR reaction are summarized in the following tables. For amplifying the N-terminal domain and linker (NTDL) regions of TCEA1 and TCEA2, the PCR product above was used as a template, the same forward primers were used, however, the reverse primers were directed to make the truncated version. They included a stop codon and plasmid overhangs.

The PCR product was loaded on a 1% agarose gel and the band of size 942 bp (TCEA1 301 codons + stop codon + plasmid overhangs) and 936 bp (TCEA2 299 codons + stop codon + plasmid overhangs), 456 bp (TCEA1-NTDL 139 codons + stop codon + plasmid overhangs), and 453 (TCEA2-NTDL 137 codons + stop codon + plasmid overhangs) were cut out, and the DNA was purified using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel 740609.250). pOG44 Flp-recombinase expression vector and a modified version of the donor plasmid pcDNA^{TM}5/FRT/TO were kindly provided by Aktas laboratory (MPIMG). The modification entails an insert encoding HBH (histidine-biotin-histidine), TEV cleavage site, and 3xFLAG tag to be joined to the N-terminus of the cloned cDNA upon ligation into the plasmid.

The plasmid was linearized by EcoRV (NEB R3195) digestion at 37°C for 15 min, run on a 1% agarose gel, cut out, and purified as above. Afterward, the linearized plasmid and the amplified cDNA with overhangs were ligated using the Gibson Assembly reaction (NEB E2611) as described by the manufacturer. Stellar chemically competent *E. coli* (Takara 636763) were transformed, single colonies were picked and expanded in 2 mL of luria broth (Invitrogen 12795027) with 1 µg/mL Carbenicillin (Sigma-aldrich C1613). The plasmid DNA was purified with a NucleoSpin Plasmid Transfection-grade kit (Machery-Nagel 740490.250). The insertions were confirmed by Sanger sequencing with a standard sequencing primer pCR3.1-BGH reverse.

Table 3: Primers and PCR reaction for cloning cDNA of TCEA1, TCEA2, TCEA1-NTDL, and TCEA2-NTDL

Primer name	**Sequence** (lower case letters correspond to plasmid overhangs)	**Function**
5'oh_TCEA1_longprimer_F	gacgatgatgacaaggatATGGAGGACGAAGTGGTCCGCTTT	Amplification of TCEA1 cDNA & plasmid overhang addition
3'oh_TCEA1_longprimer_R	ccactgtgctggattgatTCAACAGAACTTCCATCGATTTCC	
TCEA2_299_cDNA_taP2oh	gacgatgatgacaaggatATGATGGGCAAGGAAGAGGAGATT	Amplification of TCEA2 cDNA & plasmid overhang addition
3'oh_TCEA2_longprimer_R	ccactgtgctggattgatTCAGCAGAACTTCCAGCGGTTTC	
TCEA1_Nterm-tap2oh_R	ccactgtgctggattgatTCAAGAATCAGAAGTGCTTGGTGCCCG	Amplification of TCEA1-NTD (in exon 5)
TCEA2_Nterm_tap2_R	ccactgtgctggattgatTCAGGCATCACAGGTGACAGG	Amplification of TCEA2-NTD (in exon 5)

PCR reaction		**PCR cycler program**		
cDNA	400 ng	Initial denaturation	98°C	1 min
Q5 master mix (2X) (NEB M0492S)	25 µL	Denaturation	98°C	X 30
forward primer (10 µM)	2.5 µL	Annealing: TCEA1/TCEA2 TCEA1-NTD/TCEA2-NTD	50°C/72°C 72°C/72°C	
reverse primer (10 µM)	2.5 µL	Extension	72°C	
nuclease-free water	up to 50 uL	Final extension	72°C	5 min

2.3.3 Transfection of Flp-In T-Rex 293 cells

150x10^3 Flp-In T-Rex 293 cells were seeded in a 6-well plate in DMEM, supplemented with 10% FBS, 1x GlutaMAX, 5% penicillin-streptomycin, 100 µg/mL Zeocin and 15 µg/mL Blasticidin. The medium was changed in 48 hours and grown for another 24 hours. Two hours before transfection, the medium was changed to DMEM, supplemented with 10% FBS, 1x GlutaMAX, and no antibiotics were added.

The cells were transfected using Metafectene Pro transfection reagent (Biontex T040-1.0). A total of 2 µg of DNA was used per transfection; the ratio of pOG44 Flp-recombinase vector and pcDNA^{TM}5/FRT/TO-TCEA1, -TCEA2, TCEA1-NTDL, or TCEA2-NTDL was 9:1. The ratio of Metafectene to total DNA was 4:1. The transfection was carried out as instructed by the manufacturer. After 24 hours, the medium was changed: DMEM, 10% heat-inactivated FBS, P/S, and 15 µg/mL Blasticidin. The next day hygromycin selection was started by seeding the cells into 10-cm dishes in DMEM, supplemented with 1x GlutaMAX, 5% penicillin-streptomycin, heat-inactivated FBS, 15 µg/mL Blasticidin, and 100 µg/mL Hygromycin B HCl. The medium was changed for the colonies every two days for 8 days, afterwards, the isogenic colonies were combined and grown in the same composition media and cryopreserved, as described in chapter 2.1. The cell lines with the integrated gene of interest were maintained in 10% of heat-inactivated FBS, 15 µg/mL Blasticidin, 100 µg/mL Hygromycin B HCl, 1x GlutaMAX, 5% penicillin-streptomycin in DMEM without glutamine.

The tetracycline induction of HBH-FLAG-TCEA1, HBH-FLAG-TCEA2, HBH-FLAG-TCEA1-NTDL, and HBH-FLAG-TCEA-NTDL was evaluated by immunoblotting (described in 2.8) with an antibody against the FLAG tag (Sigma F3165) and an antibody against the endogenous TCEA1 (Abcam ab185947), which also detects overexpressed TCEA2.

2.4 CRISPR/Cas9 genome editing for creating monoclonal TCEA1-, TCEA2-, and double TCEA1/TCEA2 knockout cell lines

2.4.1 Design and cloning

All knockout cell lines, except TCEA1 KO exon 2, were generated by plasmid transfection. The choice of target exons to create the knockouts was based on the isoform expression data in HEK293T cells and is shown in the Results section 3.2.2. The single guides (sgRNA) were designed using CRISPOR [322]. The oligos were annealed as previously described [323] and ligated with a Bbs1-linearized Cas9 and sgRNA expression plasmid pSpCas9(BB)-2A-GFP

(PX458) (Addgene 48138) with Gibson assembly (NEB E2611S). Stellar chemically competent *E. coli* (Takara 636763) were transformed as instructed by the manufacturer. Five single colonies per construct were picked, followed by inoculation of 2 mL of luria broth (Invitrogen 12795027) with 1 µg/mL Carbenicillin (Sigma-Aldrich C1613) overnight. Plasmid DNA was purified using NucleoSpin Plasmid Transfection-grade kit (Machery-Nagel 740490) and validated with Sanger sequencing. TCEA1 KO exon 2 cell line was generated by ribonucleoprotein (RNP) complex transfection. The guides were suggested and ordered from Synthego.

Table 4: sgRNA oligos for generating TCEA1/TCEA2 knockouts

Cell line	sgRNA name	Sequence	Strand	Deletion length (bp)	Genomic location of the deletion (GRCh38/hg38)
TCEA2 KO	TCEA2-**exon1**-sg1	GCTGCGGAGGCGGGCGCGAC	-	67	Chr20: 64,063,281-64,063,347
	TCEA2-**exon1**-sg2	AGATTGCGCGGATCGCCCGG	-		
	TCEA2-**exon2**-sg1	GAGAAGGAGGACCTTCATAA	+	64	Chr20: 64,066,452-64,066,515
	TCEA2-**exon2**-sg2	CAGGTGCAGCGTGATAGGCA	+		
DKO	TCEA1-**exon1**-sg1	CGCGCCCACCCCGCTGGCAA	-	67	Chr8: 54,022,095-54,022,161
	TCEA1-**exon1**-sg2	GGTCCGCTTTGCCAAGAAGA	+		
	TCEA2-**exon2**-sg1	GAGAAGGAGGACCTTCATAA	+	64	Chr20: 64,066,452-64,066,515
	TCEA2-**exon2**-sg2	CAGGTGCAGCGTGATAGGCA	+		
TCEA1 KO	TCEA1-**exon2**-sg1	TATTTTATAGGCTGGAGCAT	+	46	Chr8: 54,010,440-54,010,485
	TCEA1-**exon2**-sg2	CACATACCTGCAGTAATTCC	-		
	TCEA1-**exon2**-sg3	TATTGGTACTGAAGATGTTT	-	didn’t cut	

2.4.2 Transfection

Early passage HEK293T cells were seeded at $400x10^5$ cells per well in 6 well-plate 24 hours before transfection, in DMEM, 10% FBS, 1x GlutaMAX, without antibiotics. 1.5 µg of plasmid DNA was mixed with the transfection reagent Metafectene Pro (Biontex T040-1.0) at the ratio of 1:4 and added to cells, as described by the manufacturer. The medium was changed after 24 hours (DMEM, 10% FBS, 1x GlutaMAX, 5% penicillin-streptomycin).

The RNP complex transfection was performed as described in Synthego Immortalized Cell Lipofection Protocol, using Cas9 protein (ThermoFisher Scientific A36499), CRISPRMAX transfection reagent (ThermoFisher Scientific CMAX00008), and Opti-MEM I reduced serum medium (Gibco 31985062).

2.4.3 Single cell sorting and cell culture handling

36 hours after transfection, the cells were prepared for single cell sorting. They were washed with PBS, trypsinized, resuspended in PBS, counted, and 0.5M EDTA (at 1:250) was added to prevent clumping. The cells were then strained through a cell strainer and the FACS facility of MPIMG sorted single GFP-positive cells into 96-well plates using the FACSAria Fusion flow cytometer, 85 nm nozzle. High genome editing efficiency was predicted in RNP complex-transfected cells, so they did not have a sorting marker. Cells were prepared as above, with addition of DAPI for 5 min. Live cells (DAPI-negative) were sorted into 96-well plates with 100 nm nozzle.

After 4 - 5 days, the plates with clones were observed under a microscope, and the wells with a single small colony were marked. The marked cells were washed with PBS (150 µL), trypsinized (30 µL) and consolidated on the same plate for easier organization and handling with a multichannel pipet. The medium was changed every two days. When the new plates were confluent, the cells were split into two additional plates for freezing. Once the wells were confluent, the cells were gently washed with PBS, trypsinized, and gently resuspended in 100 µL/well freezing medium (80% FBS, 10% DMEM, 10% DMSO). The plate was sealed with a PCR sealing sticker and parafilm, wrapped in bubble wrap, placed into a styrofoam box and stored at -80°C.

2.4.4 DNA extraction and genotyping

The plate for genotyping was washed with PBS and stored at -20°C overnight. Genomic DNA was extracted using QuickExtract DNA Extraction Solution (Lucigen: 101098). 50 µL were added per well and mixed vigorously, with scratching the bottom of the wells. The lysed cells were transferred into a 96-well PCR plate, heated at 65°C for 6 min, vortexed for 30 seconds, then heated for 2 min at 98°C and vortexed again. If the lysate was too viscous, more QuickExtract solution was mixed in. The extracted DNA was stored at -20°C.

The genotyping PCR reaction was set up as follows: 5 µL of genomic DNA, 12.5 µL of Q5 Polymerase MasterMix (NEB M0492S), 1.25 µL of 10 µM forward primer, 1.25 µL of 10 µM reverse primer, and 5 µL of water. The genotyping primers and cycler program are in the tables below. The amplicons were loaded onto a 1.5% agarose gel and separated by size with electrophoresis, at 120 V. The clones with a clear homozygous deletion were recorded. Subsequently, the PCR was repeated in a few reactions to have enough material for Sanger sequencing. The primers for sequencing are in bold in the table below.

Table 5: Knockout genotyping primers and reaction

Genotyping primers			
Primer name	Sequence	Amplicon size (WT)	Optimized T_a
GP_TCEA1KOexon1_F	CGTAAGGAAGGGGGCCTA	475 bp	51°C
GP_TCEA1KOexon1_R	GCGTGCCCTAATCCCTAAAT		
GP_TCEA1KOexon2_F	GGTGCTGTTGCTCCTTATCTG	438 bp	57°C
GP_TCEA1KOexon2_R	TGAGATTTCACTGCTACTGCC		
SP_TCEA1KOexon2 Sequencing primer	TACTGCCAACTTTAGAGATTCAGGTTTTAT	-	-
GP_TCEA2KOexon1_F	CTGGGAGTTGTGGTCCAGAG	448 bp	57.5°C
GP_TCEA2KOexon1_R	CTGCGTCCCGGTTAGTCTC		
GP_TCEA2KOexon2_F	TTCTTTTTGACCCCAGGTTG	327 bp	63°C
GP_TCEA2KOexon2_R	TCTCCTCAGGAACAGGCATT		

Genotyping PCR program			
Initial denaturation	98°C	2 min	
Denaturation	95°C	30 s	30 cycles
Annealing:	Optimized T_a	30 s	
Extension	72°C	1 min	
Final extension	72°C	8 min	

After confirming the deleted sequence, the homozygous knockout clones were quickly thawed in the 96-well plate, transferred into a 24-well plate for expansion. A pellet of 2 million cells was harvested for Western blotting (described in 2.8) for final validation. Subsequently, the clones were expanded in T75 flasks, slowly frozen in 1.5 mL cryotubes (2-4 million cells in 20% FBS, 10% DMSO, in DMEM), temporarily stored at -80°C, then transferred to a liquid nitrogen tank.

2.4.5 Off-target editing analysis

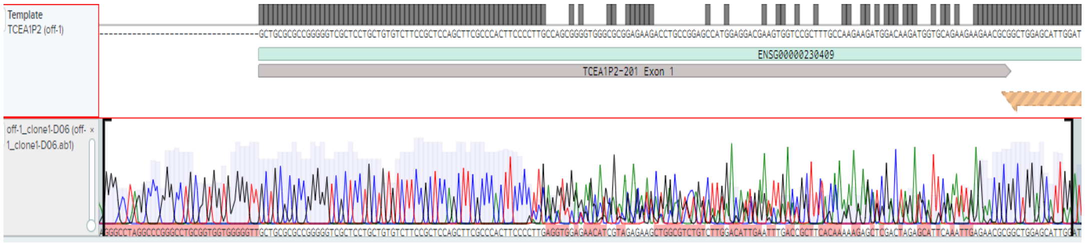
The potential off-targets were predicted and examined using CRISPOR [322]. The possible off-target editing in exons and whether the mismatches are far from the PAM site (in the seed region) are summarized in the table below. The CFD off-target scores were low, except for *TCEA1P2*, a pseudogene of *TCEA1*, which is located in an intron of *GOLGA4*. All homozygous TCEA1 knockouts with the deletion in exon 1 had a heterozygous deletion in *TCEA1P2,* as detected with Sanger sequencing. The example locus is shown below table 6 and it belongs to the clone used for targeting *TCEA2* to create a double knockout (DKO)

Table 6: Off-target editing prediction

Guide sequence with *PAM*	**MIT score**	**Minimal number of mismatches in exons**	**Coding region with mismatches only outside the seed region**	**Off-target gene & CFD off-target score**
TCEA2 knockout guides, exon 1				
AGATTGCGCGGATCGCCCGG *AGG*	99	4	One with 4 mismatches	*FGFR1* 0.05
GCTGCGGAGGCGGGCGCGAC *GGG*	69	3	None	
TCEA2 knockout guides, exon 2				
GAGAAGGAGGACCTTCATAA *AGG*	74	4	None	
CAGGTGCAGCGTGATAGGCA *TGG*	77	3	One with 4 mismatches	*TFCP2L1* 0.06
TCEA1 knockout guides, exon 2				
TATTTTATAGGCTGGAGCAT *TGG*	64	4	None	
CACATACCTGCAGTAATTCC *AGG*	79	3	One with 3 mismatches	*TK2/RP11-403P17.3* 0.12
TATTGGTACTGAAGATGTTT *TGG*	60	4	One with 4 mismatches	*KCNV1* 0.14
TCEA1 knockout guides, exon 1 (used in the DKO)				

CGCGCCCACCCCGCTGGCAA *GGG*	46	3	None with no mismatches in the seed region	*TCEA1P2* in intron of *GOLGA4* 1.00
GGTCCGCTTTGCCAAGAAGA *TGG*	41	3	Two with 4 mismatches, one with 3 mismatches	*PRDM15* 0.17 *RIMS4* 0.021 *RIMS3* 0.01

Off-target mutation in the *TCEA1P2* locus of the TCEA1 KO in exon 1 (parental clone of the double knockout)


Template
TCEA1P2 (off-1)
GCTGCGCGCCGGGGGTCGCTCCTGCTGTGTCTTCCGCTCCAGCTTCGCCCACTTCCCCTTGCCAGCGGGGTGGGCGCGGAGAAGACCTGCCGGAGCCATGGAGGACGAAGTGGTCCGCTTTGCCAAGAAGATGGACAAGATGGTGCAGAAGAAGAACGCGGCTGGAGCATTGGAT
ENSG00000230409
TCEA1P2-201 Exon 1
off-1_clone1-D06 (off-1_clone1-D06.ab1)
AGGGCCTAGGCCCGGGCCTGCGGTGGTGGGGGTTGCTGCGCGCCGGGGGTCGCTCCTGCTGTGTCTTCCGCTCCAGCTTCGCCCACTTCCCCTTGAGGTGGAGAACATCGTAGAGAAGCTGGCGTCTGTCTTGGACATTGAATTTGACCGCTTCACAAAAAGAGCTCGACTAGAGCATTCAAATTGAGAACGCGGCTGGAGCATTGGAT

2.5 Generation of degron-tagged TCEA1 and TCEA2 cell lines

Although we did not continue with experiments with the generated knock-in cell lines because of erroneous DSB repair (described in the Results section 3.2.3.), here we briefly summarize the experimental details. We used the degron tag (dTAG) system [310] and CRIS-PITCh protocol [324], which is designed to enable the microhomology-mediated end joining DNA repair pathway. Both TCEA1 and TCEA2 were tagged at the N-terminus: the tag contained a Puromycin resistance gene, a self-cleaving peptide (P2A), 2x hemagglutinin (2HA), and the degron tag FKBP12(36V).

2.5.1 Donor plasmid cloning

The dTAG DNA sequence was amplified from the donor plasmid pCRIS-PITChv2-dTAG-BRD4-puroR (the vector was provided by Georg Winter's lab (CeMM, Vienna)). The microhomology (MH) arms were designed as immediate 20 base pairs upstream (5' MH) and downstream (3' MH) from the putative Cas9 cut site. They were added to each end of the degron tag by a PCR reaction with Q5 DNA polymerase and the MH primers (table below). The PCR product was run on agarose gel and purified. Afterward, it was ligated with the donor plasmid backbone using the Gibson assembly (NEB E2611S), as instructed and the Stellar chemically competent *E. coli* (Takara 636763) were transformed as instructed by the manufacturer. DNA purification and validation were done as in 2.4.

Table 7: Primers for adding microhomology arms

Microhomology primers		**PCR reaction**	**Cycler program**
TCEA1-MH-forward	GGAGTTCCGCGTTACATAGCATCGTAC	Degradation tag: 0.5 ng	98°C - 30 s
TCEA1-MH-reverse	AGCCCACCGCCGCCTAGAGCCAAGAA	Q5 master mix: 12.5 µL	98°C - 10 s 72°C - 20 s x25 72°C - 40 s
TCEA2-MH-forward	GGAGTTCCGCGTTACATAGCATCGTAC	Primer mix (10 µM): 2.5 µL	
TCEA2-MH-reverse	AGCCCACCGCCGCCTAGAGCCTCCGA	Nuclease free water: 9 µL	72°C - 5 min

2.5.2 sgRNA design, cloning, and transfection

TCEA1 knock-in

To insert the tag as close to the N-terminus of TCEA1 as possible, we relied on an alternative PAM site (NGAN instead of NGG) because of lack of good sgRNA candidates in the first exon. We used CRISPOR [322] to design the sgRNA for a mutated Cas9 (VQR variant) and annealed the guide oligos (top: CACCGCGAAGTGGTCCGCTTTGCCA and bottom: AAACTGGCAAAGCGGACCACTTCGC), as previously described [323]. The annealed oligos were cloned into BpiI-linearized pX330A_sgPITCh (provided by Georg Winter's lab (CeMM, Vienna)/Addgene 58766), with the Quick Ligation kit (NEB M2200S), as instructed by the manufacturer, followed by the standard transformation, inoculation and endotoxin-free plasmid DNA purification as described previously. This plasmid was still necessary for cutting out the microhomologies with the dTAG from the donor plasmid.

1x10^6 early passage HEK293T cells were transfected with three plasmids: the pX330A_sgPITCh, the donor plasmid with TCEA1 microhomologies, and the Cas9-VQR variant plasmid (Addgene #65771), as described earlier.

TCEA2 knock-in

The sgRNA was also designed with CRISPOR [322]. The guide oligos (CACCGAGATTGCGCGGATGCCCGG and AAACCCGGGCGATCCGCGCAATCTC) were annealed as previously described [323]. The annealed oligos were cloned into the same BpiI-linearized pX330A_sgPITCh with the Quick Ligation kit (NEB M2200S), as instructed by the manufacturer. 1x10^6 early passage HEK293T cells were transfected with the pX330A_sgPITCh and donor plasmids as described in 2.4.

2.5.3 Puromycin selection and single clone handling

24 hours after the transfection, the medium was changed. After another 24 hours, puromycin (Gibco A1113802) was mixed into fresh medium at 1 µg/mL. The medium with puromycin was changed every day for 7 days. The resistant cells were diluted to 0.5 cells per 100 µL of medium and seeded into 96 well plates, 10 plates per each transfection. After 3-4 days, the plates were screened under a light microscope, and the wells with a single small colony were marked. The culturing, freezing, DNA extraction, and genotyping was done as in 2.3.3. The same genotyping primers (for exon 1) were used. The clones with a homozygous knock-in were further inspected by Sanger sequencing and Western blot.

2.6 Cell line characterization with flow cytometry

2.6.1 Apoptosis assay

HEK293T, TCEA1KO, TCEA2KO, and the DKO cell lines were cultured as described earlier, counted, and 1x10^6 cells were centrifuged at 1000 rpm, for 4 min, at room temperature (RT). The pellet was resuspended in 100 µL PBS. 1 µL of pre-diluted Zombi NIR dye (BioLegend 423105) (1:15 in PBS) was added to the cells and incubated for 10 min at RT, protected from light. 1 mL of FACS buffer (2 mM EDTA. 0.5% BSA in 1xPBS) was added and the cells were centrifuged at 1000 rpm, RT, for 5 min. The cell pellet was resuspended in 100 µL of FACS buffer and 5 µL of prediluted Apotracker Green (BioLegend 427401) (1:10 in PBS) was added and followed by incubation for 20 min at RT, protected from light. 1 mL of FACS buffer was added and the cells were centrifuged as above. The pellet was resuspended in 100 µL of FACS buffer, filtered through 40 µm cell strainer (PluriSelect 43-10040-40), and given to the flow cytometry facility of MPIMG to record the signal at the FITC channel for Apotracker green and the Red Laser (633 nm) for Zombie NIR dye.

2.6.2 Cell cycle analysis of unsynchronized cells

Fixing cells in ethanol

HEK293T, TCEA1KO, TCEA2KO, DKO, and independently derived single cell clones were counted. About 5x10^6 cells were centrifuged at 1000 rpm, for 3 min, at 4°C. The cell pellets were resuspended by rapid addition of -20°C 70% EtOH, while mildly vortexing, and stored at -20°C for up to two weeks.

DNA staining

On the day of FACS, the cells were centrifuged at 300 x g, for 10 min, at 4°C and resuspended in 1 mL of sterile PBS. The cells were let through a 40 µm cell strainer (PluriSelect 43-10040-40). The cells were counted by adding 10 uL to the counting chamber directly (no trypan blue needed) and using an automated cell counter (Bio-Rad 1450102). 1x10^6 cells were transferred into a new tube and centrifuged at 300 x g, for 5 min, at 4°C. The supernatant was carefully removed without losing cells. FxCycle Violet DNA stain (Invitrogen F10347) was diluted 1:1000 in PBS and 200 uL of the mix were added to each pellet. The cells were mixed thoroughly by pipetting and incubated at room temperature, for 30 min, protected from light. Afterward, the samples were cooled to 4°C for 30 min and given to the FACS facility of MPIMG for measuring.

Flow cytometry

The measurement was done using FACSAria Fusion Flow cytometer (BD Bioscience) at 4°C. The uv-450/50A voltage channel was set to 50 V, so that it matches G1 peak as precisely as possible. 20,000 events were recorded per each cell line and clone. FlowJo software was used for analysis (standard instructions by FlowJo tutorials) and visualization of the cell cycle profiles. Because the channel was set to 50 V, the profiles could be overlaid.

2.6.3 Cell synchronization and cell cycle analysis

Double thymidine block

1.5 x 10^6 HEK293T, 1.5 x10^6 TCEA1KO, 3.0 x 10^6 TCEA2KO, and 3.0 x 10^6 DKO cells were seeded into T75 flasks for semi-adhesive cell lines (Sarstedt 83.3911.302) in 12 mL or regular medium. Since the TCEA2KO and the DKO grow considerably slower than the other two, the varying seeding cell number was necessary to have the cell lines at similar confluency at the time of treatment. 7 flasks were used per each cell line, one for each time point of release (unreleased, 1 hour, 2 hours, 3 hours, 4 hours, 6 hours, and 8 hours). 24 hours after seeding, 300 µL of 100 mM Thymidine (Sigma 89270) in DMEM were added into the medium, by gently tilting the flask and carefully mixing with a serological pipet without disturbing the cells. Precisely after 16 hours, the medium was removed. The cells were carefully washed three times with PBS, prewarmed to 37°C. 2'-Deoxycytidine (Sigma D3897) was mixed into a fresh regular medium at 1:1000 and 12 mL added to the cells. After 8 hours, thymidine was added to the cells for 16 hours and washed out as above. Except the "unreleased" cells, all other samples were released with a regular medium with 2'-deoxycytidine for the corresponding time point. The "unreleased" and "released" cells were washed with PBS once, trypsinized, collected in PBS with 10% FBS, and centrifuged at 1000 rpm, for 4 min, at 4°C. The cells were fixed in 70% EtOH, stained, and analyzed with flow cytometry as described in the previous section.

2.7 Quantitative chromatin mass spectrometry

2.7.1 SILAC labeling

The SILAC medium (Thermo Scientific A33972) was prepared as instructed by the manufacturer and frozen 1x10^6 cells of each cell line (HEK293T, TCEA1KO, TCEA2KO, and DKO) were grown in the "heavy medium" for 6 doublings. Total cell lysates were prepared for testing the incorporation efficiency. 2x10^6 cells were washed with PBS and lysed in 100 µL of 8M urea (Invitrogen 15505027), 50 mM Tris-HCl pH7.5 (Jena Bioscience BU-125S), cOmplete protease inhibitor cocktail (Roche 11836170001) and PhosStop phosphatase inhibitor cocktail (Roche 4906845001), and 2 µL of Benzonase (Sigma-Aldrich E8263), for 30 min, at 4°C, rotating at 30 rpm. The lysates were given to the mass spectrometry facility for evaluation. The labeling efficiency was 99%. The labeled cells were frozen and stored in liquid nitrogen.

2.7.2 Sample setup

The experiment was done in three days, one for each sample set. In total, there were 6 biological replicates per cell line in this analysis. The sample sets were the same and included:

1) TCEA1KO-H (grown in heavy medium) to be combined with HEK293T-L (grown in light medium);
2) the label switch: HEK293T-H to be combined with TCEA1KO-L:
3) TCEA2KO-H to be combined with HEK293T-L;
4) the label switch: HEK293T-H to be combined with TCEA2KO-L;
5) the DKO-H to be combined with HEK293T-L
6) the label switch: HEK293T-H to be combined with the DKO-L

After washing with sterile PBS, the cells were trypsinized, resuspended in the same kind of medium and counted as precisely as possible. Two measurements based on two separate cell aliquots were performed. 5x10^6 cells of each condition were transferred into new 15 mL tubes and centrifuged at 200 x g, for 4 min at 4°C, washed with ice-cold PBS, and centrifuged again. The cells were resuspended in 3 mL of PBS and the H and L cells were pooled as in the setup above. After centrifuging, up to 500 µL of PBS was removed, and using a wide bore 1 mL pipet, the pellets were transferred into protein-lobind 1.5 mL tubes and centrifuged again.

2.7.3 Cell fractionation

All buffers (table 7) were freshly prepared on the same day and kept on ice. The pellets were resuspended in 150 µL of cytoplasmic lysis buffer by gently pipetting 20 times with a wide bore tip and incubated on ice for 5 min. 400 µL of the sucrose buffer was added into fresh protein-lobind tubes, the lysate was mixed again two times and carefully layered on top of the sucrose cushion. The nuclei were separated by centrifugation at 13,000 rpm, at 4°C, for 10 min. The supernatant (cytoplasmic fraction) was carefully and completely removed. The nuclei were washed with 500 µL of the nuclei wash buffer by gentle pipetting 5 times and centrifuged at 3,500 rpm, at 4°C for 1 min. The supernatant was completely removed. The nuclei were resuspended in 200 µL of glycerol buffer by pipetting 20 times. 200 µL of the nuclei lysis buffer were added and mixed by inverting the tubes 4 times and pulse-vortexing 15 s, 3 times. The nuclear lysates were incubated on ice for 2 min and centrifuged at 14,000 rpm, for 2 min, at 4°C. The supernatant was completely removed, and 50 µL of the chromatin resuspension buffer with 2 µL of Benzonase was added to the chromatin pellet. Chromatin was solubilized by shaking for 10 min, in the cold lab on a rotator and pipetting from time to time. The volume was precisely measured and the protein precipitation solution at -20°C (MilliPore 539180), at 4x the lysate volume was added and mixed by inverting 8 times. The samples were stored at -20°C. The chromatin was washed, precipitated, and dried as described by the manufacturer (MilliPore 539180) and was given to the MS facility for further same day processing.

Table 8. Buffers for cell fractionation

Cytoplasmic lysis buffer (150 µL per sample)	**Final concentration**	**Product details**
NP-40	0.15%	Thermo Scientific 28324
Tris-HCl pH7.0	10 mM	Invitrogen AM9850G
NaCl	150 mM	Invitrogen AM9760G
Protease inhibitors	1X	Roche 11836170001
Phosphatase inhibitors	1X	Roche 4906845001
Water, nuclease-free		Invitrogen 10977015
Sucrose buffer (400 µL per sample)	**Final concentration**	**Product details**
Tris-HCl pH7.0	10 mM	as above
NaCl	150 mM	as above
Sucrose	25%	Sigma-Aldrich S0389
Protease inhibitors	1X	as above
Phosphatase inhibitors	1X	as above
Water		as above
Nuclei wash buffer (500 µL per sample)	**Final concentration**	**Product details**
Triton X-100	0.1%	Sigma-Aldrich T8787

EDTA pH8.0	1 mM	Invitrogen AM9260G
Protease inhibitors	1X	as above
Phosphatase inhibitors	1X	as above
PBS	1X	Life Technologies10010-023
Glycerol buffer (200 µL per sample)	**Final concentration**	**Product details**
Tris-HCl pH8.0	20 mM	Invitrogen AM9855G
NaCl	75 mM	as above
EDTA pH8.0	0.5 mM	as above
Glycerol	50%	ThermoFisher Scientific J61059.K2
DTT (dithiothritol)	0.85 mM	ThermoFisher Scientific R0861
Protease inhibitors	1X	as above
Phosphatase inhibitors	1X	as above
Water		as above
Nuclei lysis buffer (200 µL per sample)	**Final concentration**	**Product details**
NP-40	1%	as above
HEPES pH7.5	20 mM	ThermoFisher Scientific J60712AK
NaCl	300 mM	as above
Urea	1 M	Sigma-Aldrich U5128
EDTA pH8.0	0.2 mM	as above
DTT (dithiothritol)	1 mM	as above
Protease inhibitors	1X	as above
Phosphatase inhibitors	1X	as above
Water		as above
Chromatin resuspension buffer (50 µL)	**Final concentration**	**Product details**
PBS	1X	as above
Protease inhibitors	1X	as above
Phosphatase inhibitors	1X	as above

2.7.4 Chromatin MS data processing and analysis

The MS facility processed the raw data with MaxQuant [325]. The peptides were identified based on the UniProtKB database, with a false discovery rate of 0.01. The following parameters were applied: a mass tolerance of 4.5 ppm for precursor ions; 20 ppm for fragment ions; up to two missed trypsin sites were tolerated; cysteine carbamidomethylation was set as a fixed modification, N-terminal acetylation and methionine oxidation were set as variable modifications. $^{13}C_6{}^{15}N_4$-arginine and $^{13}C_6{}^{15}N_2$-lysine were set to identify the heavy and light-labeled cells. The analysis was done in Perseus software [326]. The light to heavy (L/H) SILAC ratios were transformed as 1/x to match the H/L samples (so that all knockout cell lines are heavy and the WT is light). Then all samples were $\log_2$- transformed and filtered as follows: potential contaminants, reverse sequences, and “identified by site” were removed. Only the proteins detected in at least 70% of biological replicates were kept. The differences between the heavy-labeled (KO) and the light negative control (WT) were computed using a two-tailed one-sample Student’s t-test with Benjamini-Hochberg FDR correction of 0.05. The heatmaps of significance values were made with GraphPad Prism.

2.8 Immunoblotting

2.8.1 Sample preparation for Western blotting (WB)

Cells were trypsinized and counted using an automated cell counter. Counting was done in at least two replicates for precision. 1×10^6 cells were washed with sterile PBS, precipitated by centrifugation for 5 min at 1000 rpm at 4°C, snap frozen in liquid nitrogen, and stored at -80°C. The pellets were loosened in 37.5 µL of PBS with 1x protease and phosphatase inhibitors by pipetting, then homogenized by adding 125 units (0.5 µL) of Benzonase to each pellet and mildly shaking at 4°C for 15 min, followed by the addition of 12.5 µL of 4x ROTI WB sample loading buffer and heat denaturation at 95°C for 5 min.

Chromatin samples were prepared by following the cell fractionation part of NET-seq protocol, but without α-amanitin and SUPERase-In RNase inhibitor.

2.8.2 Protein electrophoresis and immunoblotting

The protein concentration was not measured, however, the loading volume stayed constant among conditions and cell lines. Based on the abundance of the proteins, the loading amount was between 5 and 15 µL of the lysate per well. The samples were loaded into a precast polyacrylamide gel and run at 170 V for 50 min in MOPS buffer at room temperature, but with cooling packs around the chamber. The wet transfer onto 0.2 µm nitrocellulose membrane was done at 120 V for 2 hours at 4°C, followed by 1 hour of blocking at room temperature in 5% milk in TBS-T or LI-COR Intercept blocking solution. Afterward, the membrane was incubated with a primary antibody diluted in StartingBlock T20 blocking buffer overnight at 4°C. After three 10-minute washes with TBS-T, the membrane was incubated with the appropriate IRDye 680RD secondary antibody diluted in 5% milk in TBS-T for 1 hour, washed three times, and imaged with the LICOR Odyssey Clx system.

2.8.3 WB signal quantification

The antibody signal bands were detected with the 800 nm channel, and the protein ladder - at 700 nm. The band intensity was quantified using the Image Studio Lite, without changing brightness and contrast parameters. The signal was normalized to the signal intensity of GAPDH, H2B, α-tubulin, or total protein stain by first calculating the normalization factor, which is the ratio of intensities of housekeeping proteins to the one with the maximal intensity, and then dividing the intensity of the protein of interest by that ratio.

Table 9. Materials for Western blotting

Materials	
Benzonase	Sigma E8263
cOmplete protease inhibitor cocktail	Roche 11836170001
PhosStop phosphatase inhibitor cocktail	Roche 4906845001
4x WB sample loading buffer, ROTI load	Roth K929.2
4-12% Bis-Tris NuPAGE polyacrylamide gel	ThermoFisher Scientific NP0321BOX
Precision Plus Protein Kaleidoscope prestained protein standards	Bio-Rad 1610375
0.2 µm nitrocellulose membrane	Bio-Rad 1620112
Buffers	
NuPAGE MOPS SDS running buffer, 20x	ThermoFisher Scientific NP0001
Transfer buffer	25 Mm Tris base, 190 mM glycine, 20% (v/v) methanol
1x TBS-T washing buffer	20 mM Tris, 150 mM NaCl, 1% (v/v) Triton X-100
Licor Intercept blocking solution TBS	LI-COR 92760001
Blocking solution	5% (w/v) milk in TBS-T
Antibody dilution buffer: StartingBlock T20 blocking TBS	ThermoFisher Scientific 37543
Instruments	
XCell SureLock mini gel electrophoresis system	ThermoFisher Scientific EI0001
Mini Trans-Blot Electrophoretic Transfer Cell	Bio-Rad 1703930
Odyssey CLx Infrared Imaging System	LI-COR Biosciences
Image Studio Lite	LI-COR Biosciences

Table 10. Antibodies for Western blotting

Target protein	Manufacturer and catalog number	Dilution for WB
Alpha-tubulin	Abcam ab18251	1:2000
ATR Ser428 phospho	Cell Signaling Technology 2853	1:1000
FLAG (M2 clone)	Sigma F1804	1:500
GAPDH	Ambion 326548	1:10,000
H2A.x Ser139 phospho	Cell Signaling Technology 9718	1:1000
H2B	Santa Cruz Biotechnology sc-10808	1:1000
P53 Ser15 phospho	Cell Signaling Technology 9286	1:1000
TCEA1	Abcam ab184181	1:2000
TCEA1 However, this antibody also binds overexpressed TCEA2	Abcam ab185947	1:2000
TCEA2	Aviva ARP58181	1:500
Secondary antibodies		
IRDye 680RD goat anti-mouse	LI-COR 926-68072	1:15,000
IRDye 680RD goat anti-rabbit	LI-COR 926-68071	1:15,000
IRDye 680RD goat anti-rat	LI-COR 926-68076	1:15,000

2.9 Slot blot for R-loop detection

2.9.1 Sample preparation

The slot blot was performed as previously described [327] with some modifications, which are specified below. The cells were thawed and grown for a week. All cell lines had the same number of splittings. Cells were trypsinized and counted. 2x10^6 HEK293T, TCEA1KO, TCEA2KO, the DKO cells were washed with sterile PBS, pelleted by centrifugation for 5 min at 1000 rpm at 4°C, snap frozen in liquid nitrogen, and stored at -80°C. As a positive control of increased R-loop formation, HEK293T cells were treated with camptothecin (CPT). 10 mM stock solution in DMSO was diluted in the regular medium at 1:1000. 2x10^6 HEK293T cells were resuspended in a CPT-containing medium for 10 min, at 37°C, then washed, and snap-frozen. DNA including R-loops was purified as previously described [327], but without sonication, and the DNA pellet was dissolved by pipetting and incubating at 4°C. The DNA concentration was measured using NanoDrop (Thermo Scientific™ 840274200). Having tested different loading amounts and volumes, 400 ng of DNA in 300 µL per slot was determined to be the most optimal.

2.9.2 Nucleic acid blotting

Using the slot blot apparatus, the DNA was loaded in at least three technical replicates onto a positively charged nylon membrane, previously activated by incubating it in nuclease-free water for 15 min at room temperature. As a loading control, an additional membrane was loaded on the same day for the detection of dsDNA. After loading, the membranes were dried at room temperature for 5 min, crosslinked at 0.12 J, blocked for an hour at room temperature, and incubated with the primary antibodies overnight, gently rotating at 4°C. The next day, the membranes were washed three times with 1x TBS-Tween and incubated with the secondary antibody in 5% milk in 1x TBS for an hour at room temperature, followed by additional three washes. The blot was imaged using the LICOR Odyssey Clx system.

Signal specificity of the DNA-RNA hybrid antibody was evaluated by performing RNaseH digestion of genomic DNA at 37°C overnight, carrying out a slot blot as described above, and comparing the antibody binding between RNaseH-treated and untreated samples. Although the detected bands of the treated samples were much fainter, this indicates that this antibody lot may recognize other nucleic acid species.

Table 11: Materials and instruments for nucleic acid blotting

Buffers	
Cell lysis buffer	0.5% NP-40, 80 mM KCl, 5 mM HEPES (pH 7.5) in nuclease-free water
Nuclear lysis buffer	1% SDS, 25 mM Tris-HCl (pH8.0), 5 mM EDTA in nuclease-free water
Licor Intercept blocking solution TBS	LI-COR 92760001
Blocking solution	5% (w/v) milk in TBS-T
Instruments	
Bio-Dot SF Microfiltration Apparatus	Bio-Rad 1706542
UV crosslinker	Cleaver Scientific CL508
Odyssey CLx Infrared Imaging System	LI-COR Biosciences
Image Studio Lite	LI-COR Biosciences
Antibodies	
Anti-DNA-RNA hybrid [S9.6]	Kerafast ENH001: 1:500
Anti-dsDNA	Abcam ab215896: 1:2000
Other materials	
Camptothecin (CPT)	MedChem Express HY-16560 provided by Kinkley lab, MPIMG
Hybond N+ nylon membrane	GE Healthcare Life Sciences RPN203B
RNase H	New England Biolabs M0297

2.10 Immunoprecipitation followed by mass spectrometry (IP-MS)

In general, the IP was carried out as outlined in the published protocol [328]. The details, pertinent to our cellular system and equipment, are described further.

2.10.1 Cell culture and induction of expression of epitope-tagged TCEA1, TCEA2, TCEA1-NTDL, and TCEA2-NTDL

Flp-In-TCEA1 (5 biological replicates), Flp-In-TCEA2 (5 biological replicates), Flp-In-TCEA1-NTDL (4 biological replicates), Flp-In-TCEA2-NTDL (4 biological replicates), and Flp-In T-Rex 293 (5 biological replicates) were grown to 60-70% confluency in T75 flasks. The latter cell line, not expressing any FLAG-tagged protein, served as a negative control and was exposed to the same anti-FLAG antibody as all other cell lines. 24 hours before crosslinking, 1 µg/mL Tetracycline (Thermo Fisher Scientific A39246) was added to induce the expression of epitope-tagged proteins.

2.10.2 Crosslinking

The cells were gently washed with room temperature PBS. 10 mL of methanol-free 1% formaldehyde (Thermo Scientific™ 28908) in PBS was added to the cells carefully so as not to detach them. The flask was swirled slowly to mix three times and rotated precisely for 8 min at room temperature. Formaldehyde was quenched by adding 0.5 mL of 2M glycine (Millipore Sigma G7126). The solution was removed, and the cells were washed with cold PBS twice. The cells were then scraped in 10 mL of cold PBS with cOmplete protease inhibitor cocktail (1x) and transferred into a 15 mL tube. Additional 4 mL were added to collect any remaining cells. The cells were centrifuged at 2,000 x g, for 3 min, at 4°C. The pellet was washed with 1 mL of cold PBS with cOmplete inhibitor cocktail (1x), centrifuged as above, snap-frozen in liquid nitrogen, and stored at -80°C for a few days.

2.10.3 Antibody-to-beads binding

100 µL of magnetic G-protein Dynabeads (Thermo Fisher Scientific 10003D) were used per IP sample. To efficiently process up to 5 samples at a time, two aliquots of 250 µL of beads each were washed with 1 mL filter-sterilized 0.5% BSA in PBS (BSA in PBS) four times, by precipitating the beads on a magnetic rack (Thermo Fisher Scientific, MR02), in the 4°C lab. Each aliquot of the beads was resuspended in 500 µL of filter-sterilized BSA in PBS. 15 µg per IP sample of anti-FLAG antibody (Sigma F3165) was added to the beads. The beads and antibody were rotated at room temperature for an hour, followed by 4 washes with 1 mL of BSA in PBS. The antibody-bound beads were resuspended in 100 µL in BSA in PBS per IP sample and added into each lysate. The samples were incubated overnight, in the 4°C lab, rotating, 25 rpm.

2.10.4 Lysis

The buffers are in the table on the following page. They were prepared a day in advance, however the cOmplete protease inhibitor cocktail (1x) was added right before lysing the cells. From now on, the procedure was carried out in a 4°C lab. The pellets were thawed on ice and resuspended in 10 mL of ice-cold LB1. The cells were lysed for 10 min, rotating at 36 rpm, and centrifuged at 200 x g, for 4 min, at 4°C. The pelleted nuclei were resuspended in 10 mL of ice-cold LB2 and rotated for 10 min at 4°C, at 36 rpm, and centrifuged as above. The chromatin pellet was resuspended in 350 µL of ice-cold LB3. 300 µL were transferred into a 1.5 mL TPX sonication tube (Diagenode C30010010-300) (not to exceed 20 million cells per tube) and sheared in the pre-cooled to 4°C Bioruptor Plus sonicator (Diagenode B01020001) at high intensity, 30 s on, 60 s off, for 5 cycles. 30 µL of 10% (vol/vol) Triton X-100 (Millipore Sigma T8787) in PBS was added to the samples and mixed by vortexing for 10 s. Unsheared chromatin was separated by centrifugation at 20,000 x g, 10 min, at 4°C. The supernatant (dissolved chromatin) was transferred into a protein-lobind tube with additional 2.5 µL of cOmplete protease inhibitor cocktail (50x).

2.10.5 Immunoprecipitation

Without delay, the antibody-bound beads (100 µL per IP sample) were added into lysate and mixed by gently flicking the tubes a few times. The samples were rotated overnight at 36 rpm, in the cold lab. The next day, the unbound material was removed by precipitating the beads on the magnetic rack and washing them 8 times with 1 mL of ice-cold mild RIPA buffer in the cold lab. During each wash, the beads were let to precipitate 4 times on the opposite sides of the tube by turning the tube 180°. Afterward, the beads were washed with ice-cold 1 mL 100 mM AMBIC (ammonium bicarbonate) solution two times, resuspended in 30 µL, and brought to the mass spectrometry facility of MPIMG on ice for immediate further processing. The MS facility treated the samples with trypsin, desalted using Pierce C18 tips, reconstituted in 5% acetonitrile and 2% formic acid, and sonicated for 30 s. Nanoflow reverse-phase chromatography with a C18 resin analytical column was applied to separate the peptides. The measurement was done using an orbitrap instrument.

Table 12: Buffers for IP-MS

LB1 (10 mL per IP sample)	**Final concentration**	**Product details**
HEPES-KOH pH7.5	50 mM	Thermo scientific J60712.AP
EDTA pH8.0	1 mM	Invitrogen™ AM9260G
NaCl	140 mM	Invitrogen™ AM9760G

Triton X-100	0.25%	Sigma T8787
NP-40	0.5%	Thermo Scientific™ 28324
Glycerol	10%	Thermo Scientific™ 17904
cOmplete protease inhibitor cocktail	1X	Millipore Sigma 11836170001
Water		Invitrogen 10977015
LB2 (10 mL per IP sample)	**Final concentration**	**Product details**
Tris-HCl pH8.0	10 mM	Invitrogen™ AM9856
EDTA pH8.0	1 mM	as above
NaCl	200 mM	as above
EGTA pH8.0	0.5 mM	Millipore 324626
cOmplete protease inhibitor cocktail	1X	as above
Water		as above
LB3 (300 µL per IP sample)	**Final concentration**	**Product details**
Tris-HCl pH8.0	10 mM	as above
EDTA pH8.0	1 mM	as above
Sodium deoxycholate	10 mg/mL	Thermo Scientific 89904
N-laurylsarcosine	0.5%	Sigma-Adrich L5125
cOmplete protease inhibitor cocktail	1X	as above
Water		as above
Mild RIPA buffer (8 mL per IP sample)	**Final concentration**	**Product details**
HEPES	50 mM	Sigma, H3537
NP-40	0.5%	as above
Sodium deoxycholate	0.4%	as above
EDTA pH8.0	1 mM	as above
LiCl	300 mM	Sigma-Aldrich L9650
cOmplete protease inhibitor cocktail	1X	as above
Water		as above
BSA in PBS (11 mL per IP sample)	**Final concentration**	**Product details**
BSA	5 mg/mL	Roche 10735086001
PBS	1X	Life Technologies10010-023
AMBIC solution (2.1 mL per IP sample)	**Final concentration**	**Product details**
Ammonium bicarbonate	100 mM	Thermo Scientific 393212500
Water		as above

2.10.6 IP-MS data processing and analysis

The MS facility processed the raw data with MaxQuant [325]. The peptides were identified based on the UniProtKB database, with a false discovery rate of 0.01. The following parameters were applied: a mass tolerance of 4.5 ppm for precursor ions; 20 ppm for fragment ions; up to two missed trypsin sites were tolerated; cysteine carbamidomethylation was set as a fixed modification, N-terminal acetylation and methionine oxidation were set as variable modifications. Only the proteins identified by more than 2 common peptides were included in the following analysis with Perseus software [326]. The label-free quantification (LFQ) values were $\log_2$- transformed and filtered as follows: potential contaminants, reverse sequences, and "identified by site" were removed. Only the proteins detected in at least 70% of biological replicates in at least one sample type (FLAG-protein IP vs negative control) were kept, and the missing values were imputed from normal distribution of the measured values, using the default parameters. The differences between the FLAG-protein IP and negative control IP were computed using a two-tailed two-sample student's t-test with FDR correction of 0.05.

2.10.7 IP-MS visualization and gene ontology term analysis

The findings were visualized using VolcanoseR [329]. The heatmaps of significance values were made with GraphPad Prism. The gene ontology enrichment by biological processes was done using ShinyGO 0.76.3 [330], with a FDR of ≤0.05.

2.11 Chromatin immunoprecipitation followed by sequencing

Chromatin immunoprecipitation (ChIP) protocol was based on [331], but without spike-in addition.

2.11.1 Chromatin immunoprecipitation

10 million Flp-In-TCEA1 and Flp-In-TCEA2 cells were seeded in T175 flasks and grown for 24 hours, afterward, they were treated with Tetracycline at the final concentration of 1 µg/mL for 24 hours. Two biological replicates were prepared for each cell line. An additional flask was seeded for counting. The cells were gently washed with PBS and crosslinked with 1% methanol-free formaldehyde (Thermo Scientific 28908), mixed in medium, for 8 min at room temperature, with gentle rotation. The formaldehyde medium was discarded, and the cells were rinsed with PBS with glycine (final concentration 250 mM) to quench the remaining unreacted formaldehyde. 10 mL of PBS with 1x protease inhibitor cocktail was added to the

cells, and they were detached by scraping, while being placed on ice. 50 million cells were centrifuged at 420 x g, for 4 min at 4°C.

The next steps were carried out in the 4°C lab, keeping the samples on ice. The pellet was mixed by pipetting 7 times in 10 mL of lysis buffer 1 (modified, table 12) and rotated for 20 min. The nuclei were precipitated by centrifugation at 420 x g for 10 min at 4°C, the supernatant was completely removed, and the pellet was thoroughly resuspended in 2 mL of lysis buffer 2 (modified). The nuclear fraction was split into two 1 mL AFA fiber tubes (Covaris). 20 cycles of sonication were done in an E220evolution sonicator for 20 min, on duty cycle 5%, intensity 4 and 200 cycles per burst, at ≤7°C. The unsheared heterochromatin was separated by centrifugation at 16,000 x g for 10 min at 4°C. The clear supernatant was transferred into a fresh protein lobind tube and 50 µL were taken out and saved as input. The remaining lysate was incubated with Dynabeads Protein G (100 µL per 50 million cells) overnight, which were previously incubated with the anti-FLAG antibody (Sigma F3165) (10 µg per 50 million cells) for 1.5 hours at room temperature, with rotation.

The next day, the bound beads were washed three times with washing buffers 1,2, and 3. Each wash was 45 s on a rotating wheel in the cold lab. The ChIP material was eluted from the beads twice in 100 µL of the elution buffer for 15 min at room temperature. The elution buffer was proportionally added to the input samples. ChIP material and the input were treated with 20 µg RNase A for 1 hour at 37°C at 700 rpm, followed by 160 µg Proteinase K for 2 hours at 50°C at 700 rpm. NaCl was added at the final concentration of 0.84 M, and the samples were stored at 4°C overnight. The next day, crosslinking was reversed by incubating the samples at 65°C for 5 hours at 700 rpm. The chromatin immunoprecipitated DNA fragments were purified using the ChIP DNA Clean & Concentrator kit (Zymo Research D5205). The DNA concentration was measured with Qubit HS DNA, and the input - with NanoDrop.

Table 13: Buffers for ChIP-seq

Lysis buffer 1	**Final concentration**	**Product details**
HEPES-KOH pH7.5	50 mM	Thermo scientific J60712.AP
KCl	85 mM	Sigma-Aldrich P9541
NP-40	0.5%	Thermo Scientific™ 28324
cOmplete protease inhibitor cocktail	1X	Millipore Sigma 11836170001
Water		Invitrogen 10977015
Lysis buffer 2	**Final concentration**	**Product details**
Tris-HCl pH8.0	10 mM	Invitrogen™ AM9856
EDTA pH8.0	1 mM	Invitrogen AM9260G
NaCl	150 mM	Sigma-Aldrich S3014

NP-40	1%	as above
Sodium deoxycholate	0.1%	Sigma-Aldrich SRE0046
SDS	0.1%	Sigma-Aldrich 71736
cOmplete protease inhibitor cocktail	1X	as above
Water		as above
Washing buffer 1	**Final concentration**	**Product details**
Tris-HCl pH8.0	20 mM	as above
EDTA pH8.0	2 mM	as above
NaCl	150 mM	as above
Triton X-100	1%	Sigma-Aldrich T9284
SDS	0.1%	as above
cOmplete protease inhibitor cocktail	1X	as above
Water		as above
Washing buffer 2	**Final concentration**	**Product details**
Tris-HCl pH8.0	20 mM	as above
NaCl	500 mM	as above
EDTA	2 mM	as above
Triton X-100	1%	as above
SDS	0.1 %	as above
cOmplete protease inhibitor cocktail	1X	as above
Water		as above
Washing buffer 3	**Final concentration**	**Product details**
Tris-HCl pH8.0	10 mM	as above
LiCl	250 mM	Sigma-Aldrich 62476
EDTA	1 mM	as above
NP-40	1%	as above
Sodium deoxycholate	1%	as above
cOmplete protease inhibitor cocktail	1X	as above
Water		as above
Elution buffer	**Final concentration**	**Product details**
NaHCO3	100 mM	Sigma-Aldrich S5761
SDS	1%	as above
Water		as above

2.11.2 ChIP-seq library preparation

Libraries were prepared from 10 to 15 ng of ChIP DNA and 30 ng of input DNA using the NEBNext Ultra II DNA kit according to the manufacturer's instructions with the following specified details. The adaptor was diluted 1:10 prior to adding to the ligation reaction. 8 cycles of PCR enrichment of adaptor-ligated DNA were chosen based on the ChIP DNA amounts.

After the PCR reaction cleanup with AMPure beads as described by the manufacturer, size selection was performed using an 8% TBE gel run at 120 V for 1 hour (with a 10 min pre-run) and stained with SYBR gold for 10 min at room temperature. The PCR products of 200-500 bp sizes were cut out with a new scalpel for each library and transferred into two pierced 1.5 mL DNA lobind tubes inserted into 2 mL DNA lobind tubes, centrifuged at 20,000 x g for 4 min at room temperature to grind the gel pieces to homogeneity. The libraries were eluted from the gel in the soaking buffer, shaking overnight at 1400 rpm at room temperature. The next day the DNA was further purified by centrifuging through a filter tube at 20,000 x g for 4 min at room temperature. The DNA was precipitated with GlycoBlue in isopropanol at -20°C for 4 hours, centrifuged at 21,000 x g at 4 °C for 2 hours washed with 80% ethanol. The DNA pellets were air dried to evaporate the remaining ethanol and resuspended in 10 µL of 10 mM Tris, pH8.0. The DNA concentration was measured with Qubit, and the library quality was assessed with BioAnalyzer HS DNA. The libraries were sequenced on an Illumina NovaSeq 6000 instrument in PE100 mode by the MPIMG sequencing core facility. 100 million reads were requested per library.

2.11.3 ChIP-seq data processing

The ChIP-seq data were processed by Mario Rubio. Adapters were trimmed with Cutadapt [332]. The reads were mapped to the human genome GRCh38 (release 28) with Bowtie2 [333]. PCR duplicates were marked using the markdup function from Samtools [334]. Peaks were called using the callpeak function of MACS2 suite [335]. The count average was calculated using the genomecov function of the Bedtools suite [336]. The input-normalized count coverage was computed using the bggcmp function from the MACS2 suite [335].

2.12 RNA-seq with spike-ins

2.12.1 RNA expression analysis in HEK293T, TCEA1 KO, TCEA2 KO, and the double knockout (DKO)

HEK293T, TCEA1KO, TCEA2KO, and the double knockout cells were seeded at 250x10^3 per well in 6-well dishes in four replicates per cell line and grown until they were 70-80% confluent. Total RNA was extracted from 1.5x10^6 cells using RNeasy Mini Kit as described by the manufacturer, with DNase treatment. ERCC spike in mix was added based on the number of cells, after the RNA extraction. The three biological replicates with the most consistent RNA yield were chosen for spike-in addition and library preparation. The RNA yield was within the same range for all cell lines. 1 µL of 1:10 diluted ERCC mix was added into 7µL of RNA and an additional 8 uL of RNase-free water. The RNA concentration was measured with the Qubit RNA Broad Range Assay kit on the Qubit fluorometer. The quality of RNA was assessed with the BioAnalyzer RNA 6000 Pico kit on a Bioanalyzer instrument. The sequencing facility of

MPIMG prepared the libraries with polyA enrichment and sequenced on NovaSeq2, at PE100 mode, 50 million reads per library.

Table 14: Materials for RNA-seq

Kits	
miRNeasy Mini kit	Qiagen 217004
RNeasy Mini Kit	Qiagen 74104
ERCC RNA Spike-in Mix	Invitrogen 4456740
NEBNext Multiplex Oligos for Illumina Set 2	New England BioLabs E7780S
NEBNext Ultra II Directional RNA Library Prep Kit for Illumina	New England BioLabs E77760S
Ribo Cop rRNA Depletion Module 24 preps	Lexogen M12124
Ribo Cop rRNA Depletion Probe Mix HMR V2	Lexogen M14824
Qubit RNA Broad Range Assay Kit	(Life Technologies Q10210)
BioAnalyzer RNA 6000 Pico kit	Agilent 5067-1513)
Instruments	
Qubit 3 Fluorometer	Invitrogen™ Q33216
Bioanalalyzer	Agilent G2939BA
NovaSeq2	Illumina

2.12.2 RNA-seq in TCEA2 KO upon transient TCEA2 complementation

10 uL of 1:100 diluted ERCC mix was added into the Qiazol lysates of 1x10^6 cells. RNA was purified using miRNeasy Mini kit as instructed by the manufacturer. rRNA was removed using the RiBO COP rRNA depletion kit. The efficiency of rRNA depletion was ensured with TapeStation. The sequencing libraries were prepared using the NEBNext Ultra II Directional RNA Library Preparation kit for Illumina and with dual indexing NEBNext Multiplex Oligos for Illumina and sequenced on NovaSeq2, at PE100 mode, 50 million reads per library.

2.12.3 RNA-seq data processing

The data were processed by Mario Rubio. The reads were mapped to the human genome GRCh38 (release 28) from GENCODE and ERCC92 reference (ThermoFisher) with STAR [337]. The transcript quantification was performed with Salmon [338]. Differential expression analysis was computed using DESeq2 [339]. Differentially expressed genes were defined as those with $\log_2$(KO/WT) ≥1 and an adjusted p-value ≤0.01. The over-representation analysis on the differentially expressed genes using the gene sets from Gene Ontology. Only the sets from biological process level 3 were considered. The analysis was performed by using the tool ConsensusPathDB [340, 341].

2.13 HiS-NET-seq

2.13.1 4-thiouridine (4sU) labeling

100x10^6 cells are required as input for this experiment. HEK293T, TCEA1KO, TCEA2KO, and the DKO were grown in 3-4 T175 flasks until 70-80% confluency. The cells were trypsinized, counted and 100x10^6 cells were split into two 50 mL conical tubes to have 50x10^6 cells in each tube, at 1x10^6 cells/mL. The following steps of the protocol were carried out under the red light until library preparation. After thorough mixing by pipetting, the cells were poured into fresh 50 mL tubes, containing 4sU (Glentham Life Sciences GN6085) to have the final concentration of 500 µM. The cells were inverted 20 times to ensure thorough mixing and placed into the incubator at 37°C. Precisely after 10 min, the cells were placed on ice and centrifuged at 1000 rpm, 4°C for 4 min. The spike-in NIH3T3 were labeled and processed exactly the same way. This experiment was done in two biological replicates per each cell line. Due to the time sensitivity and a large number of cells, on one day, up to two cell lines could be processed by two people (Susanne Freier) at the same time up to chromatin solubilization.

2.13.2 Cell fractionation and chromatin solubilization

The cells were resuspended in 1 mL per 50x10^6 cells of the cytoplasmic lysis buffer, incubated on ice for 3 min, and centrifuged at 500xg, 4°C, for 3 min. The pellets were washed with 2 mL of the nuclei wash buffer and centrifuged at 500xg, 4°C, for 3 min. The nuclei were resuspended in 750 µL of the glycerol buffer and lysed with 750 µL of the nuclei lysis buffer with pulse vortexing, followed by a 2 min incubation on ice. The chromatin was separated by centrifugation at 14,000 rpm, at 4°C, for 2 min and washed in 1 mL of chromatin wash buffer. The chromatin pellet was briefly spun down, and the wash buffer was completely removed. Then the pellets were softened by incubating in 1.5 mL of Trizol (Invitrogen 15596026) with added 1.5 µL of 1 M DTT and 3 µL of 0.5 M EDTA at 40°C, for 1 hour, at 1000 rpm, with vortexing for 30 s every 15 min. It was homogenized by resuspending the clumps with a 2 mL syringe with 21G, 23G, 24G, and 26G needles. Additionally, the dissolved chromatin was let through a QIAshredder spin column (Qiagen 79656) at a maximum centrifugation speed for 2 min. The corresponding samples were pooled in 15 mL tubes and stored at -80°C, protected from light. The total volume per sample is about 3 mL.

Table 15: Cell fractionation buffers for HiS-NET-seq

Cytoplasmic lysis buffer	**Final concentration**	**Product details**
NP-40	0.15%	Thermo Scientific 28324
Protease inhibitors	1X	Roche 11836170001
α-amanitin	1X	Sigma-Aldrich A2263
SUPERaseIN	1X	Invitrogen AM2694
PBS	1X	Life Technologies10010-023
Nuclei wash buffer	**Final concentration**	**Product details**
EDTA pH8.0	1 mM	Invitrogen AM9260G
Protease inhibitors	1X	as above
α-amanitin	1X	as above
SUPERaseIN	1X	as above
PBS	1X	as above
Glycerol	**Final concentration**	**Product details**
Tris-HCl pH8.0	20 mM	Invitrogen AM9855G
NaCl	75 mM	Invitrogen AM9760G
EDTA	0.5 mM	as above
Glycerol	50%	ThermoFisher Scientific J61059.K2
DTT	0.85 mM	ThermoFisher Scientific R0861
Protease inhibitors	1X	as above
α-amanitin	1X	as above
SUPERaseIN	1X	as above
Water, RNase-free		Invitrogen 10977015
Nuclei lysis buffer	**Final concentration**	**Product details**
NP-40	1%	as above
HEPES pH7.5	20 mM	ThermoFisher Scientific J60712AK
NaCl	300 mM	as above
Urea	1 M	Sigma-Aldrich U5128
EDTA	0.2 mM	as above
DTT	0.85 mM	as above
Protease inhibitors	1X	as above
α-amanitin	1X	as above
SUPERaseIN	1X	as above
Water, RNase-free		as above
Chromatin wash buffer	**Final concentration**	**Product details**
PBS	1X	as above
DTT	0.85 mM	as above
Protease inhibitors	1X	as above
α-amanitin	1X	as above
SUPERaseIN	1X	as above

2.13.3 Spike-in addition and RNA extraction from chromatin

The volume of the thawed homogenized chromatin (V_{chr}) of the HEK293T cell and the knockouts was precisely measured. The chromatin from the spike-in cells was mixed in as ⅛ of V_{chr}. The samples were added to pre-spun 15 mL MaXtract High density tubes (Qiagen, 129056). PBS was added as the ratio of V_{chr}/7.5 and chloroform was added as the ratio of V_{chr}/5. The tube was mixed by vigorously shaking for 15 s, followed by incubation for 2 min at room temperature. The nucleic acids were separated by centrifuging at 1,500xg, for 5 min, at 4°C. The upper phase was transferred into fresh 15 mL tubes (Eppendorf, EP0030122216). V_{chr}/2 of 100% isopropanol (Sigma-Aldrich, 278475) was added. The nucleic acids were precipitated for 30 min on ice, followed by centrifugation at 12,000xg, for 30 min, at 4°C. 500 µL of 80% ethanol (VWR, V1016) was carefully added to the pellet. With a wide bore tip, the pellet was transferred into a DNA lobind 1.5 mL tube. Another 500 µL of 80% ethanol was added into the other tube to get any pellet crumbs and transferred to the rest of the sample. The nucleic acids were precipitated by centrifugation at 20,000xg, for 5 min, at 4°C. The pellet was washed with 1 mL of 80% ethanol and centrifuged again. The pellet was dried, resuspended in 85 µL of nuclease-free water, and stored at -80°C overnight.

Next, 2.5 µL of DNase Turbo and 10 µL of the DNase Turbo buffer (Thermo Fisher Scientific, AM2238) were mixed into the 85 µL of the extracted nucleic acids and incubated at 37°C for 30 min in a thermomixer. Additional 2.5 µL of DNase Turbo were mixed in, and the incubation was repeated. The DNase was inactivated by 3 µL of 0.5 M EDTA. The samples were transferred into pre-spun 2 mL phase lock tubes (Qiagen,129056). The following was added to the samples: 100 µL of nuclease-free water, 200 µL of phenol chloroform isoamyl alcohol (PCI) (ROTH, A156). The aqueous phase was separated by centrifugation at 12,000xg, for 5 min, at 4°C and transferred into a 1.5 mL DNA lobind tube. The RNA was precipitated by addition of 20 µL of 0.5 M NaCl and 220 µL of 100 isopropanol, incubation on ice for 30 min, and centrifugation at 20,000xg for 30 min, at 4°C. The pellet was washed with 500 µL of 85% ethanol (at -20°C) two times, dried and resuspended in 100 µL of nuclease-free water. The RNA concentration was measured with Qubit RNA broad range. At least 150 µg of RNA is necessary to proceed with the next step.

2.13.4 Biotinylation and pull-down of 4sU-labeled RNA

To have enough material for library preparation, two biotinylation reactions, 50-100 µg each, are required. RNA is diluted in 200 µL of water and denatured at 65°C for 5 min and incubated on ice for 2 min. 3 µL of the biotin buffer and 50 µL of freshly prepared 0.1 mg/mL MTSEAbiotin-XX linker (Biotinum, cat. nr. 90066) in *N,N*-dimehtylformamide (DMF) (Sigma D4551) were added, briefly vortexed, and incubated for 30 min, at room temperature, rotating at 40 rpm. To clean up the reaction, 250 µL of PCI were added to the samples, transferred into pre-spun 2 mL phase-lock tubes, and centrifuged at 12,000xg, for 30 s, at 4°C. The aqueous phase was transferred into a new tube and mixed with 1/10 of the sample volume of 5 M NaCl, 1.1 of the sample volume of 100% isopropanol, incubated on ice for 30 min, centrifuged at 20,000xg, for 30 min, at 4°C. The pellet was washed with ice-cold 500 µL of 85% ethanol and centrifuged at 20,000xg, for 30 min, at 4°C, dried and resuspended in 50 µL of nuclease-free water.

The RNA was denatured at 65°C for 10 min and incubated on ice for 5 min. 200 µLof the µMACS streptavidin MicroBeads (Miltenyi Biotec, cat. nr. 130-092-948) were added to the RNA and incubated at room temperature for 15 min, rotating at 50 rpm. The µ columns (Miltenyi Biotec 130-042-701) were placed into the OctoMACS separator (Miltenyi 130-042-109) and 100 µL of nucleic acid equilibration buffer (provided in the kit, Miltenyi Biotec, cat. nr. 130-092-948) were applied and incubated for 15 min. The RNA-bound beads were added to the columns and the first flow-through was re-applied. The columns were washed three times with 1 mL of the pull-out-wash buffer pre-heated to 65°C, followed by three washes with the same buffer at room temperature. The RNA was eluted with 100 µL of the elution buffer with a 5-minute incubation, followed by the additional 100 µL. The RNA was concentrated using the RNA Clean and Concentrate kit (ZYMO Research, R1013) as instructed by the manufacturer, except the elution was done twice with 10 µL and 7 µL. The RNA concentration was measured with Qubit Broad Range. At least 3 µg of RNA is necessary for library preparation. In case of a low yield, the biotinylation reaction was repeated from any remaining chromatin-extracted RNA.

Table 16: Biotinylation and column washout buffers for HiS NET-seq

Biotin buffer	**Final concentration**	**Product details**
Tris-HCl pH 7.5	833 mM	Quality Biological Inc., 351006721EA
EDTA	83.3 mM	Invitrogen, AM9260G
Water, nuclease-free		Invitrogen, 10977015
Pull-out-wash buffer	**Final concentration**	**Product details**
Tris-HCl pH 7.5	100 mM	as above
EDTA	10 mM	as above
NaCl	1 M	Invitrogen, AM9760G
Tween20	0.1% (vol/vol)	Sigma-Aldrich, P9416
Water, nuclease-free		as above

2.13.5 HiS-NET-seq library preparation

The libraries were prepared by Susanne Freier. Per each cell line, three 1 µg reactions were carried out in parallel. The control (HEK293T) and the knockouts (TCEA1 KO, TCEA2 KO, DKO) were prepared as one set, on the same day.

Barcode ligation and RNA fragmentation

RNA was denatured at 70°C for 3 min and incubated on ice for 2 min. A DNA barcode, comprising a unique molecular identifier (decamer) and an internal 8 nt-long index (5‘-rApp/(N)10CTGTAGGCACCATCAAT/3‘-ddC), was ligated to the 3' end of the 4sU-labeled RNA transcripts. Into 1 µg of RNA, 1x T4 RNA ligase buffer, 200 U of truncated T4 RNA ligase 2 (NEB, M0239S), 1 µg of barcode DNA linker, PEG8000, 20% (vol/vol), DMSO, 10% (vol/vol), and up to 20 µL water were mixed in and incubated at 37°C for 3 hours. 2 µL of the fragmentation solution (NEB, E6150S) were added, mixed and incubated in a cycler at 95°C for 10 min. 2 µL of the provided STOP solution were thoroughly mixed in. The ligated RNA was cleaned up using the RNA Clean and concentrator kit (ZYMO Research, R1013) as described in the manual and was eluted twice with 10 µL of water (total V=20 µL).

RNA size selection

20 µL of the 2x TBE-urea buffer were added to the samples, followed by denaturation at 70°C for 3 min and incubation on ice for 2 min. The RNA was size-separated by electrophoresis in a 15% TBE-Urea gel (Invitrogen, EC6885BOX) at 200 V for 70 min, with gel pre-run for 15 min before sample loading. The gels were stained with 1x SYBR Gold (Invitrogen, SS1194) for 5 min, and, using the 20 bp DNA ladder (Takara, 3420A), the RNA between 60 and 140 nt- long were cut out on a blue light transilluminator. RNA was extracted from the gel using ZR-small RNA gel recovery kit (Zymo Research, R1070), as instructed by the manufacturer, and eluted in 12 µL, the eluate was re-applied and eluted again.

Reverse transcription

The RNA was mixed with a ToToRo_NET-seq reverse primer (table 16) and 10 mM dNTPs, denatured at 80°C for 2 min, then annealed at 65°C for 5 min, and cooled on ice for 2 min. From the reverse transcription kit (ThermoFisher Scientific, 18090010), 3.5 µL of SuperSCript IV Transcriptase 5x buffer, 0.85 µL of DTT, 170 U of SuperSCript IV, and 0.85 µL of RNase OUT (Invitrogen, 10777019) were thoroughly mixed with the RNA and incubated at at 55°C for 20 min. The RNA was degraded by alkaline hydrolysis: addition of 1.8 µL of 1 M NaOH and incubation at 98°C for 20 min, followed by neutralization with 1.8 µL of 1 M HCl.

Single-stranded (ss) DNA size selection and extraction

17 µL of 2x TBE-urea buffer were added to the samples and 10 µL - to the prepared 10 µL of 20 bp DNA ladder (Takara, 3420A). Both the samples and the ladder were denatured at 95°C for 3 min, loaded in 10% TBE-Urea gel (Invitrogen EC6875BOX), and run at 200 V for 65 min, with a 15 min pre-run prior to sample loading. 100-180 nt-long ssDNA was cut out from the gel and homogenized by centrifugation at 10,000xg at room temperature through a 0.5 mL DNA lobind tube, with two holes pierced with a needle, placed into a whole 1.5 mL lobind tube. 400 µL of RNA recovery buffer (from kit Zymo Research, R1070) were added to the slurry and incubated at 65°C, at 1,400 rpm, for 15 min. The samples were frozen at -80°C for 5 min and incubated at 65°C for 5 min. The slurry was transferred into Zymo-SpinTM III-F filter (Zymo Research C1057-50) inserted into a collection tube and spun at 5,000xf for 6 min. The flow-through was transferred into a Zymo-SpinTM IIICG Column (Zymo Research, C1006-50-G) in a collection tube and centrifuged at 2,000xg for 30 s. The flow-through volume was measured and twice as much RNA MAX buffer was added. The mixture was loaded into a Zymo-SpinTM IC Column (Zymo Research, C1004-50) and centrifuged at 12,000xg for 30 s. The flow-through was discarded, the remaining samples were re-loaded and spun down again. 400 µL of RNA Prep Buffer was added, followed by a centrifugation at 12,000xg for 1 min. 800 µL of RNA Wash Buffer was added to the column followed by the same centrifugation. The wash step was repeated with 400 µL of the same buffer. The column was dried by centrifugation at the same speed, for 2 min. The column was transferred into a fresh DNA lobind tube. 16 µL of DNase/RNase-free water was applied directly on the filter and incubated for 2 min at room temperature. The ssDNA was eluted by centrifugation at 10,000xg for 1 min.

cDNA circularization and clean-up

Circularization was done as described by the manufacturer (Lucgien, CL411K), except half the amount of $MnCl_2$ was added. The circularized cDNA was purified using the RNA clean and concentrator 5’ kit (Zymo Research, R1013) and eluted in 16 µL.

Library amplification

In order to minimize the amount of redundant PCR duplicates, a test PCR was performed to determine the optimal number of cycles (6-10). After the PCR, the samples were run on the 4% E-Gel EX Agarose-Gel (Invitrogen) for 15 min. The PCR reaction was repeated with the optimal number of cycles, usually 8-9 and run on a gel for 25 min. The amplified libraries were cut out and extracted from gel NucleoSpin Gel & PCR clean-up (Macherey-Nagel). The DNA concentration was measured with Qubit (HS DNA kit). The quality of libraries was assessed with BioAnalyzer (HS DNA kit). The average size of the libraries should be at about 180 nt. The libraries were given to the sequencing facility of MPIMG for sequencing on NovaSeq 6000 (Illumina).

Table 17: HiS-NET-seq library amplification test PCR

PCR reaction			**Cycler program**		
Phusion high-fidelity (HF) DNA polymerase	0.18 µL	NEB, M0530S	Denature	98°C	30 s
5x buffer	3.6 µL		Cycles 6-10	98°C	10 s
10 mM dNTPs	0.36 µL	ThermoFisher R0194		60°C	10 s
10 mM oNTI231 primer	0.9 µL	CAAGCAGAAGACGGCATACGA		72°C	5 s
10 mM 10-ToToRo	0.9 µL	AATGATACGGCGACCACCGAGATCTACACGATCGGAAGAGCACACGTCTGAACTCCAGTCACCCAACATTTCCGACGATCATTGATGGTGCCTA*C*A*G			
Water	Up to 18 µL	Invitrogen, 10977015			

2.13.6 HiS-NET-seq data processing and computational analyses

The data were processed by Mario Rubio. In summary, the adapter sequence was trimmed with Cutadapt [332]. PCR duplicates were removed using unique molecular identifiers (UMI) with Starcode [342] The reads were mapped to the combined reference genome consisting of the human genome GRCh38 (release 28) and the mouse genome GRCm38 (release M18) using STAR aligner [337]. The UMIs were removed and the mapping position of the 5' end reads was recorded for uniquely mapped reads. Multimapping reads were removed. The reads that aligned to the 3' ends of introns and exons were removed as splicing intermediates. The following short RNAs were also removed: 5S, 7SK, HY1, HY2, HY3, HY4, HY5, U1, U2, U3, U4, U5, U6, U7, U8, U13, U14, U17, tRNA, Y-RNA, antisense RNA, guide RNA, miRNA, telomerase RNA, vault RNA, miscRNA, ncRNA, rRNA, sRNA, scRNA, scaRNA, snRNA, snoRNA, LSU-sRNA, Rnase MRP, Rnase P, and SSU rRNA.

Spike-in normalization was based on the median-of-ratios method [343] as implemented in the DESeq2 package [339] to determine the factor reflecting the content of spike-in mouse reads in each sample. The normalization factors were computed using the genes from the mouse genome GRCm38 (release M18). We implemented a variation of the method described in [344] and selected for normalization only the 10% of mouse genes with the lowest coefficient of variation (standard deviation divided by mean) computed from the raw counts considering all the samples in our experiment.

Active genes were assigned as follows: 1) active transcript isoforms were defined as those with the number of mapped reads <0 in at least one of the WT replicates and with the average of 2 biological replicates of TPM >=1; 2) active genes were defined by grouping the transcript isoforms per gene and defining the gene TSS as the earliest TSS of all corresponding transcript isoforms, while the gene pA site was defined as the latest pA site of all corresponding transcript isoforms; 3) the genes on chromosomes Y and M were excluded, 4) the nascent transcription was quantified for each gene in TPMs and filtered out the genes with <=0.25 TPM in all the samples and <=5 TPM on average across all the samples.

Differential Pol II occupancy among the cell lines was calculated using DESeq2 at the promoter-proximal region (defined as TSS +300 bp) and the gene body (defined as +300 bp from TSS until the PAS) of active genes, ≥1 kb in length. The log2(KO/WT) in the promoter-proximal region was plotted against log2(KO/WT) in the gene body to obtain the Pol II pausing matrix.

Chapter 3

RESULTS

3.1 TFIIS diversified into paralogs as early as the evolution of jawless vertebrates

During metazoan evolution, the transcription elongation factor S-II (TFIIS) remained remarkably conserved and diversified into a few paralogs that were experimentally validated in human, mouse and frog cells [284]. We wanted to get more insight into TFIIS evolution by estimating when during evolution the diversification into paralogs had taken place. Taking advantage of the currently available whole genome data of numerous organisms in the Ensembl genome browser, we examined the presence of TFIIS genes in various taxonomic groups and summarized our observations in a cladogram (Figure 7). Our search suggested that TFIIS paralogs, however not a complete set of four, were present in the common ancestor of vertebrates, suggesting that their emergence could be a consequence of the two whole genome duplication events that as recent analysis revealed, preceded the ancestral vertebrate [345]. Surprisingly though, four TFS paralogs were recently found in the archaea species *Sulfolobus solfataricus* [346], indicating that TFIIS evolutionary history is more complicated than expected, since TFIIS may have its own evolutionary history in at least some archaeal phyla. Tracking TFIIS evolution across phyla of archaea and eukarya will require a deeper phylogenetic analysis, so our cladogram is meant to convey the presence or absence of the homologs and paralogs in model organisms with a focus on vertebrate evolution.

Unexpectedly, we found that *TCEA1* was not present in early vertebrates. For validation, we narrowed down further the search on the transition from invertebrates to jawless vertebrates (cyclostomes) and to jawed vertebrates (gnathostomes). We scanned the model organism genomes of cephalochordates and tunicates, with *TCEA1*-, *TCEA2*-, and *TCEA3*- calculated cDNA of elephant shark as input. This species was chosen as input because its genome is well annotated and it belongs to cartilaginous fish class, which is between the jawless vertebrates and Osteichthyes. We confirmed that tunicates and cephalochordates have one TFIIS gene. Analysis of the cyclostomata, which comprises the extant hagfish and lamprey, revealed the presence of *TCEA2,* but not *TCEA1.* In the inshore hagfish (*Eptatretus burgeri*) genome, two genes, *TCEA2* and *TCEANC,* were detected. The lamprey genomes of *Petromyzon marinus* and *Lethenteron camtschaticum* contain *TCEA3*, in addition to *TCEA2* and *TCEANC*. The four paralogs, *TCEA1, TCEA2, TCEA3, and TCEANC,* were found in the two extant gnathostome taxa, Chondrichthyes (cartilaginous fishes) and Osteichthyes (bony fishes). The latter encompass at least 99% of all living jawed vertebrate species [347].

Taken together, our analysis indicates that TFIIS diverged into TCEA paralogs during early vertebrate evolution: some paralogs exist in primitive vertebrates. *TCEA1* appears to have evolved later, in jawed vertebrates.

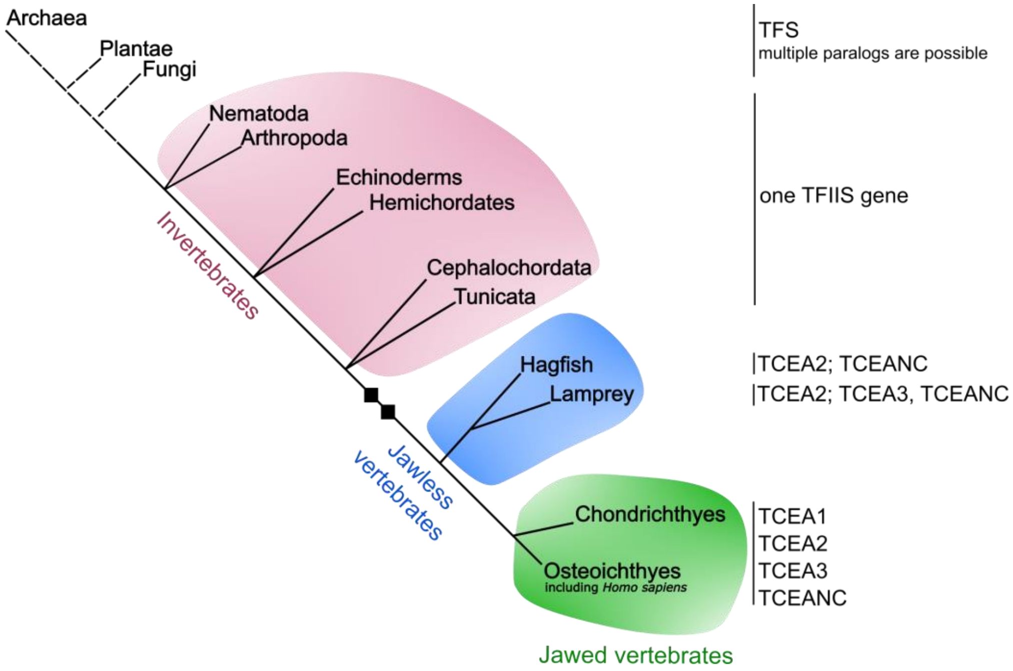


Figure 7: Cladogram illustrating the presence of known TFIIS genes.
Phylogenetic arrangement was done based on cDNA homology and in the context of vertebrate lineage. The dashed branches indicate that those genomes have TFIIS homologs, however, the presence or absence of paralogs needs to be determined in a large number of taxonomic groups. Whole genome duplication events are marked as black squares.

3.2 Creating cellular models for the functional analysis of TCEA1 and TCEA2

3.2.1 Inducible expression of epitope-tagged TCEA1 and TCEA2

We set out to determine the interaction partners of TCEA1 and TCEA2. A paralog-unique interaction partner would potentially point to the unique function of each paralog. However, immunoprecipitation-based detection of TCEA1 and TCEA2 is challenging for two reasons: 1) TCEA1 and TCEA2 are very similar in their structure and amino acid sequence (Figure 8) and, currently, there is no optimal antibody for efficient, paralog-specific immunoprecipitation and 2) TCEA2 is rather lowly expressed in HEK293T cells (Figure 10A). To circumvent these limitations, we chose the Flp-In T-Rex system (described in chapter 2.3.1) and generated stable cell lines that can be induced to express TCEA1 and TCEA2 (Figure 9A & B). The epitope tag was added at the N-terminus because the C-terminal domain directly fits into the active site of Pol II [262] and any perturbation could disturb the RNA cleavage. We have also investigated the interactomes of the N-terminal domain with the unconserved and unstructured linker of TCEA1 and TCEA2 to gain insight into the function of this domain.

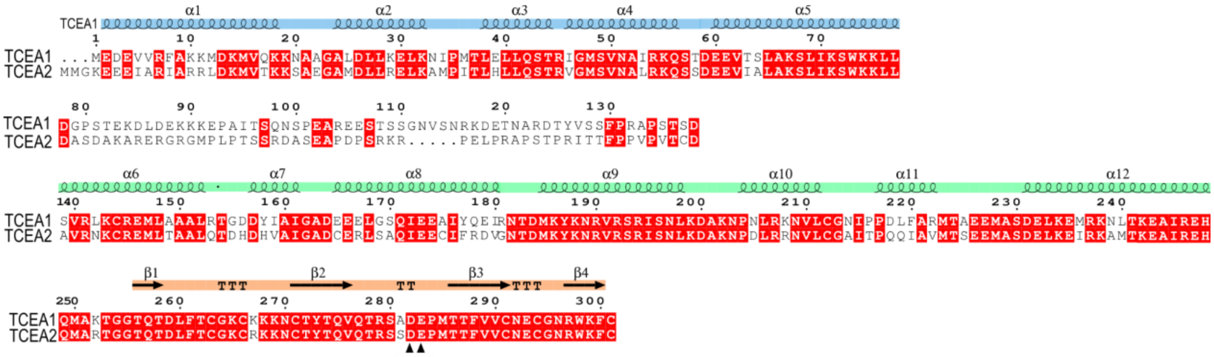

α1
α2
α3
α4
α5
TCEA1
1
10
20
30
40
50
60
70
TCEA1 ...MEDEVVRFAKKMDKMVQKKNAAGALDLLKELKNIPMTLELLQSTRIGMSVNAIRKQSTDEEVTSLAKSLIKSWKKLL
TCEA2 MMGKEEEIARIARRLDKMVTKKSAEGAMDLLRELKAMPITLHLLQSTRVGMSVNALRKQSSDEEVIALAKSLIKSWKKLL
80
90
100
110
20
130
TCEA1 DGPSTEKDLDEKKKEPAITSQNSPEAREESTSSGNVSNRKDETNARDTYVSSFPRAPSTSD
TCEA2 DASDAKARERGRGMPLPTSSRDASEAPDPSRKR.....PELPRAPSTPRITTFPPVPVTCD
α6
α7
α8
α9
α10
α11
α12
140
150
160
170
180
190
200
210
220
230
240
TCEA1 SVRLKCREMLAAALRTGDDYIAIGADEEELGSQIEEAIYQEIRNTDMKYKNRVRSRISNLKDAKNPNLRKNVLCGNIPPDLFARMTAEEMASDELKEMRKNLTKEAIREH
TCEA2 AVRNKCREMLTAALQTDHDHVAIGADCERLSAQIEECIFRDVGNTDMKYKNRVRSRISNLKDAKNPDLRRNVLCGAITPQQIAVMTSEEMASDELKEIRKAMTKEAIREH
β1
TTT
β2
TT
β3
TTT
β4
250
260
270
280
290
300
TCEA1 QMAKTGGTQTDLFTCGKCKKKNCTYTQVQTRSADEPMTTFVVCNECGNRWKFC
TCEA2 QMARTGGTQTDLFTCGKCRKKNCTYTQVQTRSSDEPMTTFVVCNECGNRWKFC
▲▲

Figure 8: Amino acid sequence alignment of human TCEA1 and TCEA2.
The alignment was done in Clustal W [288] and analyzed in ESPript 3.0 [289]. The predicted secondary structures are denoted as loops for α-helix and an arrow for a β-sheet, turns with the letter T. The domain architecture was added based on TCEA1 structures available in the Protein Data Bank. The N-terminal, central, and C-terminal domains are denoted in blue, green, and orange bars, respectively. The region between the N-terminal and central domain is unstructured. The two acidic amino acids (D and E), responsible for RNA cleavage, are marked with black triangles. The identical residues are in red blocks.

Construct design for the expression of epitope-tagged TCEA1, TCEA2, TCEA1-NTDL, and TCEA2-NTDL

Based on our RNA-seq data in HEK293T cells (Figure 10A), we determined which transcript isoforms are predominantly expressed. The cDNA of *TCEA1* isoform ENST00000521604 encoding the 301 amino acid protein (34 kDa) and *TCEA2* isoform ENST00000343484 encoding the protein of 299 amino acids (33.6 kDa) were chosen for overexpression. To generate the cell lines with inducible expression of the most dissimilar part of TCEA1 and TCEA2, we chose to express the N-terminal domain and the subsequent disordered part of the proteins that we refer to as “NTDL”: the first 139 amino acids (15.4 kDa) and 137 amino acids (15.1 kDa) of TCEA1 and TCEA2, respectively. At the N-terminus of each protein, we included two affinity tags: HBH (6xHistidine-Biotin-6xHistidine) and 3xFLAG, separated by a TEV protease cleavage site. The total size of the added tags is 13 kDa.

Tetracycline-mediated induction of TCEA1, TCEA2, TCEA1-NTDL, and TCEA2-NTDL

We tested the level of overexpression of each protein in isogenic populations, by performing a time course experiment of Tetracycline treatment and evaluated the expression with Western blot against the FLAG epitope and the endogenous TCEA1. However, the latter antibody turned out to be unspecific to TCEA1, as it had cross-reacted with the overexpressed TCEA2 (Figure 9C & D). We found that a moderate dose of Tetracycline (1 µg/mL) is sufficient to induce the expression of all constructs in 6 hours.

Interestingly, the expression of FLAG-TCEA2 is noticeably higher than that of FLAG-TCEA1. The reason for this is puzzling because the coding region is of the same length and is inserted into the same locus. Possibly, a higher codon adaptation index of 0.84 compared to 0.73 of TCEA1 made the expression of TCEA2 more efficient (the index was calculated using GenScript Rare Codon Analysis Tool). Fine-tuning the overexpression level of four constructs by titration of Tetracycline is challenging and would introduce additional variabilities. For our immunoprecipitation experiments, we chose to treat all cell lines with Tetracycline for 24 hours because a good level of expression is consistent among all cell lines and the expression levels increased within physiological range: approximately 1.4-fold for FLAG-TCEA1, 1.6-fold for both TCEA1-NTDL and TCEA2-NTDL, and 3-fold for FLAG-TCEA2 compared to the endogenous TCEA1/2 (quantified WB in supplemental figure S1).

In summary, this cellular system is reliable for expressing epitope-tagged full length TCEA1 and TCEA2, as well as the truncated versions, enabling paralog-specific immunoprecipitation.

We used this system to determine the interaction partners of TCEA1 and TCEA2 and gained insight into the potential function of their NTDL by performing IP-MS (immunoprecipitation with mass spectrometry). We have also performed ChIP-seq (chromatin immunoprecipitation followed by sequencing) to determine where in gene regions TCEA1/TCEA2 bind Pol II to relieve backtracking (sections 3.3 and 3.4).

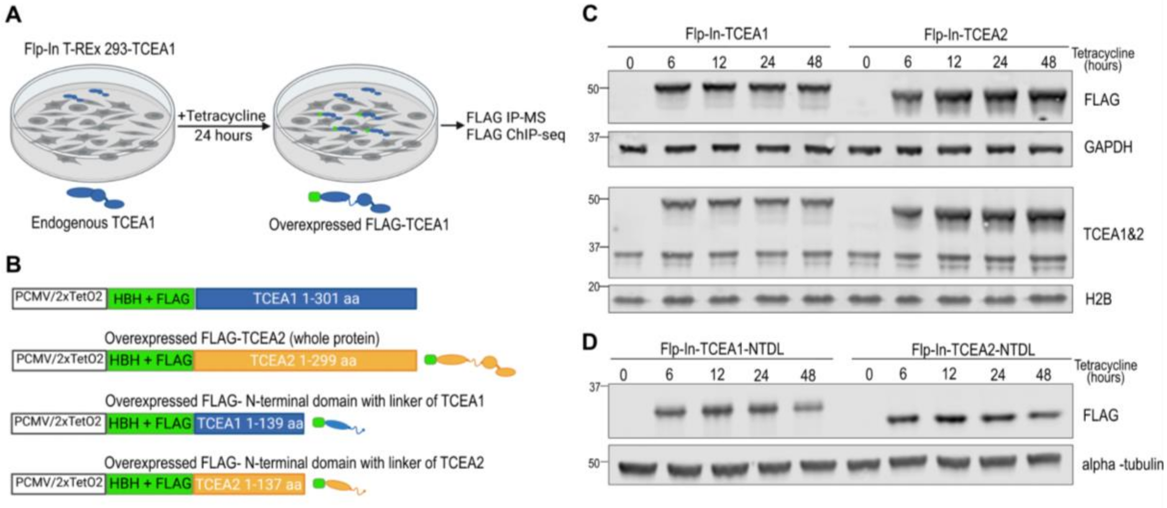


Figure 9: Inducible expression of epitope-tagged TCEA1, TCEA2, and the N-terminal domain and linker (NTDL) of TCEA1 and TCEA2

A. Schematic of tetracycline induction of FLAG-TCEA1 in Flp-In TREx-293-TCEA1 cells; **B.** Tetracycline-inducible locus in Flp-In TREx-293 cells with the integrated cDNA constructs for the expression of the whole TCEA1 and TCEA2 proteins and the NTDL of TCEA1 and TCEA2, all with histidine-biotin-histidine and 3xFLAG tags; **C.** Western blots showing the time course of Tetracycline induction of expression. The blots are against FLAG and endogenous TCEA1 and TCEA2. The latter shows the expression of the tagged TCEA1, the endogenous TCEA1, as well as cross-reactivity of the antibody with the overexpressed TCEA2. **D.** Western blots against FLAG showing the expression of the tagged TCEA1-NTDL and TCEA2-NTDL at the varying duration of tetracycline treatment.

3.2.2 Generating the *TCEA1*-, *TCEA2*-, and double *TCEA1 and TCEA2* knockouts

We determined the active transcript isoforms in HEK293T with RNA-seq (Figure 10A). The aim of our design was to completely eliminate all transcript isoforms as close to the first common exon, as possible and to make a small frameshift-causing deletion to minimize disruption of any potential *cis*-regulatory elements. At first, we targeted Cas9 to make a deletion, under 70 base pairs, in the first exon, in which the RNA-seq signal of all potential isoforms of TCEA1 and TCEA2 was present. Having genotyped the single clones with PCR to find homozygotes and confirmed them with Sanger sequencing, we were not able to validate the lack of the protein with Western blot due to the lack of reliable paralog-specific antibodies. To validate the knockouts, we generated additional cell lines with the CRISPR/Cas9 target in the second actively transcribed exon and kept searching for more specific antibodies. The design of the deletions for all knockouts is shown in Figure 10, and the Sanger and Western blot genotyping is shown in Figure 11. For all the generated knockouts, we preserved at least 3 independent homozygous clones to be able to validate the phenotypes.

Generating the TCEA1 knockout (TCEA1 KO) cell line

Targeting *TCEA1* for mutagenesis is challenging because one of its pseudogenes (*TCEA1P2*) is almost identical to the exon sequence of the *TCEA1* gene. To achieve a proper knockout of *TCEA1*, we had to carry out two CRISPR experiments. In the first experiment, we aimed to make a deletion that would eliminate any possibility of expression of a truncated protein-coding isoform. We targeted the first expressed exon, and the homozygous single clones had a 67 base pair deletion (Figure 10B). However, the sgRNA was not unique to the gene and also matched the pseudogene *TCEA1P2*, located in an intron of *GOLGA4*. Having genotyped multiple homozygous clones for the off-target mutagenesis, we found that all of them also had a heterozygous mutation in the pseudogene (shown in 2.4.5). Although the mutation was in an intron and would likely be spliced out, we generated an additional cell line without off-target editing, a TCEA1 knockout with a deletion (46 bp) in exon 2 (Figure 10B). The genotyping for the new single-cell knockout (TCEA1exon2 KO) is shown in Figure 11A. On the genotyping Western blot against TCEA1, the TCEA1 band, present in the unmodified HEK293T cells, is missing in the TCEA1 KO (Figure 11D). The intensity of unspecific bands is increased in the knockout because the main target is absent. *In silico* off-target analysis did not point to probable mutations (Table 6). We used the new clone for the RNA-seq and HiS-NET-seq experiments (sections 3.6 and 3.7). The growth and cell cycle analyses were done in both TCEA1exon1 KO and TCEA1exon2 KO (section 3.5).

Generating the TCEA2 knockout cell line

At the *TCEA2* locus, there are potentially three protein-coding isoforms that can be expressed (Figure 10A). Transcript ENST00000343484 encodes 299 amino acids, which constitute the complete three-domain TFIIS protein. Its first exon is clearly expressed, so we chose to target it for deletion (67 base pairs) (Figure 10C). Despite the design, this deletion did not disrupt the open reading frame of another expressed short protein-coding isoform

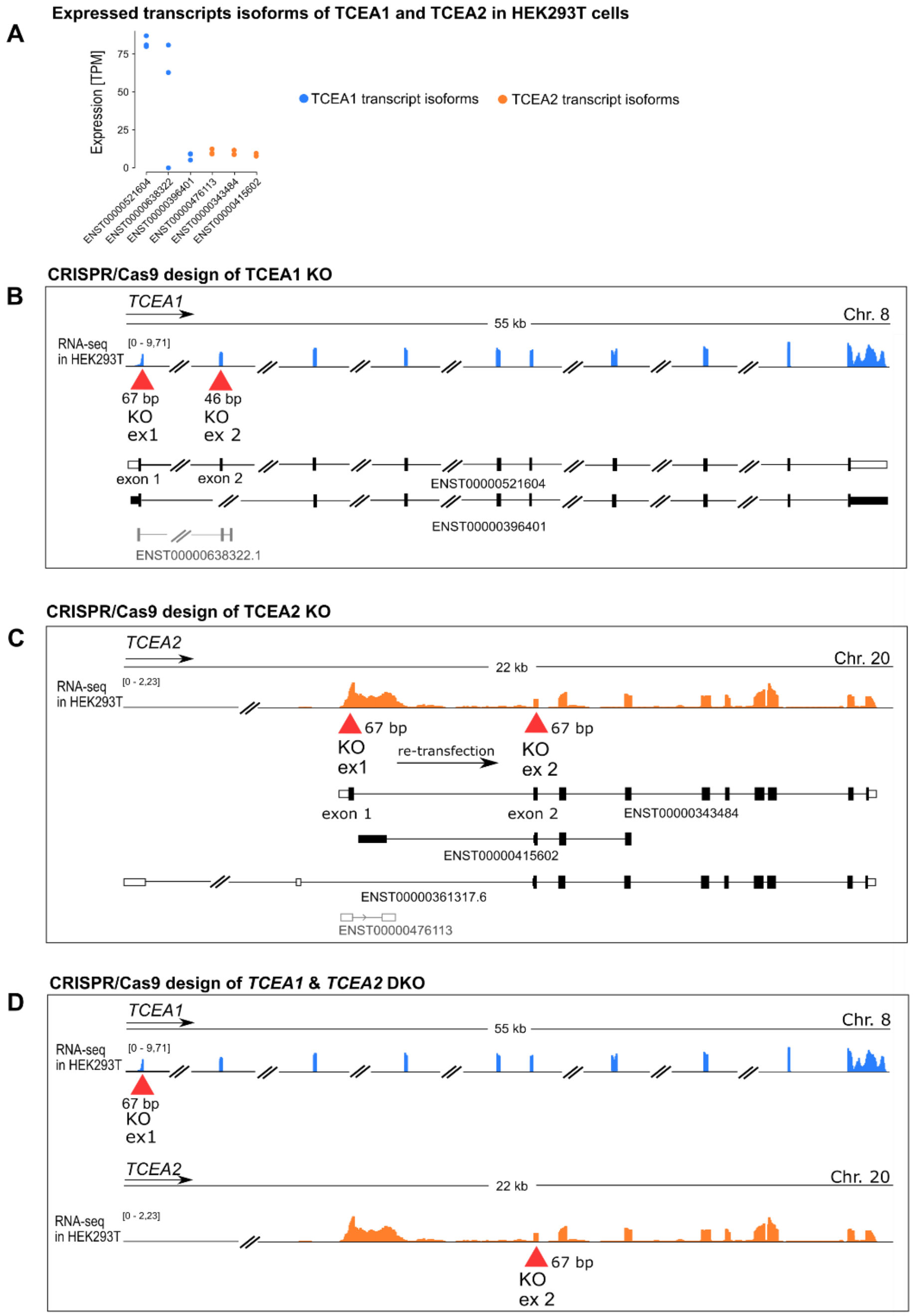


Figure 10: CRISPR/Cas9 design of the TCEA knockout cell lines.
A. Expression of all active transcript isoforms of *TCEA1* and *TCEA2* in HEK293T cells detected with RNA-seq; **B-D** The protein-coding isoforms are in black, the protein-non-coding isoforms are in gray. Red triangles point to the exons in which deletions were made with CRISPR; **B.** mRNA of *TCEA1* is illustrated with the blue RNA-seq track. TCEA1 KOex1 is the first generated cell line, which has an off-target mutation. An additional TCEA1 KO with the deletion in exon 2 was generated; **C.** mRNA of *TCEA2* is illustrated with the orange RNA-seq track. TCEA2 KOex1 is the first generated cell line, an additional cell line was generated by re-transfecting the TCEA2 KOex1 cell line to make a deletion in exon2 to ensure a complete knockout; **D.** DKO was generated by transfecting the TCEA1 KOex1 cell line to make a deletion in exon 2 of TCEA2.

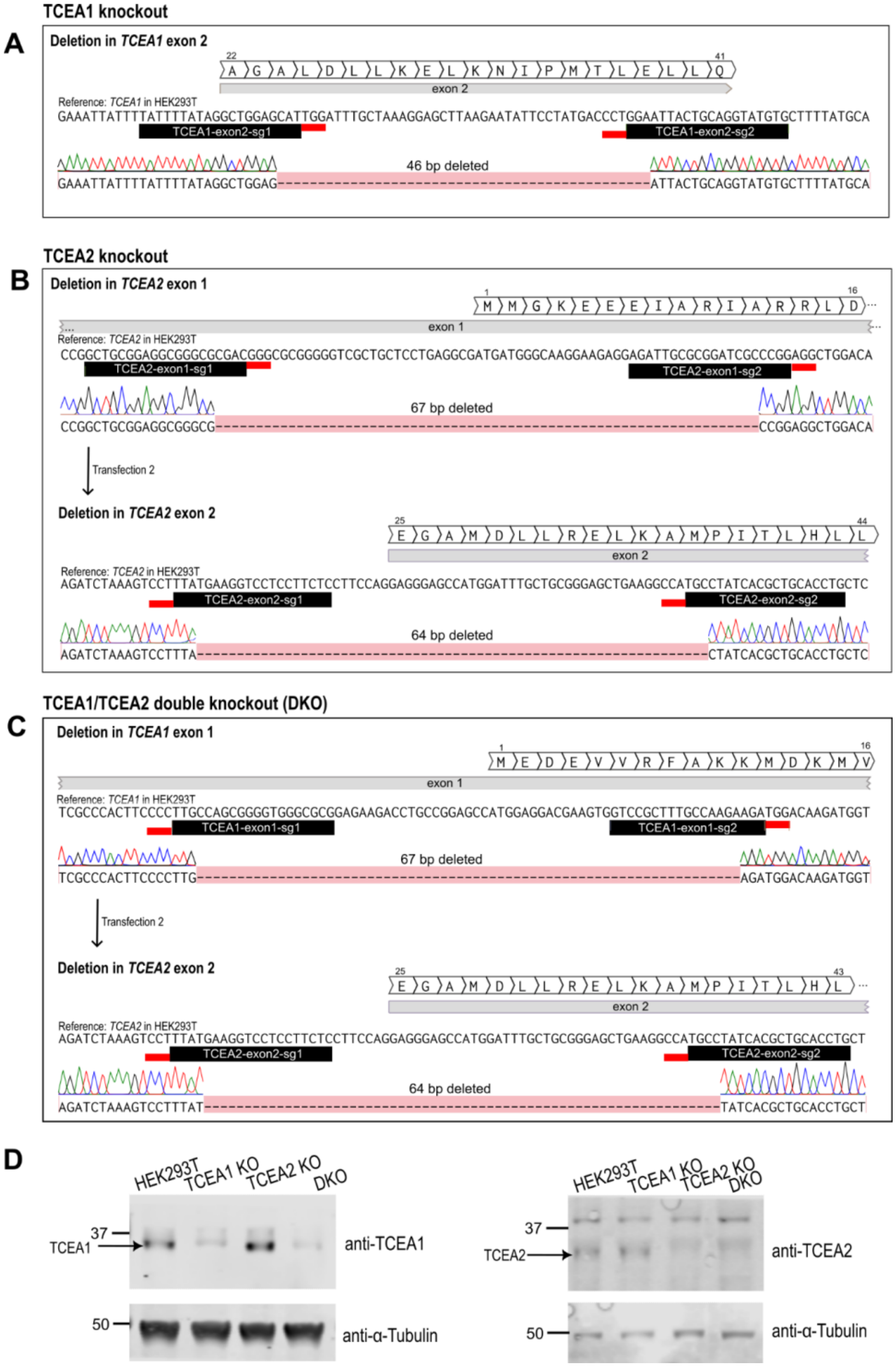

A
TCEA1 knockout
Deletion in TCEA1 exon 2
22
41
A G A L D L L K E L K N I P M T L E L L Q
exon 2
Reference: TCEA1 in HEK293T
GAAATTATTTTATTTTATAGGCTGGAGCATTGGATTTGCTAAAGGAGCTTAAGAATATTCCTATGACCCTGGAATTACTGCAGGTATGTGCTTTTATGCA
TCEA1-exon2-sg1
TCEA1-exon2-sg2
46 bp deleted
GAAATTATTTTATTTTATAGGCTGGAG
ATTACTGCAGGTATGTGCTTTTATGCA
B
TCEA2 knockout
Deletion in TCEA2 exon 1
1
16
M M G K E E E I A R I A R R L D
exon 1
Reference: TCEA2 in HEK293T
CCGGCTGCGGAGGCGGGCGCGACGGGCGCGGGGGTCGCTGCTCCTGAGGCGATGATGGGCAAGGAAGAGGAGATTGCGCGGATCGCCCGGAGGCTGGACA
TCEA2-exon1-sg1
TCEA2-exon1-sg2
67 bp deleted
CCGGCTGCGGAGGCGGGCG
CCGGAGGCTGGACA
Transfection 2
Deletion in TCEA2 exon 2
25
44
E G A M D L L R E L K A M P I T L H L L
exon 2
Reference: TCEA2 in HEK293T
AGATCTAAAGTCCTTTATGAAGGTCCTCCTTCTCCTTCCAGGAGGGAGCCATGGATTTGCTGCGGGAGCTGAAGGCCATGCCTATCACGCTGCACCTGCTC
TCEA2-exon2-sg1
TCEA2-exon2-sg2
64 bp deleted
AGATCTAAAGTCCTTTA
CTATCACGCTGCACCTGCTC
C
TCEA1/TCEA2 double knockout (DKO)
Deletion in TCEA1 exon 1
1
16
M E D E V V R F A K K M D K M V
exon 1
Reference: TCEA1 in HEK293T
TCGCCCACTTCCCCTTGCCAGCGGGGTGGGCGCGGAGAAGACCTGCCGGAGCCATGGAGGACGAAGTGGTCCGCTTTGCCAAGAAGATGGACAAGATGGT
TCEA1-exon1-sg1
TCEA1-exon1-sg2
67 bp deleted
TCGCCCACTTCCCCTTG
AGATGGACAAGATGGT
Transfection 2
Deletion in TCEA2 exon 2
25
43
E G A M D L L R E L K A M P I T L H L
exon 2
Reference: TCEA2 in HEK293T
AGATCTAAAGTCCTTTATGAAGGTCCTCCTTCTCCTTCCAGGAGGGAGCCATGGATTTGCTGCGGGAGCTGAAGGCCATGCCTATCACGCTGCACCTGCT
TCEA2-exon2-sg1
TCEA2-exon2-sg2
64 bp deleted
AGATCTAAAGTCCTTTAT
TATCACGCTGCACCTGCT
D
HEK293T
TCEA1 KO
TCEA2 KO
DKO
37
TCEA1
anti-TCEA1
50
anti-α-Tubulin
HEK293T
TCEA1 KO
TCEA2 KO
DKO
37
TCEA2
anti-TCEA2
50
anti-α-Tubulin

Figure 11: Genotyping of homozygous TCEA1-, TCEA2- and TCEA1/TCEA2 knockouts. Genotyping of single-cell derrived knockout clones by Sanger sequencing of PCR amplicons with deletions in **A**. *TCEA1;* **B**. *TCEA2;* **C**. both *TCEA1* and *TCEA2.* **D**. Genotyping Western blot against TCEA1 and TCEA2.

(ENST00000415602), encoding 82 amino acids (Figure 10C). Having genotyped and found homozygous knockouts, we could not confirm the knockout on the protein level, because of the lack of a specific anti-TCEA2 antibody, at that time. Furthermore, another isoform (ENST00000361317.6), could potentially minimally contribute to expression, but could not be distinguished because of its two skipped exons far upstream of all other isoforms. Therefore, we have targeted the second exon for a deletion (64 base pairs) (Figures 10C, 11B) in a subsequent transfection of the homozygous TCEA2 exon 1 KO clone to ensure that all potential isoforms would have a deletion in the coding sequence. To decrease the intensity of unspecific bands on the Western blot, the loading amount was reduced, and it can be seen that TCEA2 signal is missing in the TCEA2 KO (Figure 11D). We used the TCEA2 knockout with deletions in the first and second exons for RNA-seq and HiS-NET-seq experiments (sections 3.6 and 3.7). The growth and cell cycle analyses were done in two independent single-cell clones of TCEA2 exon 1 KO and TCEA2 exons1 & 2 KO (section 3.5).

Generating the double TCEA1/TCEA2 knockout (DKO) cell line

A homozygous TCEA1 KO clone, with a deletion in exon 1, was re-transfected to make a deletion in exon 2 of TCEA2 (Figure 10D). A homozygous clone was used for RNA-seq and HiS-NET-seq experiments (Chapters 3.6 and 3.7). The growth and cell cycle analyses were done in two independent single-cell DKO clones (Chapter 3.5).

3.2.3 Attempts to generate degron-tagged TCEA1 and TCEA2 cell lines

Prior to generating knockouts, our initial strategy was the rapid targeted protein degradation [310] that would have enabled gene loss experiment while avoiding the indirect adaptive effects of a knockout. Unfortunately, tagging both proteins for degradation did not work as designed and this section describes our attempts. No genomic data were generated in these cell lines.

In this degron system, the degradation tag is cloned into the gene of interest and is expressed endogenously. The knock-in cells are treated with a PROTAC degrader, dTAG [310], which binds to the tagged target protein and the E3 ubiquitin ligase cereblon (CRBN) of the cellular ubiquitin-protease system, thereby bringing the target protein close to CRBN [348, 349]. Remarkably, the tagged protein gets almost completely degraded within a few hours, which enables functional analysis without adaptive effects.

For both proteins, we chose to insert the degradation tag at the N-terminus because the C-terminal domain is essential for RNA cleavage activity (described in 1.2.2). The cassette also contained a puromycin resistance gene for the selection of tagged clones and a P2A signal to separate its translation from the rest of the protein. The translated tagged TCEA1/TCEA2 was designed to lose the first 9/11 amino acids and begin with a 2xHA epitope and the FKBP12^{36V} degradation tag [310]. Using the CRIS-PITCh approach, based on microhomology-assisted DNA repair, was advantageous because cloning of the long homology arms was not necessary (Figure 12A) [324]. Despite our efforts, we were unable to create proper knock-ins using this approach, as the microhomology-mediated repair turned out to be error-prone and did not work exactly as designed.

Degron tagging of TCEA1

Out of 800 single clones, we found 4 homozygous knock-ins that expressed the tagged TCEA1 on the protein level detected by anti-HA and anti-TCEA1 (cross-reactive antibody with TCEA2) Western blots (Figure 12B). Although all homozygous clones had a single PCR product of the correct insertion size, they all had differentially repaired alleles, as evidenced by a mixed Sanger signal at the site of the tag integration (Figure 12C). We chose not to proceed with further experiments in these cells because it is unclear how each allele is affected.

Degron tagging of TCEA2

The design of the degradation tag integration was similar to that of TCEA1 (Figure 12A). Cas9 was targeted to the first exon of the main isoform encoding the whole protein (ENST00000343484) (Figure 10A). Having genotyped 500 clones, we did not find any homozygotes, but we further analyzed the heterozygotes that had a deletion causing a premature stop codon in the other allele (Figure 13B). Sanger sequencing revealed that the integration occurred almost as designed, but with frame non-shifting mutations (Figure 13C). The Sanger signal is slightly mixed in the reaction with a forward primer (upper track of Figure 13C), however, the reaction with a reverse primer revealed the correct insertion (lower track of figure 13C).

Based on the genotype, we expected to see a tagged TCEA2 of ~50 kDa (TCEA2 without 11 amino acids is ~32.6 kDa and the 2HA+degron tag is 17 kDa) and no band of the endogenous TCEA2 on a Western blot against the endogenous TCEA2. This antibody is rather unspecific as can be seen by multiple bands (Figure 13D) even when TCEA2 is highly overexpressed in

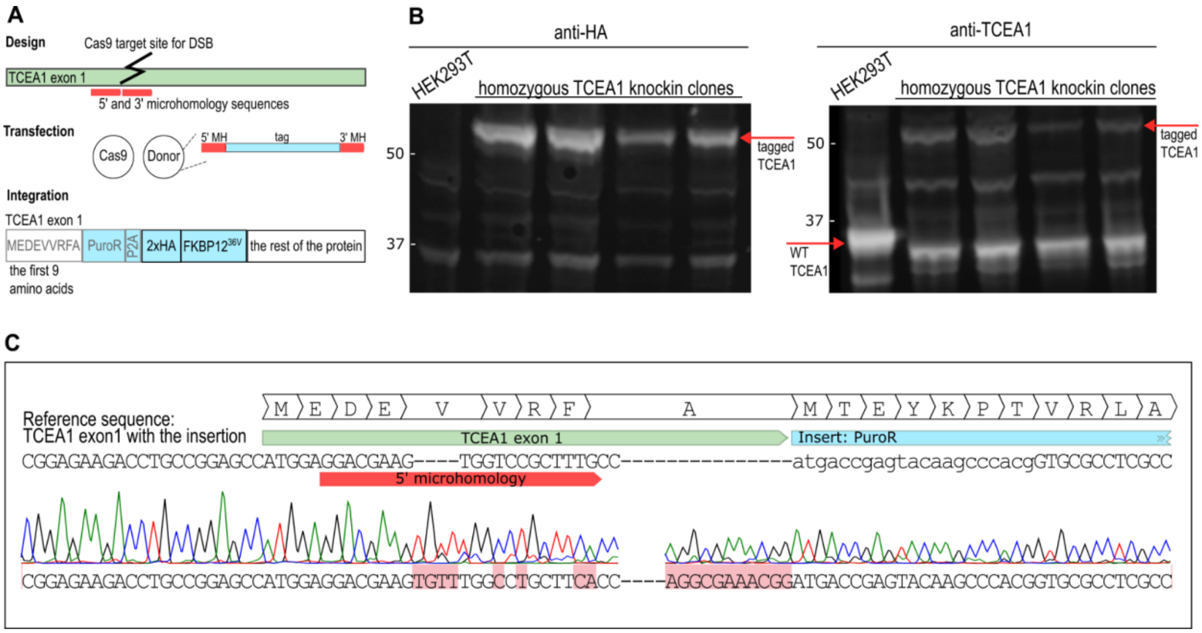

A
Design
Cas9 target site for DSB
TCEA1 exon 1
5' and 3' microhomology sequences
Transfection
Cas9
Donor
5' MH
tag
3' MH
Integration
TCEA1 exon 1
MEDEVVRFA
PuroR
P2A
2xHA
FKBP12^{36V}
the rest of the protein
the first 9 amino acids
B
anti-HA
HEK293T
homozygous TCEA1 knockin clones
50
37
tagged TCEA1
anti-TCEA1
HEK293T
homozygous TCEA1 knockin clones
50
37
tagged TCEA1
WT TCEA1
C
M E D E V V R F A M T E Y K P T V R L A
Reference sequence:
TCEA1 exon1 with the insertion
TCEA1 exon 1
Insert: PuroR
CGGAGAAGACCTGCCGGAGCCATGGAGGACGAAG----TGGTCCGCTTTGCC---------------atgaccgagtacaagcccacgGTGCGCCTCGCC
5' microhomology
CGGAGAAGACCTGCCGGAGCCATGGAGGACGAAGTGTTTGGCCTGCTTCACC----AGGCGAAACGGATGACCGAGTACAAGCCCACGGTGCGCCTCGCC

Figure 12: Genotyping of homozygous degron-tagged TCEA1.
A. Schematic of the dTAG integration into TCEA1 exon1 with PiTCH system. The part in gray is translated separately from the rest of the protein; **B**. Genotyping Western blot showing the detection of HA epitope in the homozygous TCEA1 knock-ins and the lack thereof in the unmodified HEK293T cells. The WT TCEA1 is 34 kDa, the tag is 17 kDa, and the tagged TCEA1 is 51 kDa. In the second blot, TCEA1 is detected as a tagged version at 51 kDa, and not detected at the correct size (34 kDa) of the WT TCEA1. The appearance of unspecific bands is increased in the knock-ins (a similar effect is seen in KO western blots in the previous section). The Western blot was done for initial genotyping, not a relative comparison, so there is no loading control, as we wanted to see the whole span of sizes at that step of cell line validation; **C**. An example of Sanger sequencing of one homozygous clone at the site of the integration showing that the insertion happened with indels differentially in both alleles, hence the mixed signal.

our previously described Flp-In T-Rex system, used here as a control. As a negative control, a homozygous TCEA2 knockout clone was included to confirm the absence of endogenous TCEA2 in the knock-in cells. Unexpectedly, the tagged TCEA2 was not detected. It is also not detected with the anti-HA antibody, but instead, there is a puzzling band above 25 kDa. There are two scenarios that could explain this protein. The insertion may have caused a change in splicing of the first exon of the main isoform, ENST00000343484, resulting in fusion of the inserted tag with the immediately following first exon of another isoform, ENST00000415602. It is a short transcript, encoding 82 amino acids (~9 kDa). Together with the tag, this protein would be ~26 kDa. What remains perplexing is the absence of the full length, tagged TCEA2 and a considerably high expression level of the new short protein. Another possibility is that the self-cleaving peptide (P2A) was not used properly, which resulted in the expression of a puromycin resistance gene fused with the rest of the protein and subsequently got cleaved off

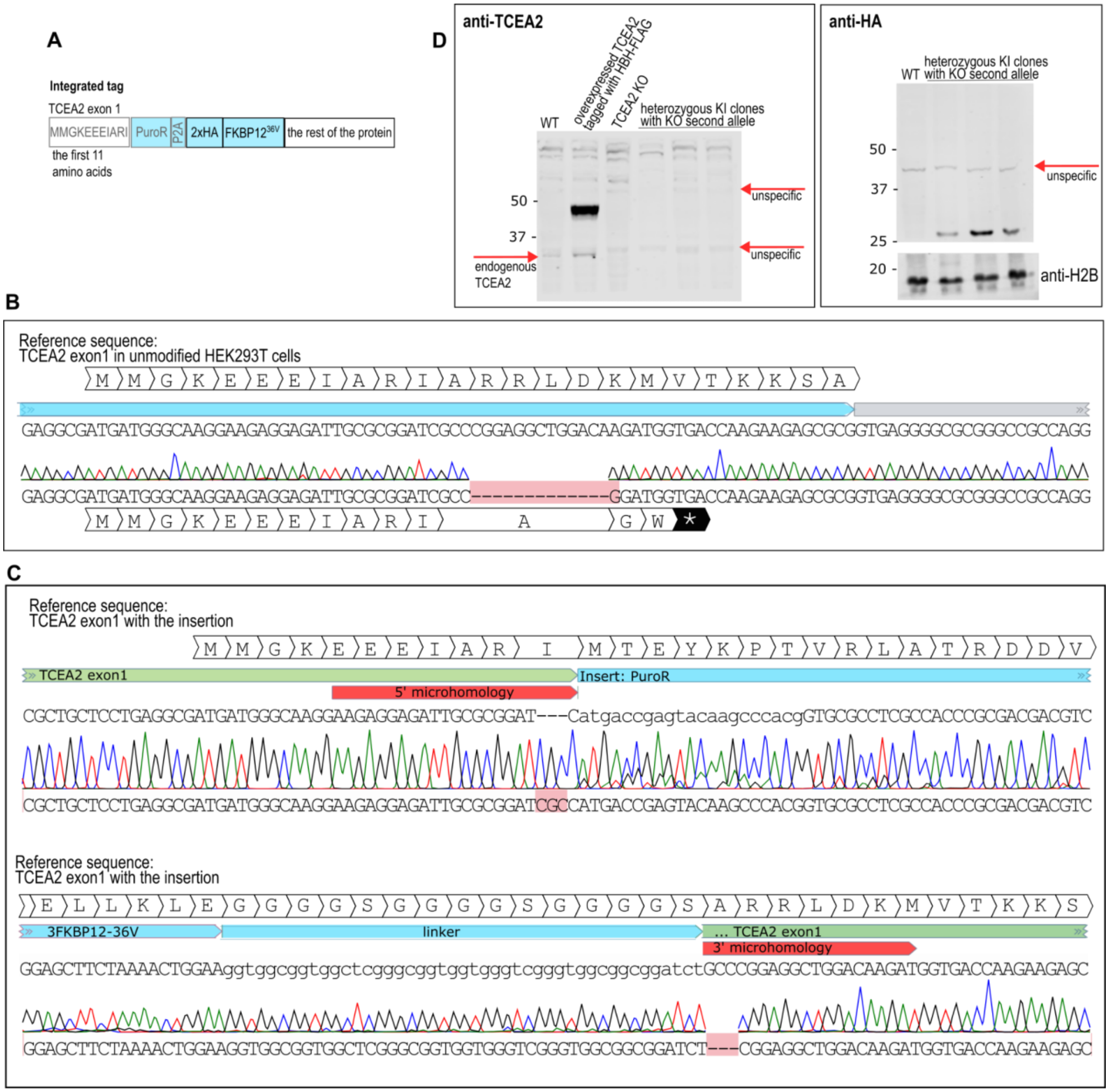

A
Integrated tag
TCEA2 exon 1
MMGKEEEIARI
PuroR
P2A
2xHA
FKBP12^{36V}
the rest of the protein
the first 11 amino acids
D
anti-TCEA2
WT
overexpressed TCEA2 tagged with HBH-FLAG
TCEA2 KO
heterozygous KI clones with KO second allele
50
37
unspecific
unspecific
endogenous TCEA2
anti-HA
WT
heterozygous KI clones with KO second allele
50
37
25
20
unspecific
anti-H2B
B
Reference sequence:
TCEA2 exon1 in unmodified HEK293T cells
M M G K E E E I A R I A R R L D K M V T K K S A
GAGGCGATGATGGGCAAGGAAGAGGAGATTGCGCGGATCGCCCGGAGGCTGGACAAGATGGTGACCAAGAAGAGCGCGGTGAGGGGCGCGGGCCGCCAGG
GAGGCGATGATGGGCAAGGAAGAGGAGATTGCGCGGATCGCC-------------GGATGGTGACCAAGAAGAGCGCGGTGAGGGGCGCGGGCCGCCAGG
M M G K E E E I A R I A G W *
C
Reference sequence:
TCEA2 exon1 with the insertion
M M G K E E E I A R I M T E Y K P T V R L A T R D D V
TCEA2 exon1
Insert: PuroR
5' microhomology
CGCTGCTCCTGAGGCGATGATGGGCAAGGAAGAGGAGATTGCGCGGAT---Catgaccgagtacaagcccacg GTGCGCCTCGCCACCCGCGACGACGTC
CGCTGCTCCTGAGGCGATGATGGGCAAGGAAGAGGAGATTGCGCGGATCGCCATGACCGAGTACAAGCCCACGGTGCGCCTCGCCACCCGCGACGACGTC
Reference sequence:
TCEA2 exon1 with the insertion
E L L K L E G G G G S G G G G S G G G G S A R R L D K M V T K K S
3FKBP12-36V
linker
... TCEA2 exon1
3' microhomology
GGAGCTTCTAAAACTGGAAggtggcggtggctcgggcggtggtgggtcgggtggcggcggatctGCCCGGAGGCTGGACAAGATGGTGACCAAGAAGAGC
GGAGCTTCTAAAACTGGAAGGTGGCGGTGGCTCGGGCGGTGGTGGGTCGGGTGGCGGCGGATCT---CGGAGGCTGGACAAGATGGTGACCAAGAAGAGC

Figure 13: Genotyping of degron-tagged TCEA2.
A. Schematic of the integrated dTAG into TCEA2 exon1. The part in gray is translated separately from the rest of the protein; **B**. An example of Sanger sequencing of an untagged, but mutated allele of a heterozygously tagged clone; **C**. An example of Sanger sequencing of the allele with the insertion of the same heterozygous clone; **D**. Genotyping Western blots against TCEA2 and HA. The former blot also includes a homozygous TCEA2 KO, as a negative control. An overexpressed FLAG-HBH-TCEA2 was included to help decrease the nonspecific signal. The blot shows that TCEA2 is not detected in the heterozygous knock-in (KI) with the knocked-out second allele, as well as in the TCEA2 homozygous KO. The anti-HA blot reveals an unexpectedly tagged small protein, right above 25 kDa.

downstream of the HA tag. The detected band, in this case, would be puromycin resistance protein and HA tag, also amounting to ~26 kDa.

In summary, tagging TCEA2 for degradation in the first exon was not feasible. Integrating the tag in more downstream exons can be worthwhile; however, with extensive validation because the addition of the degron tag can potentially diminish a rather low expression level of TCEA2 and interfere with the discovered interactions that the N-terminal domain mediates (described in 3.3). The TCEA1 knock-in cell line can be of use, but should be further validated and characterized in terms of TCEA1 expression level, off-target genome editing, and cell growth.

3.3 TCEA1 and TCEA2 interact with multiple RNA processing factors

3.3.1 TCEA1 and TCEA2 have similar interactomes

Immunoprecipitation of the endogenous TCEA1 and TCEA2 in HEK293T cells is challenging because of the poor specificity of the currently available antibodies and the low expression level of TCEA2. We attempted to determine their interactomes with the aim of finding paralog-unique interaction partners and learning of other potential interactors besides the expected Pol II subunits.

Using the inducible overexpression of FLAG-tagged TCEA1 and TCEA2 in Flp-In T-Rex 293 (described in 3.2.1), we gained insight into their interactomes with crosslink-assisted immunoprecipitation followed by mass spectrometry (IP-MS). With this approach, we detect the spatial associations of the bait protein with other proteins, rather than only the direct physical protein-protein interactions. We chose to include a short formaldehyde crosslinking step to stabilize as many transient interactions as possible and chromatin solubilization by sonication instead of nuclease digestion to preserve RNA-mediated interactions. The expression of FLAG-TCEA1 and FLAG-TCEA2 was induced for 24 hours with tetracycline, then the cells were crosslinked and lysed down to nucleic fraction for immunoprecipitation with an anti-FLAG antibody. IP-MS with the same antibody was done in the host unmodified Flp-In T-Rex 293 cells, as a negative control. We inferred protein abundance using the label-free quantification method (LFQ), which considers only the proteins with a minimum of two detected peptides per sample [350].

In the FLAG-TCEA1 IP-MS, 83 proteins were robustly detected in at least 3 out of 5 biological replicates and 30 were significantly enriched according to a student's T-test, with the p-value ≤0.05 and $\log_2$ (fold change) ≥1. There were more robustly detected interactors of FLAG-TCEA2 (96 proteins), 44 of which were significantly enriched. The full list of significant interactors can be found in supplemental table 1.

As expected, we found the bait proteins, TCEA1 and TCEA2, as the most enriched (Figure 14A & B). Upon examination of the interactomes, we noticed that most enriched proteins were RNA-processing factors, specifically splicing factors, such LSM3, SRSF1, and SRSF3 and heterogeneous nuclear ribonucleoproteins (hnRNPs), such as HNRNPA1, HNRNP2AB1, and HNRNPU. Interestingly, in the FLAG-TCEA2 IP, we found TCEA1 as an interactor, however TCEA2 was absent in the FLAG-TCEA1 IP. The TCEA2-unique peptides likely escaped detection because of the low expression level of the endogenous TCEA2 and overall lower detection sensitivity in FLAG-TCEA1 IP-MS caused by a lower overexpression of the bait protein (Figure 9C). Having omitted the probable contaminants, according to the Crapome database [351], e.g. Vimentin, Lamin B, and TRIM28, we compared the list of interactors of FLAG-TCEA1 and FLAG-TCEA2. All significant interaction partners of FLAG-TCEA1 were detected in FLAG-TCEA2 IP-MS (Figure 14C), indicating that the two paralogs have the affinity to the same proteins. Gene ontology enrichment analysis was done on the list of the significant TCEA2 interactors and confirmed that the interactome consists of RNA processing regulators (Figure 14D). The detected hnRNPs and splicing factors are shown in Figure 14E. All seemingly TCEA2-specific interactors are found in the FLAG-TCEA1 IP-MS, but with lower significance.

It was surprising to find RNA-processing factors as the predominant interactors of TCEA1 and TCEA2 and the interaction with Pol II as relatively less abundant. The Pol II subunit, RPB3, was also robustly detected, however under the statistical threshold of the T-test. We expected some factors of the elongation complex to co-immunoprecipitate with TCEA1 and TCEA2 based on the structure of the elongation complex [149, 352, 353]. Having examined the proteins that were filtered out because they were found in less than three biological replicates, we found more Pol II subunits and additional factors of the transcription elongation complex: NELF-E, PAF1, SPT4, SPT6 and others and showed their abundance as unique peptide counts in Figure 14F. Interactions with histones were also significant (Supplemental table 1), possibly reflecting the involvement of TCEA1 and TCEA2 during transcription through the nucleosome.

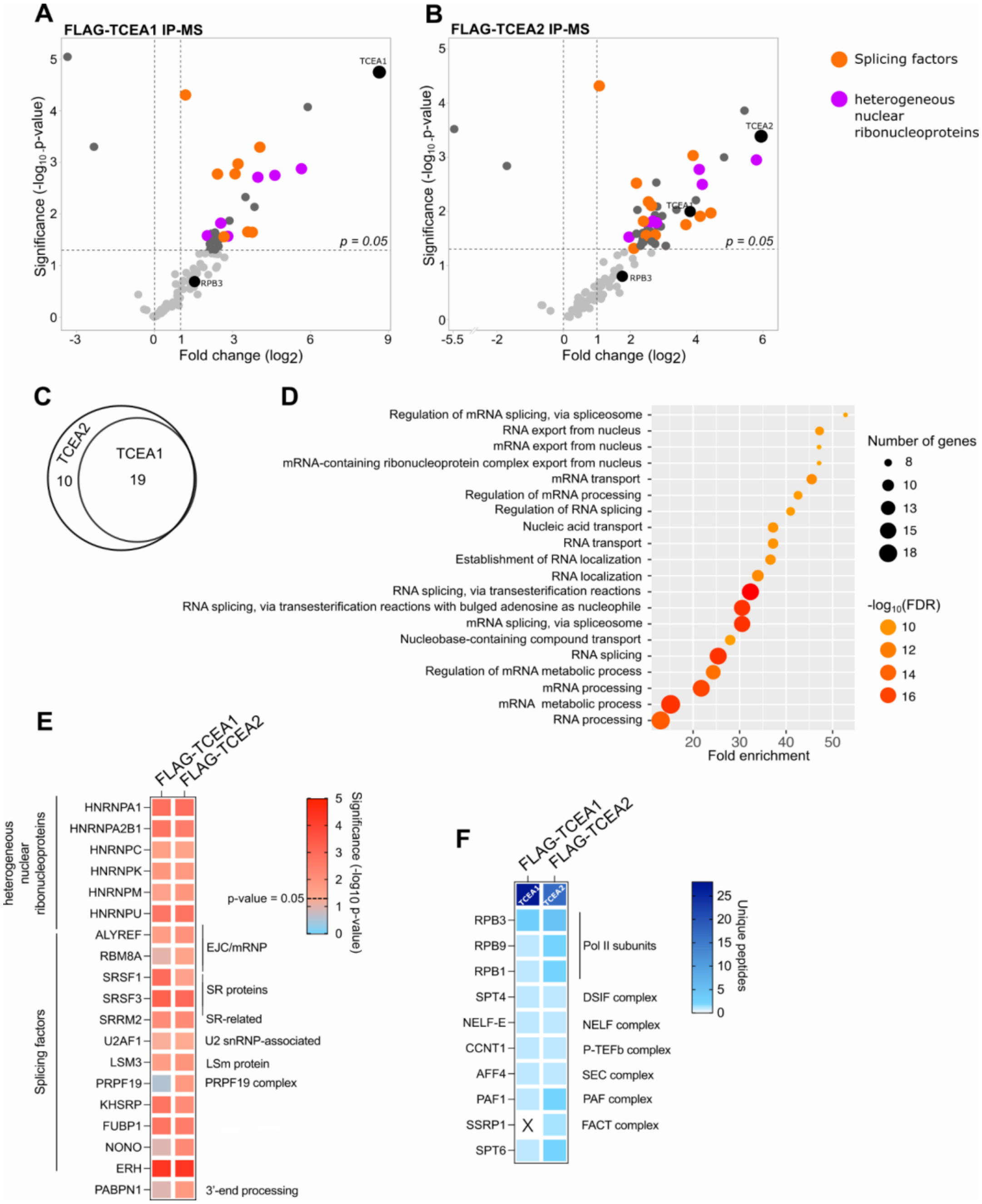


Figure 14: TCEA1 and TCEA2 associate with numerous RNA-processing factors.
A. Volcano plot of proteins that have co-immunoprecipitated with FLAG-TCEA1 and were detected with mass spectrometry; **B.** Volcano plot of proteins that have co-immunoprecipitated with FLAG-TCEA2 and were detected with mass spectrometry; **C.** Venn diagram of significant interactors of FLAG-TCEA1 and FLAG-TCEA2; **D.** Significant (FDR >0.05) gene ontology terms of robustly detected (in at least 3 out of 5 biological replicates and with the significance of -log$_{10}$(p-value) >1.3) FLAG-TCEA2 interactors; **E**. List of RNA-processing factors that were robustly detected in either FLAG-TCEA1 or FLAG-TCEA2 IP-MS The color denotes the significance; **F.** The expected interactions with the elongation complex were below the significance threshold of the T-test and are shown as the average unique peptide count. "X" means that a protein was not detected.

In summary, FLAG-TCEA1 and FLAG-TCEA2 predominantly interact with RNA-processing factors. The low abundance of the elongation complex and the high enrichment of RNA-processing factors suggest the following scenario. In our system, both TCEA1 and TCEA2 are overexpressed 1.4- and 3- fold, respectively, therefore, it is likely that in the IP, we pull down both TCEA1/TCEA2 engaged in assisting backtracked Pol II and those that are in excess in the nucleus, the latter may be the ones, associating with the spliceosome and hnRNPs. Such a substantial stoichiometric difference can also indicate that the recruitment of TCEA1/TCEA2 to Pol II is a rather rare or transient event, in contrast to more abundant interactions with RNA processing factors. Additionally, the low detection of other elongating factors can also mean that TCEA1 and TCEA2 do not interact with them directly.

3.3.2 TCEA1 and TCEA2 interact with RNA processing factors via their N-terminal domains

Not having found the unique interactors of TCEA1 and TCEA2, we attempted to increase the detection sensitivity by overexpressing the most dissimilar part of TCEA1 and TCEA2, the first 139 and 137 amino acids, respectively, constituting the N-terminal domain and the subsequent flexible linker (NTDL) (Figures 8 & 9). Furthermore, we aimed to better understand the function of the NTDL, which is still unclear and known not to be required for RNA cleavage [250]. The expression of FLAG-TCEA1-NTDL and FLAG-TCEA2-NTDL was induced for 24 hours, followed by crosslinking and immunoprecipitation with an anti-FLAG antibody, exactly as in the experiment with the whole TCEA1 and TCEA2 proteins.

85 proteins were robustly detected in 3 out 4 biological replicates in the FLAG-TCEA1-NTDL IP-MS, 40 of which were significantly enriched (p-value ≤0.05 and $\log_2$(fold change) ≥1). In FLAG-TCEA2-NTDL IP-MS, 52 proteins were robustly detected, 23 of which were significantly enriched. The volcano plots reflect all robustly detected proteins and many of them belong to the categories of hnRNPs and splicing factors (Figure 15A & B). Interestingly, no Pol II subunits were detected among significant and nonsignificant proteins.

We found that the TCEA1- and TCEA2- NTDL interactors substantially overlap with each other, as well as with the interactors of the whole proteins (Figure 15C), confirming our previously found interactions with RNA processing factors. In this IP-MS experiment, the list of the RNA-processing factors extended and included additional hnRNPs, splicing factors and previously not observed RNA helicases (Figure 15D). The unique interactors of TCEA1-NTDL are RBM14 and RBMX, both are RNA processing factors that also play a role in genome integrity [354, 355]. Although the splicing factor LSM3 was not detected in TCEA2-NTDL IP, it was found in the whole TCEA1 and TCEA2 IP-MS (Figure 14E). SF3B2 is not a TCEA1-selective interactor either as it was also detected in the whole TCEA1 and TCEA2 IP-MS, but below the significance threshold. An RNA DEAD-box helicase DDX3X was found as a unique interaction partner of TCEA2-NTDL, despite overall lower IP efficiency. This helicase is among those that regulate RNA metabolism and genome integrity (reviewed in [356]). It was not detected in the whole TCEA1/TCEA2 IP-MS.

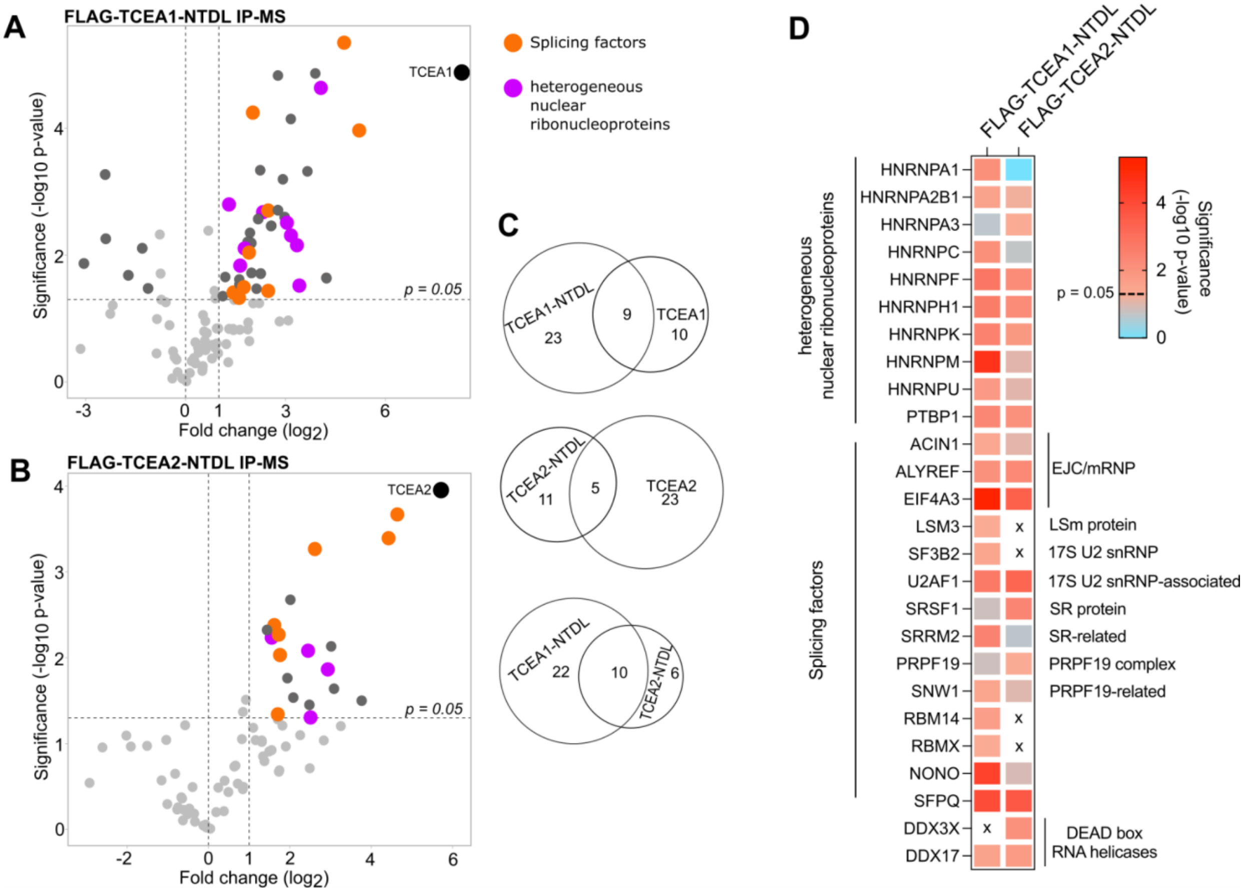


Figure 15: The N-terminal domain and linker (NTDL) of TCEA1 and TCEA2 interact with RNA-processing factors.

A. Volcano plot of proteins that have co-immunoprecipitated with FLAG-TCEA1-NTDL and were detected with mass spectrometry; **B.** Volcano plot of proteins that have co-immunoprecipitated with FLAG-TCEA2-NTDL and were detected with mass spectrometry; **C.** Venn diagrams of significant interactors of FLAG-TCEA1-NTDL and the whole FLAG-TCEA1 protein, FLAG-TCEA2-NTDL and the whole FLAG-TCEA2 protein, and FLAG-TCEA1-NTDL and FLAG-TCEA2-NTDL; the likely contaminants were omitted for this comparison; **D.** List of RNA-processing factors that were robustly detected (in at least 3 out of 4 biological replicates and with the significance of $-\log_{10}$(p-value) ≥1.3 in either FLAG-TCEA1-NTDL or FLAG-TCEA2-NTDL IP-MS. The color denotes the significance. “X” means that the protein was not detected.

Altogether, our findings indicate that TCEA1 and TCEA2 physically associate with multiple RNA processing factors via their NTDL. Overexpression of only the NTDL increased the sensitivity enabling the detection of more significant RNA processing factors and putative unique interaction partners, however, their direct interactions remain to be verified by a reverse IP-MS or co-IP Western blot.

3.3.3 The localization to chromatin of the TCEA1- and TCEA2- interacting RNA processing factors is not severely impacted in the knockouts

TFIIS was found to have a direct impact on Pol II processivity [154], which is kinetically coupled to RNA splicing [217]. Based on the determined physical associations of TCEA1/TCEA2 with splicing factors and hnRNPs, we wondered whether, besides alleviating backtracked Pol II, TCEA1/TCEA2 have a role in mediating RNA processing by recruiting the RNA processing factors that were found as significantly enriched in the IP-MS.

We estimated the abundance of these interactors in the quantitative chromatin mass spectrometry (MS) data set that we generated in the TCEA1 KO, TCEA2 KO, DKO and the unmodified HEK293T cells using SILAC (stable isotope labeling with amino acids in cell culture). We found that most of the detected TCEA1/TCEA2 interactors did not change in abundance at chromatin in the knockouts, relative to unmodified cells (Figure 16). Mild changes in abundance are observed in TCEA1 KO and TCEA2 KO, with the most displacement of factors in the DKO. The DKO showed a significant, but mild reduction in abundance of HNRNPC, HNRNPF, ALYREF, EIF4A3, RBM8A, and RBM14. In TCEA1 KO, only two factors were displaced (HNRNPA3 and HNRNPF), as well as in TCEA2 KO (HNRNPA2B1 and RBM14). Unexpectedly the potentially TCEA1-unique interactor, RBM14, is found as slightly more abundant in the TCEA1 KO, however it gets displaced from chromatin in TCEA2 KO and more severely in the DKO. Interestingly, some factors become more abundant, especially in the TCEA2 KO.

Based on the observations of only marginal changes, both increase and reduction in abundance of the interactors, and inconsistent changes among knockouts, we propose that the role of TCEA1 and TCEA2 in transcription has an effect on chromatin composition, however, they are not directly involved in recruitment, but rather affiliate with numerous RNA processing factors in the nucleus, potentially conveniently proximal to Pol II.

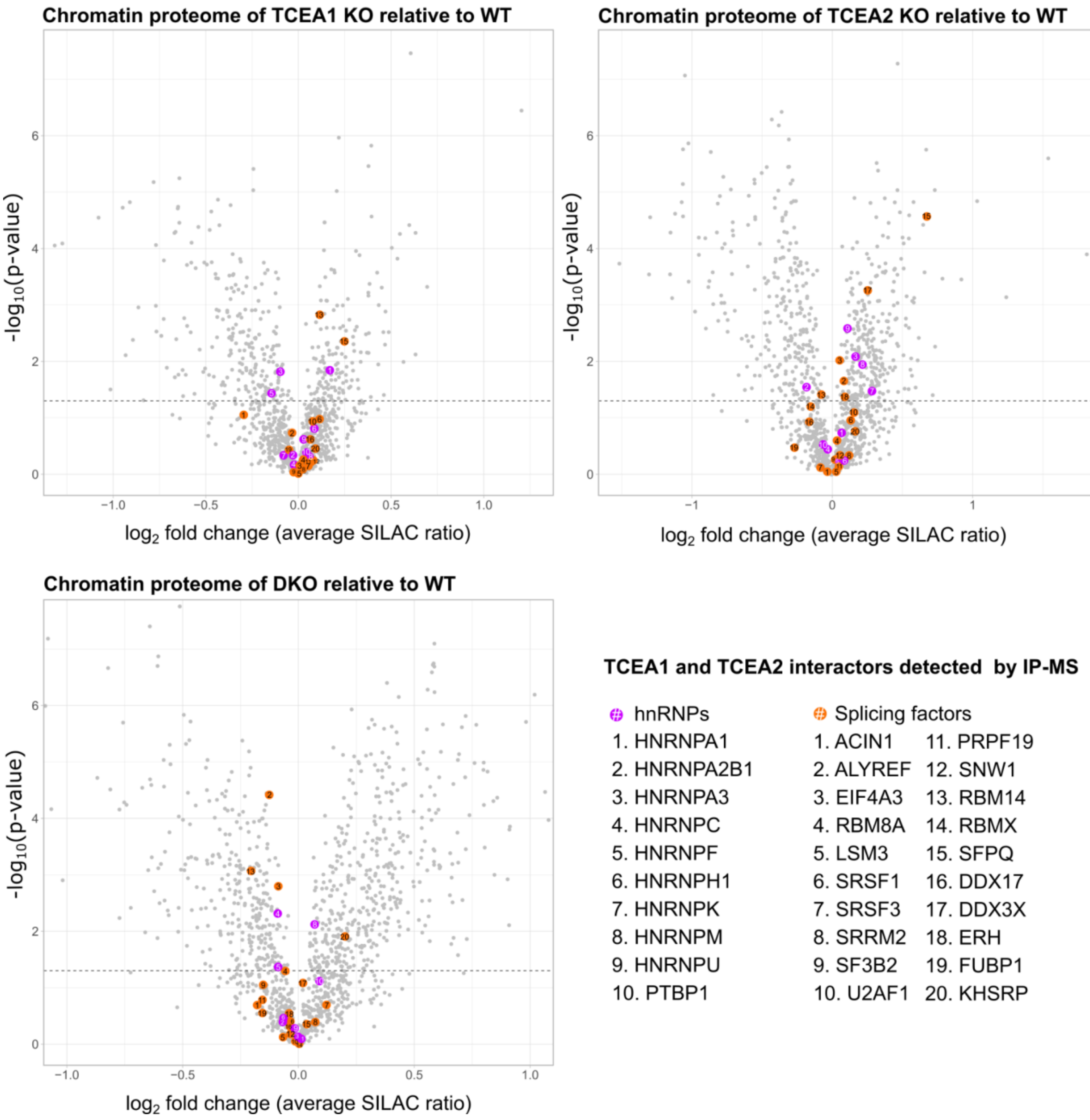


Figure 16: The abundance at chromatin of co-immunoprecipitated RNA-processing factors with TCEA1 and TCEA2 is not severely impacted in the TCEA1 KO, TCEA2 KO, and DKO.

Volcano plots showing the quantitatively determined (based on normalized SILAC ratios) change in chromatin proteomes of knockouts relative to unmodified HEK293T (WT). Average SILAC ratio <0 corresponds to the proteins that are displaced from chromatin in the knockout; average SILAC ratio >0 denotes the proteins that became more abundant at chromatin in the knockout. Significant proteins are above the dashed line (p ≤0.05).

3.4 Genome-wide occupancy profiling of TCEA1 and TCEA2 reveals that the paralogs bind to Pol II during early transcription elongation

It was previously shown that TFIIS predominantly occupies the promoter-proximal regions of genes in mouse embryonic stem cells [357] and in human breast cancer cells [154]. In the first study, an affinity tag was endogenously added to TCEA1 for chromatin immunoprecipitation with DNA sequencing (ChIP-seq) and TCEA1 was found at 5’ pause sites of Pol II- and Pol III- transcribed genes. In the other study, ChIP-seq was done to profile overexpressed affinity-tagged mouse TCEA1 and it revealed that TCEA1 localizes not only at the 5’ pause sites, but also in the region downstream of polyadenylation sites (PAS). Although these studies provide insight into the predominant localization of TCEA1-relieved backtracking, the binding of TCEA2 remains unknown. We aimed to understand whether TCEA2 can compete with TCEA1 and compare the binding patterns of TCEA1 and TCEA2 to other elongation factors.

We chose to use the same inducible FLAG-TCEA1 and FLAG-TCEA2 overexpression cell lines for FLAG ChIP-seq because chromatin immunoprecipitation (ChIP) of endogenous TCEA1 and TCEA2 would be challenging for two reasons. First, TCEA2 is lowly expressed, and there is no optimal antibody for ChIP. Our attempt with the available TCEA2-specific antibody resulted in no enrichment over input (data not shown). Second, both TCEA1 and TCEA2 are not highly abundant. Furthermore, ChIP efficiency is expected to be lower because they associate with DNA through their interaction with Pol II, as there is no evidence that TCEA1 or TCEA2 directly bind DNA *in vivo*. The overexpression system enabled us to increase the detection efficiency and gain insight into the binding affinity of each paralog. The FLAG ChIP-seq experiments were done without exogenous genome spike-ins because the focus was on whether TCEA1 and TCEA2 can occupy the same sites, rather than quantitative comparison of binding intensities, since the level of overexpression of FLAG-TCEA1 and FLAG-TCEA2 is not the same (Figure 9).

3.4.1 TCEA1 and TCEA2 mostly bind protein-coding genes and enhancers

Thousands of binding sites were robustly detected in both biological replicates of FLAG-TCEA1 (14129) and FLAG-TCEA2 (17763) ChIP-seq. In FLAG-TCEA2 ChIP-seq, ~20% more binding sites were detected, likely due to a higher FLAG-TCEA2 expression level enhancing ChIP efficiency. The majority of the binding sites (11391) were bound by FLAG-TCEA1 and FLAG-TCEA2, indicating that they both associate with Pol II at the same sites

(Figure 17A). As expected, most of the binding sites were found at protein-coding genes and annotated enhancers (Figure 17B), both transcribed by Pol II. Interestingly, some TCEA1/TCEA2 binding was also detected at Pol III-transcribed genes, which is in accordance with previous reports, suggesting that TFIIS plays a general role in Pol III transcription [357, 358]. It is somewhat surprising because Pol III has an intrinsic TFIIS-like subunit (Rpc11), also containing the highly conserved motif of aspartic and glutamic acids in the Zinc finger, which enhances the transcript cleaving property [359]. Some binding was also detected at lncRNA genes and miRNA genes.

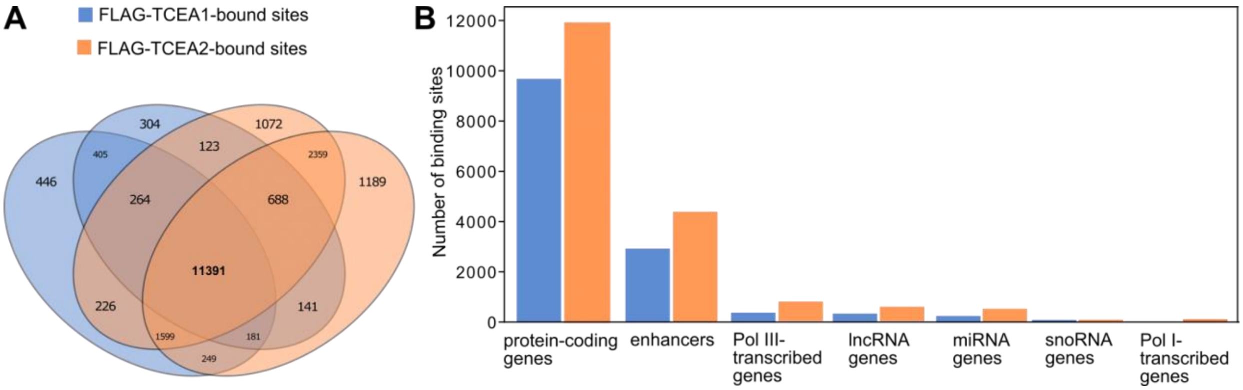


Figure 17: TCEA1 and TCEA2 can bind the same sites, predominantly at protein-coding genes
A. Venn diagram showing an overlap between FLAG-TCEA1 and FLAG-TCEA2 binding sites, in two biological replicates of each cell line. **B**. Average distribution of binding sites of FLAG-TCEA1 and FLAG-TCEA2 at various genomic regions.

3.4.2 TCEA1 and TCEA2 occupancy profiles correlate with Pol II occupancy

Next, we compared the meta gene binding profiles and observed a clear similarity between the profiles of FLAG-TCEA1, FLAG-TCEA2 and Pol II (Figure 18). In this analysis, we included actively transcribed genes in HEK293T cells, based on the RNA-seq signal. The most TCEA1/TCEA2 binding occurs in the promoter-proximal region (300 bp downstream of transcription start site (TSS)), some enrichment is observed in the termination zone (500 bp downstream of the PAS), and a slightly lower enrichment - in gene bodies (300 bp after TSS until PAS). TCEA1/TCEA2 also bind Pol II transcribing the other strand, as can be seen as a peak, right upstream of the TSS at the single gene examples (Figure 19). Arranging all binding sites in the order of FLAG-TCEA1 enrichment intensity revealed the relationship between Pol II and FLAG-TCEA1/TCEA2 occupancy: the more polymerases are transcribing a given gene, the more TCEA1/TCEA2 will be found at that gene (heatmaps in Figure 18).

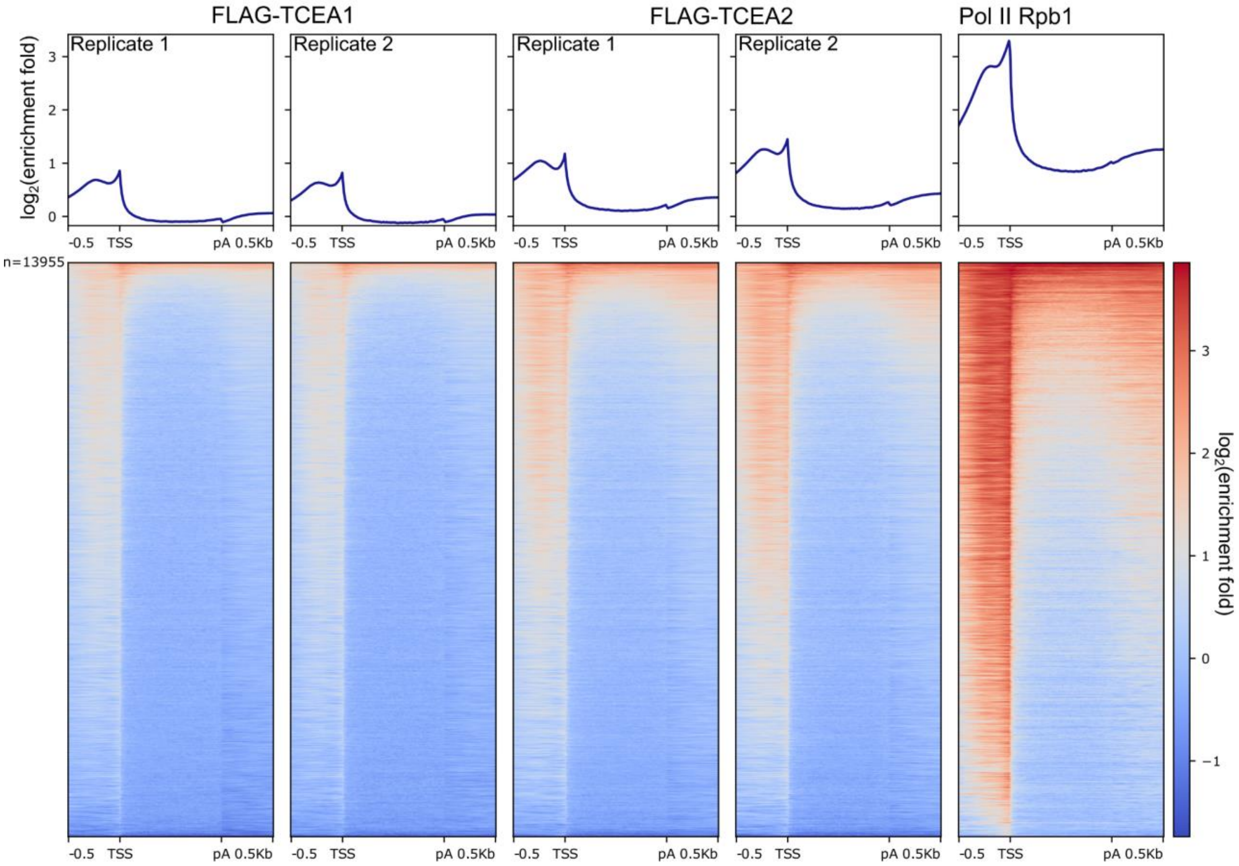


Figure 18: TCEA1 and TCEA2 bind to Pol II most predominantly at the promoter-proximal region. Metagene binding profiles of our FLAG-TCEA1, FLAG-TCEA2, and available Rpb1 of Pol II [360] ChIP-seq in HEK293T cells. Occupancy at 13955 actively transcribed genes in HEK293T cells across all heatmaps is ordered based on the highest occupancy of FLAG-TCEA1. The y-axis of the metagene plots and the heatmap color reflect the $\log_2$(enrichment fold to background). TSS is the transcription start site, pA is the polyadenylation site. The length of the region between the TSS and pA (gene body) is proportionally condensed.

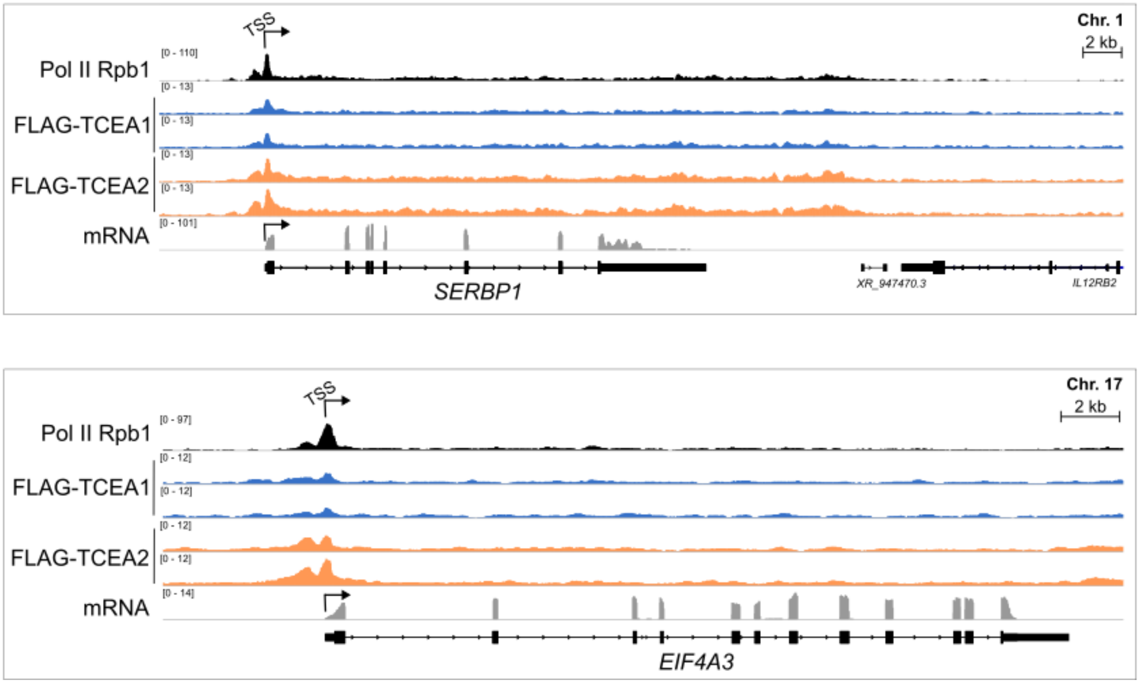


Figure 19: Single gene examples of TCEA1/TCEA2 binding.
Single locus examples of ChIP-seq tracks of Pol II Rpb1 in HEK293T cells [360] and induced FLAG-TCEA1 and FLAG-TCEA2 Flp-In T-Rex 293 cells and RNA-seq in HEK293T cells.

3.4.3 TCEA1 and TCEA2, on average, bind Pol II before other elongation factors and prior to the +1 nucleosome

Our ChIP-seq experiment indicated that TCEA1 and TCEA2 are predominantly early elongation regulators. Next, we wanted to better understand TCEA1/2 binding relative to other classical elongation factors that are required for promoter-proximal pause, pause release, and transcription through the +1 nucleosome. We integrated the meta gene occupancy of publicly available ChIP-seq data of Pol II RPB1 [360], NELF-A [361], LEO1 (PAF complex), and SPT6 [362], all done in HEK293 cells. We also included a chromatin accessibility profile based on ATAC-seq data in the same cell line [363].

The comparison of the meta-gene binding enrichment revealed that, on average, TCEA1 and TCEA2 bind Pol II during early elongation, as can be seen by the overlap of the curves. Interestingly, TCEA1/2 appear to bind Pol II prior to the promoter-proximal pause, denoted by the NELF-A peak, which is slightly more upstream of the +1 nucleosome. According to structural studies, TFIIS is unable to bind and activate paused Pol II when the latter is bound by the NELF complex [95] and the factors are expected to compete with each other for binding. More analyses are required to elucidate the temporal binding of TFIIS and NELF at human genes, for example, a gene-by-gene comparison of NELF, TCEA1, and TCEA2 occupancy and the +1 nucleosome location, with methods enabling higher resolution.

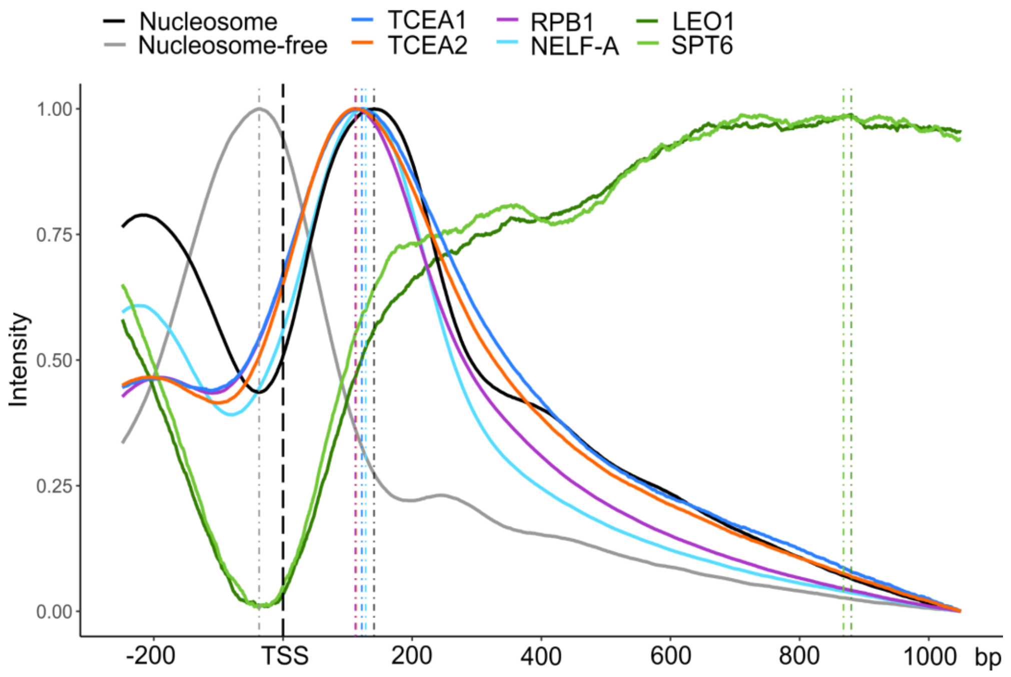


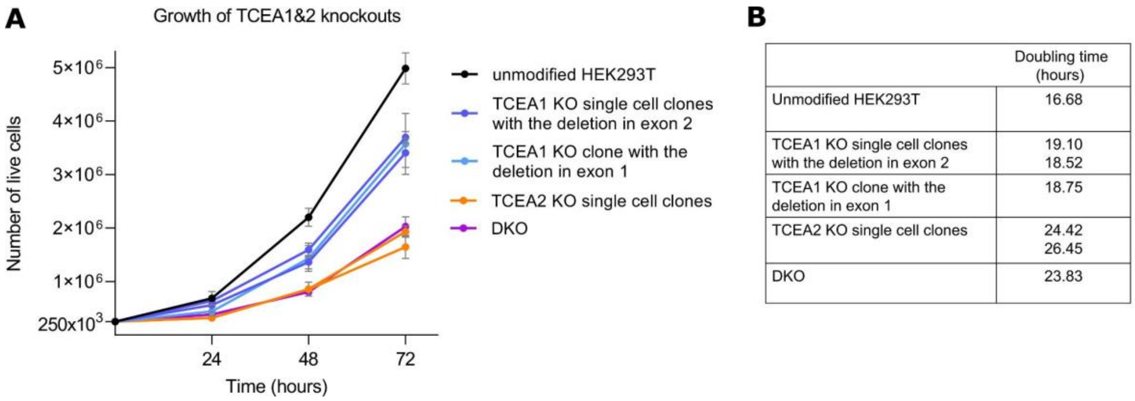
Figure 20: TCEA1 and TCEA2 bind Pol II prior to the promoter-proximal pause and the +1 nucleosome
Integrated metagene plots based on ChIP-seq experiments in HEK293 cells, showing the binding pattern of TCEA1, TCEA2, total Pol II [360], and other elongation factors (LEO1 [362], NELF-A [361], SPT6 [362]) in the context of accessible chromatin determined with ATAC-seq [363]. Peak maxima are marked by the dashed lines. The intensities were scaled proportionally.

In summary, our TCEA1 and TCEA2 ChIP-seq experiment revealed that the paralogs predominantly bind Pol II genes, however, they are also found at Pol III genes. TCEA1/2 have can bind Pol II at the same genomic sites in the overexpression system. A higher Pol II occupancy appears to correlate with higher TCEA1 and TCEA2 binding. The binding pattern of the TCEA paralogs is distinct from other elongation factors: they bind Pol II during early elongation before the promoter-proximal pause, possibly shortly after initiation. Their binding is also detected on the antisense Pol II.

3.5 The loss of TCEA1, TCEA2 and both TCEA paralogs led to a slow growth phenotype and DNA damage response

3.5.1 Characterization of TCEA1 knockout (KO), TCEA2 KO, and the double TCEA1 and TCEA2 knockout (DKO) cell lines

Although our experiments in the overexpression system point to an almost complete functional redundancy of the paralogs, the deletion of the lower expressed TCEA2 resulted in more dramatic phenotypic changes than TCEA1 deletion. During single clone expansion, it became apparent that all mutants were growing considerably slower than the unmodified HEK293T cells. This was not entirely surprising because of the recent studies in mammalian cells. TCEA1 knockdown declined cell proliferation in breast, lung, and pancreatic cancer cell lines [293]. Furthermore, overexpression of the inactivated TFIIS mutant in Flp-In T-Rex 293 cells reduced the growth rate and cell viability [308]. However, until now, the specific effect of TCEA2 has not been investigated.



	Doubling time (hours)
Unmodified HEK293T	16.68
TCEA1 KO single cell clones with the deletion in exon 2	19.10 18.52
TCEA1 KO clone with the deletion in exon 1	18.75
TCEA2 KO single cell clones	24.42 26.45
DKO	23.83

Figure 21. Deletion of TCEA2 impedes cell proliferation more severely than TCEA1 deletion.
A. Growth curve illustrating that all knockouts have a slow growth phenotype, but to different extents. This growth analysis was done in 4 - 6 biological replicates and in separately derived single cell clones. In addition, TCEA1 KOs, denoted in bright blue and light blue, were generated by targeting different exons. The light blue curve refers to the parental cell line of the DKO. **B**. Cell doubling time.

The growth analysis was performed by seeding all cell lines at the same concentration in identical plates and counting the cells after 24, 48, and 72 hours. To verify the phenotype, we have included additional single cell-derived knockout cell lines and the clone with the deletion in exon 1 and the off-target deletion in *TCEA1P1*, since it is the parental cell line of the DKO (double knockout of TCEA1 and TCEA2). Although TCEA1 is the higher expressed paralog, unexpectedly, TCEA2 KO (knockout) cell line grew strikingly slower than TCEA1 KO and at the same pace as the DKO (Figure 21). Interestingly, the DKO was tolerable to cells, indicating that the intrinsic RNA cleavage of Pol II may be sufficient to overcome transcription arrest to some extent.

Furthermore, by examining the cells under a microscope, we found some peculiar features that were more frequent in the TCEA2 KO and DKO cell lines (Figure 22). Larger than usual, possibly, senescent cells are sometimes observed in HEK293T culture, but there are noticeably more of them in the TCEA1 KO and TCEA2 KO. These abnormally large circular multinuclear aggregates are the most frequent in the DKO. The increased presence of such cells is indicative of either a defect in mitosis or increased cell cycle arrest.

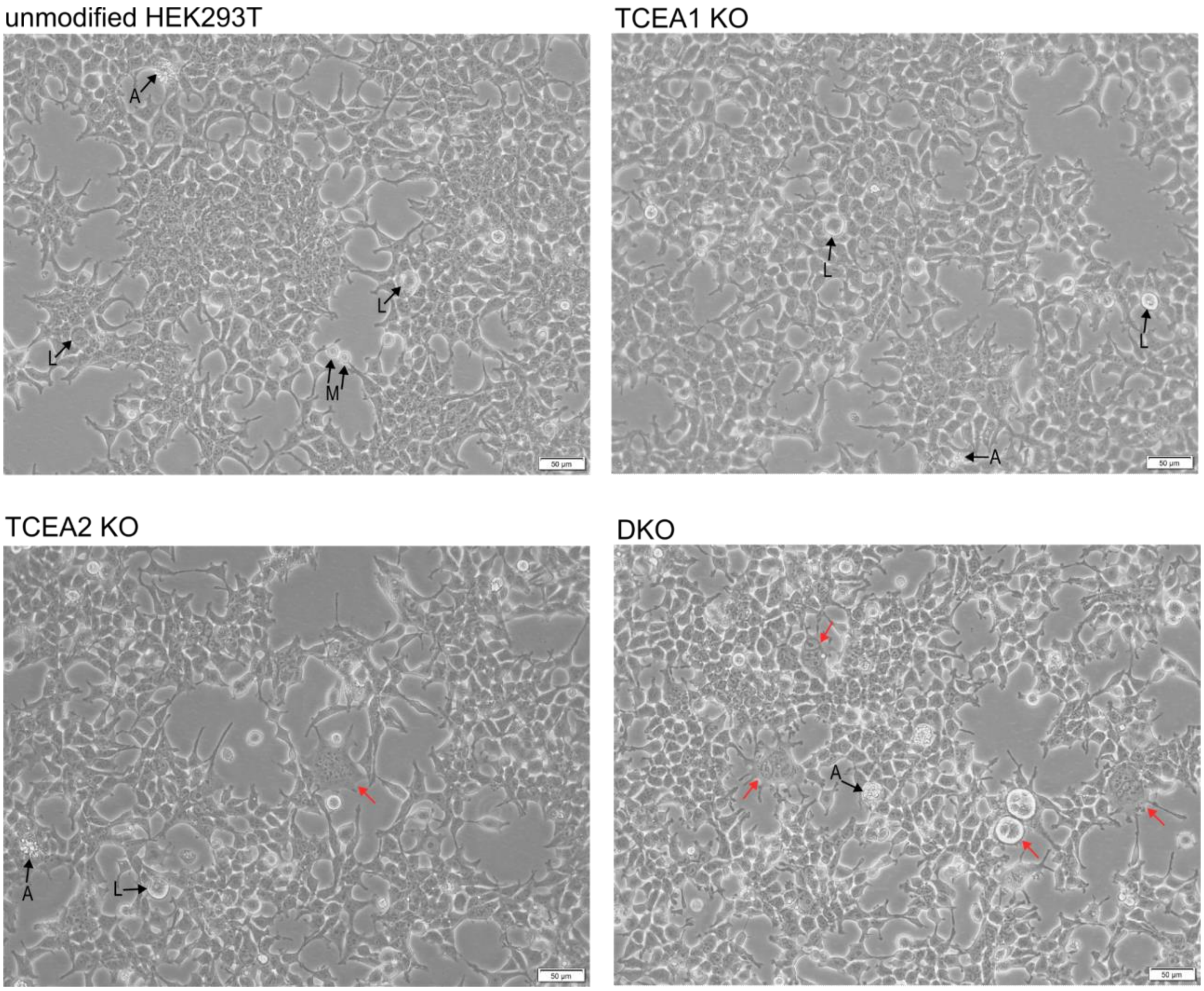


Figure 22. Cell culture abnormalities are observed in the TCEA2 KO and the DKO.
Microscope photos at 20X magnification of HEK293T, TCEA1 KO, TCEA2 KO, and the DKO. Normal events are shown with black arrows: “M” denotes mitotic cells, “A” denotes late apoptotic cells, “L” stands for larger than usual cells. Abnormal events are shown with red arrows: multinuclear aggregates and extra-large cells.

The cell viability of all cell lines was always above 90% and we did not observe unusually many detached cells that would have suggested a high number of apoptotic or necrotic cells. To further characterize the knockout cell lines, we evaluated apoptosis by flow cytometry. Live cells, grown normally and at the same density, were stained with Apotracker Green, which detects the presence of phosphatidylserine residues on the cell surface, indicating the onset and progression of apoptosis, and Zombie NIR, which only penetrates damaged membranes of dead cells or those in late apoptosis and necrosis (Figure 23). We defined the cells undergoing apoptosis as those with Apotracker green signal intensity >10^3. Dead cells, including late apoptotic and necrotic cells, were defined as those with Zombie NIR signal intensity >10^3. All cell lines showed similar percentages of viable, apoptotic, and dead cells. No changes in caspase-3 activity were detected (Supplemental figure S4).

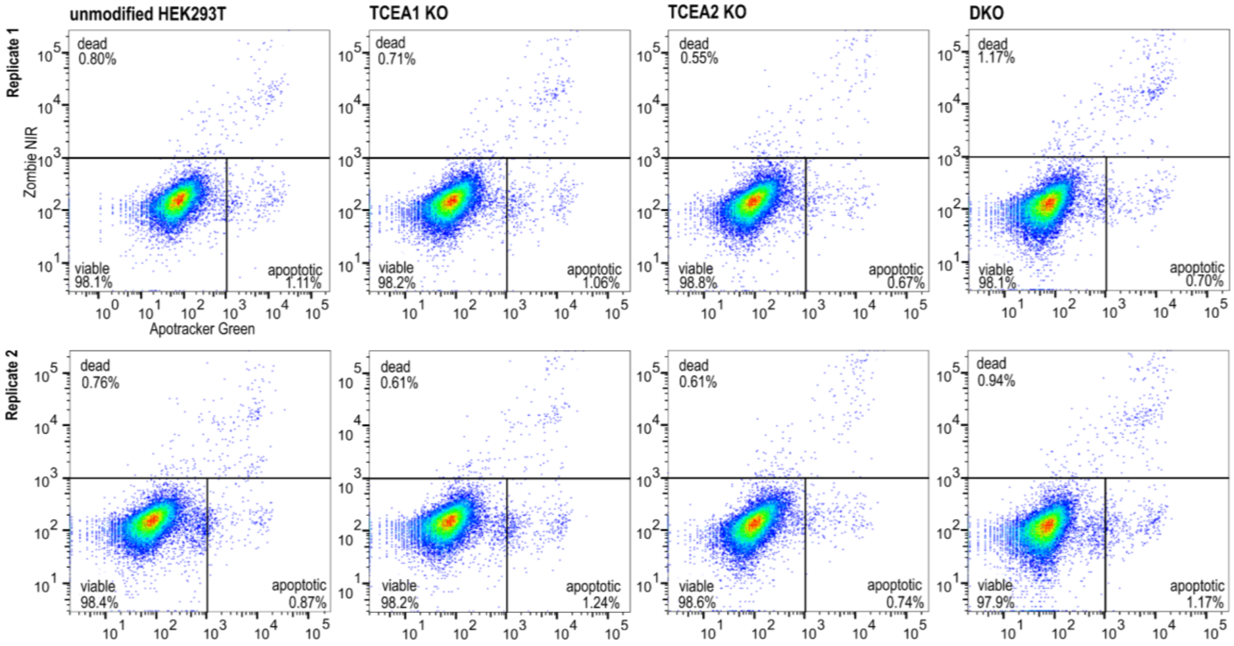


Figure 23. Deletion of TCEA1 and TCEA2 does not induce apoptosis.
Apoptosis assay of HEK293T, TCEA1 KO, TCEA2, and DKO using flow cytometry after staining live cells with Apotracker Green and Zombie NIR. The assay was done in two biological replicates per cell line, shown in rows 1 and 2, and 20,000 cells were analyzed in each measurement. The x-axis of the scatter plots represents the signal intensity of Apotracker Green, reflecting the amount of phosphatidylserine residues. The y-axis reflects the permeability of cells to Zombie NIR.

3.5.2 The loss of TCEA1 and TCEA2 has an impact on cell cycle

To further understand the slow growth phenotype, we performed cell cycle analysis in asynchronous and synchronized populations by analyzing DNA content with FxCycle Violet DNA stain and flow cytometry. In asynchronous cells, we found only a slight difference between the knockouts and the control. A mild decrease, more prominent in the TCEA2 KO and DKO clones, in the number cells in G0/G1 state can be seen in the cell cycle profiles (Figure 24A). However, the quantification, based on Dean-Jett-Fox model [364], did not show any change in the proportion of cells in each phase within each cell line (Figure 24B).

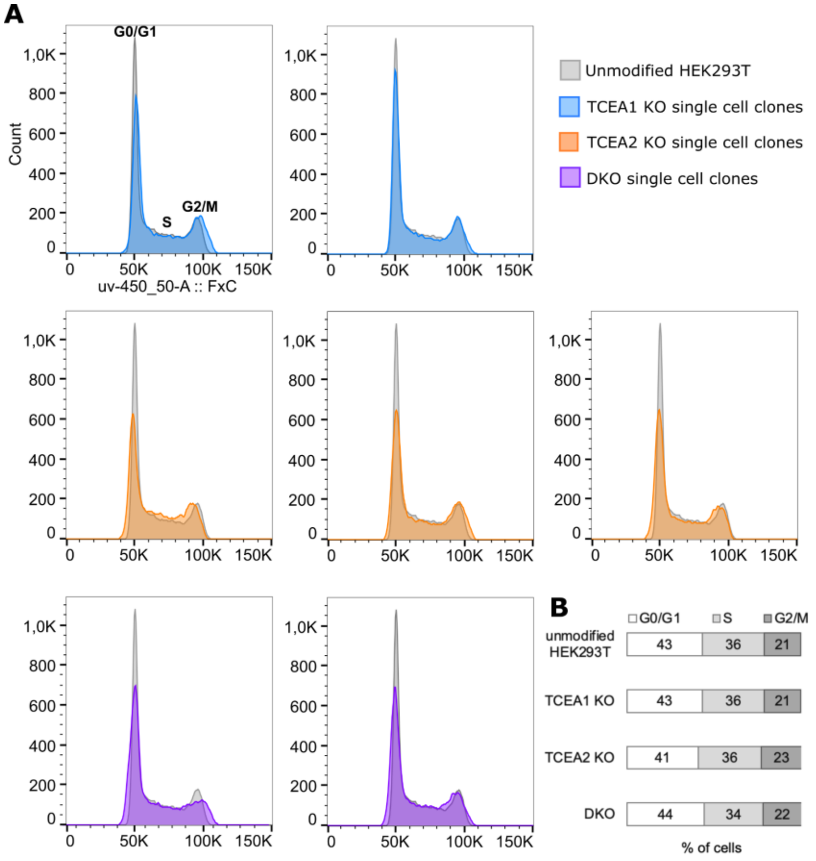


Figure 24. Cell cycle analysis in non-synchronized cells indicates subtle differences in cell cycle phases among the knockouts.

A. Cell cycle phases profiled with FxCycle Violet DNA stain in HEK293T, two single cell TCEA1 KO clones, three single cell TCEA2 KO clones, and two single cell DKO clones. The histogram indicates the DNA content, which relates to the cell cycle phases as: $G_0/G_1 \propto$ 1X DNA (signal intensity 50K), S $\propto$ >1X DNA (signal intensity between 50K and 100K), and $G_2/M \propto$ 2X DNA (signal intensity 100K). The x-axis corresponds to the DNA stain signal intensity. The y-axis is proportional to the cell number. **B**. Quantified percentage of cells in G_0/G_1, S, and G_2/M indicates minor differences among the knockouts. The percentage was calculated using the Dean-Jett-Fox model, then averaged among the single cell clones.

To examine this change closer, we synchronized the cells by a double thymidine block and analyzed their cell cycle progression. Excess thymidine allosterically inhibits ribonucleoside diphosphate reductase, thereby depleting deoxycytidine triphosphate (dCTP) pool [365], essential for DNA synthesis. Upon addition of excess thymidine, DNA replication is blocked, and the cells are arrested at the G1/S boundary [366]. Similar to a published protocol [367], we added thymidine to the cells for 16 hours, followed by a wash-out and release in regular medium for 8 hours, then blocked again for 16 hours to increase the percentage of synchronized cells. Then the cells were released into the S phase, harvested at different time points, and stained with FxCycle. The DNA content was analyzed with flow cytometry (Figure 25A).

We observed that TCEA2 KO and the DKO progress through S phase considerably slower than TCEA1 KO, which is slightly delayed, compared to the unmodified HEK293T cells (WT) (Figure 25B). After one hour of release, the cells begin to visibly shift towards G2 in the WT, this shift is observed in TCEA1 KO after 2 hours and not seen until 4 and 3 hours after the release of TCEA2 KO and the DKO, respectively. A prominent G1 peak, indicating that some cells have passed the S and G2 phases, divided, and re-entered G1, appeared between 2 and 3 hours in WT, after 3 hours in the TCEA1 KO, between 4 and 6 hours in the TCEA2 KO, and after 4 hours in the DKO.

In summary, all knockouts experience a prolonged S phase, but to different extents. The TCEA1 KO is mildly affected, and its progression through S phase is slower for about an hour than the WT. The TCEA2 KO experiences the most replication stress delaying S phase procession by about 2 hours. Surprisingly, the profile of the DKO is somewhat between the TCEA1 KO and TCEA2 KO. The prolonged S phase indicates that the cells experience some stress during DNA replication. One of the causes of replication stress is DNA damage, which we investigated further.

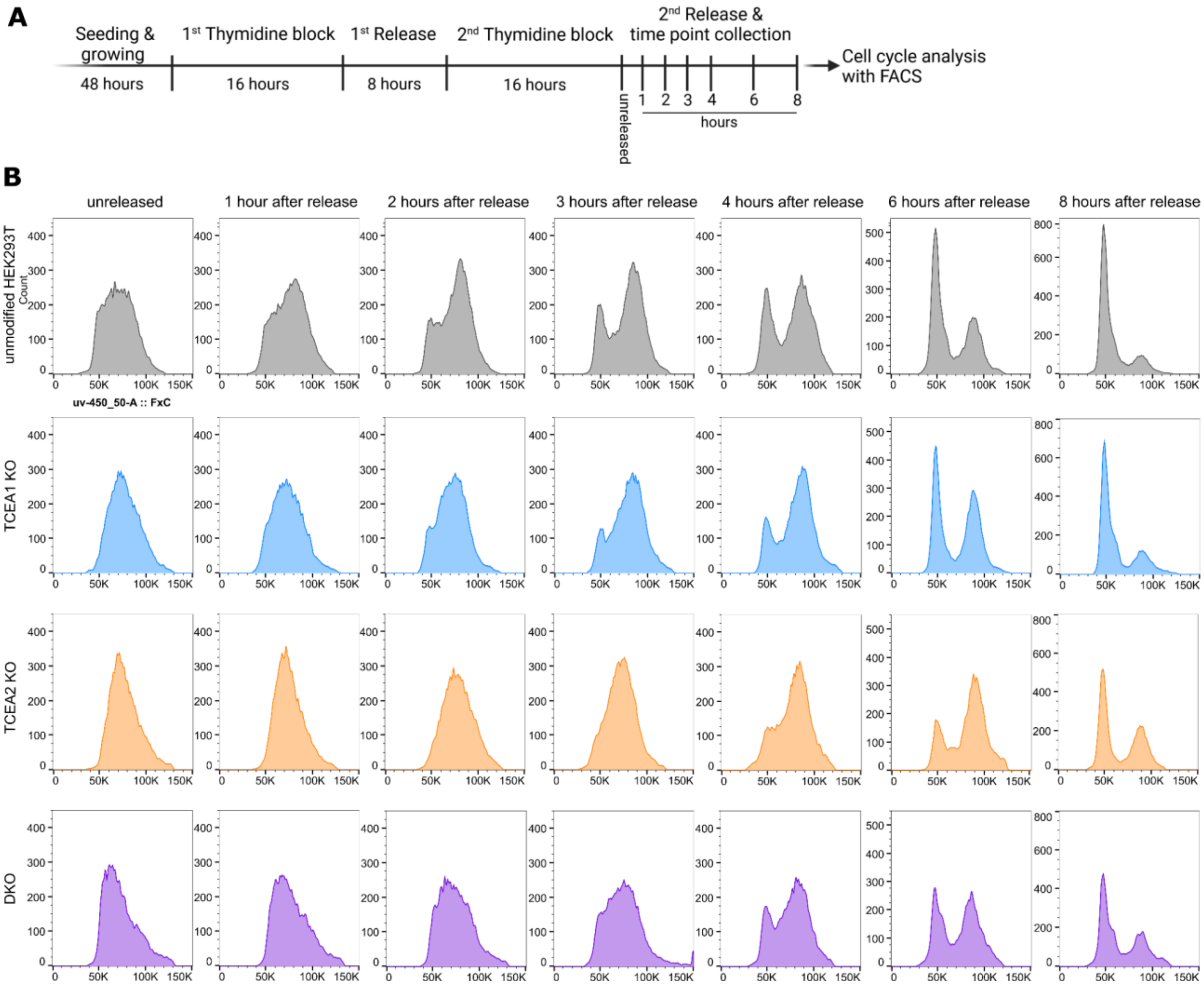


Figure 25. Deletion of TCEA2 causes a clear deceleration of S phase progression.
A. Schematic illustrating cell synchronization in late G_1/early S phases using the double thymidine block method, **B.** Cell cycle progression, after the release from the late G_1/early S block, is profiled in HEK293T, TCEA1 KO, TCEA2 KO, and DKO. The histogram indicates the amount of DNA content, which relates to the cell cycle phases as $G_0/G_1 \propto$ 1X DNA (signal intensity 50K), 1X <S <2X DNA (signal intensity between 50K and 100K), and $G_2/M \propto$ 2X DNA (signal intensity 100K). The x-axis corresponds to the DNA stain (FxCycle) signal intensity. The y-axis is proportional to the cell number.

3.5.3 Activated DNA damage response in the TCEA knockouts

Previous study revealed that overexpression of inactivated TCEA1 mutant caused aberrant R-loops and subsequent DNA damage [368]. We searched for evidence of DNA damage response (DDR) in chromatin fractions of all knockouts by Western blotting against some classical proteins involved in the DDR (Figure 26). As a positive control for both DNA damage and replication stress, we treated unmodified HEK293T cells with camptothecin (CPT), which is a potent inhibitor of DNA topoisomerase TOP1. It binds TOP1-DNA complex and prevents re-ligation of DNA, which causes DNA replication stress via DNA damage and inhibits cell cycle progression (reviewed in [369]. We included additional independently derived clones, but with a smaller passage number, which potentially contributes to variabilities on the Western blots.

The most upstream and central regulators of DDR in mammalian cells are the kinases ATM (Ataxia-telangiectasia mutated) and ATR (ATM- and Rad3-Related) (reviewed in [370]). ATM mostly gets activated by DNA double strand breaks (DSBs), in contrast to ATR, which is activated in response to a broader spectrum of DNA damage including DSBs and replication stress [370]. We found that activated ATR (ATR-Ser428p) level was elevated in the TCEA2 KO and CPT-treated cells. Additionally, a phosphorylation of P53 at Ser15, which can be done by many kinases including ATR (reviewed in [371]), was the most striking indicator of DDR. It was highly increased in the TCEA2 knockout and the DKO. The DNA damage marker, H2A.x, was also mildly elevated in the knockout cells.

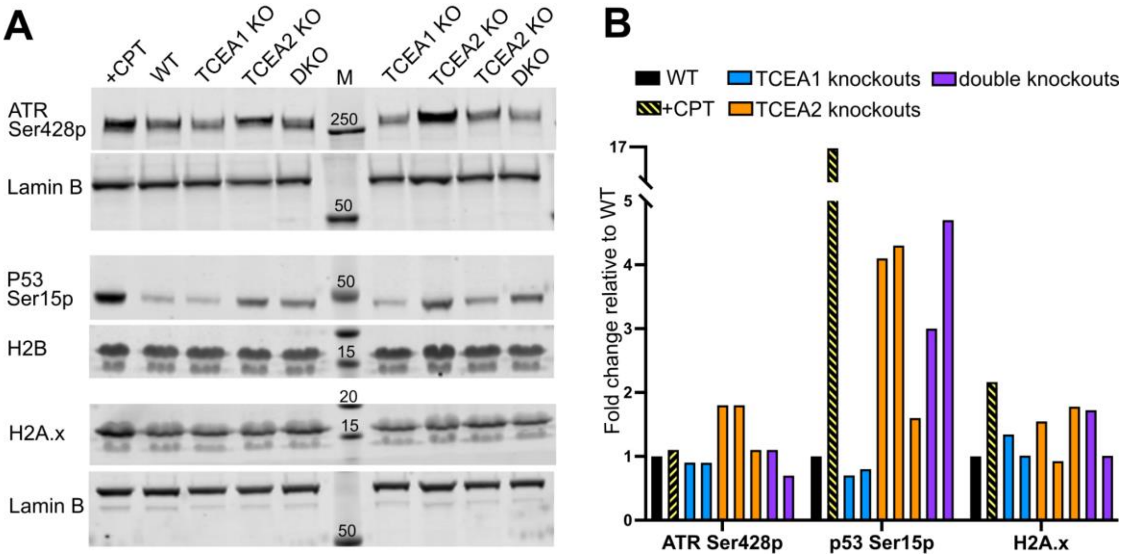


Figure 26. Indication of DNA damage in the TCEA2 KO and the DKO.
A. Western blot against DNA damage response markers in multiple knockout clones. +CPT refers to unmodified HEK293T cells treated with CPT to induce DNA damage. Lamin B and H2B are the loading controls; **B**. Quantified western blot signal normalized to loading control, shown as fold change to WT.

Next, we wondered whether the DNA damage was caused by aberrant R-loops in our knockouts, as it was shown that anterior to Pol II R-loops form excessively upon stabilization of arrested Pol II with the overexpressed inactivated TCEA1 mutant [308]. R-loops are 100 - 2,000 base pair RNA-DNA hybrids that are involved in multiple physiological processes as positive and negative regulators and, additionally, prompt DNA damage [372]. To evaluate R-loop formation, we performed a slot blot with the S9.6 antibody against DNA-RNA hybrids, using dsDNA as a loading control (Figure 27). The specificity of the S9.6 antibody is a subject of debate. Therefore, we treated the samples with RNase H overnight and found that the antibody also detected other nucleic acid species (Figure 27C).

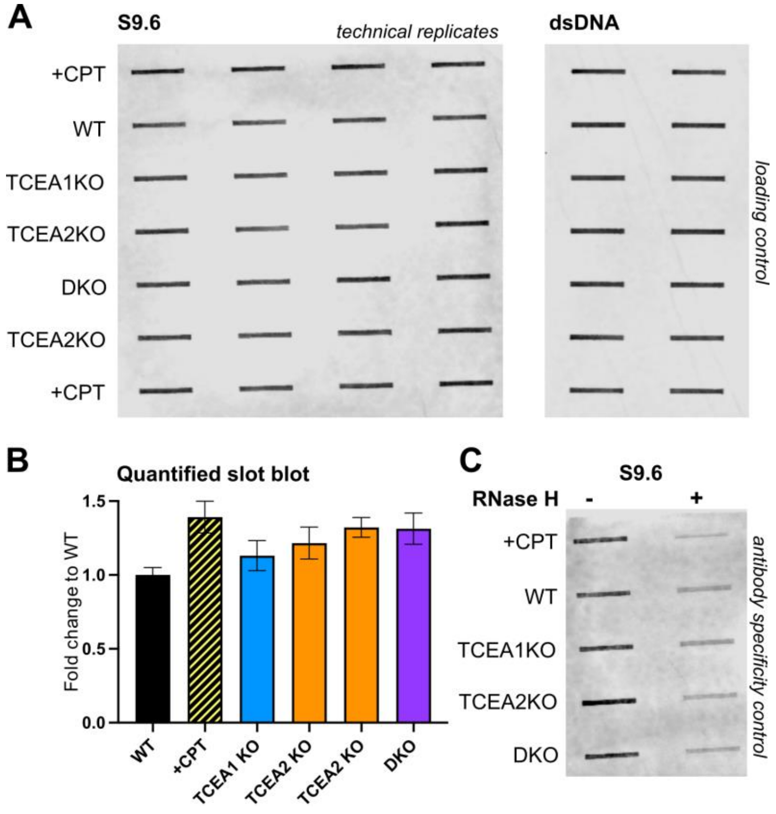


Figure 27. R-loop level is moderately increased in the TCEA knockouts.
A. Slot blot against DNA-RNA hybrids with S9.6 antibody in 4 technical replicates per cell line. CPT treatment was used to induce R-loop formation in the WT and served here as a positive control in 8 technical replicates. Two independently derived TCEA2 KO clones were tested. dsDNA was used as a loading control in 2 technical replicates; **B**. Quantified slot blot normalized to loading control, shown as fold change to WT; **C**. S9.6 slot blot with overnight RNaseH treated and untreated samples indicates that the S9.6 antibody does not have the optimal specificity for DNA-RNA hybrids.

Quantification of band intensities revealed only a mild increase in R-loops, which was more prominent in the TCEA2 KO and the DKO. Although we expected to observe a greater R-loop accumulation, we propose that, in the knockout system, the immediate consequences of arrested Pol II are more transient because its intrinsic nuclease activity is not perturbed, in contrast to the system where backtracked Pol II is immobilized by an overexpressed mutant TFIIS [308]. Additionally, it is plausible that the cells adapted to alleviate R-loop accumulation.

3.5.4 Additional DNA damage repair factors are enriched and lost from chromatin

To further understand the DNA damage response in the knockouts, we explored our quantitative SILAC chromatin mass spectrometry datasets. Gene ontology term analysis revealed a “DNA repair” term enrichment and it was present among both increased and decreased proteins in all knockout cell lines. We have inspected the proteins in this category, cross compared the changes across knockouts and grouped them based on their roles, as well as added their expression change determined by RNA-seq (Figure 28). As long as a protein was significant and changed by at least 20% in at least one knockout, it was included in the list and was also searched in the other knockouts. Because not all proteins were significantly changed in all datasets, we have still included the ones below the thresholds and added the information of adjusted p-value in the figure.

In our setup, SILAC chromatin MS does not provide the information about post-translational modifications, however, it provides an IP-independent quantitative evidence of recruitment or displacement of proteins, relative to their abundance in unmodified HEK293T cells. In general, the most changes in DDR protein abundance are found in the DKO. A lot of these changes are unique to the DKO, indicating that the absence of both TCEA1 and TCEA2 stresses the cells and drives the most severe phenotypic changes.

Recruited DNA damage repair factors in all knockouts

Although the well-known R-loop resolving factors, such as RNase H, SETX, DHX9, XRN2 [372] were not detected, we found an enrichment of four proteins involved in transcription-coupled nucleotide excision repair (TC-NER), consistently among all knockouts: ELOF1, USP7, TRIP12, and XAB2. In the current view, TC-NER machinery indirectly recognizes DNA damage by sensing stalled Pol II via proteins CSA and CSB (not detected), together with some general NER factors [373]. Ubiquitin-specific protease 7 (USP7) regulates the stability of CSB by deubiquitinating it [374], while TRIP12 targets USP7 for polyubiquitination and proteasome degradation [375]. Two recent studies elucidated the role of the transcription elongation factor

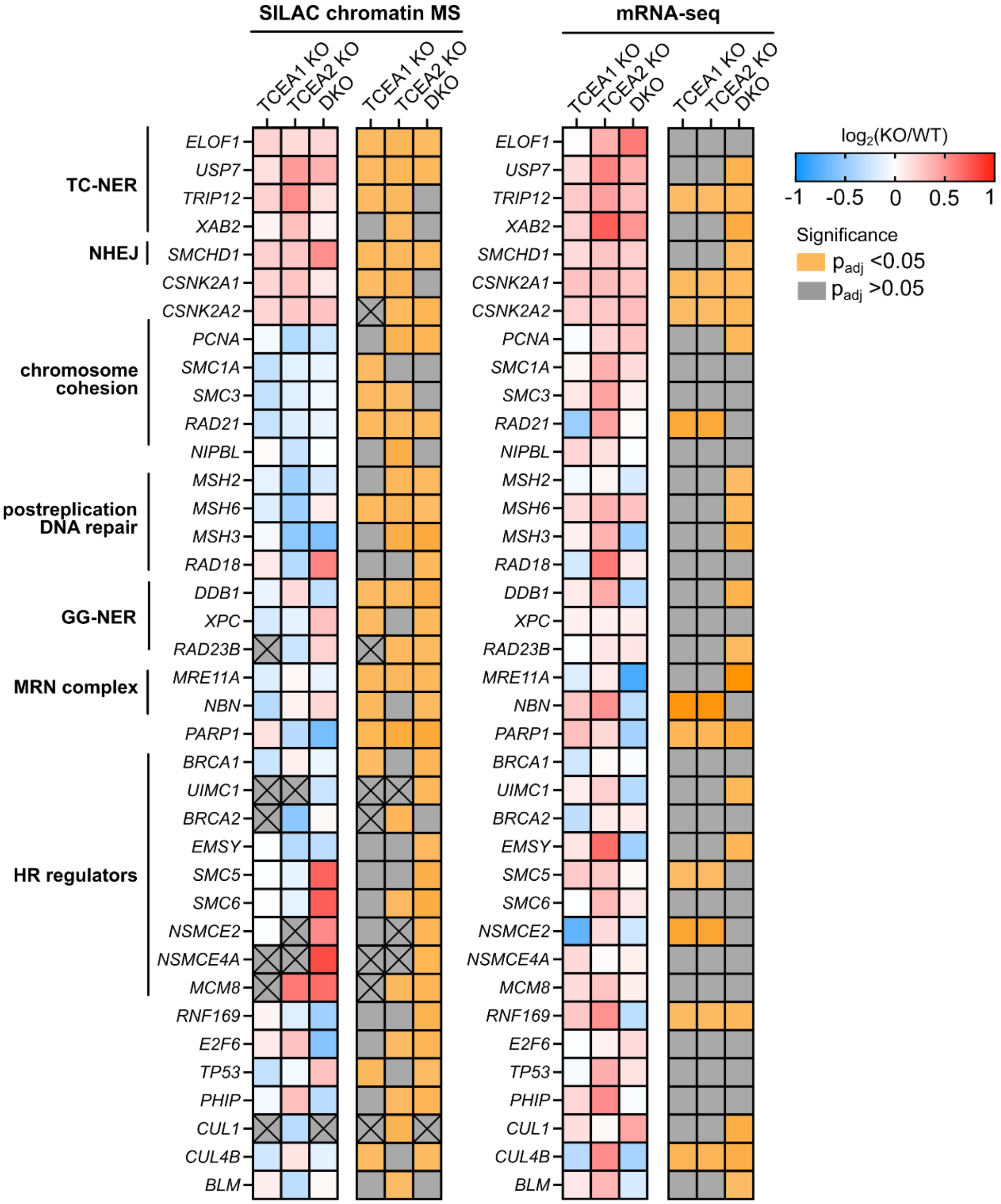


Figure 28. Chromatin MS reveals changes in abundance of proteins involved in DNA repair in the TCEA1 KO, TCEA2 KO, and the DKO.

Abundance of proteins involved in DNA repair and their gene transcription levels are compared among the knockouts and shown as $\log_2$(SILAC ratio of KO to WT) and $\log_2$(fold change of KO to WT), respectively. Statistically significant factors are denoted with an orange square, the ones below the threshold (adjusted p-value >0.05) are in gray. “X” denotes that the protein was not detected. TC-NER: transcription-coupled nucleotide excision repair, NHEJ: nonhomologous end joining, GG-NER: global genome nucleotide excision repair, HR: homologous recombination.

ELOF1 in TC-NER: 1) it promotes Pol II ubiquitylation [376], 2) it can prevent R-loops via a CSB-independent pathway, and 3) facilitates the recruitment of UVSSA and TFIIH for repair [377]. This regulatory mechanism may be activated due to the DNA damage or by backtracked Pol II in the absence of TCEA1 and TCEA2 or both.

SMCHD1 (Structural maintenance of chromosomes flexible hinge domain-containing protein 1) is consistently upregulated in all knockouts. This protein has a few crucial roles: it mediates chromosome architecture during development [378-381]; it was found at sites of DNA damage promoting nonhomologous end joining (NHEJ) [382, 383], and it activates ATM-mediated DNA repair at uncapped telomeres [384]. The enrichment of this factor in the knockouts suggests that it is implicated in DDR in our knockout cells.

Casein kinase 2 protein family, including CSNK2A1 (CK2α) and CSNK2A2 (CK2α′), is also upregulated at both protein and transcription levels. CK2 is essential for numerous regulatory pathways, including cell cycle progression [385]. For example, it interacts with the FACT complex and phosphorylates P53, thereby promoting cell cycle arrest [386], which we have observed (Figure 26).

Displaced DNA damage repair factors

Overall, the proteins that were displaced from chromatin across all knockouts can be grouped as chromosome cohesion and post-replication DNA repair factors. Deregulation of these factors can explain the prolonged S phase of the knockouts. Some proteins and transcripts were detected below the significance thresholds, so only the significantly detected proteins are described further.

Proliferating cell nuclear antigen (PCNA) is an essential factor in DNA replication and repair and is depleted from chromatin in the TCEA2 KO and DKO, despite being upregulated at the mRNA level in the DKO, potentially reflecting a problem with recruitment and stability. After replication, DNA damage or incorrect base pairing is carefully screened by the cell's mismatch repair (MMR) system. Some MMR factors are also displaced in the knockouts. Msh2-Msh6 and Msh2-Msh3 heterodimers recognize small (1-2 base pairs) mismatches and larger (15 nt) insertion deletions, respectively [387]. Furthermore, we find a consistent displacement of RAD21, a DSB repair protein and a subunit of the cohesin complex, which is indispensable for post-replicative DNA repair and proper chromosome segregation [388], in all the knockouts. Interestingly, the *RAD21* gene is upregulated in the TCEA2 KO, suggesting that

RAD21 is not properly recruited or stabilized, despite being elevated at the mRNA level. Additional components of the cohesin complex (SMC1a and SMC3) and a cohesin-loading factor NIPBL are also lost from chromatin with variable statistical significance.

Knockout-specific deregulation of DNA damage repair factors

Global genome nucleotide excision repair (GG-NER) is a NER subpathway, which in contrast to TC-NER, is activated directly by erroneous base pairing causing DNA helix distortions (reviewed in [373]). A few factors involved in GG-NER were differentially abundant at chromatin in the knockouts. For example, DDB1 (Damage specific DNA damage binding protein) is a part of the UV-DDB complex, which senses helix distortions [373], and is depleted in the DKO both at the protein and transcript levels. In the TCEA1 KO and TCEA2 KO, DDB1 is mildly displaced and enriched respectively, however their transcription levels are not clear because they are below the significance threshold. Interestingly, XPC, which is a key DNA damage sensor of NER, and its stabilizing interactor, RAD23B [373], are both enriched only in the DKO.

MRE11 (*MRE11A*) and Nibrin (*NBN*) are components of the MRN (MRE11/Rad50/NBS1) complex, which is a central regulator of DDR. It recognizes DNA damage and recruits and activates ATM [389] and poly(ADP-ribose) polymerase 1 (PARP-1, *PARP1*) [390]. MRE11 and NBN are less abundant in the TCEA1 KO, but not changed in the TCEA2 KO. In the DKO, MRE11 is transcriptionally downregulated and the protein is depleted from chromatin, however NBN is slightly enriched in chromatin.

PARP-1 is a key regulator of single and double strand break repair, NER, chromatin structure and stabilizer or replication forks [391]. Interestingly, the transcription of PARP-1 is upregulated in the TCEA1 KO and TCEA2 KO, however its recruitment to chromatin is compromised in the TCEA2 KO, while slightly enhanced in the TCEA1 KO. This suggests that PARP-1 recruitment may be dependent directly or indirectly on TCEA2. In the DKO, it is reduced at transcription and chromatin protein levels.

In the DKO, we observe a significant enrichment of factors involved in homologous recombination (HR). For example, the structural maintenance of chromosome complex SMC5-SMC6 mediates DNA damage repair and chromosome segregation (reviewed in [392]). Besides many other subunits, the complex also includes an E3 SUMO ligase subunit, NSMCE2 and a kleisin subunit NSMCE4A [392], which are also enriched in the DKO chromatin. Additionally, a helicase implicated in HR-mediated DNA repair, MCM8 [393], is enriched in the chromatin of TCEA2 KO and the DKO (it is not detected in TCEA1 KO).

In summary, there is a deregulation of multiple DDR proteins of various pathways that possibly contribute and/or reflect the growth defects of the knockouts. The DDR factors are deregulated at the level of mRNA and protein abundance at chromatin. However, we found a clear depletion of the proteins, responsible for genome integrity during and after the DNA replication phase, specifically, the cohesin complex and other post-replication repair proteins and a consistent upregulation of the TC-NER factors.

3.6 TCEA1 and TCEA2 are general transcription factors and their loss globally impacts the transcriptional output of cells

3.6.1 RNA-seq reveals that TCEA1 and TCEA2 deletions globally affect the transcriptional output of the cells

We have expected global changes in transcription in the TCEA1 KO and DKO because our ChIP-seq revealed that FLAG-TCEA1 and FLAG-TCEA2 bind at thousands of actively transcribed protein-coding genes (Figures 17&18) and backtracking had been previously shown to occur ubiquitously in human cells [154]. Sheridan et al. had shown that the increased backtracking, provoked by the overexpressed RNA cleavage-inhibiting TCEA1 mutant, caused striking changes in transcription [154]. They observed defects in the 5' pause release, a decreased transcription elongation rate, and earlier transcript termination, so we expected considerable changes in the steady state RNA in the knockouts. We have also hypothesized that TCEA1 is the prevailing transcription elongation factor over TCEA2 because it is higher expressed in HEK293T cells. To test these hypotheses, we performed RNA-seq of polyadenylated RNA transcripts of unmodified HEK293T (WT), the TCEA1 KO, TCEA2 KO, and DKO. This experiment included spike-in RNA, which enabled us to quantitatively compare gene expression across all conditions.

The differential expression (DE) analysis was performed using the DESeq2 package [339] with a rather strict threshold. We defined the differentially expressed genes as those whose number of transcripts per million base pairs (TPM) had changed by at least a factor of 2 ($\log_2$fold change ≥1) with the statistical significance of adjusted p-value ≤0.01 (Figure 29A). Unexpectedly, TCEA2 KO affected almost twice as many genes as the TCEA1 KO. In the TCEA2 KO, the transcription of 954 genes was changed (533 genes were downregulated and

421 genes were upregulated), in contrast to 486 differentially expressed genes (232 were downregulated and 254 genes were upregulated) in the TCEA1 KO. The DKO had the most profound effect on gene expression, changing the transcriptional output of 1213 genes (804 genes were downregulated and 409 genes were upregulated) (Figure 29B).

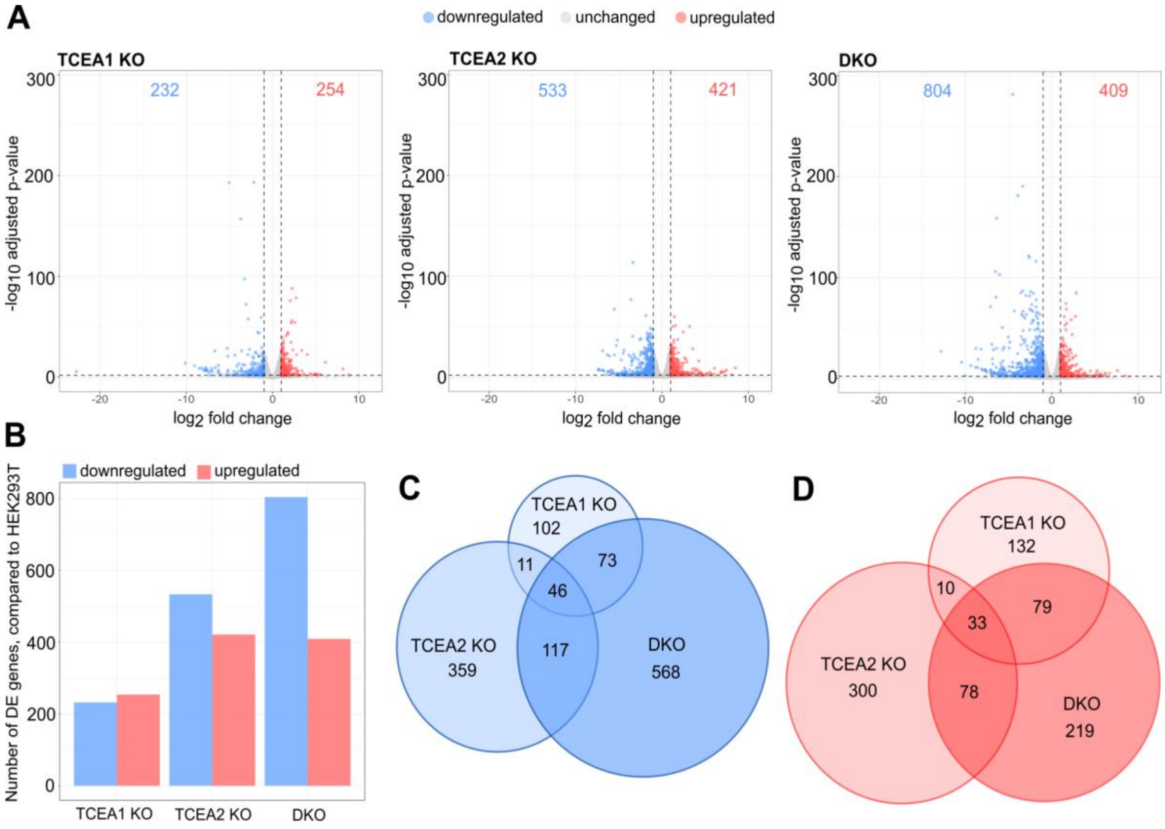


Figure 29: TCEA1/2 deletions cause changes in the steady state RNA levels of numerous genes. A. Volcano plots based on spike-in normalized RNA-seq data in TCEA1 KO, TCEA2 KO, DKO, and HEK293T cells as a control. RNA-seq was performed in three biological replicates per condition. The x-axis reflects $\log_2$ fold change in expression between the knockout and control unmodified HEK293T. The y-axis reflects the statistical significance, as a $-\log_{10}$ transformed FDR-adjusted p-value. Differentially expressed genes are defined as those with $\log_2$ fold change ≥1 and the adjusted p-value ≤0.01. The number in blue refers to the number of genes with reduced expression. The number in red refers to the number of genes with increased expression. **B.** A bar plot showing the numbers differentially up- and downregulated genes in the knockouts. **C.** A Venn diagram of downregulated genes in all knockouts. **D.** A Venn diagram of upregulated genes in all knockouts.

As expected, we observed a reduction in the transcription of numerous genes in the TCEA1 KO, TCEA2 KO, and DKO. Although the overlap of the downregulated genes is substantial, the affected genes are not the same across the conditions (Figure 29C). There are two likely reasons for this observation: 1) the implemented statistical and fold change thresholds eliminated slightly changed genes in the knockouts and 2) accumulating indirect effects, which are unavoidable in a knockout system. The latter may also be reflected in the upregulated genes.

3.6.2 The slow growth phenotype of the knockouts is reflected at the mRNA level

To get a general idea of the genes affected by the TCEA1, TCEA2, and double TCEA1/TCEA2 deletions, we performed gene ontology (GO) enrichment analysis on significantly upregulated and downregulated genes in all knockouts. Using ConsensusPathDB [340], we performed an over-representation analysis on the lists of DE genes with detected transcripts in HEK293T and all knockouts serving as the background. In accordance with the detected massive changes in transcription, we found that numerous biological processes were significantly enriched in all knockouts and summarized the results as top 50 enriched GO terms (Figures 30 and 31).

The downregulated genes mostly associate with the morphogenesis of many organs and, in general, reflect the phenotypic HEK293 characteristics: the enriched GO terms associate with the development of organs of both mesodermal and ectodermal origins. HEK293 cells are usually described as kidney cells with the expression of some neuron-specific genes [394]. A recent transcriptomic analysis determined that the cells were likely isolated from an embryonic adrenal precursor structure [395]. While the adrenal cortex is derived from intermediate mesoderm, the medulla of the adrenal gland arises from neural crest cells, which are derived from ectoderm [396]. This, supposedly, contributes to the neuronal-like phenotype of HEK293 cells. The largest number of the enriched GO terms is in the DKO, followed by TCEA2 KO, and TCEA1 KO, and could explain the observed differential extent of transcriptional changes and the growth defect.

From the GO term analysis of upregulated genes, we can estimate the adaptive, secondary effects. A broad range of biological processes is enriched. Interestingly, some gene categories can explain the alterations in cellular growth. For example, “G1 to G0 transition” (cell cycle arrest) and “cellular response to redox state” [397]. The upregulated genes belonging to these

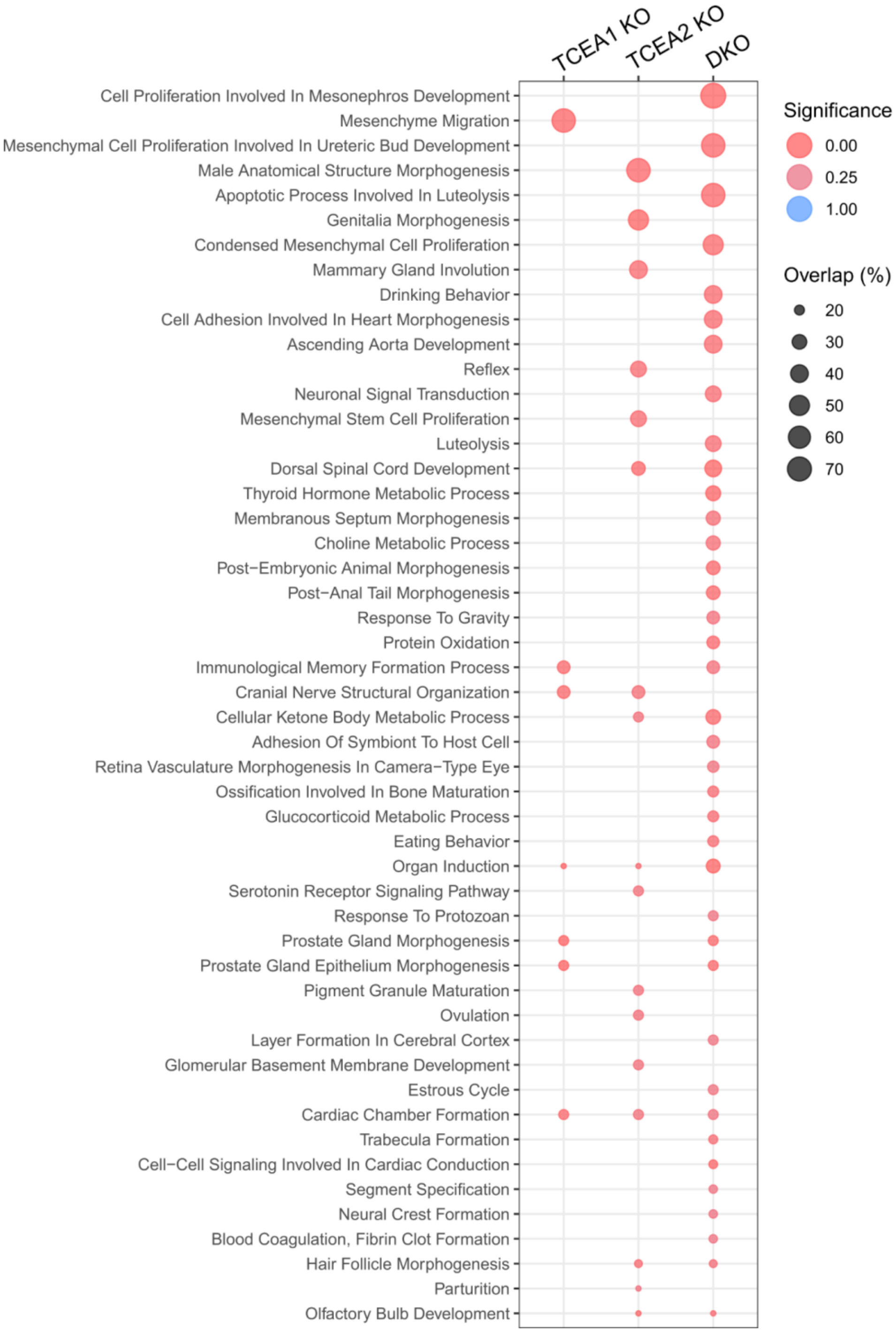


Figure 30. Gene ontology term analysis of downregulated genes in TCEA1 KO, TCEA2 KO and DKO.

Top 50 enriched GO terms arranged based on percentage overlap between the number of downregulated genes belonging to a term and all genes belonging to that term in the background list, consisting of all expressed genes in HEK293T and all knockouts. Significance is $\log_{10}$(q value) and is denoted by the color: red is for high significance, pink is for lenient significance, and blue is for not significant enrichment.

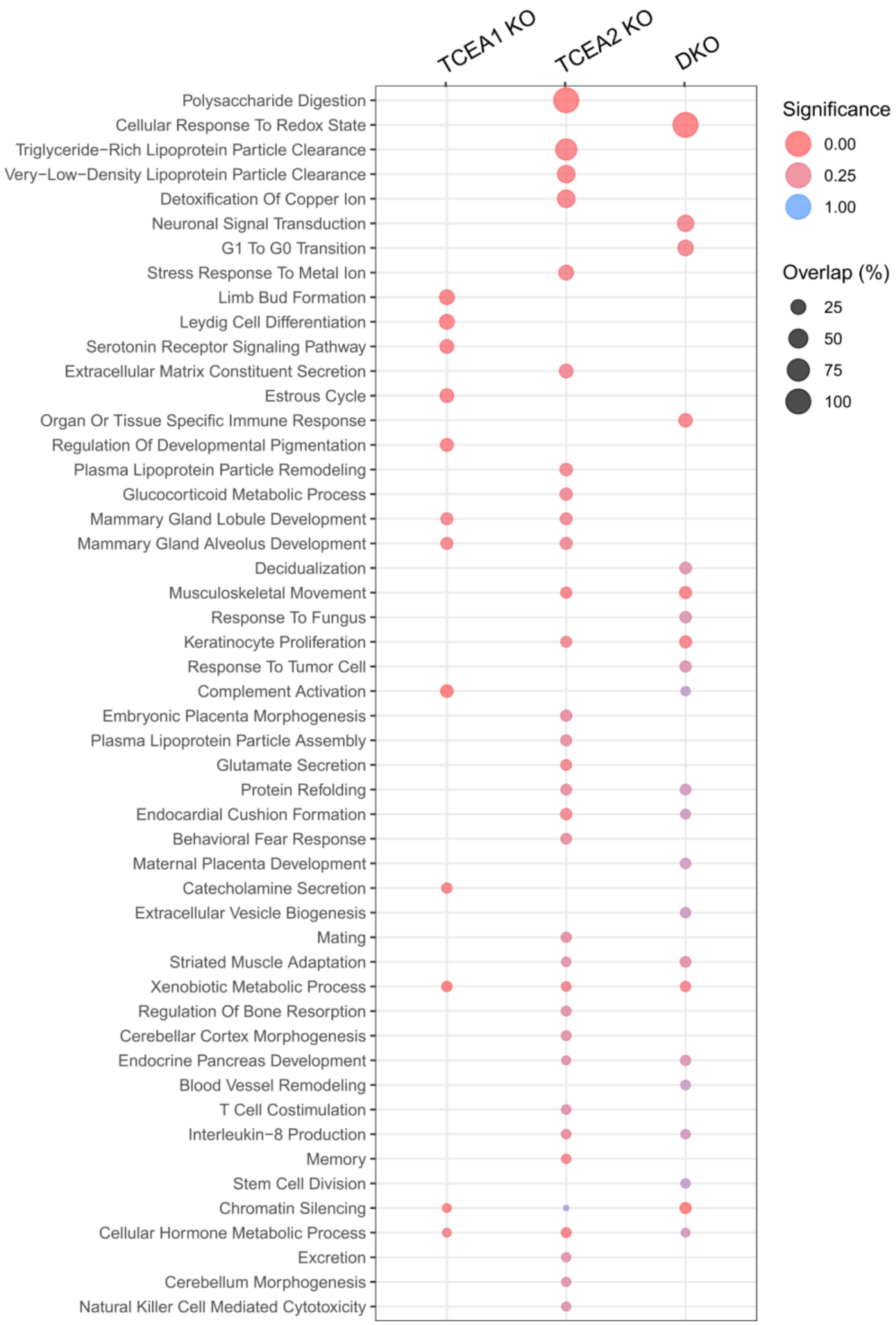


Figure 31. Gene ontology term analysis of upregulated genes in TCEA1 KO, TCEA2 KO and DKO.

Top 50 enriched GO terms are arranged based on percentage overlap between the number of upregulated genes belonging to a term and all genes belonging to that term in the background list, consisting of all expressed genes in HEK293T and all knockouts. Significance is $\log_{10}$(q value) and is denoted by the color: red is for high significance, pink is for lenient significance, and blue is for not significant enrichment.

terms are *SMPD3* which mediates lipid metabolism [398], *CYP27B1,* which catalyzes vitamin D activation in kidney and extraskeletal tissues [399], and *VASN* which was suggested to attenuate TGF-β signaling by direct binding [400].

In summary, both downregulated and upregulated genes are involved in multiple processes that can explain the observed slow growth phenotype, as many of the gene categories comprise proliferation- and development- related genes. Many enriched GO terms match across the knockouts, cross-validating the effects of TCEA1 and TCEA2 deletions.

3.6.3 Transient TCEA2 complementation in the TCEA2 knockout cell line did not reverse the aberrant transcription, however, suggested a few putative TCEA2-target genes

Although TCEA1 is the higher expressed elongation factor, the deletion of TCEA2 impacts the transcriptional output of twice as many genes (Figure 29). We hypothesized that if TCEA2 were a more potent elongation factor than TCEA1, potentially, because of the subtle differences in the NTDL, supplementing TCEA2 back into the TCEA2 KO would reverse the transcriptional changes, at least partially. To test this hypothesis, we performed a TCEA2 gene complementation experiment in the TCEA2 KO via transient transfection with a FLAG-TCEA2 expression plasmid and evaluated the changes in gene expression with RNA-seq.

We were interested in the immediate TCEA2-responsive genes, so we chose the earliest time points, 8 and 12 hours, at which TCEA2 was observed at the protein level (Figure 34B) and harvested the cells for RNA-seq. The negative control was the TCEA2 KO cells transfected with only Lipofectamine. The experiment was done in three biological replicates for each time point, each replicate was separately transfected, however, one replicate in each time point was an outlier and was omitted from the differential expression (DE) analysis.

We did not observe a global increase in transcription, but found only 8 DE genes after 8 hours of complementation and 15 DE genes after 12 hours of complementation (Figure 34C). Interestingly, almost all downregulated genes play a role in growth or apoptosis. For example, the only gene that was affected in both complementation durations is *ASAH2,* a ceramidase regulating the sphingolipid pathway. Ceramidases cleave ceramides to produce sphingosine-1 phosphate and free fatty acids. The homeostasis between ceramides and sphingosines is of high importance because of their involvement in cell signaling and cell fate decisions [401, 402]. Ceramides were shown to mediate stress response and apoptosis [403], in contrast to sphingosines, which stimulate growth and proliferation [404]. Furthermore, *ASAH2* was reported to repress the P53 pathway to suppress ferroptosis [405], regulated cell death caused by lipid peroxidation and iron accumulation [406]. Among the upregulated genes, there are also some genes that can potentially link TCEA2 with regulation of cell proliferation, for example, *ATF3* and *FAM3C.*

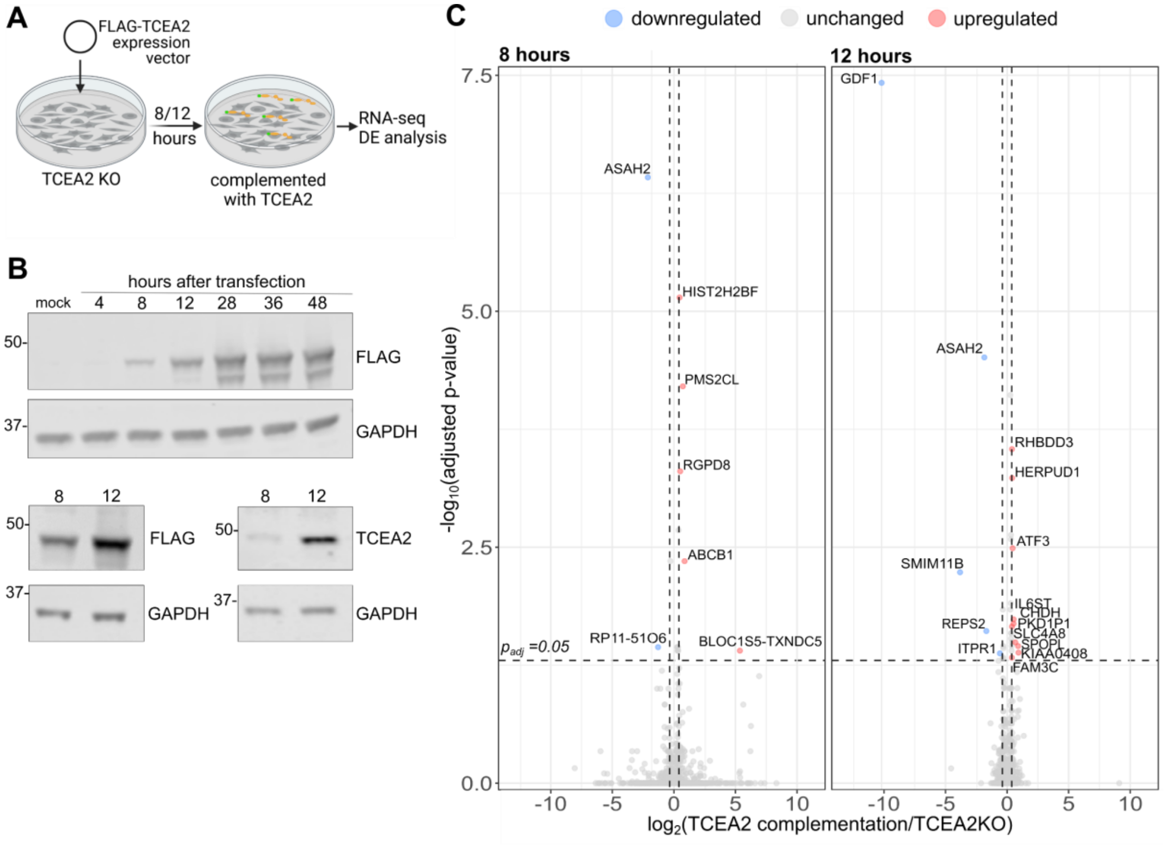


Figure 32. TCEA2 complementation for 8 and 12 hours changes the expression of few genes. **A.** Experimental setup of the TCEA2 complementation experiment. TCEA2 KO is transiently transfected with FLAG-TCEA2 cDNA-containing expression plasmid and after 8 and 12 hours harvested for RNA-seq. **B**. Western blots showing the expression level of FLAG-TCEA2 after 4 - 48 hours after transfection. On the upper FLAG Western blot, a degradation product, seen as the lower band, was caused by sample preparation with expired protease inhibitors. Sample preparation with fresh protease inhibitors eliminated the artifact (lower blots). **C**. Volcano plot showing the DE genes upon 8 and 12 hour - TCEA2 complementation. The x-axis corresponds to the log2 fold change of TCEA2-complemented TCEA2 KO to uncomplemented TCEA2 KO. The dash lines indicate the change in expression by 30% (log_{10}(1.3)). Significance is shown on y-axis as -log_{10} of adjusted p-value. TCEA2 expression was omitted from the volcano plot, as it has a very low adjusted p value of $1x10^{-300}$ for both time points and log_2 fold change of 9.05 and 10.6, for 8 and 12 hours, respectively).

Next, we compared the TCEA2-complementation-sensitive genes in terms of the change in their expression compared to unmodified HEK293T cells (WT) (Figure 33) and found that TCEA2 complementation could reverse the effect of the knockout on some genes. Eight genes were repressed in the TCEA2 KO compared to WT and were upregulated when TCEA2 was added back. Conversely, five genes were upregulated in the TCEA2 KO, compared to WT and were downregulated upon TCEA2 complementation. The expression of the rest was changed in one direction.

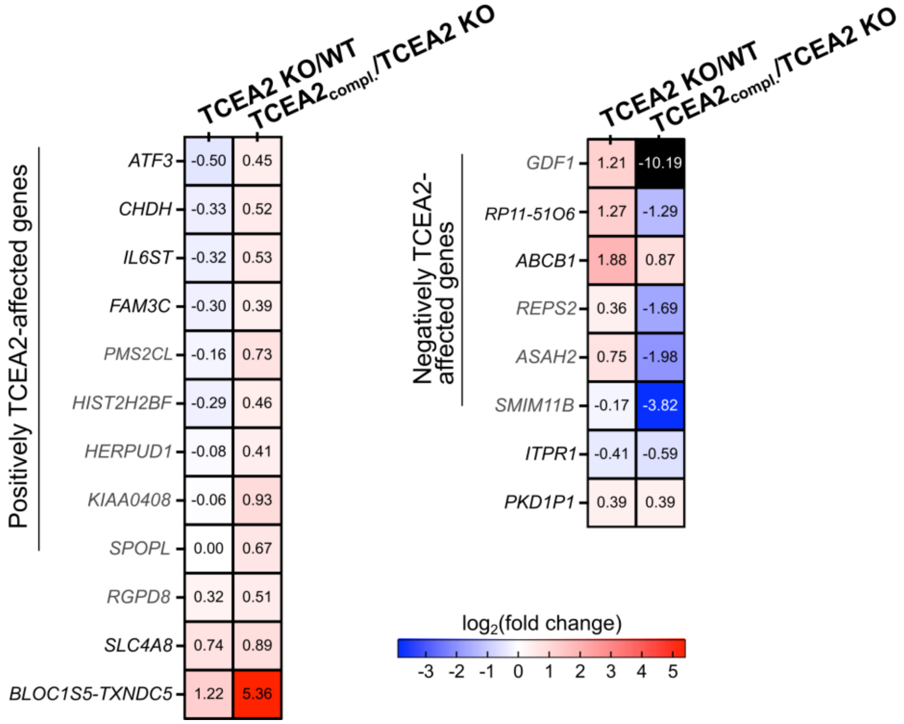


Figure 33. TCEA2 complementation reverses aberrant transcription of some genes.
Comparison of the changes in transcriptional output of the genes that were differentially expressed upon TCEA2 complementation for both 8 and 12 hours to their expression in unmodified HEK293T cells (WT) as a log$_2$ fold change. The genes in gray had a $-\log_{10}(p_{adj})<1.3$ in RNA-seq of WT vs TCEA2 KO.

In summary, a short TCEA2 complementation did not restore the global transcription deregulation caused by the loss of TCEA2. Possibly, the adaptive effects in the TCEA2 KO cell line had been established for too long to be globally corrected by a short gene complementation.

3.7 HiS-NET-seq reveals that Pol II accumulates in the promoter-proximal regions in the absence of TCEA1 and TCEA2 on majority of genes

HiS-NET-seq (High-sensitive nascent transcript sequencing) is based on NET-seq [89] and SI-NET-seq [113], which capture transcriptionally engaged Pol II with its associated nascent RNA at a single nucleotide resolution and with strand specificity. HiS-NET-seq entails an additional step: a short metabolic labeling with 4-thiouridine (4sU) of cells in culture (Figure 6), which enables enrichment and sequencing of only the nascent RNA with an incorporated 4sU, enhancing the number of informative reads [311]. As SI-NET-seq, this protocol also includes the addition of mouse cells as spike-in control, enabling quantitative comparisons among cell lines in terms of Pol II occupancy and transcript abundance. We performed HiS-NET-seq in unmodified HEK293T (WT), the TCEA1 KO, TCEA2 KO, and double TCEA1 and TCEA2 KO (DKO) in two biological replicates. The computational analysis is still ongoing and here we show our preliminary results.

The metagene analysis of ~6500 active genes revealed a clear increase in transcriptionally engaged Pol II abundance at the promoter-proximal region (PPR) in all TCEA knockouts (Figure 34). More specifically, the accumulation of Pol II occurs within the first 200 nucleotides downstream of the TSS, which indicates that TCEA1 and TCEA2 regulate early elongation, as expected based on our ChIP-seq observation (Figure 18). Our results are also in accordance with a recent study, demonstrating that Pol II is prone to backtracking particularly at the 5’ end of genes in the mutant TCEA1 overexpression system [154]. Intriguingly, despite the lower abundance, TCEA2 deletion also had a considerable effect, indicating that TCEA1 alone is not sufficient to relieve all backtracked Pol II. The double knockout also had a striking increase in Pol II density in the PPR. Additionally, there is an upwards shift in magnitude in Pol II occupancy downstream of the promoter-proximal peak, which points to an elongation defect that Pol II experiences in the gene body in the absence of each and both TCEA paralogs.

The upregulation in Pol II occupancy signal can be interpreted as prolonged pausing at regular, scheduled pause sites, pause release defect, or an upregulation in gene transcription. We have performed further bioinformatic analyses in an attempt to disentangle the complexity of this phenotype. We computed the pausing index, which compares Pol II density in the PPR to the signal in the gene body (GB) (Supplemental figure S3). We obtained a positive pausing index for all knockout cell lines, indicating that there is either an impaired pause release at the

PPR or a reduction in transcription in the gene body. To further examine the effect of the TCEA1 and TCEA2 deletions, we calculated the change in Pol II density as $\log_2$(KO/WT) at the PPR and at the GB and presented the changes as a two-dimensional matrix, which we refer to as the pausing matrix (Figure 34D).

The matrix concurrently shows the changes in Pol II density at the PPR and GB of all KO-affected genes and comprises four quadrants. This analysis does not discriminate between an increase in signal intensity and prolonged pausing, but reflects the changes in Pol II occupancy. Quadrant I represents the genes that have a signal decrease at the PPR and a signal increase at the GB and consists of a rather small number of genes (100, 138, and 129 in the TCEA1 KO, TCEA2 KO, and DKO, respectively), indicating that this transcriptional change is not common. However, quadrant II contains a large proportion of genes (3143, 1994, and 2063 in the TCEA1 KO, TCEA2 KO, and DKO, respectively), at which the signal is increased at both PPR and GB, indicating a whole gene upregulation and imaginably reflecting the adaptive effects of the knockouts. However, possibly, some genes of this quadrant experience excessive Pol II pausing at both regions, as prolonged pauses contribute to the overall signal intensity. Quadrant III represents the genes that are downregulated at both PPR and in the GB. This quadrant does not contain a large number of genes (163, 685, and 534 in the TCEA1 KO, TCEA2 KO, and DKO, respectively). The genes in this quadrant appear to be generally downregulated. Interestingly, TCEA2 and the DKO have a larger number of genes that are generally downregulated and some of them can potentially be selectively regulated by TCEA2. Quadrant IV comprises a large fraction of genes (2671, 1325, and 2445 in the TCEA1 KO, TCEA2 KO, and DKO, respectively) and represents the genes that have an increased Pol II occupancy at the PPR and a decreased Pol II occupancy at the GB, indicating an impaired Pol II release from PPR pause or a processivity defect, defined as the inability of Pol II to travel through the whole gene.

In a gene example from this quadrant, which is also bound by TCEA1/TCEA2 in the inducible overexpression system, we can observe a profound increase in signal intensity within 200-250 nucleotides downstream of the TSS in all three knockouts (Figure 35). The pausing pattern of the knockouts is distinct from that of the WT: in the knockouts, the pausing site appears to contain two enrichment sites, the higher peak occurring more downstream. The processivity defect is not reflected on the mRNA level of the first exon.

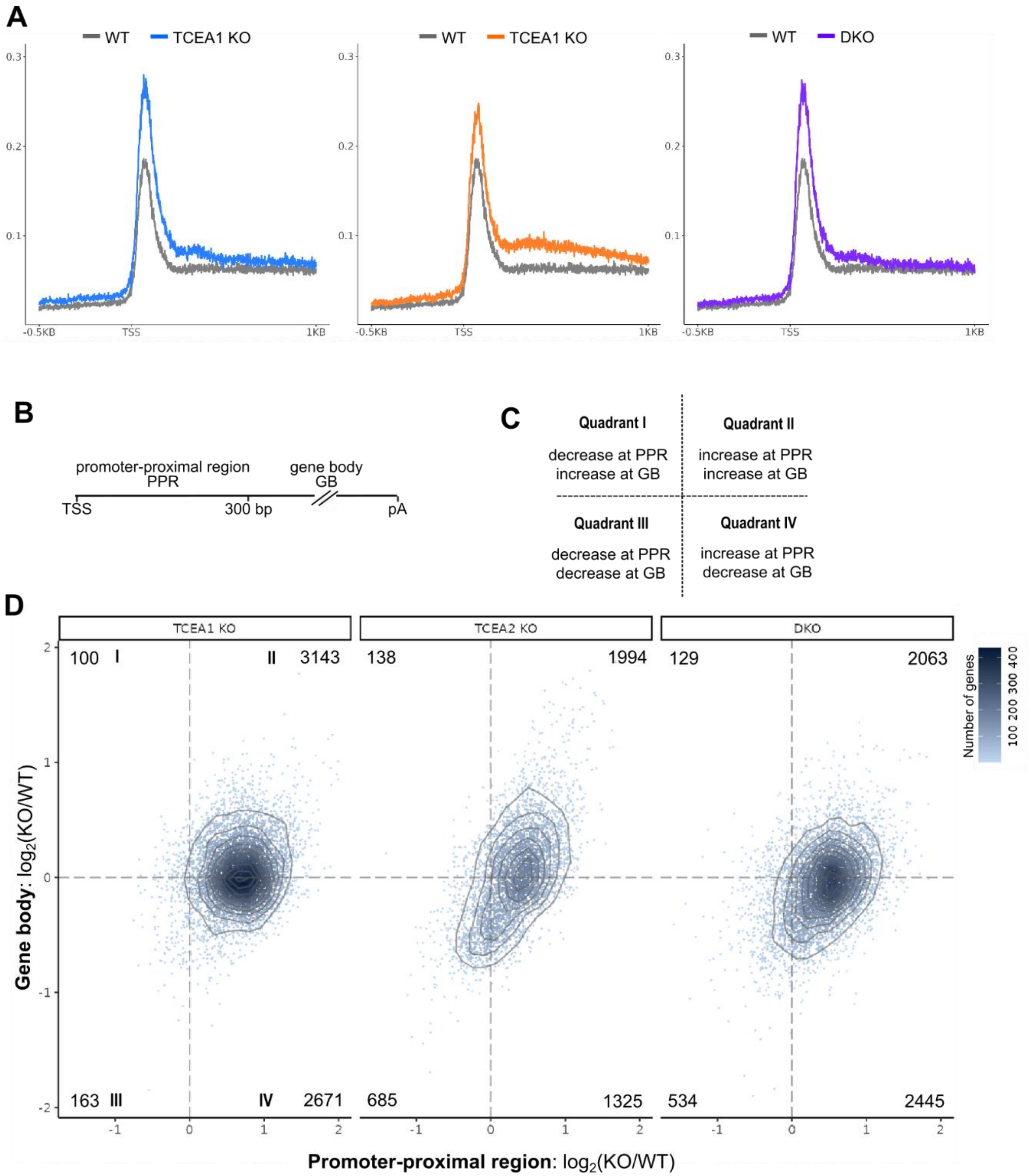


Figure 34. Pol II accumulates at the promoter-proximal region in the absence of TCEA1 and TCEA2.

A. Spike-in normalized metagene analysis of HiS-NET-seq signal in the TCEA1 KO, TCEA2 KO, and DKO compared to unmodified HEK293T cells (WT) in the first 1000 bp of genes, shown as an average of two biological replicates; **B**. Definition of the promoter-proximal region and the gene body, **C**. schematic summary of the pausing matrix; **D**. Pausing matrix showing the change in Pol II occupancy at the promoter-proximal region (x-axis) and the gene body (y-axis) as $\log_2$(fold change of all differentially expressed genes in each knockout compared to the WT). This plot includes all differentially expressed genes, no p-value filtering was applied. I, II, III, IV refer to quadrants, the numbers refer to the number of genes in each quadrant.

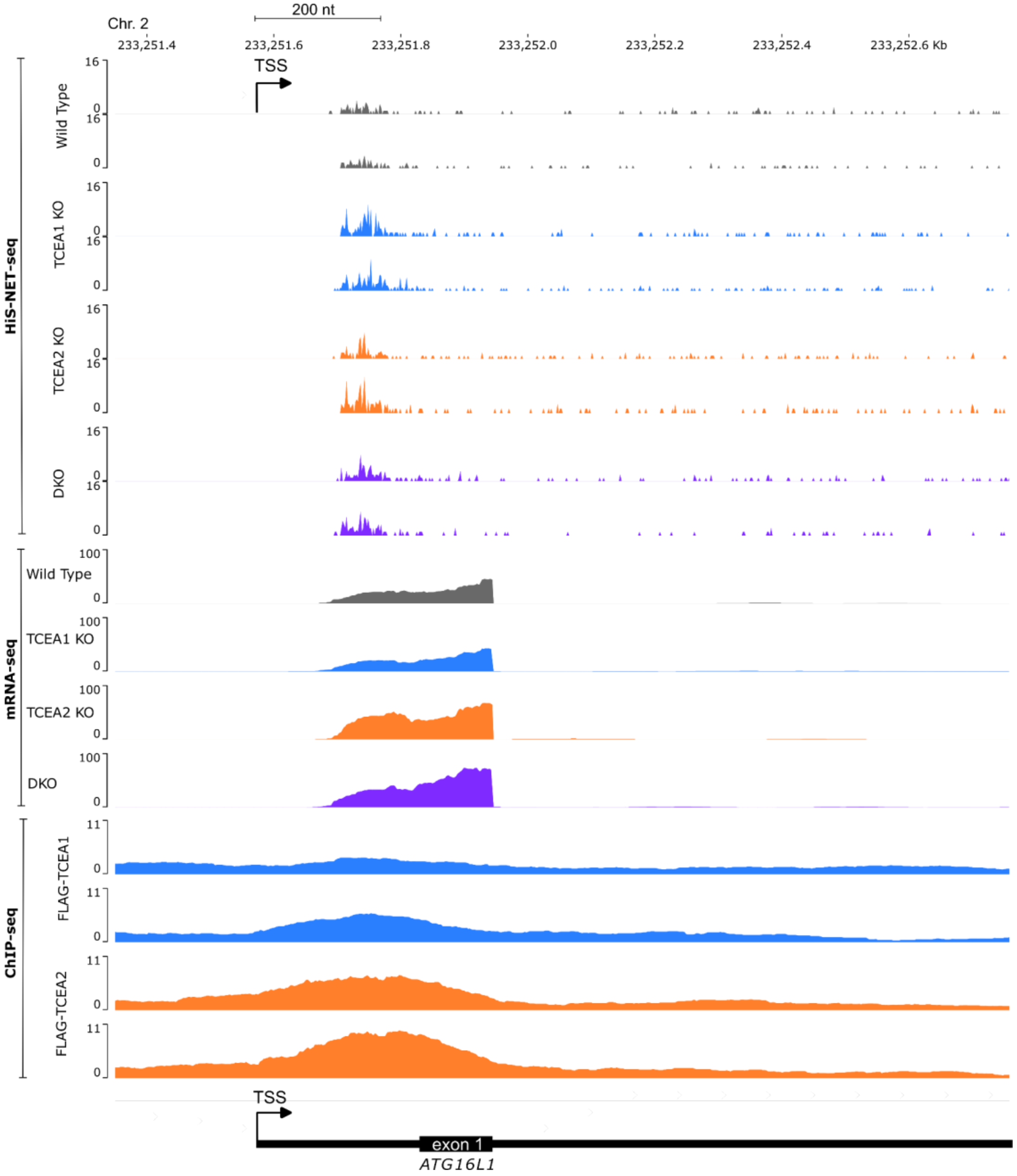


Figure 35. Representative example of a gene from quadrant IV.
HiS-NET-seq profiles at *ATF16L1* gene in unmodified HEK293T cells (Wild Type, gray), TCEA1 KO (blue), TCEA2 KO (orange), and the double TCEA1/TCEA2 KO (DKO, purple), Spike-in (SI)-normalized. Two biological replicates are shown. mRNA-seq profiles, SI-normalized, are shown as an average of three biological replicates. FLAG-TCEA1 and FLAG-TCEA2 ChIP-seq in the inducible expression cellular system are included here to show that *ATF16L1* is bound by TCEA1 and TCEA2 in close-to-physiological conditions. TSS: transcription start site.

In summary, the HiS-NET-seq experiment revealed that TCEA deletions lead to a global accumulation of Pol II at the promoter-proximal region. The pausing matrix demonstrated that there are two major transcriptional scenarios: 1) an accumulation of Pol II at the promoter-proximal region followed by a decrease in Pol II density at the gene body and 2) an overall increase in Pol II occupancy at the promoter-proximal region and gene body.

Chapter 4

DISCUSSION

4.1 TFIIS diversification into paralogs

In the late 1990s, a few laboratories identified a family of TFIIS genes in mouse, frog, and human cells and provided the first clues about their expression patterns and conservation among vertebrates [280, 284]. Taking advantage of the current whole genome databases, we revisited the question of TFIIS diversification. In general, our search confirmed that TCEA paralogs emerged in a common ancestor of jawless and jawed vertebrates, however, possibly, as not a complete set of four.

Although in model plant organisms, we detected only one TFIIS-like gene (*TFS*), we cannot rule out the possibility of multiple paralogs in some clades because whole genome duplications (WGDs) are known to occur more frequently in plants than in vertebrates [407]. Additionally, not a lot is known about the evolution of fungi, protists, and archaea. Our search identified only one TFIIS-like gene in a few model organisms, however, a few archaea species were recently reported to have up to four TFS paralogs [346] and six TFIIS paralogs were found in a protist *Paramecium tetraurelia* [408], therefore it is plausible that TFIIS has its own duplication history within some phyla.

Having focused our research on the animal kingdom, we found that invertebrate animals have one TFIIS gene and vertebrates have more than one TFIIS gene. Based on coding DNA sequence similarity, we determined that two ancient extant jawless fishes, hagfish and lamprey, have two and three paralogous TFIIS genes, respectively. This observation suggests that the TFIIS gene was duplicated more than 400 million years ago, at a perplexing time point for evolutionary biologists and paleontologists. After many years of debates, thanks to the recent whole-genome studies [345], the “1R-2R hypothesis,” proposing that two consecutive WGDs occurred in early vertebrate lineage more than 450 million years ago, is now firmly established. Possibly, the first or second round of WGD gave rise to the paralogous genes in cyclostomes (lamprey and hagfish). However, the paralogs could have also emerged due to small-scale duplications.

The phylogenetic relationship between hagfish and lamprey is still a subject of debate. Our cDNA-based screening approach revealed a different number of TFIIS paralogs in each. In the lamprey genome, we detected three paralogs, *TCEA2*, *TCEA3*, and *TCEANC*, while in hagfish only two, *TCEA2* and *TCEANC,* were present. The absence of *TCEA3* in the hagfish genome requires verification because its genome assembly is presently at a scaffold level. Additionally, a gene loss event can also explain the absence of *TCEA3*.

Surprisingly, although *TCEA1* is the prevalently expressed and most studied paralog in mammalian cells, we did not detect it in the two available lamprey species and one hagfish species. The absence of *TCEA1* in the current genome assemblies of hagfish and lamprey is puzzling and can be explained by a few scenarios: 1) the genome assemblies for both lamprey and hagfish are incomplete and currently presented as scaffolds, not chromosomes, so it is possible that many genes have not been identified and mapped yet; 2) *TCEA1* could have been present in the common ancestor of jawless and jawed vertebrates, but was lost in a gene loss event; 3) in the case of the sea lamprey, *Petromyzon marinus,* the paralogs can potentially be deleted during the regulated massive genome deletions in somatic cells that occur during embryonic development [409], and 4) *TCEA1* appeared in the lineage of jawed vertebrates only, and *TCEA2* is the "older" paralog. Although we cannot be certain about the point in evolution when *TCEA1* has emerged, we found that *TCEA2* and *TCEANC* are ancient types of TFIIS, and their preservation throughout evolution suggests that their function may be as pivotal as that of the most studied paralog, TCEA1.

4.2 TCEA1 and TCEA2 physically associate with multiple RNA-processing factors via their N-terminal domains

TCEA1 and TCEA2 are structurally similar proteins, especially in their C-terminal and central domains. The N-terminal part is conserved to a lesser extent and has two distinct regions: a conserved and structured domain and an unstructured unconserved linker (Figure 8 and S2). Detection of unique interacting partners by immunoprecipitation followed by mass spectrometry (IP-MS) of individual paralogs was limited because of the lack of specific antibodies, optimal for IP. We overcame this limitation by creating HEK293-derivative cell lines with inducible epitope-tagged versions of TCEA1 or TCEA2. In our experimental design, we aimed to gain insight into the nuclear environment of TCEA1 and TCEA2 and find potential unique interaction partners.

Structural studies have resolved TFIIS binding to Pol II with high resolution [262-265]. As expected, our IP-MS experiments revealed interactions with Pol II subunits and elongation complex factors, including PAF1, SPT4, and SPT6, but with low detection efficiency. Unexpectedly, NELF-E was also detected, although weakly. The NELF complex, specifically NELF-A-NELF-C module, was shown to preclude TFIIS binding to Pol II [95]. NELF-A and NELF-C were not detected in our IP-MS. NELF-E is not required for stabilization of paused Pol II, but it can bind RNA hairpin structures [410]. The detection of NELF-E in our TCEA1/TCEA2 IP-MS can be explained by preserved RNA-mediated interactions.

Unexpectedly, the majority of the detected interactions were with RNA splicing factors and splicing regulating heterogeneous nuclear ribonucleoproteins (hnRNPs), with high statistical significance. This finding indicates that TCEA1/2 have an affinity to multiple RNA processing factors and potentially provides information about how TCEA1 and TCEA2 behave in the nucleus when they are not actively rescuing Pol II. In our system, TCEA1 or TCEA2 are in excess and, imaginably, we detect two populations: TCEA1/2 bound to elongating Pol II and TCEA1/2 not engaged in the rescue but bound to some RNA processing factors. The latter is likely in a higher proportion because of the observed higher stoichiometric ratio compared to the elongation complex (EC) components. Another possibility is that the low detection of interactions with the EC reflects the transient nature of TCEA1/TCEA2 binding to backtracked Pol II. Additionally, some splicing factors can directly interact with Pol II and these interactions may also be present in our data.

Intriguingly, some of the detected splicing factors, including SRSF1, SRSF3, SRRM2, are known to localize to nuclear speckles (reviewed in [411]), suggesting that TCEA1 and TCEA2 may also associate with the splicing granules. Interestingly, the interactions were also detected with even higher enrichment in the experiment with the overexpression of truncated TCEA1/2 proteins, consisting of only the N-terminal domain and flexible linker (NTDL). This suggests a possibility that TCEA1/TCEA2 may physically associate with nuclear speckles, while not being actively bound to the elongation complex, because no EC components were detected in the NTDL IP-MS.

We found that the interactomes of TCEA1 and TCEA2 were nearly identical suggesting that both paralogs have a similar affinity to the same RNA processing factors in our cellular system. Narrowing down our point of view to the most dissimilar parts, the NTDL, helped identify the potential unique interactors of each paralog.

4.2.1 Putative unique interactors of TCEA1 and TCEA2

RNA binding proteins, RBM14 and RBMX, were detected in only TCEA1-NTDL IP-MS, whereas an RNA DEAD-box helicase DDX3X was found exclusively in TCEA2-NTDL IP-MS. RBM14 is an intrinsically disordered ribonucleoprotein and plays a role in DNA repair [354, 412] and alternative splicing regulation of some genes [413, 414]. RBMX, also known as hnRNPG, is thought to promote alternative splicing by binding an m^{6}A (N^6-methyladenosine) - modified nascent RNA and phosphorylated Pol II, subsequently facilitating the recruitment of splicing factors [415, 416]. DDX3X is an RNA helicase, which has a wide array of biological functions: promoting transcription, splicing, RNA export, and translation [417].

Although the detected putative unique interactors are interesting and may further strengthen the link between elongating Pol II and co-transcriptional splicing, more experiments are required to verify these observations. Another limitation, besides the potential secondary effects of protein overexpression, is the inability to express the paralogs at exactly the same level under similar conditions. The variability in bait protein expression affects the IP efficiency. Therefore, we cannot be certain about the exclusivity of a given unique interactor because it could have also been detected if the abundance of the other bait protein had been higher. Performing this experiment with SILAC labeling will enable a better quantification method. Additionally, reverse IP-MS, co-IP-Western blot, and TCEA1/TCEA2 affinity chromatography analysis will help confirm the interactions. Furthermore, in our crosslinking-assisted capture without RNase treatment, we may also be detecting potentially unspecific RNA-mediated interactions, performing a native IP-MS and RNase treatment can potentially improve the detection sensitivity of direct interactors.

4.2.2 N-terminal domains of TCEA1 and TCEA2 may mediate interactions with multiple RNA-processing factors

Our TCEA1/2-NTDL IP-MS experiments confirmed that the full-length paralogs interact with multiple RNA-processing factors and revealed that the interactions are mediated by the N-terminal domain. Neither Pol II subunits nor transcription elongation factors were detected as interactors of this domain, suggesting that it may not be necessary for the elongation function in human cells as was earlier demonstrated in yeast [250]. The exact role of the N-terminal domain of TFIIS in transcription in human cells is still unclear, however, a few non-contradictory functions were brought forward over the years. The NTD of TFIIS was proposed to play a role in transcription initiation, to serve as an interaction platform for intrinsically disordered nuclear proteins, and to be necessary for nuclear import of TFIIS. Our discovery of interactions with RNA processing factors provides additional insight into the nuclear environment of TCEA1 and TCEA2 in human cells.

A few studies demonstrated that the N-terminal part of TFIIS may function in early transcription events, independently from the RNA cleavage-stimulating C-terminal domain. A study, based on affinity chromatography, found that TFIIS can interact with Pol II holoenzyme that is transcription-competent Pol II bound by initiation factors, including TFIIB, TFIID, TFIIE, and TFIIF. Interestingly, the authors also showed that the N-terminal part of TFIIS is required for the interaction with Pol II holoenzyme, while the C-terminal part of TFIIS binds free Pol II [255]. In contrast, in our cellular system, Pol II subunits were not detected as interactors of NTDL. Another study showed that the NTD, together with the central domain and the linker, but not the C-terminal domain of TFIIS, was necessary for the efficient formation of the pre-initiation complex (PIC) in vitro [256]. Additionally, TFIIS involvement in transcription initiation was observed in yeast: a Mediator complex component, Med13, and a SAGA complex component, Spt8, were co-purified with TFIIS and the interactions were proposed to occur via the NTD [257]. In our whole TCEA1/TCEA2 IP-MS and NTDL IP-MS, we did not detect transcription initiation factors and the Mediator components, even below the statistical threshold. Possibly, TCEA1/2 are not involved in transcription initiation in human cells, however, the measurement may have been not sensitive enough, and, therefore, interactions with initiation factors cannot be completely ruled out. Another explanation is that, in our system, we overexpress not only the structured part of the domain, but also the unstructured linker and, maybe, the affinity of the RNA processing factors to this linker is much stronger and they may physically hinder binding to Pol II or the PIC. Whether TFIIS is a regulator of initiation in human cells is still an intriguing question and is an opportunity for new analyses with global approaches for profiling transcription initiation in NTD mutants.

Previous reports showed that the structured part of NTD of TFIIS shares a structural similarity with other nuclear proteins, including Elongin A [245, 246, 259], CRSP70/MED26 [260], PPP1R10, PIBP, IWS1 [246, 261], LEDGF, and HRP2 [246]. Furthermore, the conserved NTD motif consisting of 5 α-helices recently has gained a spotlight as it was proposed to serve as a protein interaction platform for selective binding of transcription regulators with intrinsically disordered regions [246]. The disruption of such interaction was demonstrated with IWS1 and SPT6, which resulted in a substantial elongation defect and led to a hypothesis that such structural motifs may function in the spatial arrangement of factors in the nucleus, thereby regulating transcription elongation [246]. The role of NTD of TFIIS in the context of this hypothesis has not been investigated yet and presents a compelling question. Additionally, investigating whether the interactions occur between the IDRs of the detected RNA-processing factors and the disordered linker of TCEA1/2 will be insightful.

The NTD of TFIIS was also demonstrated to be necessary for nuclear import. TFIIS lacks a classical nuclear localization signal and hypothetically gets imported into the nucleus thanks to its ability to interact with proteins targeted to the nucleus [261]. Potentially, the nuclear import of TCEA1/2 is facilitated by the detected RNA processing factors. In our experimental design, we focus on the nuclear fraction; it would be interesting to test whether the binding also occurs in the cytoplasm with immunofluorescence microscopy.

Having discovered highly abundant interactions with RNA processing factors, we wanted to understand whether TCEA1/2 are involved in the recruitment of the splicing factors and other splicing regulators, hnRNPs. We hypothesized that if TCEA1/2 recruit some RNA processing factors to the nascent RNA, we would observe dissociation of these factors from chromatin in our quantitative chromatin MS performed in the knockout cell lines, TCEA1 KO, TCEA2 KO, and the double knockout (DKO). Overall, we found marginal changes in chromatin localization of some RNA processing factors, with the most effect observed in the DKO. Since we did not observe any factor consistently change across all knockouts, we propose that TCEA1/2, likely, do not act as the direct recruiters for the RNA processing factors.

Altogether, our IP-MS experiments indicated that TCEA1 has a putative unique affinity to RBM14 and RBMX, while TCEA2 appears to associate with DDX3X uniquely. We found that both TCEA1 and TCEA2 associate with multiple RNA processing factors in our inducible overexpression system and a hypothetical explanation of our results is shown in the figure below. Further analyses, including native IP-MS, RNase treatment, reverse IP-MS, and IP-MS in SILAC-labeled cells, can elucidate the interaction partners of TCEA1 and TCEA2 with more precision.

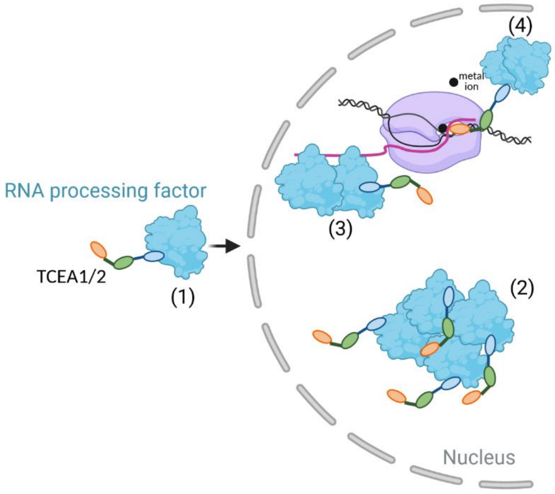


Figure 36: **A schematic explanation of the detection of interactions of TCEA1 and TCEA2 in the inducible overexpression system.** TCEA1/2, via their N-terminal domains (blue oval), bind RNA processing factors (1) for nuclear import; (2) when TCEA1/2 are not actively involved in the rescue of arrested Pol II; (3) because the RNA-mediated interactions may be preserved in our experiment, an RNA processing factor can be bound by TCEA1/2 directly and/or it is bound by another splicing processing factor in the vicinity of EC with TCEA1/2, and (4) at the same time as TCEA1/2 are relieving arrested Pol II.

4.3 TCEA1 and TCEA2 associate with Pol II predominantly during early elongation, at the promoter-proximal region

Previous studies in mammalian cells revealed that TCEA1 binding occurs predominantly at the 5' pause sites of genes [154, 357], which suggests that these regions are most prone to backtracking, however, it is still debatable whether TCEA1 directly functions in Pol II release from promoter-proximal pause into productive elongation. TCEA2 activity in transcription elongation is unknown. The inducible epitope-tagged TCEA1/2 overexpression system enabled us to profile the genomic binding pattern of TCEA1 and TCEA2, however with a limitation: the level of induced expression of TCEA2 is twice as high compared to TCEA1. This affects ChIP efficiency and explains why more binding sites are found in the TCEA2 ChIP-seq. However, a large overlap between TCEA1- and TCEA2- occupied sites suggests that the paralogs can compete for binding under the overexpression conditions. This observation also suggests that TCEA1 and TCEA2 are functionally redundant in Pol II rescue. In the future, if the antibodies that can discriminate between endogenous TCEA1 and TCEA2, suitable for chromatin immunoprecipitation, become available, ChIP-seq or CUT&RUN, the latter rendering a higher resolution, with spike-in normalization will elucidate the preferential binding sites, specific to each paralog.

We determined that the majority of TCEA1/2 binding occurs at protein-coding genes and enhancers, which is expected because these genomic regions are transcribed by Pol II. TCEA1/2 binding was also detected at Pol III-transcribed genes. Previously, TFIIS was also shown to play a role in Pol III transcription, which is somewhat unexpected because Pol III has a subunit, Rpc11, with intrinsic transcript cleavage activity [357, 358]. However, TFIIS was proposed to play a role in Pol III transcription start site selection rather than stimulating RNA cleavage. It would be of interest to further investigate the impact of TCEA1 and TCEA2 deletions on Pol III transcription.

The binding profiles of TCEA1 and TCEA2 are similar to the pausing pattern of Pol II: the paralogs are predominantly found at the promoter-proximal regions (within the first 300 base pairs downstream of TSS), less in the gene body, and the occupancy somewhat increases in the termination zones. Similar observations were made with ChIP-seq of overexpressed TCEA1 in human breast cancer cells [154]. Interestingly, TCEA1/2 also bind Pol II engaged in antisense transcription, pointing to their affinity to any Pol II engaged in the early phase of elongation.

What is still unclear is whether TCEA1/2 are recruited to Pol II at the point of the observed enrichment in the promoter-proximal region and remain associated with it throughout the transcription cycle. Hypothetically, there are three scenarios that are not mutually exclusive: 1) TCEA1/2 are recruited to initiating Pol II and travel with it from the transcription start sites until the end of the transcription cycle, therefore the observed landscape reflects all pausing events, including transient pause, backtracking and arrest, of TCEA1/2-bound Pol II, 2) the promoter-proximal regions and the termination zones are especially prone to backtracking and arrest and TCEA1/2 get recruited there only when Pol II is locked in an inactive conformation, and 3) TCEA1/2 are involved during or immediately after initiation, but then they dissociate and bind Pol II only when it is in the arrested state at the gene body.

The first scenario is supported by the evidence that TFIIS plays a putative role in transcription initiation [255-257, 418, 419] and the observed peak immediately upstream of the TSS, corresponding to TCEA1/2-bound Pol II transcribing in the antisense direction in our data. Additionally, a previous study found that, upon arrest-inducing NTP depletion treatment in yeast, TFIIS recruitment was found to only slightly increase, suggesting that most TFIIS is recruited regardless of the severe transcriptional stress [358]. Possibly, TCEA1/2 are generally recruited during early elongation and additionally to the arrested Pol II more downstream.

Early transcription at the promoter-proximal region is regulated by a growing list of transcription factors. We aimed to understand where TCEA1/2 bind, relative to other classical elongation factors and the +1 nucleosome, which is an intrinsic barrier for elongating Pol II, where TFIIS is expected to alleviate backtracking (reviewed in [79]). By overlaying our ChIP-seq data with published ChIP-seq of other elongation factors in HEK293T cells, we found that the binding pattern of TCEA1 and TCEA2 is distinct from that of other elongation factors and that the paralogs may be required before Pol II encounters the +1 nucleosome. The TCEA1/TCEA2 enrichment in binding is slightly upstream of NELF and noticeably more upstream than SPT6 and PAF. NELF binding to Pol II is thought to be mutually exclusive to the binding of PAF and TFIIS based on structural studies [95, 100]. This is also observed in the ChIP-seq profiles of NELF and PAF (LEO1), while the temporal binding of TCEA1/2 and NELF is not clear, as TCEA1/2 appear to bind Pol II prior to NELF. A more thorough analysis is necessary to elucidate whether TCEA1/2 are displaced from Pol II by NELF and then can re-bind Pol II when NELF dissociates. For example, a gene-by-gene comparison of NELF and TCEA1/2 binding intensity and binding site determination with higher resolution can reveal the relationship between the elongation factors.

4.4 TCEA1 and TCEA2 loss leads to a defect in cell growth and induces DNA damage response

Previously, it was demonstrated that a complete loss of TCEA1 in mouse embryos was detrimental to the growth of hematopoietic cells in fetal liver and resulted in embryonic death [290]. Additionally, TCEA1 knockdown by RNA interference reduced the cell growth and proliferation of human breast, lung, and pancreatic cancer cells [293]. Therefore, the growth defect of the TCEA1 KO was not entirely surprising to us. However, the reduction of growth of TCEA1 KO cells was rather mild in comparison to the TCEA2 KO cells, which was unexpected because TCEA2 is more lowly expressed than TCEA1 in HEK293T cells. The double TCEA1/2 knockout (DKO) cells were viable and the DKO growth rate was as slow as that of the TCEA2 KO. These observations suggested that the presence of TCEA1 alone must be insufficient for proper transcription regulation and that TCEA2 may have an independent role in regulating cell cycle, which prompted a more thorough cell cycle analysis.

We determined that the slow growth was not accompanied by the exposure of phosphatidylserine residues on the cell surface (Figure 23), which is a signature of apoptotic cells [420-423]. We also did not detect changes in pro-caspase 3 level and any cleaved caspase 3 (Supplemental figure S4). HEK293T cells were immortalized via transformation with adenovirus 5 [424, 425] to insert the genes encoding E1A/E1B that affect the cell cycle control pathways and inhibit apoptosis [426, 427]. Additionally, HEK293T cells express Simian virus 40 large tumor antigen (SV40T), which inhibits P53 [428, 429]. This can explain why there is no prominent apoptosis. Another possibility is that the knockout cells experience irreversible cell cycle arrest. A possible evidence of senescence is the increased frequency of abnormally large cells with merged multiple nuclei [430, 431], most apparent in the DKO culture (Figure 22). Additionally, the gene ontology term “G1 to G0 transition” is significantly enriched in the analysis of upregulated genes in the DKO. It will be of interest to assay for senescence markers, such as the presence of β-galactosidase activity and p21.

4.4.1 The slow growth phenotype is likely caused by replication stress

Cell synchronization revealed that the key difference in growth among the cell lines is the duration of the DNA replication phase, it is distinctly longer in the TCEA2 KO and the DKO than in the TCEA1 KO (Figure 25). There are two potential causes of this phenotype: 1) the cells are not readily granted to pass the G1 to S checkpoint and 2) the DNA replication itself is taking longer. DNA damage is a likely cause of both delays, however, in our knockout cell lines, the mRNA level of many genes is altered, therefore the effect on the factors involved in DNA damage repair and cell cycle regulation is probable and can contribute to the complexity of the phenotype. In our attempt to better understand the extent of the TCEA1/2 loss on well-known DNA repair factors, we showed that many factors are indeed deregulated at the mRNA and protein levels, latter as the relative abundance at chromatin (Figure 28). It is of note that the DNA damage response cascades are rather complex and fine-tuned by an interplay of numerous stabilizing and activating post-translational modifications (PTMs). A least the classical pathways should be investigated systematically at the levels of nascent RNA, steady state RNA, protein abundance, and activated protein isoforms to be able to elucidate the causes of the slowed cell cycle progression of the knockouts.

4.4.2 R-loops as the cause of DNA damage in the TCEA knockouts

Recently, TFIIS was shown to play a role in genome stability by preventing R-loop accumulation [308]. Zatreanu et al. inhibited Pol II cleavage activity by overexpression of inactivated mutant TFIIS and found that the cells grew slower and had increased levels of DNA damage markers. The cause of that phenotype was attributed to the transcriptional stress that provoked R-loop accumulation: mutant TFIIS (TCEA1) induced Pol II backtracking and arrest and led to the formation of R-loops, anterior to Pol II, provoking the breaks in the DNA strand. We were curious whether TCEA1 and TCEA2 loss also induced aberrant R-loop formation, and, surprisingly, we did not detect a striking increase in the R-loop levels, but they were only mildly elevated in all knockouts (Figure 27). Importantly, the specificity of the only available antibody against DNA-RNA hybrids (S9.6) is questionable as we detected its binding to other nucleic acids besides DNA-RNA hybrids. Therefore, such subtle differences in R-loop levels in our experiment should be interpreted with caution. We hypothesize that the lack of a clear increase in R-loops may be due to the differences in the cellular systems. In the knockouts, the intrinsic nuclease property of Pol II is not affected and, possibly, it is sufficient to cleave the RNA in some cases, otherwise the DKO cells would have had a genome-wide shutdown of nascent RNA synthesis and would not be viable. In contrast, the mutant TFIIS binds exactly like the endogenous TFIIS and, imaginably, disables any translocation and prolonging the backtracking and arrest states.

The slight elevation of R-loops in the TCEA knockouts can potentially contribute to the DNA damage and trigger the DNA damage response. Additional evidence of the transcription stress in the knockouts is a consistent recruitment of a few DNA repair factors (ELOF1, TRIP12, USP7, XAB2) (Figure 28), involved in transcription-coupled nucleotide excision repair (TC-NER), the DNA damage repair pathway which is activated specifically during active transcription, when Pol II is stalled at DNA lesions [373]. It will be of interest to investigate TC-NER in the TCEA knockouts with additional methods.

4.4.3 Tumor suppressor P53 activation in TCEA2-absent cells

Apoptosis, transient cell cycle arrest at the G1/S point, and senescence can all be induced by the tumor suppressor P53 (reviewed in [432]). We detected a clear increase in P53 phosphorylated at serine 15 (P53-Ser15p) in the TCEA2 knockout and the DKO cells. In unstressed cells, P53 is maintained at a low level by ubiquitination ligase MDM2, which monoubiquitinates and transports it from the nucleus for polyubiquitination and proteasomal degradation. P53 is stabilized by many PTMs in response to various cellular stresses, such as DNA damage, hypoxia, and replicative senescence (reviewed in [371]). P53-Ser15p is generally thought to indicate an ongoing DNA damage response because this phosphorylation is deposited by the kinases, such as ATM and ATR, that get rapidly activated upon recognition of DNA double strand breaks and other types of DNA damage, respectively [371]). We tested a classical marker for DNA damage, a phosphorylation of histone H2A (H2A.X), and found that it is only mildly increased in the knockouts, however with some variability among single cell clones (Figure 26). The immunoblotting of ATM and ATR was not conclusive because the changes in the ATR phosphoisoform are very subtle and ATM immunoblotting was not successful. More thorough investigation of the P53 pathway is needed to elucidate why it is constitutively phosphorylated. For example, the levels of other activated kinases, such as DNA-PK, SMG1, mTOR, Chk1 and destabilizing proteins MDM2 and MDMX should be evaluated. Furthermore, P53-Ser15p is not essential for the stabilization, but it is a precursor for a series of further phosphorylations [371], therefore other P53 phosphoisoforms should be investigated to determine whether the tumor suppressor is fully functional. Next, the transcriptional targets of P53 should also be examined. In general, this analysis is rather convoluted because 1) the P53 target genes can also be affected at the transcriptional level by the loss of TCEA1 and TCEA2 and 2) activation of P53 target genes in HEK293T cells can be impaired because SV40T was shown to bind the DNA-binding surface of P53, potentially interfering with its transactivation function [428, 429, 433, 434]. P53 ChIP-seq in the knockouts vs. unmodified HEK293T will be informative of its activity and has the potential to determine the cell type-specific P53 target genes, in addition to the expected universal ones.

Because P53-Ser15p is increased in TCEA2-absent cells and not in the TCEA1 KO, there is an intriguing possibility that TCEA2 is a specific transcriptional regulator of one or more factors mediating P53 activation. This is the case for TCEA3, which was shown to regulate transcription of specific target genes in various cell types, leading to both growth and differentiation promoting and suppressive effects depending on the cell state (described in 1.3.3). Integrative analysis of our RNA-seq and HiS-NET-seq data has the potential to provide a clue about the target genes of TCEA2, which, expectedly, would be the ones with a reduction in nascent RNA and mRNA in the TCEA2 KO and DKO, but not in the TCEA1 KO. The relevance of TCEA2 to cellular growth should be further investigated in other cell lines. Our discovery of TCEA2 as a potential regulator of the P53 pathway will hopefully prompt future research in the context of cancer.

4.5 TCEA1 and TCEA2 regulate early transcription elongation

HiS-NET-seq profiled the occupancy of transcriptionally engaged Pol II and revealed a clear genome-wide increase in Pol II occupancy within the first 200 nucleotides downstream of the transcription start site of thousands of actively transcribed genes (Figure 34). The promoter-proximal region harbors a key regulatory step, the promoter-proximal pause and release, during which Pol II is prepared and committed to productive elongation (described in section 1.1.2). This regulatory checkpoint is controlled by a growing list of elongation factors. Our finding confirms that TCEA1 and TCEA2 are also involved at that genomic site and alleviate transcription through the +1 nucleosome and, possibly, also more upstream of it, as our ChIP-seq metagene plot overlaid with nucleosome positioning suggests (Figure 20). More profound analysis is needed to clarify whether TCEA1 and TCEA2 are involved only during the transcription of the nucleosome-bound DNA or whether they also stimulate the release from the “scheduled” NELF-mediated pause.

Impaired backtracking relief is predicted to increase Pol II HiS-NET-seq signal as, imaginably, polymerases would be “stuck” for a longer time point and accumulate there, provided that transcription initiation was not impaired. In another TFIIS study, the expression of mutant TFIIS caused a strong reduction in global run-on sequencing (GRO-seq) signal, indicating a decrease in actively elongating polymerases and, in contrast to our observation, did not lead to an increase in Pol II occupancy, as profiled by Pol II ChIP-seq and mNET-seq, supposedly, due to premature termination of backtracked polymerases [154]. It is not clear why our observations differ in terms of Pol II occupancy, but there are two possible explanations: 1) as an adaptive reaction of the knockouts, many genes got upregulated and the activated transcription is reflected in the metagene plot as an increase in signal at the promoter-proximal regions, at least to some extent, and 2) mechanistic differences of the depletion and inhibition systems may impact Pol II turnover differently, such that upon inhibition of RNA cleavage, polymerases are completely immobilized, then ubiquitinated and removed, while in the knockouts, the intrinsic nuclease activity may be sufficient to slowly work thorough the pause-predisposing site, so most of polymerases are not removed as readily. Furthermore, there is a fundamental difference in the nascent transcription technologies, GRO-seq and HiS-NET-seq, that explains why we observed an increase in signal at the 5’ ends of genes and Sheridan et al. found the opposite, although both methods enrich labelled nascent RNAs and precisely discern the location of Pol II active site. GRO-seq informs about the elongation-competent, RNA synthesizing polymerases, omitting backtracked polymerases as they cannot be restarted in the run-on reaction (reviewed in [435, 436]). Whereas, HiS-NET-seq profiles all 4sU- labeled chromatin-associated Pol II, in both actively transcribing, paused, and backtracked states, as long as at least one nucleotide has been incorporated. The metabolic labeling in HiS-NET-seq is implemented to enhance the sensitivity of nascent RNA sequencing of lowly transcribed regions, but the kinetics of nascent RNA synthesis are not measured.

In all TCEA knockouts, we find two major transcriptional scenarios 1) an increase in Pol II occupancy at the promoter-proximal regions, followed by a reduced occupancy at the gene body and 2) an overall increase at the promoter-proximal region and gene bodies (Figure 34). The first scenario potentially reflects an inefficient elongation through the promoter-proximal pause sites and the +1 nucleosomes, resulting in fewer polymerases entering the gene body. Additionally, the occupancy decrease in the gene body can be observed in the case of defected Pol II processivity: Pol II does not complete the full transcript because it is arrested and removed earlier in the knockouts. The second scenario likely reflects the total upregulation of gene expression as adaptive effects, unavoidable in knockout systems. However, possibly, there are also genes that are both activated only in the knockouts and have a lot of pause sites in the gene body. A more profound bioinformatic analysis is needed to elucidate the cases with promoter-proximal escape impairment and Pol II processivity defect. Additionally, an integrative analysis of HiS-NET-seq and mRNA-seq will be of interest to determine how the observed changes at the nascent RNA level affect the steady state mRNA. Lastly, it will be interesting to compare the affected genes in terms of the *cis*-regulatory elements at and in the vicinity of promoters, such as core promoter elements and known pause motifs.

Our next analysis entails investigation of the uniquely affected genes per knockout, as they have the potential to reveal the specific target genes of TCEA1 and TCEA2. Interestingly, in the TCEA2 KO and the DKO, we detect a population of genes that is generally downregulated at both the promoter-proximal region and gene body (Figure 34A, quadrant III). Some studies proposed that TFIIS is also involved in transcription initiation [255-257], therefore the genes that are severely downregulated at the nascent RNA level are the candidate target genes, whose initiation is controlled by TCEA2. Hypothetically, in HEK293T cells, some developmental gene is governed by a transcription factor that acts as a co-factor for TCEA2 with higher affinity than TCEA1.

Additionally, it will be of interest to follow up on the observations made by Sheridan et al. and Zatreanu et al. Inhibition of RNA cleavage activity was shown to reduce the elongation rate at gene bodies [154]. Furthermore, the processivity of longer than average genes was greatly reduced and the splicing of last exons was affected [308]. It will be interesting to determine the length of the genes and transcripts of each quadrant of the pausing matrix (Figure 34B). We used a similar alternative splicing analysis tool and found the splicing of the first exon as the most frequent alternative splicing effect (Supplemental figure S5), we will investigate the splicing patterns in more detail in the future. Lastly, Sheridan et al. observed a termination defect and proposed that slow elongation in the termination zone is advantageous for XRN2-mediated Pol II eviction, resulting in “earlier” Pol II removal [154]. Using HiS-NET-seq data, we will compare the terminating polymerases to better understand the potential role of TCEA1 and TCEA2 in transcription termination.

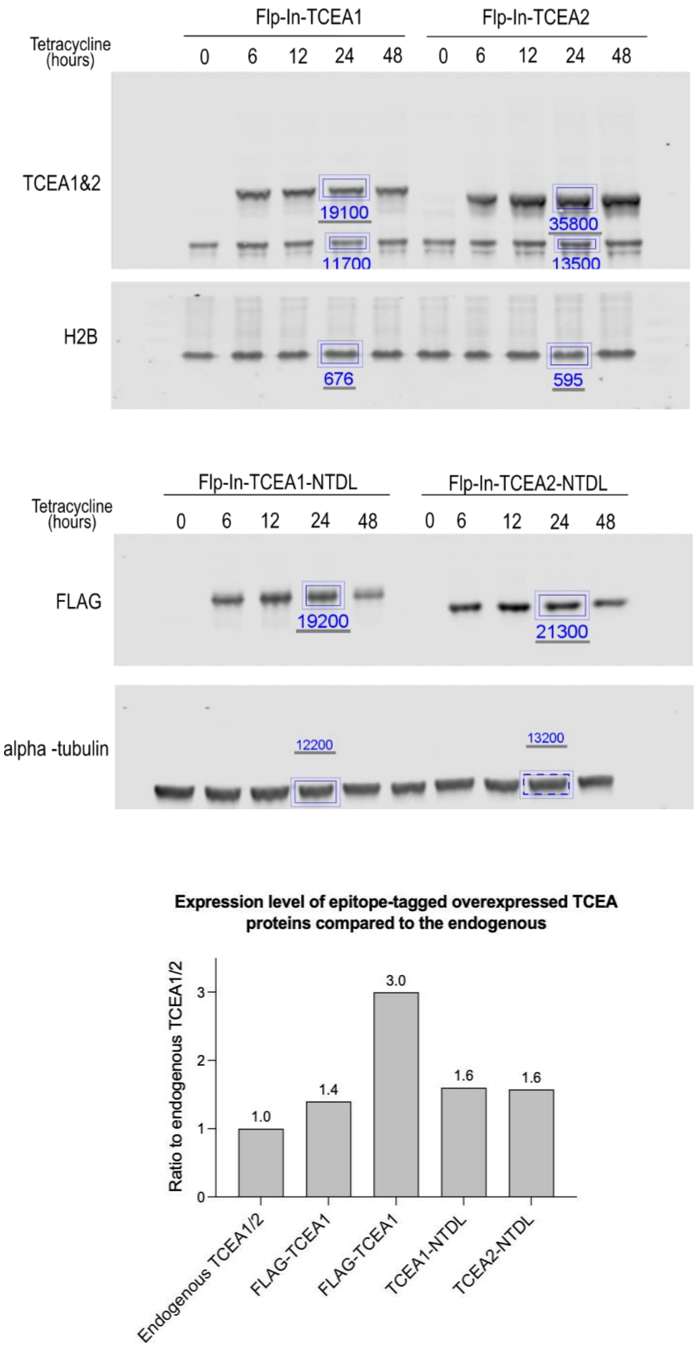
Using our HiS-NET-seq data, we will also be able to analyze the impact of the loss of TCEA1 and TCEA2 on enhancer transcription. Since we detect TCEA1 and TCEA2 binding at many enhancers (Figure 17), it will be of great interest to investigate whether increased backtracking at those regions has an effect on target genes and enhancer RNA levels.

4.6 Ongoing experiments

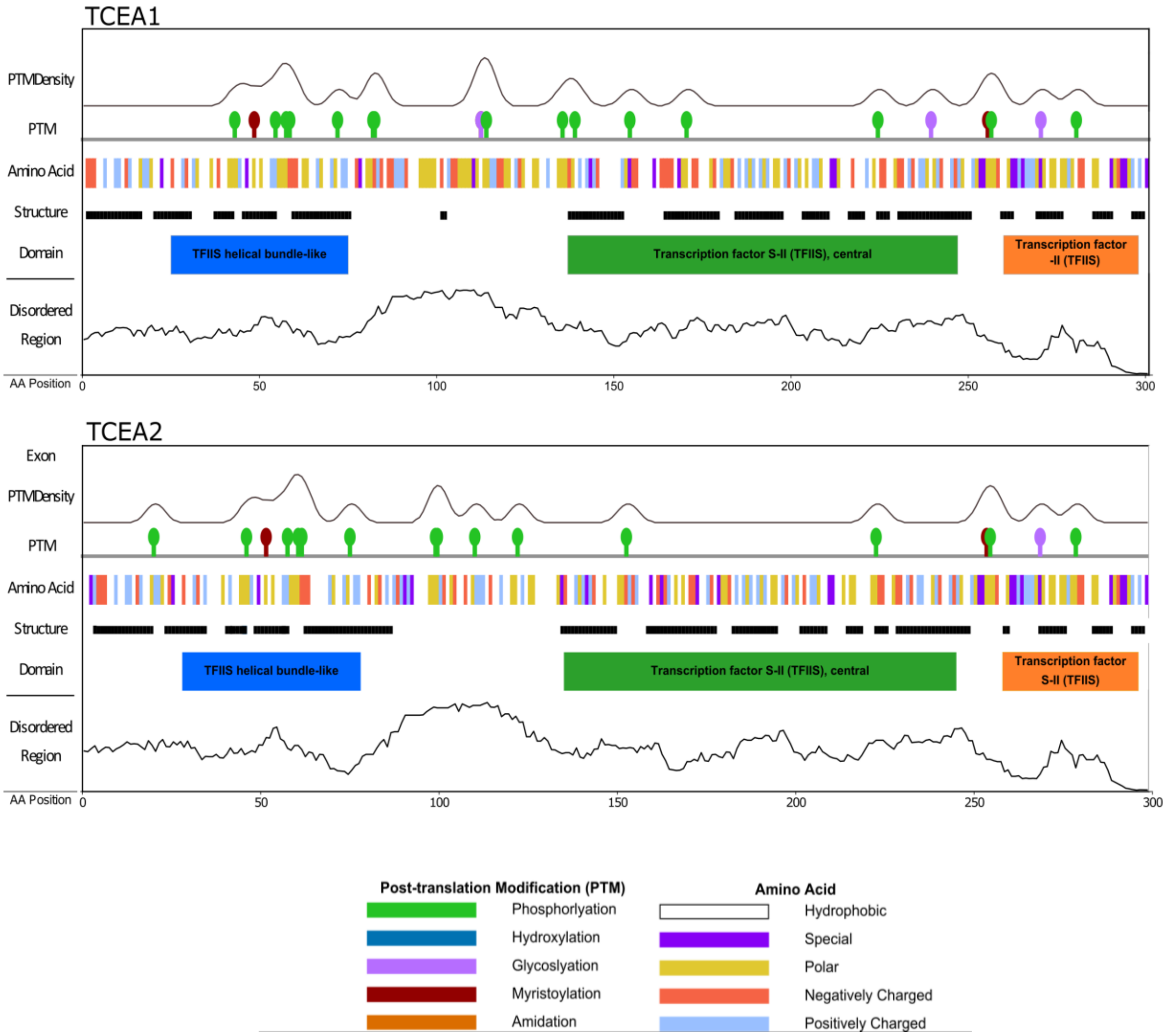
We have shown that TCEA2 loss leads to a reduction in cell proliferation in HEK293T cells, likely, via activation of the P53 pathway. In our ongoing work, we are investigating the oncogenic potential of TCEA2. Curiously, a previous study revealed that TCEA2 interacts with BRCA1 and that this interaction has a physiological effect on cell growth, furthermore, the *TCEA2* locus was reported as amplified in breast and ovarian tumors [297] (described in section 1.3.2). TCEA1 is also relevant in breast cancer, as its knockdown affected the growth of breast cancer cells more severely than noncancerous breast cells because of a differential impact on estrogenic, c-myc, and P53 pathways [293]. We have set out to characterize TCEA2 in the context of cancerous and noncancerous breast cells.

In addition to the role of TCEA2 in cancer cell proliferation, we are also interested in its potential role in neurons because this paralog is rather well expressed in cerebellum, as seen in the GTEx database. Previous studies have highlighted the roles of TCEA1 and TCEA3 in a few differentiation systems (described in sections 1.3.1 and 1.3.3), therefore, we are curious whether TCEA2 may also function in neurodifferentiation.

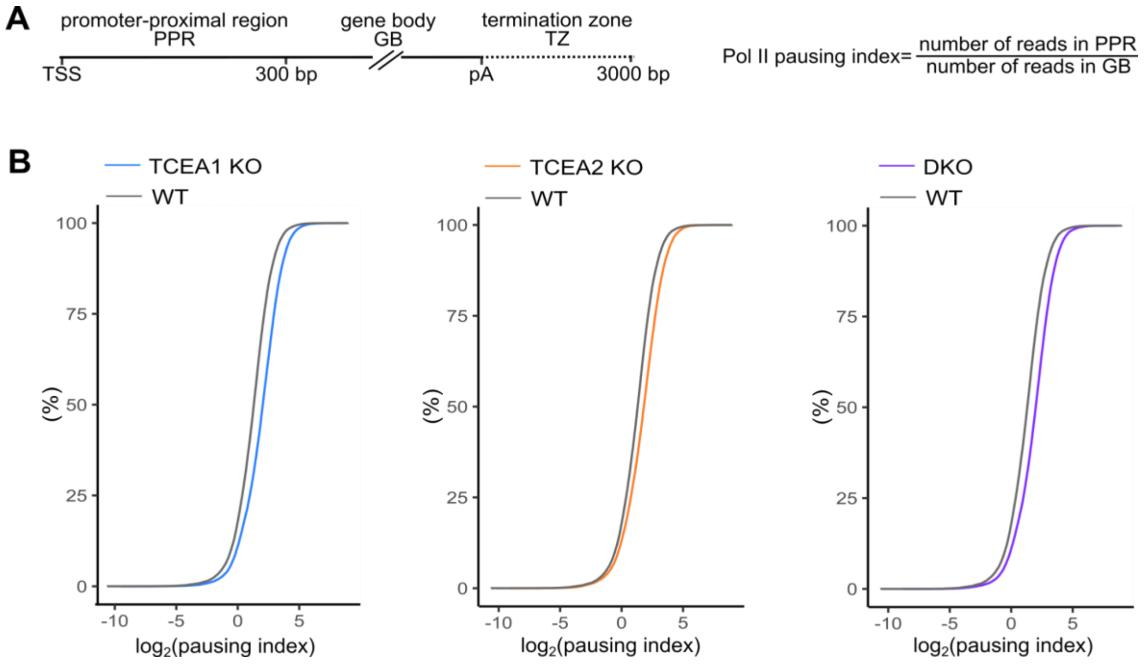
Supplemental data



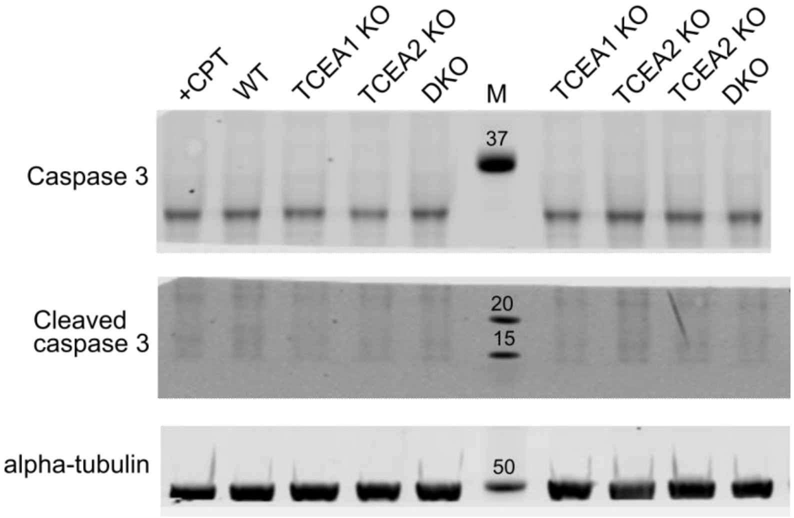
Supplemental figure S1: Comparison of Tetracycline-induced overexpression of epitope-tagged TCEA1, TCEA2, TCEA1-NTDL, and TCEA2-NTDL to endogenous TCEA1. This quantification is based on one biological replicate and refers to the 24-hour Tetracycline treatment (Figure 8). Signal was normalized to loading control either H2B or α-tubulin.



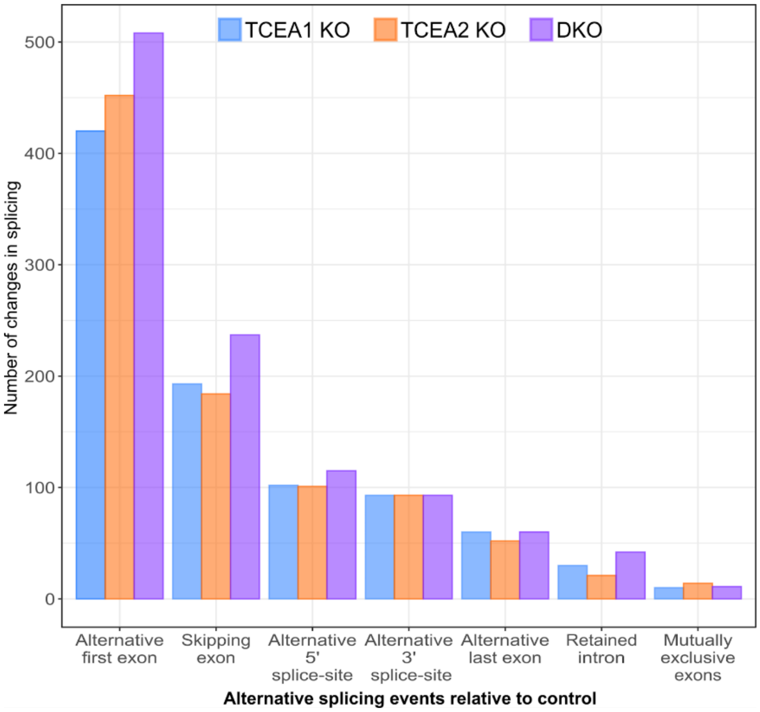
Supplemental figure S2: Domain architecture of TCEA1 and TCEA2, disordered regions, and predicted post-translational modifications, determined with IsoTV [437].



Supplemental figure S3. The pausing index is increased in the knockouts.
A. Schematic of gene regions and the pausing index formula
B. Pausing indices of all actively transcribed genes are displayed as cumulative distribution function.



Supplemental figure S4. No change in caspase 3 activity in the TCEA knockouts.
Western blot against caspase 3 and cleaved caspase 3 (activated) in the TCEA 1 and 2 knockout independent single cell clones, unmodified HEK293T (WT) and CPT-treated HEK293T cells. Alpha-tubulin is the loading control.



Supplemental figure S5. Changes in splicing patterns in the TCEA knockouts compared to control unmodified HEK293T cells, based on three biological replicates, determined by mRNA transcript quantification with SUPPA [438].

Supplemental Table 1: Significant interactors of FLAG-TCEA1, FLAG-TCEA2, FLAG-TCEA1-NTLD, and FLAG-TCEA2-NTDL

FLAG-TCEA1 IP-MS		
Gene name	**T-test Difference**	**-log (Student's T-test p-value)**
TCEA1	8.64	4.75
TRIM28	6.75	5.16
NPM1	5.89	4.07
HNRNPA1	5.65	2.88
HNRNPA2B1	4.63	2.75
SRSF3	4.05	3.30
HNRNPU	3.98	2.71
SRRM2	3.85	2.14
ALYREF	3.76	1.65
LSM3	3.59	1.65
SERBP1	3.50	2.33
SRSF1	3.21	2.97
KHSRP	3.10	2.78
SAFB	2.88	1.87
PPIA	2.82	1.57
HNRNPM	2.66	1.55
HNRNPK	2.55	1.82
EIF5A; EIF5AL1; EIF5A2	2.50	1.55
Mutlimapped histones	2.48	1.51
VIM	2.46	1.38
FUBP1	2.42	2.77
HIST1H1C	2.33	1.31
MATR3	2.33	1.64
MDC1	2.28	1.49
LMNB1	2.26	1.32
STMN1	2.20	1.62
RPS21	2.15	1.32
ACTG1	2.13	1.42
HNRNPC	2.03	1.58
ERH	1.19	4.31
FLAG-TCEA2 IP-MS		
Gene name	**T-test Difference**	**-log (Student's T-test p-value)**
TRIM28	6.50	4.96
TCEA2	5.94	3.39
HNRNPA1	5.80	2.95
NPM1	5.44	3.86
SERBP1	4.83	3.00
ALYREF	4.43	1.97
HNRNPA2B1	4.17	2.50
LSM3	4.11	1.91
HNRNPU	4.08	2.78
SRRM2	3.98	2.21
SRSF3	3.89	3.03
TCEA1	3.81	2.00

PABPN1	3.68	1.76
LMNB1	3.40	2.03
MYL6	3.08	1.36
SAFB	2.97	1.92
Mutlimapped histones	2.92	1.72
NPM3	2.82	1.73
STMN1	2.82	2.09
RPS28	2.79	2.53
Mutlimapped histones	2.75	1.40
LARP1	2.75	1.93
TPM3	2.72	1.56
SRSF1	2.70	1.55
HNRNPM	2.69	1.82
PRDX1	2.65	1.44
KHSRP	2.64	2.11
PPIA	2.62	1.62
NONO	2.55	2.18
HNRNPK	2.54	1.77
MDC1	2.51	1.64
EIF5A; EIF5AL1; EIF5A2	2.48	1.56
PRPF19	2.46	1.79
RBM8A	2.39	1.49
RPS21	2.39	1.43
ACTG1	2.38	1.60
HIST1H1C	2.32	1.37
RBBP4	2.22	2.03
MATR3	2.19	1.59
FUBP1	2.19	2.52
U2AF1; U2AF1L4	2.11	1.32
TUBA1A; TUBA1B; TUBA1C; TUBA3E	1.97	1.53
HNRNPC	1.97	1.52
ERH	1.08	4.32
FLAG-TCEA1-NTDL IP-MS		
Gene name	**T-test Difference**	**-log (Student's T-test p-value)**
TCEA1	8.29626	4.87918
SFPQ	5.21431	3.96591
EIF4A3	4.76099	5.34998
Mutlimapped histones	4.22852	1.6411
HNRNPM	4.06406	4.63706
SAFB	3.89699	4.86491
RPL6	3.65949	3.3209
HNRNPA2B1	3.41039	1.51521
HNRNPC	3.34101	2.15745
PTBP1	3.16544	2.32224
TRIM28	3.16139	4.14681
HNRNPK	3.03499	2.52263
RPS14	2.9801	2.59702
Mutlimapped histones	2.9219	3.19607

MATR3	2.77524	2.70822
RPL4	2.77401	4.83189
SRRM2	2.57448	2.46211
SF3B2	2.489	1.4279
U2AF1; U2AF1L4	2.39751	2.69875
HNRNPH1	2.31777	2.67842
HIST1H4A	2.25778	1.70969
MIF	2.24159	3.33885
NCL	2.18281	2.57023
RPS11	2.13096	1.47047
NONO	2.02003	4.24733
RPS27A; UBC; UBB; UBA52	1.98132	1.71927
RPS7	1.9679	2.19401
NPM1	1.93425	2.35108
RPS17	1.87765	2.2057
ALYREF	1.85845	2.07546
HNRNPF	1.31749	2.79858
RBM14	1.19188	1.65533
LARP1	1.11768	1.35282
LSM3	0.89414	1.31711
BCLAF1	0.865654	1.43587
LMNB1	0.684148	2.38577
FLAG-TCEA2-NTDL IP-MS		
Gene name	**T-test Difference**	**-log (Student's T-test p-value)**
TCEA2	5.71488	3.95183
SFPQ	4.64091	3.66864
EIF4A3	4.42625	3.39171
Mutlimapped histones	3.76525	1.4989
SAFB	3.08539	1.63841
RPL4	3.01649	2.13165
HNRNPK	2.92914	1.86247
U2AF1; U2AF1L4	2.61456	3.2657
HNRNPA3	2.50342	1.3001
TRIM28	2.4805	1.45038
PTBP1	2.42348	2.07927
Mutlimapped histones	2.08494	1.53513
HIST1H4A	2.0132	2.6745
DDX17	1.93928	1.7616
DDX3X	1.74863	2.03304
PRPF19	1.70918	1.33898
ALYREF	1.69987	2.25671
SRSF1	1.61895	2.37474
HNRNPF	1.58675	2.2381
POLDIP3; PDIP46	1.44117	2.32265
ATXN2L	0.920931	1.50973
RBM17	0.848335	1.36777
RBM17	0.848335	1.36777

References

1. Kim, T.K., et al., *Widespread transcription at neuronal activity-regulated enhancers.* Nature, 2010. **465**(7295): p. 182-7.
2. De Santa, F., et al., *A large fraction of extragenic RNA pol II transcription sites overlap enhancers.* PLoS Biol, 2010. **8**(5): p. e1000384.
3. Statello, L., et al., *Gene regulation by long non-coding RNAs and its biological functions.* Nat Rev Mol Cell Biol, 2021. **22**(2): p. 96-118.
4. Nojima, T. and N.J. Proudfoot, *Mechanisms of lncRNA biogenesis as revealed by nascent transcriptomics.* Nat Rev Mol Cell Biol, 2022. **23**(6): p. 389-406.
5. O'Brien, J., et al., *Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation.* Front Endocrinol (Lausanne), 2018. **9**: p. 402.
6. Svejstrup, J.Q., *The RNA polymerase II transcription cycle: cycling through chromatin.* Biochim Biophys Acta, 2004. **1677**(1-3): p. 64-73.
7. Conaway, R.C. and J.W. Conaway, *General initiation factors for RNA polymerase II.* Annu Rev Biochem, 1993. **62**: p. 161-90.
8. Kornberg, R.D., *The molecular basis of eukaryotic transcription.* Proc Natl Acad Sci U S A, 2007. **104**(32): p. 12955-61.
9. Pal, M., A.S. Ponticelli, and D.S. Luse, *The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II.* Mol Cell, 2005. **19**(1): p. 101-10.
10. Murakami, K., et al., *Formation and fate of a complete 31-protein RNA polymerase II transcription preinitiation complex.* J Biol Chem, 2013. **288**(9): p. 6325-32.
11. Haberle, V. and A. Stark, *Eukaryotic core promoters and the functional basis of transcription initiation.* Nat Rev Mol Cell Biol, 2018. **19**(10): p. 621-637.
12. Carninci, P., et al., *Genome-wide analysis of mammalian promoter architecture and evolution.* Nat Genet, 2006. **38**(6): p. 626-35.
13. Rach, E.A., et al., *Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level.* PLoS Genet, 2011. **7**(1): p. e1001274.
14. Bernstein, B.E., et al., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells.* Cell, 2006. **125**(2): p. 315-26.
15. Knezetic, J.A. and D.S. Luse, *The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro.* Cell, 1986. **45**(1): p. 95-104.
16. Lorch, Y., J.W. LaPointe, and R.D. Kornberg, *Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones.* Cell, 1987. **49**(2): p. 203-10.
17. Lorch, Y. and R.D. Kornberg, *Chromatin-remodeling for transcription.* Q Rev Biophys, 2017. **50**: p. e5.
18. Cramer, P., *Organization and regulation of gene transcription.* Nature, 2019. **573**(7772): p. 45-54.
19. Bernstein, B.E., et al., *Global nucleosome occupancy in yeast.* Genome Biol, 2004. **5**(9): p. R62.
20. Lee, C.K., et al., *Evidence for nucleosome depletion at active regulatory regions genome-wide.* Nat Genet, 2004. **36**(8): p. 900-5.
21. Ozsolak, F., et al., *High-throughput mapping of the chromatin structure of human promoters.* Nat Biotechnol, 2007. **25**(2): p. 244-8.
22. Yuan, G.C., et al., *Genome-scale identification of nucleosome positions in S. cerevisiae.* Science, 2005. **309**(5734): p. 626-30.
23. Schones, D.E., et al., *Dynamic regulation of nucleosome positioning in the human genome.* Cell, 2008. **132**(5): p. 887-98.
24. Boeger, H., et al., *Nucleosomes unfold completely at a transcriptionally active promoter.* Mol Cell, 2003. **11**(6): p. 1587-98.
25. Reinke, H. and W. Horz, *Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter.* Mol Cell, 2003. **11**(6): p. 1599-607.
26. Lomvardas, S. and D. Thanos, *Modifying gene expression programs by altering core promoter chromatin architecture.* Cell, 2002. **110**(2): p. 261-71.
27. Fuda, N.J., M.B. Ardehali, and J.T. Lis, *Defining mechanisms that regulate RNA polymerase II transcription in vivo.* Nature, 2009. **461**(7261): p. 186-92.

28. Utley, R.T., et al., *Transcriptional activators direct histone acetyltransferase complexes to nucleosomes.* Nature, 1998. **394**(6692): p. 498-502.
29. Reinberg, D., et al., *The RNA polymerase II general transcription factors: past, present, and future.* Cold Spring Harb Symp Quant Biol, 1998. **63**: p. 83-103.
30. Rimel, J.K. and D.J. Taatjes, *The essential and multifunctional TFIIH complex.* Protein Sci, 2018. **27**(6): p. 1018-1037.
31. Buratowski, S., et al., *Five intermediate complexes in transcription initiation by RNA polymerase II.* Cell, 1989. **56**(4): p. 549-61.
32. Albright, S.R. and R. Tjian, *TAFs revisited: more data reveal new twists and confirm old ideas.* Gene, 2000. **242**(1-2): p. 1-13.
33. Bleichenbacher, M., S. Tan, and T.J. Richmond, *Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes.* J Mol Biol, 2003. **332**(4): p. 783-93.
34. Hoiby, T., et al., *A facelift for the general transcription factor TFIIA.* Biochim Biophys Acta, 2007. **1769**(7-8): p. 429-36.
35. Thomas, M.C. and C.M. Chiang, *The general transcription machinery and general cofactors.* Crit Rev Biochem Mol Biol, 2006. **41**(3): p. 105-78.
36. Kostrewa, D., et al., *RNA polymerase II-TFIIB structure and mechanism of transcription initiation.* Nature, 2009. **462**(7271): p. 323-30.
37. Chen, H.T. and S. Hahn, *Mapping the location of TFIIB within the RNA polymerase II transcription preinitiation complex: a model for the structure of the PIC.* Cell, 2004. **119**(2): p. 169-80.
38. Bushnell, D.A., et al., *Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms.* Science, 2004. **303**(5660): p. 983-8.
39. Sainsbury, S., J. Niesser, and P. Cramer, *Structure and function of the initially transcribing RNA polymerase II-TFIIB complex.* Nature, 2013. **493**(7432): p. 437-40.
40. Cabart, P., et al., *Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes.* Proc Natl Acad Sci U S A, 2011. **108**(38): p. 15786-91.
41. Plaschka, C., et al., *Transcription initiation complex structures elucidate DNA opening.* Nature, 2016. **533**(7603): p. 353-8.
42. He, Y., et al., *Near-atomic resolution visualization of human transcription promoter opening.* Nature, 2016. **533**(7603): p. 359-65.
43. Schilbach, S., et al., *Structures of transcription pre-initiation complex with TFIIH and Mediator.* Nature, 2017. **551**(7679): p. 204-209.
44. Ohkuma, Y. and R.G. Roeder, *Regulation of TFIIH ATPase and kinase activities by TFIIE during active initiation complex formation.* Nature, 1994. **368**(6467): p. 160-3.
45. Allen, B.L. and D.J. Taatjes, *The Mediator complex: a central integrator of transcription.* Nat Rev Mol Cell Biol, 2015. **16**(3): p. 155-66.
46. Plaschka, C., et al., *Architecture of the RNA polymerase II-Mediator core initiation complex.* Nature, 2015. **518**(7539): p. 376-80.
47. Tsai, K.L., et al., *Mediator structure and rearrangements required for holoenzyme formation.* Nature, 2017. **544**(7649): p. 196-201.
48. Nozawa, K., T.R. Schneider, and P. Cramer, *Core Mediator structure at 3.4 A extends model of transcription initiation complex.* Nature, 2017. **545**(7653): p. 248-251.
49. Chen, X., et al., *Structures of the human Mediator and Mediator-bound preinitiation complex.* Science, 2021. **372**(6546).
50. Baek, H.J., Y.K. Kang, and R.G. Roeder, *Human Mediator enhances basal transcription by facilitating recruitment of transcription factor IIB during preinitiation complex assembly.* J Biol Chem, 2006. **281**(22): p. 15172-81.
51. Johnson, K.M., et al., *TFIID and human mediator coactivator complexes assemble cooperatively on promoter DNA.* Genes Dev, 2002. **16**(14): p. 1852-63.
52. Esnault, C., et al., *Mediator-dependent recruitment of TFIIH modules in preinitiation complex.* Mol Cell, 2008. **31**(3): p. 337-46.
53. Nair, D., Y. Kim, and L.C. Myers, *Mediator and TFIIH govern carboxyl-terminal domain-dependent transcription in yeast extracts.* J Biol Chem, 2005. **280**(40): p. 33739-48.
54. Eick, D. and M. Geyer, *The RNA polymerase II carboxy-terminal domain (CTD) code.* Chem Rev, 2013. **113**(11): p. 8456-90.

55. Harlen, K.M. and L.S. Churchman, *The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain.* Nat Rev Mol Cell Biol, 2017. **18**(4): p. 263-273.
56. Zaborowska, J., S. Egloff, and S. Murphy, *The pol II CTD: new twists in the tail.* Nat Struct Mol Biol, 2016. **23**(9): p. 771-7.
57. Buratowski, S., *Progression through the RNA polymerase II CTD cycle.* Mol Cell, 2009. **36**(4): p. 541-6.
58. Ghosh, A. and C.D. Lima, *Enzymology of RNA cap synthesis.* Wiley Interdiscip Rev RNA, 2010. **1**(1): p. 152-72.
59. Wong, K.H., Y. Jin, and K. Struhl, *TFIIH phosphorylation of the Pol II CTD stimulates mediator dissociation from the preinitiation complex and promoter escape.* Mol Cell, 2014. **54**(4): p. 601-12.
60. Czudnochowski, N., C.A. Bosken, and M. Geyer, *Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition.* Nat Commun, 2012. **3**: p. 842.
61. Rimel, J.K., et al., *Selective inhibition of CDK7 reveals high-confidence targets and new models for TFIIH function in transcription.* Genes Dev, 2020. **34**(21-22): p. 1452-1473.
62. Kim, T.K., R.H. Ebright, and D. Reinberg, *Mechanism of ATP-dependent promoter melting by transcription factor IIH.* Science, 2000. **288**(5470): p. 1418-22.
63. Wang, W., M. Carey, and J.D. Gralla, *Polymerase II promoter activation: closed complex formation and ATP-driven start site opening.* Science, 1992. **255**(5043): p. 450-3.
64. Fishburn, J., et al., *Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation.* Proc Natl Acad Sci U S A, 2015. **112**(13): p. 3961-6.
65. Dienemann, C., et al., *Promoter Distortion and Opening in the RNA Polymerase II Cleft.* Mol Cell, 2019. **73**(1): p. 97-106 e4.
66. Tirode, F., et al., *Reconstitution of the transcription factor TFIIH: assignment of functions for the three enzymatic subunits, XPB, XPD, and cdk7.* Mol Cell, 1999. **3**(1): p. 87-95.
67. Gill, J., *Boleslaw Gutowski 1888-1966.* Acta Physiol Pol, 1987. **38**(3): p. 237-45.
68. Hampsey, M., *Molecular genetics of the RNA polymerase II general transcriptional machinery.* Microbiol Mol Biol Rev, 1998. **62**(2): p. 465-503.
69. Li, C., B. Lenhard, and N.M. Luscombe, *Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome.* Genome Res, 2018. **28**(5): p. 676-688.
70. Revyakin, A., et al., *Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching.* Science, 2006. **314**(5802): p. 1139-43.
71. Kapanidis, A.N., et al., *Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism.* Science, 2006. **314**(5802): p. 1144-7.
72. Fazal, F.M., et al., *Real-time observation of the initiation of RNA polymerase II transcription.* Nature, 2015. **525**(7568): p. 274-7.
73. Holstege, F.C., U. Fiedler, and H.T. Timmers, *Three transitions in the RNA polymerase II transcription complex during initiation.* EMBO J, 1997. **16**(24): p. 7468-80.
74. Hieb, A.R., et al., *An 8 nt RNA triggers a rate-limiting shift of RNA polymerase II complexes into elongation.* EMBO J, 2006. **25**(13): p. 3100-9.
75. Luse, D.S., *Promoter clearance by RNA polymerase II.* Biochim Biophys Acta, 2013. **1829**(1): p. 63-8.
76. Winkelman, J.T. and R.L. Gourse, *Open complex DNA scrunching: A key to transcription start site selection and promoter escape.* Bioessays, 2017. **39**(2).
77. Steitz, T.A. and J.A. Steitz, *A general two-metal-ion mechanism for catalytic RNA.* Proc Natl Acad Sci U S A, 1993. **90**(14): p. 6498-502.
78. Dangkulwanich, M., et al., *Complete dissection of transcription elongation reveals slow translocation of RNA polymerase II in a linear ratchet mechanism.* Elife, 2013. **2**: p. e00971.
79. Noe Gonzalez, M., D. Blears, and J.Q. Svejstrup, *Causes and consequences of RNA polymerase II stalling during transcript elongation.* Nat Rev Mol Cell Biol, 2021. **22**(1): p. 3-21.
80. Nudler, E., et al., *The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase.* Cell, 1997. **89**(1): p. 33-41.
81. Gilmour, D.S. and J.T. Lis, *RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells.* Mol Cell Biol, 1986. **6**(11): p. 3984-9.

82. Rougvie, A.E. and J.T. Lis, *The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged.* Cell, 1988. **54**(6): p. 795-804.

83. Muse, G.W., et al., *RNA polymerase is poised for activation across the genome.* Nat Genet, 2007. **39**(12): p. 1507-11.

84. Parkerton, T.F., et al., *Guidance for evaluating in vivo fish bioaccumulation data.* Integr Environ Assess Manag, 2008. **4**(2): p. 139-55.

85. Zeitlinger, J., et al., *RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo.* Nat Genet, 2007. **39**(12): p. 1512-6.

86. Nechaev, S., et al., *Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila.* Science, 2010. **327**(5963): p. 335-8.

87. Kwak, H., et al., *Precise maps of RNA polymerase reveal how promoters direct initiation and pausing.* Science, 2013. **339**(6122): p. 950-3.

88. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.* Science, 2008. **322**(5909): p. 1845-8.

89. Mayer, A., et al., *Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution.* Cell, 2015. **161**(3): p. 541-554.

90. Nojima, T., et al., *Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing.* Cell, 2015. **161**(3): p. 526-540.

91. Krebs, A.R., et al., *Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters.* Mol Cell, 2017. **67**(3): p. 411-422 e4.

92. Erickson, B., et al., *Dynamic turnover of paused Pol II complexes at human promoters.* Genes Dev, 2018. **32**(17-18): p. 1215-1225.

93. Steurer, B., et al., *Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II.* Proc Natl Acad Sci U S A, 2018. **115**(19): p. E4368-E4376.

94. Gressel, S., et al., *CDK9-dependent RNA polymerase II pausing controls transcription initiation.* Elife, 2017. **6**.

95. Vos, S.M., et al., *Structure of paused transcription complex Pol II-DSIF-NELF.* Nature, 2018. **560**(7720): p. 601-606.

96. Saba, J., et al., *The elemental mechanism of transcriptional pausing.* Elife, 2019. **8**.

97. Yamaguchi, Y., H. Shibata, and H. Handa, *Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond.* Biochim Biophys Acta, 2013. **1829**(1): p. 98-104.

98. Wada, T., et al., *DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs.* Genes Dev, 1998. **12**(3): p. 343-56.

99. Bernecky, C., J.M. Plitzko, and P. Cramer, *Structure of a transcribing RNA polymerase II-DSIF complex reveals a multidentate DNA-RNA clamp.* Nat Struct Mol Biol, 2017. **24**(10): p. 809-815.

100. Vos, S.M., et al., *Structure of activated transcription complex Pol II-DSIF-PAF-SPT6.* Nature, 2018. **560**(7720): p. 607-612.

101. Missra, A. and D.S. Gilmour, *Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex.* Proc Natl Acad Sci U S A, 2010. **107**(25): p. 11301-6.

102. Pei, Y. and S. Shuman, *Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5.* J Biol Chem, 2002. **277**(22): p. 19639-48.

103. Hu, S., et al., *SPT5 stabilizes RNA polymerase II, orchestrates transcription cycles, and maintains the enhancer landscape.* Mol Cell, 2021. **81**(21): p. 4425-4439 e6.

104. Aoi, Y., et al., *SPT5 stabilization of promoter-proximal RNA polymerase II.* Mol Cell, 2021. **81**(21): p. 4413-4424 e5.

105. Yamaguchi, Y., et al., *NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation.* Cell, 1999. **97**(1): p. 41-51.

106. Core, L. and K. Adelman, *Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation.* Genes Dev, 2019. **33**(15-16): p. 960-982.

107. Vos, S.M., et al., *Architecture and RNA binding of the human negative elongation factor.* Elife, 2016. **5**.

108. Gilchrist, D.A., et al., *Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation.* Cell, 2010. **143**(4): p. 540-51.

109. Core, L.J., et al., *Defining the status of RNA polymerase at promoters.* Cell Rep, 2012. **2**(4): p. 1025-35.
110. Aoi, Y., et al., *NELF Regulates a Promoter-Proximal Step Distinct from RNA Pol II Pause-Release.* Mol Cell, 2020. **78**(2): p. 261-274 e5.
111. Marshall, N.F. and D.H. Price, *Purification of P-TEFb, a transcription factor required for the transition into productive elongation.* J Biol Chem, 1995. **270**(21): p. 12335-8.
112. Altendorfer, E., Y. Mochalova, and A. Mayer, *BRD4: a general regulator of transcription elongation.* Transcription, 2022. **13**(1-3): p. 70-81.
113. Arnold, M., et al., *A BRD4-mediated elongation control point primes transcribing RNA polymerase II for 3'-processing and termination.* Mol Cell, 2021. **81**(17): p. 3589-3603 e13.
114. Muhar, M., et al., *SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis.* Science, 2018. **360**(6390): p. 800-805.
115. Zheng, B., et al., *Acute perturbation strategies in interrogating RNA polymerase II elongation factor function in gene expression.* Genes Dev, 2021. **35**(3-4): p. 273-285.
116. Price, D.H., *P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II.* Mol Cell Biol, 2000. **20**(8): p. 2629-34.
117. Luo, Z., C. Lin, and A. Shilatifard, *The super elongation complex (SEC) family in transcriptional control.* Nat Rev Mol Cell Biol, 2012. **13**(9): p. 543-7.
118. He, N., et al., *HIV-1 Tat and host AFF4 recruit two transcription elongation factors into a bifunctional complex for coordinated activation of HIV-1 transcription.* Mol Cell, 2010. **38**(3): p. 428-38.
119. Lin, C., et al., *AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia.* Mol Cell, 2010. **37**(3): p. 429-37.
120. Smith, E., C. Lin, and A. Shilatifard, *The super elongation complex (SEC) and MLL in development and disease.* Genes Dev, 2011. **25**(7): p. 661-72.
121. Zhou, Q., T. Li, and D.H. Price, *RNA polymerase II elongation control.* Annu Rev Biochem, 2012. **81**: p. 119-43.
122. Cho, S., et al., *Acetylation of cyclin T1 regulates the equilibrium between active and inactive P-TEFb in cells.* EMBO J, 2009. **28**(10): p. 1407-17.
123. Sanso, M., et al., *P-TEFb regulation of transcription termination factor Xrn2 revealed by a chemical genetic screen for Cdk9 substrates.* Genes Dev, 2016. **30**(1): p. 117-31.
124. Huang, K.L., et al., *Integrator Recruits Protein Phosphatase 2A to Prevent Pause Release and Facilitate Transcription Termination.* Mol Cell, 2020. **80**(2): p. 345-358 e9.
125. Yamada, T., et al., *P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation.* Mol Cell, 2006. **21**(2): p. 227-37.
126. Cortazar, M.A., et al., *Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism.* Mol Cell, 2019. **76**(6): p. 896-908 e4.
127. Pavri, R., et al., *Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5.* Cell, 2010. **143**(1): p. 122-33.
128. Rahl, P.B., et al., *c-Myc regulates transcriptional pause release.* Cell, 2010. **141**(3): p. 432-45.
129. Zumer, K., et al., *Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo.* Mol Cell, 2021. **81**(15): p. 3096-3109 e8.
130. Van Oss, S.B., C.E. Cucinotta, and K.M. Arndt, *Emerging Insights into the Roles of the Paf1 Complex in Gene Regulation.* Trends Biochem Sci, 2017. **42**(10): p. 788-798.
131. Hou, L., et al., *Paf1C regulates RNA polymerase II progression by modulating elongation rate.* Proc Natl Acad Sci U S A, 2019. **116**(29): p. 14583-14592.
132. Sdano, M.A., et al., *A novel SH2 recognition mechanism recruits Spt6 to the doubly phosphorylated RNA polymerase II linker at sites of transcription.* Elife, 2017. **6**.
133. Narain, A., et al., *Targeted protein degradation reveals a direct role of SPT6 in RNAPII elongation and termination.* Mol Cell, 2021. **81**(15): p. 3110-3127 e14.
134. Bortvin, A. and F. Winston, *Evidence that Spt6p controls chromatin structure by a direct interaction with histones.* Science, 1996. **272**(5267): p. 1473-6.
135. Ehara, H., et al., *Structural basis of nucleosome disassembly and reassembly by RNAPII elongation complex with FACT.* Science, 2022. **377**(6611): p. eabp9466.

136. Kristjuhan, A. and J.Q. Svejstrup, *Evidence for distinct mechanisms facilitating transcript elongation through chromatin in vivo.* EMBO J, 2004. **23**(21): p. 4243-52.

137. Schwabish, M.A. and K. Struhl, *Asf1 mediates histone eviction and deposition during elongation by RNA polymerase II.* Mol Cell, 2006. **22**(3): p. 415-22.

138. Izban, M.G. and D.S. Luse, *Transcription on nucleosomal templates by RNA polymerase II in vitro: inhibition of elongation with enhancement of sequence-specific pausing.* Genes Dev, 1991. **5**(4): p. 683-96.

139. Izban, M.G. and D.S. Luse, *Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates.* J Biol Chem, 1992. **267**(19): p. 13647-55.

140. Kireeva, M.L., et al., *Nucleosome remodeling induced by RNA polymerase II: loss of the H2A/H2B dimer during transcription.* Mol Cell, 2002. **9**(3): p. 541-52.

141. Belotserkovskaya, R., et al., *FACT facilitates transcription-dependent nucleosome alteration.* Science, 2003. **301**(5636): p. 1090-3.

142. Kimura, H. and P.R. Cook, *Kinetics of core histones in living human cells: little exchange of H3 and H4 and some rapid exchange of H2B.* J Cell Biol, 2001. **153**(7): p. 1341-53.

143. Thiriet, C. and J.J. Hayes, *Replication-independent core histone dynamics at transcriptionally active loci in vivo.* Genes Dev, 2005. **19**(6): p. 677-82.

144. Kulaeva, O.I., et al., *Mechanism of chromatin remodeling and recovery during passage of RNA polymerase II.* Nat Struct Mol Biol, 2009. **16**(12): p. 1272-8.

145. Weber, C.M., S. Ramachandran, and S. Henikoff, *Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase.* Mol Cell, 2014. **53**(5): p. 819-30.

146. Chen, Z., et al., *High-resolution and high-accuracy topographic and transcriptional maps of the nucleosome barrier.* Elife, 2019. **8**.

147. Bondarenko, V.A., et al., *Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II.* Mol Cell, 2006. **24**(3): p. 469-79.

148. Kujirai, T., et al., *Structural basis of the nucleosome transition during RNA polymerase II passage.* Science, 2018. **362**(6414): p. 595-598.

149. Farnung, L., S.M. Vos, and P. Cramer, *Structure of transcribing RNA polymerase II-nucleosome complex.* Nat Commun, 2018. **9**(1): p. 5432.

150. Kireeva, M.L., et al., *Nature of the nucleosomal barrier to RNA polymerase II.* Mol Cell, 2005. **18**(1): p. 97-108.

151. Hendrix, D.A., et al., *Promoter elements associated with RNA Pol II stalling in the Drosophila embryo.* Proc Natl Acad Sci U S A, 2008. **105**(22): p. 7762-7.

152. Chan, C.L. and R. Landick, *Dissection of the his leader pause site by base substitution reveals a multipartite signal that includes a pause RNA hairpin.* J Mol Biol, 1993. **233**(1): p. 25-42.

153. Watts, J.A., et al., *cis Elements that Mediate RNA Polymerase II Pausing Regulate Human Gene Expression.* Am J Hum Genet, 2019. **105**(4): p. 677-688.

154. Sheridan, R.M., et al., *Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5' Pause Release, Termination, and Transcription Elongation Rate.* Mol Cell, 2019. **73**(1): p. 107-118 e4.

155. Gajos, M., et al., *Conserved DNA sequence features underlie pervasive RNA polymerase pausing.* Nucleic Acids Res, 2021. **49**(8): p. 4402-4420.

156. Pomerantz, R.T. and M. O'Donnell, *The replisome uses mRNA as a primer after colliding with RNA polymerase.* Nature, 2008. **456**(7223): p. 762-6.

157. Hamperl, S., et al., *Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses.* Cell, 2017. **170**(4): p. 774-786 e19.

158. Takeuchi, Y., T. Horiuchi, and T. Kobayashi, *Transcription-dependent recombination and the role of fork collision in yeast rDNA.* Genes Dev, 2003. **17**(12): p. 1497-506.

159. Prado, F. and A. Aguilera, *Impairment of replication fork progression mediates RNA polII transcription-associated recombination.* EMBO J, 2005. **24**(6): p. 1267-76.

160. Teloni, F., et al., *Efficient Pre-mRNA Cleavage Prevents Replication-Stress-Associated Genome Instability.* Mol Cell, 2019. **73**(4): p. 670-683 e12.

161. Hobson, D.J., et al., *RNA polymerase II collision interrupts convergent transcription.* Mol Cell, 2012. **48**(3): p. 365-74.

162. Kamieniarz-Gdula, K., et al., *Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination.* Mol Cell, 2019. **74**(1): p. 158-172 e9.
163. Gregersen, L.H. and J.Q. Svejstrup, *The Cellular Response to Transcription-Blocking DNA Damage.* Trends Biochem Sci, 2018. **43**(5): p. 327-341.
164. Chiou, Y.Y., et al., *RNA polymerase II is released from the DNA template during transcription-coupled repair in mammalian cells.* J Biol Chem, 2018. **293**(7): p. 2476-2486.
165. Selby, C.P. and A. Sancar, *Cockayne syndrome group B protein enhances elongation by RNA polymerase II.* Proc Natl Acad Sci U S A, 1997. **94**(21): p. 11205-9.
166. Xu, J., et al., *Structural basis for the initiation of eukaryotic transcription-coupled DNA repair.* Nature, 2017. **551**(7682): p. 653-657.
167. Caron, P., J. van der Linden, and H. van Attikum, *Bon voyage: A transcriptional journey around DNA breaks.* DNA Repair (Amst), 2019. **82**: p. 102686.
168. Machour, F.E. and N. Ayoub, *Transcriptional Regulation at DSBs: Mechanisms and Consequences.* Trends Genet, 2020. **36**(12): p. 981-997.
169. Shanbhag, N.M., et al., *ATM-dependent chromatin changes silence transcription in cis to DNA double-strand breaks.* Cell, 2010. **141**(6): p. 970-81.
170. Awwad, S.W., et al., *NELF-E is recruited to DNA double-strand break sites to promote transcriptional repression and repair.* EMBO Rep, 2017. **18**(5): p. 745-764.
171. Skourti-Stathaki, K., N.J. Proudfoot, and N. Gromak, *Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination.* Mol Cell, 2011. **42**(6): p. 794-805.
172. Cohen, S., et al., *Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations.* Nat Commun, 2018. **9**(1): p. 533.
173. Pankotai, T., et al., *DNAPKcs-dependent arrest of RNA polymerase II transcription in the presence of DNA breaks.* Nat Struct Mol Biol, 2012. **19**(3): p. 276-82.
174. Caron, P., et al., *WWP2 ubiquitylates RNA polymerase II for DNA-PK-dependent transcription arrest and repair at DNA breaks.* Genes Dev, 2019. **33**(11-12): p. 684-704.
175. Proudfoot, N.J., *Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut.* Science, 2016. **352**(6291): p. aad9926.
176. Connelly, S. and J.L. Manley, *A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II.* Genes Dev, 1988. **2**(4): p. 440-52.
177. Hirose, Y. and J.L. Manley, *RNA polymerase II is an essential mRNA polyadenylation factor.* Nature, 1998. **395**(6697): p. 93-6.
178. Shi, Y., S. Chan, and G. Martinez-Santibanez, *An up-close look at the pre-mRNA 3'-end processing complex.* RNA Biol, 2009. **6**(5): p. 522-5.
179. Zarudnaya, M.I., et al., *Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures.* Nucleic Acids Res, 2003. **31**(5): p. 1375-86.
180. Eaton, J.D. and S. West, *Termination of Transcription by RNA Polymerase II: BOOM!* Trends Genet, 2020. **36**(9): p. 664-675.
181. Sun, Y., K. Hamilton, and L. Tong, *Recent molecular insights into canonical pre-mRNA 3'-end processing.* Transcription, 2020. **11**(2): p. 83-96.
182. Zhang, H., F. Rigo, and H.G. Martinson, *Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site.* Mol Cell, 2015. **59**(3): p. 437-48.
183. Zhang, Y., et al., *Structural Insights into the Human Pre-mRNA 3'-End Processing Machinery.* Mol Cell, 2020. **77**(4): p. 800-809 e6.
184. Chan, S.L., et al., *CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing.* Genes Dev, 2014. **28**(21): p. 2370-80.
185. Clerici, M., et al., *Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex.* Nat Struct Mol Biol, 2018. **25**(2): p. 135-138.
186. Sun, Y., et al., *Molecular basis for the recognition of the human AAUAAA polyadenylation signal.* Proc Natl Acad Sci U S A, 2018. **115**(7): p. E1419-E1428.
187. Helmling, S., A. Zhelkovsky, and C.L. Moore, *Fip1 regulates the activity of Poly(A) polymerase through multiple interactions.* Mol Cell Biol, 2001. **21**(6): p. 2026-37.
188. Mandel, C.R., et al., *Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease.* Nature, 2006. **444**(7121): p. 953-6.

189. Callebaut, I., et al., *Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family.* Nucleic Acids Res, 2002. **30**(16): p. 3592-601.
190. Bienroth, S., W. Keller, and E. Wahle, *Assembly of a processive messenger RNA polyadenylation complex.* EMBO J, 1993. **12**(2): p. 585-94.
191. Eckmann, C.R., C. Rammelt, and E. Wahle, *Control of poly(A) tail length.* Wiley Interdiscip Rev RNA, 2011. **2**(3): p. 348-61.
192. Schonemann, L., et al., *Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33.* Genes Dev, 2014. **28**(21): p. 2381-93.
193. Nagaike, T., et al., *Transcriptional activators enhance polyadenylation of mRNA precursors.* Mol Cell, 2011. **41**(4): p. 409-18.
194. Xiang, K., et al., *Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex.* Nature, 2010. **467**(7316): p. 729-33.
195. Murthy, K.G. and J.L. Manley, *The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation.* Genes Dev, 1995. **9**(21): p. 2672-83.
196. Gilmartin, G.M. and J.R. Nevins, *An ordered pathway of assembly of components required for polyadenylation site recognition and processing.* Genes Dev, 1989. **3**(12B): p. 2180-90.
197. Fong, N. and D.L. Bentley, *Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD.* Genes Dev, 2001. **15**(14): p. 1783-95.
198. Kaneko, S., et al., *The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3' processing and transcription termination.* Genes Dev, 2007. **21**(14): p. 1779-89.
199. Tian, B. and J.L. Manley, *Alternative polyadenylation of mRNA precursors.* Nat Rev Mol Cell Biol, 2017. **18**(1): p. 18-30.
200. Kim, S., et al., *Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation.* Genes Cells, 2010. **15**(9): p. 1003-13.
201. Brown, K.M. and G.M. Gilmartin, *A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im.* Mol Cell, 2003. **12**(6): p. 1467-76.
202. Zhu, Y., et al., *Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation.* Mol Cell, 2018. **69**(1): p. 62-74 e4.
203. Dettwiler, S., et al., *Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization.* J Biol Chem, 2004. **279**(34): p. 35788-97.
204. Kubo, T., et al., *Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3'-UTRs.* Nucleic Acids Res, 2006. **34**(21): p. 6264-71.
205. Martin, G., et al., *Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length.* Cell Rep, 2012. **1**(6): p. 753-63.
206. Gruber, A.R., et al., *Cleavage factor Im is a key regulator of 3' UTR length.* RNA Biol, 2012. **9**(12): p. 1405-12.
207. Barilla, D., B.A. Lee, and N.J. Proudfoot, *Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 2001. **98**(2): p. 445-50.
208. West, S. and N.J. Proudfoot, *Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination.* Nucleic Acids Res, 2008. **36**(3): p. 905-14.
209. Cotten, M., et al., *Specific contacts between mammalian U7 snRNA and histone precursor RNA are indispensable for the in vitro 3' RNA processing reaction.* EMBO J, 1988. **7**(3): p. 801-8.
210. Dominski, Z., X.C. Yang, and W.F. Marzluff, *The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing.* Cell, 2005. **123**(1): p. 37-48.
211. Sullivan, K.D., M. Steiniger, and W.F. Marzluff, *A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs.* Mol Cell, 2009. **34**(3): p. 322-32.
212. Sun, Y., et al., *Structure of an active human histone pre-mRNA 3'-end processing machinery.* Science, 2020. **367**(6478): p. 700-703.
213. Kolev, N.G. and J.A. Steitz, *Symplekin and multiple other polyadenylation factors participate in 3'-end maturation of histone mRNAs.* Genes Dev, 2005. **19**(21): p. 2583-92.

214. Brugiolo, M., L. Herzel, and K.M. Neugebauer, *Counting on co-transcriptional splicing.* F1000Prime Rep, 2013. **5**: p. 9.

215. Oesterreich, F.C., et al., *Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II.* Cell, 2016. **165**(2): p. 372-381.

216. Drexler, H.L., K. Choquet, and L.S. Churchman, *Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores.* Mol Cell, 2020. **77**(5): p. 985-998 e8.

217. Herzel, L., et al., *Splicing and transcription touch base: co-transcriptional spliceosome assembly and function.* Nat Rev Mol Cell Biol, 2017. **18**(10): p. 637-650.

218. Wahl, M.C., C.L. Will, and R. Luhrmann, *The spliceosome: design principles of a dynamic RNP machine.* Cell, 2009. **136**(4): p. 701-18.

219. Cvitkovic, I. and M.S. Jurica, *Spliceosome database: a tool for tracking components of the spliceosome.* Nucleic Acids Res, 2013. **41**(Database issue): p. D132-41.

220. Will, C.L. and R. Luhrmann, *Spliceosome structure and function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).

221. Nojima, T., et al., *RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing.* Mol Cell, 2018. **72**(2): p. 369-379 e4.

222. Harlen, K.M., et al., *Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue.* Cell Rep, 2016. **15**(10): p. 2147-2158.

223. Gu, B., D. Eick, and O. Bensaude, *CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo.* Nucleic Acids Res, 2013. **41**(3): p. 1591-603.

224. Kwak, H. and J.T. Lis, *Control of transcriptional elongation.* Annu Rev Genet, 2013. **47**: p. 483-508.

225. Alexander, R.D., et al., *Splicing-dependent RNA polymerase pausing in yeast.* Mol Cell, 2010. **40**(4): p. 582-93.

226. de la Mata, M., et al., *A slow RNA polymerase II affects alternative splicing in vivo.* Mol Cell, 2003. **12**(2): p. 525-32.

227. Howe, K.J., C.M. Kane, and M. Ares, Jr., *Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae.* RNA, 2003. **9**(8): p. 993-1006.

228. Dujardin, G., et al., *How slow RNA polymerase II elongation favors alternative exon skipping.* Mol Cell, 2014. **54**(4): p. 683-90.

229. Fong, N., et al., *Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate.* Genes Dev, 2014. **28**(23): p. 2663-76.

230. Guo, Y.E., et al., *Pol II phosphorylation regulates a switch between transcriptional and splicing condensates.* Nature, 2019. **572**(7770): p. 543-548.

231. Sekimizu, K., et al., *Purification of a factor from Ehrlich ascites tumor cells specifically stimulating RNA polymerase II.* Biochemistry, 1976. **15**(23): p. 5064-70.

232. Rappaport, J., et al., *Purification and functional characterization of transcription factor SII from calf thymus. Role in RNA polymerase II elongation.* J Biol Chem, 1987. **262**(11): p. 5227-32.

233. Reinberg, D. and R.G. Roeder, *Factors involved in specific transcription by mammalian RNA polymerase II. Transcription factor IIS stimulates elongation of RNA chains.* J Biol Chem, 1987. **262**(7): p. 3331-7.

234. Reines, D., *Elongation factor-dependent transcript shortening by template-engaged RNA polymerase II.* J Biol Chem, 1992. **267**(6): p. 3795-800.

235. Izban, M.G. and D.S. Luse, *The RNA polymerase II ternary complex cleaves the nascent transcript in a 3'----5' direction in the presence of elongation factor SII.* Genes Dev, 1992. **6**(7): p. 1342-56.

236. Wang, D. and D.K. Hawley, *Identification of a 3'-->5' exonuclease activity associated with human RNA polymerase II.* Proc Natl Acad Sci U S A, 1993. **90**(3): p. 843-7.

237. Reines, D., M.J. Chamberlin, and C.M. Kane, *Transcription elongation factor SII (TFIIS) enables RNA polymerase II to elongate through a block to transcription in a human gene in vitro.* J Biol Chem, 1989. **264**(18): p. 10799-809.

238. Thomas, M.J., A.A. Platas, and D.K. Hawley, *Transcriptional fidelity and proofreading by RNA polymerase II.* Cell, 1998. **93**(4): p. 627-37.

239. Izban, M.G. and D.S. Luse, *The increment of SII-facilitated transcript cleavage varies dramatically between elongation competent and incompetent RNA polymerase II ternary complexes.* J Biol Chem, 1993. **268**(17): p. 12874-85.

240. Izban, M.G. and D.S. Luse, *SII-facilitated transcript cleavage in RNA polymerase II complexes stalled early after initiation occurs in primarily dinucleotide increments.* J Biol Chem, 1993. **268**(17): p. 12864-73.
241. Gu, W. and D. Reines, *Variation in the size of nascent RNA cleavage products as a function of transcript length and elongation competence.* J Biol Chem, 1995. **270**(51): p. 30441-7.
242. Zhang, C. and Z.F. Burton, *Transcription factors IIF and IIS and nucleoside triphosphate substrates as dynamic probes of the human RNA polymerase II mechanism.* J Mol Biol, 2004. **342**(4): p. 1085-99.
243. Ishibashi, T., et al., *Transcription factors IIS and IIF enhance transcription efficiency by differentially modifying RNA polymerase pausing dynamics.* Proc Natl Acad Sci U S A, 2014. **111**(9): p. 3419-24.
244. Morin, P.E., et al., *Elongation factor TFIIS contains three structural domains: solution structure of domain II.* Proc Natl Acad Sci U S A, 1996. **93**(20): p. 10604-8.
245. Booth, V., et al., *Structure of a conserved domain common to the transcription factors TFIIS, elongin A, and CRSP70.* J Biol Chem, 2000. **275**(40): p. 31266-8.
246. Cermakova, K., et al., *A ubiquitous disordered protein interaction module orchestrates transcription elongation.* Science, 2021. **374**(6571): p. 1113-1121.
247. Olmsted, V.K., et al., *Yeast transcript elongation factor (TFIIS), structure and function. I: NMR structural analysis of the minimal transcriptionally active region.* J Biol Chem, 1998. **273**(35): p. 22589-94.
248. Awrey, D.E., et al., *Yeast transcript elongation factor (TFIIS), structure and function. II: RNA polymerase binding, transcript cleavage, and read-through.* J Biol Chem, 1998. **273**(35): p. 22595-605.
249. Qian, X., et al., *Novel zinc finger motif in the basal transcriptional machinery: three-dimensional NMR studies of the nucleic acid binding domain of transcriptional elongation factor TFIIS.* Biochemistry, 1993. **32**(38): p. 9944-59.
250. Nakanishi, T., et al., *Structure-function relationship of yeast S-II in terms of stimulation of RNA polymerase II, arrest relief, and suppression of 6-azauracil sensitivity.* J Biol Chem, 1995. **270**(15): p. 8991-5.
251. Nakanishi, T., et al., *Purification, gene cloning, and gene disruption of the transcription elongation factor S-II in Saccharomyces cerevisiae.* J Biol Chem, 1992. **267**(19): p. 13200-4.
252. Koyama, H., et al., *Transcription elongation factor S-II maintains transcriptional fidelity and confers oxidative stress resistance.* Genes Cells, 2003. **8**(10): p. 779-88.
253. Qian, X., et al., *Structure of a new nucleic-acid-binding motif in eukaryotic transcriptional elongation factor TFIIS.* Nature, 1993. **365**(6443): p. 277-9.
254. Jeon, C., H. Yoon, and K. Agarwal, *The transcription factor TFIIS zinc ribbon dipeptide Asp-Glu is critical for stimulation of elongation and RNA cleavage by RNA polymerase II.* Proc Natl Acad Sci U S A, 1994. **91**(19): p. 9106-10.
255. Pan, G., T. Aso, and J. Greenblatt, *Interaction of elongation factors TFIIS and elongin A with a human RNA polymerase II holoenzyme capable of promoter-specific initiation and responsive to transcriptional activators.* J Biol Chem, 1997. **272**(39): p. 24563-71.
256. Kim, B., et al., *The transcription elongation factor TFIIS is a component of RNA polymerase II preinitiation complexes.* Proc Natl Acad Sci U S A, 2007. **104**(41): p. 16068-73.
257. Wery, M., et al., *Members of the SAGA and Mediator complexes are partners of the transcription elongation factor TFIIS.* EMBO J, 2004. **23**(21): p. 4232-42.
258. Cermakova, K., V. Veverka, and H.C. Hodges, *The TFIIS N-terminal domain (TND): a transcription assembly module at the interface of order and disorder.* Biochem Soc Trans, 2023. **51**(1): p. 125-135.
259. Shilatifard, A., et al., *An RNA polymerase II elongation factor encoded by the human ELL gene.* Science, 1996. **271**(5257): p. 1873-6.
260. Ryu, S., et al., *The transcriptional cofactor complex CRSP is required for activity of the enhancer-binding protein Sp1.* Nature, 1999. **397**(6718): p. 446-50.
261. Ling, Y., A.J. Smith, and G.T. Morgan, *A sequence motif conserved in diverse nuclear proteins identifies a protein interaction domain utilised for nuclear targeting by human TFIIS.* Nucleic Acids Res, 2006. **34**(8): p. 2219-29.

262. Kettenberger, H., K.J. Armache, and P. Cramer, *Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage.* Cell, 2003. **114**(3): p. 347-57.
263. Kettenberger, H., K.J. Armache, and P. Cramer, *Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS.* Mol Cell, 2004. **16**(6): p. 955-65.
264. Wang, D., et al., *Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution.* Science, 2009. **324**(5931): p. 1203-6.
265. Cheung, A.C. and P. Cramer, *Structural basis of RNA polymerase II backtracking, arrest and reactivation.* Nature, 2011. **471**(7337): p. 249-53.
266. Bar-Nahum, G., et al., *A ratchet mechanism of transcription elongation and its control.* Cell, 2005. **120**(2): p. 183-93.
267. Wang, D., et al., *Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis.* Cell, 2006. **127**(5): p. 941-54.
268. Weilbaecher, R.G., et al., *Intrinsic transcript cleavage in yeast RNA polymerase II elongation complexes.* J Biol Chem, 2003. **278**(26): p. 24189-99.
269. Sosunov, V., et al., *Unified two-metal mechanism of RNA synthesis and degradation by RNA polymerase.* EMBO J, 2003. **22**(9): p. 2234-44.
270. Cramer, P., et al., *Architecture of RNA polymerase II and implications for the transcription mechanism.* Science, 2000. **288**(5466): p. 640-9.
271. Hausner, W., U. Lange, and M. Musfeldt, *Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase.* J Biol Chem, 2000. **275**(17): p. 12393-9.
272. Laptenko, O., et al., *Transcript cleavage factors GreA and GreB act as transient catalytic components of RNA polymerase.* EMBO J, 2003. **22**(23): p. 6322-34.
273. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics.* Annu Rev Genet, 2005. **39**: p. 309-38.
274. Hirashima, S., et al., *Molecular cloning and characterization of cDNA for eukaryotic transcription factor S-II.* J Biol Chem, 1988. **263**(8): p. 3858-63.
275. Marshall, T.K., H. Guo, and D.H. Price, *Drosophila RNA polymerase II elongation factor DmS-II has homology to mouse S-II and sequence similarity to yeast PPR2.* Nucleic Acids Res, 1990. **18**(21): p. 6293-8.
276. Plant, K.E., A. Hair, and G.T. Morgan, *Genes encoding isoforms of transcription elongation factor TFIIS in Xenopus and the use of multiple unusual RNA processing signals.* Nucleic Acids Res, 1996. **24**(18): p. 3514-21.
277. Xu, Q., et al., *Cloning and identification of testis-specific transcription elongation factor S-II.* J Biol Chem, 1994. **269**(4): p. 3100-3.
278. Chen, H.C., L. England, and C.M. Kane, *Characterization of a HeLa cDNA clone encoding the human SII protein, an elongation factor for RNA polymerase II.* Gene, 1992. **116**(2): p. 253-8.
279. Park, H., et al., *Characterization of the gene encoding the human transcriptional elongation factor TFIIS.* Gene, 1994. **139**(2): p. 263-7.
280. Weaver, Z.A. and C.M. Kane, *Genomic characterization of a testis-specific TFIIS (TCEA2) gene.* Genomics, 1997. **46**(3): p. 516-9.
281. Umehara, T., et al., *Isolation and characterization of a cDNA encoding a new type of human transcription elongation factor S-II.* Gene, 1995. **167**(1-2): p. 297-302.
282. Kanai, A., et al., *Heterogeneity and tissue-specific expression of eukaryotic transcription factor S-II-related protein mRNA.* J Biochem, 1991. **109**(5): p. 674-7.
283. Yoo, O.J., et al., *Cloning, expression and characterization of the human transcription elongation factor, TFIIS.* Nucleic Acids Res, 1991. **19**(5): p. 1073-9.
284. Labhart, P. and G.T. Morgan, *Identification of novel genes encoding transcription elongation factor TFIIS (TCEA) in vertebrates: conservation of three distinct TFIIS isoforms in frog, mouse, and human.* Genomics, 1998. **52**(3): p. 278-88.
285. Cunningham, F., et al., *Ensembl 2022.* Nucleic Acids Res, 2022. **50**(D1): p. D988-D995.
286. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this thesis were obtained from the GTEx Portal in summer 2020.
287. UniProt, C., *UniProt: the Universal Protein Knowledgebase in 2023.* Nucleic Acids Res, 2023. **51**(D1): p. D523-D531.

288. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

289. Robert, X. and P. Gouet, *Deciphering key features in protein structures with the new ENDscript server.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W320-4.

290. Ito, T., et al., *Transcription elongation factor S-II is required for definitive hematopoiesis.* Mol Cell Biol, 2006. **26**(8): p. 3194-203.

291. Saso, K., et al., *Identification of a novel tissue-specific transcriptional activator FESTA as a protein that interacts with the transcription elongation factor S-II.* J Biochem, 2003. **133**(4): p. 493-500.

292. Yang, T., et al., *TCEA1 regulates the proliferative potential of mouse myeloid cells.* Exp Cell Res, 2018. **370**(2): p. 551-560.

293. Hubbard, K., et al., *Knockdown of TFIIS by RNA silencing inhibits cancer cell proliferation and induces apoptosis.* BMC Cancer, 2008. **8**: p. 133.

294. You, S., et al., *Abnormal expression of YEATS4 associates with poor prognosis and promotes cell proliferation of hepatic carcinoma cell by regulation the TCEA1/DDX3 axis.* Am J Cancer Res, 2018. **8**(10): p. 2076-2087.

295. Ito, T., et al., *Spermatocyte-specific expression of the gene for mouse testis-specific transcription elongation factor S-II.* FEBS Lett, 1996. **385**(1-2): p. 21-4.

296. Nakata, A., et al., *GRIP1tau, a novel PDZ domain-containing transcriptional activator, cooperates with the testis-specific transcription elongation factor SII-T1.* Genes Cells, 2004. **9**(11): p. 1125-35.

297. Hill, S.J., et al., *Systematic screening reveals a role for BRCA1 in the response to transcription-associated DNA damage.* Genes Dev, 2014. **28**(17): p. 1957-75.

298. Park, K.S., et al., *Transcription elongation factor Tcea3 regulates the pluripotent differentiation potential of mouse embryonic stem cells via the Lefty1-Nodal-Smad2 pathway.* Stem Cells, 2013. **31**(2): p. 282-92.

299. Cha, Y., et al., *Tcea3 regulates the vascular differentiation potential of mouse embryonic stem cells.* Gene Expr, 2013. **16**(1): p. 25-30.

300. Xu, X.Q., et al., *Global expression profile of highly enriched cardiomyocytes derived from human embryonic stem cells.* Stem Cells, 2009. **27**(9): p. 2163-74.

301. Zhu, Y., et al., *Effect of TCEA3 on the differentiation of bovine skeletal muscle satellite cells.* Biochem Biophys Res Commun, 2017. **484**(4): p. 827-832.

302. Kazim, N., A. Adhikari, and J. Davie, *The transcription elongation factor TCEA3 promotes the activity of the myogenic regulatory factors.* PLoS One, 2019. **14**(6): p. e0217680.

303. Cha, Y., et al., *TCEA3 binds to TGF-beta receptor I and induces Smad-independent, JNK-dependent apoptosis in ovarian cancer cells.* Cell Signal, 2013. **25**(5): p. 1245-51.

304. Liao, J.M., et al., *TFIIS.h, a new target of p53, regulates transcription efficiency of pro-apoptotic bax gene.* Sci Rep, 2016. **6**: p. 23542.

305. Li, J., et al., *TCEA3 Attenuates Gastric Cancer Growth by Apoptosis Induction.* Med Sci Monit, 2015. **21**: p. 3241-6.

306. Adelman, K., et al., *Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS.* Mol Cell, 2005. **17**(1): p. 103-12.

307. Churchman, L.S. and J.S. Weissman, *Nascent transcript sequencing visualizes transcription at nucleotide resolution.* Nature, 2011. **469**(7330): p. 368-73.

308. Zatreanu, D., et al., *Elongation Factor TFIIS Prevents Transcription Stress and R-Loop Accumulation to Maintain Genome Stability.* Mol Cell, 2019. **76**(1): p. 57-69 e9.

309. Schwalb, B., et al., *TT-seq maps the human transient transcriptome.* Science, 2016. **352**(6290): p. 1225-8.

310. Nabet, B., et al., *The dTAG system for immediate and target-specific protein degradation.* Nat Chem Biol, 2018. **14**(5): p. 431-441.

311. Bressin, A., et al., *High-sensitive nascent transcript sequencing reveals direct BRD4-specific control of widespread enhancer and target gene transcription*. In revision.

312. Cunningham, F., et al., *Ensembl 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D745-D751.

313. Clark, K., et al., *GenBank.* Nucleic Acids Res, 2016. **44**(D1): p. D67-72.

314. Potter, S.C., et al., *HMMER web server: 2018 update.* Nucleic Acids Res, 2018. **46**(W1): p. W200-W204.

315. Katoh, K., et al., *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res, 2002. **30**(14): p. 3059-66.

316. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.

317. Andrews, B.J., et al., *The FLP recombinase of the 2 micron circle DNA of yeast: interaction with its target sequences.* Cell, 1985. **40**(4): p. 795-803.

318. Goldberg, M.F., *Retinal vaso-occlusion in sickling hemoglobinopathies.* Birth Defects Orig Artic Ser, 1976. **12**(3): p. 475-515.

319. Postle, K., T.T. Nguyen, and K.P. Bertrand, *Nucleotide sequence of the repressor gene of the TN10 tetracycline resistance determinant.* Nucleic Acids Res, 1984. **12**(12): p. 4849-63.

320. Hillen, W. and C. Berens, *Mechanisms underlying expression of Tn10 encoded tetracycline resistance.* Annu Rev Microbiol, 1994. **48**: p. 345-69.

321. Hillen, W., et al., *Control of expression of the Tn10-encoded tetracycline resistance genes. Equilibrium and kinetic investigation of the regulatory reactions.* J Mol Biol, 1983. **169**(3): p. 707-21.

322. Concordet, J.P. and M. Haeussler, *CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens.* Nucleic Acids Res, 2018. **46**(W1): p. W242-W245.

323. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nat Protoc, 2013. **8**(11): p. 2281-2308.

324. Sakuma, T., et al., *MMEJ-assisted gene knock-in using TALENs and CRISPR-Cas9 with the PITCh systems.* Nat Protoc, 2016. **11**(1): p. 118-33.

325. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.

326. Tyanova, S., et al., *The Perseus computational platform for comprehensive analysis of (prote)omics data.* Nat Methods, 2016. **13**(9): p. 731-40.

327. Ramirez, P., et al., *R-Loop Analysis by Dot-Blot.* J Vis Exp, 2021(167).

328. Mohammed, H., et al., *Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes.* Nat Protoc, 2016. **11**(2): p. 316-26.

329. Goedhart, J. and M.S. Luijsterburg, *VolcaNoseR is a web app for creating, exploring, labeling and sharing volcano plots.* Sci Rep, 2020. **10**(1): p. 20560.

330. Ge, S.X., D. Jung, and R. Yao, *ShinyGO: a graphical gene-set enrichment tool for animals and plants.* Bioinformatics, 2020. **36**(8): p. 2628-2629.

331. Baluapuri, A., et al., *MYC Recruits SPT5 to RNA Polymerase II to Promote Processive Transcription Elongation.* Mol Cell, 2019. **74**(4): p. 674-687 e11.

332. MARTIN, M. *Cutadapt removes adapter sequences from high-throughput sequencing reads*. 2011. **v. 17, n. 1,**, DOI: 10.14806/ej. 17.1. 200.

333. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

334. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

335. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.

336. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

337. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

338. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nat Methods, 2017. **14**(4): p. 417-419.

339. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

340. Kamburov, A., et al., *ConsensusPathDB--a database for integrating human functional interaction networks.* Nucleic Acids Res, 2009. **37**(Database issue): p. D623-8.

341. Kamburov, A., et al., *ConsensusPathDB: toward a more complete picture of cell biology.* Nucleic Acids Res, 2011. **39**(Database issue): p. D712-7.

342. Zorita, E., P. Cusco, and G.J. Filion, *Starcode: sequence clustering based on all-pairs search.* Bioinformatics, 2015. **31**(12): p. 1913-9.

343. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.

344. Carmona, R., et al., *Automated identification of reference genes based on RNA-seq data.* Biomed Eng Online, 2017. **16**(Suppl 1): p. 65.

345. Holland, L.Z. and D. Ocampo Daza, *A new look at an old question: when did the second whole genome duplication occur in vertebrate evolution?* Genome Biol, 2018. **19**(1): p. 209.

346. Fouqueau, T., et al., *The transcript cleavage factor paralogue TFS4 is a potent RNA polymerase inhibitor.* Nat Commun, 2017. **8**(1): p. 1914.

347. Brazeau, M.D. and M. Friedman, *The origin and early phylogenetic history of jawed vertebrates.* Nature, 2015. **520**(7548): p. 490-7.

348. Winter, G.E., et al., *DRUG DEVELOPMENT. Phthalimide conjugation as a strategy for in vivo target protein degradation.* Science, 2015. **348**(6241): p. 1376-81.

349. Kronke, J., et al., *Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells.* Science, 2014. **343**(6168): p. 301-5.

350. Cox, J., et al., *Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.* Mol Cell Proteomics, 2014. **13**(9): p. 2513-26.

351. Mellacheruvu, D., et al., *The CRAPome: a contaminant repository for affinity purification-mass spectrometry data.* Nat Methods, 2013. **10**(8): p. 730-6.

352. Xu, Y., et al., *Architecture of the RNA polymerase II-Paf1C-TFIIS transcription elongation complex.* Nat Commun, 2017. **8**: p. 15741.

353. Filipovski, M., et al., *Structural basis of nucleosome retention during transcription elongation.* Science, 2022. **376**(6599): p. 1313-1316.

354. Jang, Y., et al., *Intrinsically disordered protein RBM14 plays a role in generation of RNA:DNA hybrids at double-strand break sites.* Proc Natl Acad Sci U S A, 2020. **117**(10): p. 5329-5338.

355. Zheng, T., et al., *RBMX is required for activation of ATR on repetitive DNAs to maintain genome stability.* Cell Death Differ, 2020. **27**(11): p. 3162-3176.

356. Cargill, M., R. Venkataraman, and S. Lee, *DEAD-Box RNA Helicases and Genome Stability.* Genes (Basel), 2021. **12**(10).

357. Carriere, L., et al., *Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells.* Nucleic Acids Res, 2012. **40**(1): p. 270-83.

358. Ghavi-Helm, Y., et al., *Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription.* Genes Dev, 2008. **22**(14): p. 1934-47.

359. Chedin, S., et al., *The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIS-like subunit and is important for transcription termination.* Genes Dev, 1998. **12**(24): p. 3857-71.

360. Liang, K., et al., *Targeting Processive Transcription Elongation via SEC Disruption for MYC-Induced Cancer Therapy.* Cell, 2018. **175**(3): p. 766-779 e17.

361. Saldi, T., N. Fong, and D.L. Bentley, *Transcription elongation rate affects nascent histone pre-mRNA folding and 3' end processing.* Genes Dev, 2018. **32**(3-4): p. 297-308.

362. Tellier, M., et al., *CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation.* Nucleic Acids Res, 2020. **48**(14): p. 7712-7727.

363. Karabacak Calviello, A., et al., *Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling.* Genome Biol, 2019. **20**(1): p. 42.

364. Fox, M.H., *A model for the computer analysis of synchronous DNA distributions obtained by flow cytometry.* Cytometry, 1980. **1**(1): p. 71-7.

365. Bjursell, G. and P. Reichard, *Effects of thymidine on deoxyribonucleoside triphosphate pools and deoxyribonucleic acid synthesis in Chinese hamster ovary cells.* J Biol Chem, 1973. **248**(11): p. 3904-9.

366. Schvartzman, J.B., D.B. Krimer, and J. Van't Hof, *The effects of different thymidine concentrations on DNA replication in pea-root cells synchronized by a protracted 5-fluorodeoxyuridine treatment.* Exp Cell Res, 1984. **150**(2): p. 379-89.

367. Chen, G. and X. Deng, *Cell Synchronization by Double Thymidine Block.* Bio Protoc, 2018. **8**(17).

368. Naama, J.K., et al., *Prevention of immune precipitation by purified components of the alternative pathway.* Clin Exp Immunol, 1985. **60**(1): p. 169-77.
369. Pommier, Y., *Topoisomerase I inhibitors: camptothecins and beyond.* Nat Rev Cancer, 2006. **6**(10): p. 789-802.
370. Marechal, A. and L. Zou, *DNA damage sensing by the ATM and ATR kinases.* Cold Spring Harb Perspect Biol, 2013. **5**(9).
371. Lavin, M.F. and N. Gueven, *The complexity of p53 stabilization and activation.* Cell Death Differ, 2006. **13**(6): p. 941-50.
372. Crossley, M.P., M. Bocek, and K.A. Cimprich, *R-Loops as Cellular Regulators and Genomic Threats.* Mol Cell, 2019. **73**(3): p. 398-411.
373. Marteijn, J.A., et al., *Understanding nucleotide excision repair and its roles in cancer and ageing.* Nat Rev Mol Cell Biol, 2014. **15**(7): p. 465-81.
374. Zhu, Q., et al., *USP7-mediated deubiquitination differentially regulates CSB but not UVSSA upon UV radiation-induced DNA damage.* Cell Cycle, 2020. **19**(1): p. 124-141.
375. Liu, X., et al., *Trip12 is an E3 ubiquitin ligase for USP7/HAUSP involved in the DNA damage response.* FEBS Lett, 2016. **590**(23): p. 4213-4222.
376. van der Weegen, Y., et al., *ELOF1 is a transcription-coupled DNA repair factor that directs RNA polymerase II ubiquitylation.* Nat Cell Biol, 2021. **23**(6): p. 595-607.
377. Geijer, M.E., et al., *Publisher Correction: Elongation factor ELOF1 drives transcription-coupled repair and prevents genome instability.* Nat Cell Biol, 2021. **23**(7): p. 809.
378. Nozawa, R.S., et al., *Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway.* Nat Struct Mol Biol, 2013. **20**(5): p. 566-73.
379. Jansz, N., et al., *Smchd1 regulates long-range chromatin interactions on the inactive X chromosome and at Hox clusters.* Nat Struct Mol Biol, 2018. **25**(9): p. 766-777.
380. Wang, C.Y., et al., *SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X.* Cell, 2018. **174**(2): p. 406-421 e25.
381. Gdula, M.R., et al., *The non-canonical SMC protein SmcHD1 antagonises TAD formation and compartmentalisation on the inactive X chromosome.* Nat Commun, 2019. **10**(1): p. 30.
382. Coker, H. and N. Brockdorff, *SMCHD1 accumulates at DNA damage sites and facilitates the repair of DNA double-strand breaks.* J Cell Sci, 2014. **127**(Pt 9): p. 1869-74.
383. Tang, M., et al., *Structural maintenance of chromosomes flexible hinge domain containing 1 (SMCHD1) promotes non-homologous end joining and inhibits homologous recombination repair upon DNA damage.* J Biol Chem, 2014. **289**(49): p. 34024-32.
384. Vancevska, A., et al., *SMCHD1 promotes ATM-dependent DNA damage signaling and repair of uncapped telomeres.* EMBO J, 2020. **39**(7): p. e102668.
385. Roffey, S.E. and D.W. Litchfield, *CK2 Regulation: Perspectives in 2021.* Biomedicines, 2021. **9**(10).
386. Keller, D.M., et al., *A DNA damage-induced p53 serine 392 kinase complex contains CK2, hSpt16, and SSRP1.* Mol Cell, 2001. **7**(2): p. 283-92.
387. Pal, A., H.M. Greenblatt, and Y. Levy, *Prerecognition Diffusion Mechanism of Human DNA Mismatch Repair Proteins along DNA: Msh2-Msh3 versus Msh2-Msh6.* Biochemistry, 2020. **59**(51): p. 4822-4832.
388. Cheng, H., N. Zhang, and D. Pati, *Cohesin subunit RAD21: From biology to disease.* Gene, 2020. **758**: p. 144966.
389. Kobayashi, J., *Molecular mechanism of the recruitment of NBS1/hMRE11/hRAD50 complex to DNA double-strand breaks: NBS1 binds to gamma-H2AX through FHA/BRCT domain.* J Radiat Res, 2004. **45**(4): p. 473-8.
390. Huang, R.X. and P.K. Zhou, *DNA damage response signaling pathways and targets for radiotherapy sensitization in cancer.* Signal Transduct Target Ther, 2020. **5**(1): p. 60.
391. Ray Chaudhuri, A. and A. Nussenzweig, *The multifaceted roles of PARP1 in DNA repair and chromatin remodelling.* Nat Rev Mol Cell Biol, 2017. **18**(10): p. 610-621.
392. Aragon, L., *The Smc5/6 Complex: New and Old Functions of the Enigmatic Long-Distance Relative.* Annu Rev Genet, 2018. **52**: p. 89-107.
393. Griffin, W.C. and M.A. Trakselis, *The MCM8/9 complex: A recent recruit to the roster of helicases involved in genome maintenance.* DNA Repair (Amst), 2019. **76**: p. 1-10.

394. Shaw, G., et al., *Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells.* FASEB J, 2002. **16**(8): p. 869-71.

395. Lin, Y.C., et al., *Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations.* Nat Commun, 2014. **5**: p. 4767.

396. Gruenwald, P., *Embryonic and postnatal development of the adrenal cortex, particularly the zona glomerulosa and accessory nodules.* Anat Rec, 1946. **95**: p. 391-421.

397. Menon, S.G. and P.C. Goswami, *A redox cycle within the cell cycle: ring in the old with the new.* Oncogene, 2007. **26**(8): p. 1101-9.

398. Shamseddine, A.A., M.V. Airola, and Y.A. Hannun, *Roles and regulation of neutral sphingomyelinase-2 in cellular and pathological processes.* Adv Biol Regul, 2015. **57**: p. 24-41.

399. Saponaro, F., A. Saba, and R. Zucchi, *An Update on Vitamin D Metabolism.* Int J Mol Sci, 2020. **21**(18).

400. Ikeda, Y., et al., *Vasorin, a transforming growth factor beta-binding protein expressed in vascular smooth muscle cells, modulates the arterial response to injury in vivo.* Proc Natl Acad Sci U S A, 2004. **101**(29): p. 10732-7.

401. Simons, K. and E. Ikonen, *Functional rafts in cell membranes.* Nature, 1997. **387**(6633): p. 569-72.

402. Van Brocklyn, J.R. and J.B. Williams, *The control of the balance between ceramide and sphingosine-1-phosphate by sphingosine kinase: oxidative stress and the seesaw of cell survival and death.* Comp Biochem Physiol B Biochem Mol Biol, 2012. **163**(1): p. 26-36.

403. Perry, D.K., L.M. Obeid, and Y.A. Hannun, *Ceramide and the regulation of apoptosis and the stress response.* Trends Cardiovasc Med, 1996. **6**(5): p. 158-62.

404. Spiegel, S., et al., *Roles of sphingosine-1-phosphate in cell growth, differentiation, and death.* Biochemistry (Mosc), 1998. **63**(1): p. 69-73.

405. Zhu, H., et al., *Asah2 Represses the p53-Hmox1 Axis to Protect Myeloid-Derived Suppressor Cells from Ferroptosis.* J Immunol, 2021. **206**(6): p. 1395-1404.

406. Li, J., et al., *Ferroptosis: past, present and future.* Cell Death Dis, 2020. **11**(2): p. 88.

407. Van de Peer, Y., S. Maere, and A. Meyer, *The evolutionary significance of ancient genome duplications.* Nat Rev Genet, 2009. **10**(10): p. 725-32.

408. Maliszewska-Olejniczak, K., et al., *TFIIS-Dependent Non-coding Transcription Regulates Developmental Genome Rearrangements.* PLoS Genet, 2015. **11**(7): p. e1005383.

409. Smith, J.J., et al., *Programmed loss of millions of base pairs from a vertebrate genome.* Proc Natl Acad Sci U S A, 2009. **106**(27): p. 11212-7.

410. Pagano, J.M., et al., *Defining NELF-E RNA binding in HIV-1 and promoter-proximal pause regions.* PLoS Genet, 2014. **10**(1): p. e1004090.

411. Galganski, L., M.O. Urbanek, and W.J. Krzyzosiak, *Nuclear speckles: molecular organization, biological function and role in disease.* Nucleic Acids Res, 2017. **45**(18): p. 10350-10368.

412. Kai, M., *Roles of RNA-Binding Proteins in DNA Damage Response.* Int J Mol Sci, 2016. **17**(3): p. 310.

413. Li, J., et al., *Rbm14 maintains the integrity of genomic DNA during early mouse embryogenesis via mediating alternative splicing.* Cell Prolif, 2020. **53**(1): p. e12724.

414. Zhou, L.T., et al., *A novel role of fragile X mental retardation protein in pre-mRNA alternative splicing through RNA-binding protein 14.* Neuroscience, 2017. **349**: p. 64-75.

415. Liu, N., et al., *N6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein.* Nucleic Acids Res, 2017. **45**(10): p. 6051-6063.

416. Zhou, K.I., et al., *Regulation of Co-transcriptional Pre-mRNA Splicing by m(6)A through the Low-Complexity Protein hnRNPG.* Mol Cell, 2019. **76**(1): p. 70-81 e9.

417. Mo, J., et al., *DDX3X: structure, physiologic functions and cancer.* Mol Cancer, 2021. **20**(1): p. 38.

418. Prather, D.M., E. Larschan, and F. Winston, *Evidence that the elongation factor TFIIS plays a role in transcription initiation at GAL1 in Saccharomyces cerevisiae.* Mol Cell Biol, 2005. **25**(7): p. 2650-9.

419. Guglielmi, B., et al., *TFIIS elongation factor and Mediator act in conjunction during transcription initiation in vivo.* Proc Natl Acad Sci U S A, 2007. **104**(41): p. 16062-7.

420. Fadok, V.A., et al., *Exposure of phosphatidylserine on the surface of apoptotic lymphocytes triggers specific recognition and removal by macrophages.* J Immunol, 1992. **148**(7): p. 2207-16.

421. Koopman, G., et al., *Annexin V for flow cytometric detection of phosphatidylserine expression on B cells undergoing apoptosis.* Blood, 1994. **84**(5): p. 1415-20.

422. Martin, S.J., et al., *Early redistribution of plasma membrane phosphatidylserine is a general feature of apoptosis regardless of the initiating stimulus: inhibition by overexpression of Bcl-2 and Abl.* J Exp Med, 1995. **182**(5): p. 1545-56.

423. Zullig, S., et al., *Aminophospholipid translocase TAT-1 promotes phosphatidylserine exposure during C. elegans apoptosis.* Curr Biol, 2007. **17**(11): p. 994-9.

424. Graham, F.L., et al., *Characteristics of a human cell line transformed by DNA from human adenovirus type 5.* J Gen Virol, 1977. **36**(1): p. 59-74.

425. Louis, N., C. Evelegh, and F.L. Graham, *Cloning and sequencing of the cellular-viral junctions from the human adenovirus type 5 transformed 293 cell line.* Virology, 1997. **233**(2): p. 423-9.

426. Berk, A.J., *Recent lessons in gene expression, cell cycle control, and cell biology from adenovirus.* Oncogene, 2005. **24**(52): p. 7673-85.

427. Sha, J., et al., *E1A interacts with two opposing transcriptional pathways to induce quiescent cells into S phase.* J Virol, 2010. **84**(8): p. 4050-9.

428. Sheppard, H.M., et al., *New insights into the mechanism of inhibition of p53 by simian virus 40 large T antigen.* Mol Cell Biol, 1999. **19**(4): p. 2746-53.

429. Lilyestrom, W., et al., *Crystal structure of SV40 large T-antigen bound to p53: interplay between a viral oncoprotein and a cellular tumor suppressor.* Genes Dev, 2006. **20**(17): p. 2373-82.

430. Sherwood, S.W., et al., *Defining cellular senescence in IMR-90 cells: a flow cytometric analysis.* Proc Natl Acad Sci U S A, 1988. **85**(23): p. 9086-90.

431. Fernandez Larrosa, P.N., et al., *RAC3 more than a nuclear receptor coactivator: a key inhibitor of senescence that is downregulated in aging.* Cell Death Dis, 2015. **6**(10): p. e1902.

432. Vousden, K.H. and C. Prives, *Blinded by the Light: The Growing Complexity of p53.* Cell, 2009. **137**(3): p. 413-31.

433. Deppert, W., T. Steinmayer, and W. Richter, *Cooperation of SV40 large T antigen and the cellular protein p53 in maintenance of cell transformation.* Oncogene, 1989. **4**(9): p. 1103-10.

434. Pipas, J.M. and A.J. Levine, *Role of T antigen interactions with p53 in tumorigenesis.* Semin Cancer Biol, 2001. **11**(1): p. 23-30.

435. Wissink, E.M., et al., *Nascent RNA analyses: tracking transcription and its regulation.* Nat Rev Genet, 2019. **20**(12): p. 705-723.

436. Adelman, K. and J.T. Lis, *Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans.* Nat Rev Genet, 2012. **13**(10): p. 720-31.

437. Annaldasula, S., M. Gajos, and A. Mayer, *IsoTV: processing and visualizing functional features of translated transcript isoforms.* Bioinformatics, 2021. **37**(18): p. 3070-3072.

438. Alamancos, G.P., et al., *Leveraging transcript quantification for fast computation of alternative splicing profiles.* RNA, 2015. **21**(9): p. 1521-31.

Contributions

Mario Rubio processed all ChIP-seq, HiS-NET-seq, and RNA-seq, analyzed the data with input from Dr. Andreas Mayer and me. Together, we prepared figures 18, 20, 29, 30, 31, 32, 34, 35, S3, and S5. Martyna Gajos performed the initial analyses of RNA-seq and ChIP-seq experiments and, together, we prepared figures 5, 17, S2, and S5. Susanne Freier prepared the HiS-NET-seq libraries. All mass spectrometry analyses and raw data processing was done by the mass spectrometry facility of MPIMG. All sequencing was done by the sequencing facility of MPIMG. All cell sorting was performed by the FACS facility of MPIMG.

Relevant publications

The manuscript based on this thesis is currently in preparation.

Altendorfer, E.*, Mochalova, Y.*, and A. Mayer, *BRD4: a general regulator of transcription elongation.* Transcription, 2022. **13**(1-3): p. 70-81.

(*: equal contribution)

Appendix

Abbreviations

A	adenine
ATP	Adenosine triphosphate
bp	base pair
C	cytosine
cDNA	complementary DNA
circDNA	circularized DNA
ChIP	chromatin immunoprecipitation
ChIP-seq	chromatin immunoprecipitation and DNA sequencing
CRISPR	clustered regularly interspaced short palindromic repeats
CTD	C-terminal domain
DE	differential expression
DDR	DNA damage response
DKO	double knockout
DRN	5,6-dichlorobenzimidazole 1-β-D-ribofuranoside
dNTP	deoxy-nucleoside triphosphate
DSB	DNA double strand break
dTAG	degradation tag
EC	elongation complex
FC	fold change
FDR	false discovery rate
G	guanine
GB	gene body
GO	gene ontology
GTF	general transcription factors
HA	hemagglutinin
HBH	6xHistidine-Biotin-6xHistidine
HiS-NET-seq	High-sensitive native elongating transcript sequencing
hnRNP	heterogeneous nuclear ribonucleoproteins
IDR	intrinsically disordered region
IP	immunoprecipitation
kbp	kilobase pair
KI	knock-in
KO	knockout
LFQ	label-free quantification

mRNA	messenger RNA
MS	mass spectrometry
NELF	Negative elongation factor
NET-seq	Native elongating transcript sequencing
nt	nucleotide
NTD	N-terminal domain
NTDL	N-terminal domain and subsequent linker of TCEA1 or TCEA2
NTP	nucleoside triphosphate
PIC	pre-initiation complex
PPR	promoter-proximal region
pA	polyadenylation
PAS	polyadenylation site
PCR	polymerase chain reaction
Pol II	RNA polymerase II
PROTAC	proteolysis targeting chimera
PuroR	Puromycin resistance gene
P-TEFb	positive transcription elongation factor b
RT	room temperature or reverse transcription
SDS	sodium dodecyl sulfate
SI	spike-in
SILAC	stable isotope labeling by amino acids in cell culture
T	thymine
T_a	annealing temperature
TCEA1	Transcription Elongation Factor A1
TCEA2	Transcription Elongation Factor A2
TCEA3	Transcription Elongation Factor A3
TCEANC	Transcription Elongation Factor A N-Terminal and Central Domain Containing
TF	transcription factor
TSS	transcription start site
U	uracil
UMI	unique molecular identifier
WB	Western blot
WT	wild-type (unmodified)
4sU	4-thiouridine

List of figures

List of tables