PhD Thesis

Investigation of the Impact of the RNA Exosome on Cellular Transcriptome and Function

Mengjun Wu

Main Supervisor: Prof. Dr. Albin Sandelin
Co-supervisor: Prof. Dr. Torben Heick Jensen

March 7, 2021
To BS
Abstract

RNA degradation is an integral part in RNA metabolism and plays an important role in determining cellular steady state RNA levels. Here, we focus on an RNA degradation machine – the RNA exosome. It is a highly conserved 3’-5’ ribonucleolytic protein complex and a main player in eukaryotic nuclear RNA turnover. In the nucleus, its substrates are composed of a variety of species, including a multitude of long non-coding RNAs (lncRNAs), such as PROMoter uPstream Transcripts (PROMPTs), enhancer RNAs (eRNAs) and several stable nuclear RNAs such as ribosomal RNAs (rRNAs), and small nuclear RNAs (snRNAs). In the nucleoplasm of mammalian cells, the substrate specificity of the RNA exosome is achieved through two major adaptors – the nuclear exosome targeting (NEXT) complex and the poly(A) exosome targeting (PAXT) connection.

In this thesis, we investigated the regulation and functions of the RNA exosome from three aspects: the role of the RNA exosome in shaping transcriptome derived from protein-coding (pc) genes, the contribution of molecular features to the RNA exosome degradation pathways, and the function of the RNA exosome in embryonic stem cell (ESC) development.

Employing several complementary genome-wide techniques, we identified and characterized a number of exosome-sensitive transcripts produced within pc genes. We identified two types of genes that utilize a single annotated transcription start site (TSS), where the first type produces mainly exosome-sensitive full-length transcripts, and the second type mainly prematurely terminated transcripts. Many genes of the former type are immediate early genes and encode transcription factors; whereas many genes of the latter type are in head-to-head configurations with other pc genes, and likely transcribed due to strong transcription initiation of the genes on the upstream opposite strand. For genes with multiple active TSSs, the TSSs producing exosome-sensitive transcripts only have minor contributions to the overall gene expression, and the produced transcripts are often prematurely terminated. Our results revealed a complex sense transcription landscape within pc genes shaped by the RNA exosome, and suggest the nuclear RNA exosome plays
Abstract

a role in regulating the expression of some important pc genes.

Using machine learning approaches, we delineated molecular determinants of nuclear exosome degradation pathways. We found the molecular features of the transcript end site (TES) to be most predictive for distinguishing the NEXT and PAXT pathways, while TSS related features were only found to be distinct for NEXT targets but not able to distinguish PAXT from non-exosome targets.

By knockout of the PAXT component ZFC3H1, we observed impaired mouse ESC differentiation. In Zfc3h1−/− cells, besides known PAXT substrates, many polycomb repressive complex 2 (PRC2)-repressed genes were upregulated, accompanied with a decreased PRC2 binding and reduced H3K27me3 levels. Integrity of the PRC2 complex was observed to be decreased with increased levels of nonspecific RNA bound to PRC2. These results underscore the importance of controlling nuclear RNA levels during ESC development and suggest a potential way to regulate transcription by bulk RNA.
Dansk Resumé

Nedbrydning af RNA er en central del af RNA-stofskiftet og spiller en vigtig rolle i at regulere ligevægtsmængden af RNA i cellen. Her fokuserer vi på en bestemt mekanisme for nedbrydning af RNA, kaldet RNA exosomet (the RNA exosome). RNA exosomet er et konserveret 3’-5’ ribonukleolytisk proteinkompleks. I cellekerne er substraterne for RNA exosomet en række forskellige klasser af RNA, deriblandt ustabile RNA som long non-coding RNAs (lncRNAs), PROMoter uPstream Transcripts (PROMPTs), enhancer RNAs (eRNAs), og stabile RNA som ribosomal RNAs (rRNAs) og small nuclear RNAs (snRNAs). I den mammale cellekerne styres substratspecificiteten for RNA exosomet via to adapterer: Det cellekerne-rettede exosom kompleks (the nuclear exosome targeting (NEXT) complex) og den poly(A) exosom-rettede forbindelse (poly(A) exosome targeting (PAXT) connection).

I denne PhD-afhandling undersøgte vi, hvordan RNA exosomet fungerer og er reguleret fra tre forskellige perspektiver: Dets rolle i omsætning af mRNA transskriptommet, hvordan molekylære strukturer af RNA har indflydelse på nedbrydning, og dets betydning for differentiering af embryonale stamceller (ESC).

Vi brugte flere komplimentære genomiske teknikker til at identificere og karakterisere exosom-følsomme transkripter, der bliver produceret fra protein-kodende gener. Vi identificerede to typer af gener, der benytter et enkelt transkriptions startsted (Transcription Start Site, TSS). Den første type producerer primært exosom-følsomme og fuld-længe transkripter, og den anden type producerer primært for tidligt terminerede transkripter. Gener af den første type ofte tidlige responsgener og transkriptionsfaktorer. Gener fra den anden type er ofte modsatrettede med et nabogen: Det er sandsynligvis den stærke initiering af transkription ved dette nabogen, der driver deres transkription. For gener der bruger mere end et TSS, bidrager de TSSs, der producerer exosom-følsomme transkripter, kun i mindre grad til den overordnede ekspression fra gener. Disse transkripter er derudover ofte for tidligt terminerede. Til sammen kaster vores resultater lys over den vigtige rolle, RNA exosomet spiller i den komplekse regulering af transkription fra protein-
Dansk Resumé

Kodende gener, og tyder på at RNA exosomet kan være en faktor i at regulere ekspressionen fra vigtige protein-kodende gener.

Vi brugte maskinlæring til at finde den molekylære basis for nedbrydning af RNA i cellekernen via exosomet. Vi fandt, at den molekylære struktur ved transkriptions slutstedet (Transcription End Site, TES) er mest prædiktivt for at skelne mellem NEXT og PAXT nedbrydning. Derimod er den molekylære struktur ved TSS kun vigtigt for NEXT, men ikke vigtig for at adskille PAXT fra RNA, der ikke bliver nedbrudt via exosomet.

Vi observerede, at en knockout af PAXT-komponenten ZFC3HI hæmmer differntiering af murine ESCs. I Zfc3h1<sup>-/-</sup> celler er der, udover kendte PAXT-substrater, en opregulering af mange polycomb repressive complex 2 (PRC2) gener. Dette er forbundet med en nedregulering af PRC2 binding og H3K27me3 niveauer. Vi så, at stabiliteten af PRC2 komplekset falder i takt med stigende niveauer af ikke-specifik binding af RNA til PRC2. Tilsammen viser disse resultater vigtigheden af korrekt regulering af RNA i cellekernen for differentiering af ESCs, og peger på en mulig ny mekanisme for at regulere transkription via totalt RNA.
Publications included in this thesis

I Mengjun Wu, Evdoxia Karadoulama, Marta Lloret-Llinares, Jerome Olivier Rouviere, Christian Skov Vaagensø, Martin Moravec, Bingnan Li, Jingwen Wang, Guifen Wu, Maria Gockert, Vicent Pelechano, Torben Heick Jensen, Albin Sandelin.

The RNA exosome shapes the expression of key protein-coding genes

II Mengjun Wu, Manfred Schmid, Torben Heick Jensen, Albin Sandelin.

Identifying determinants of nuclear RNA exosome degradation pathways using machine learning approaches
Manuscript in Preparation


A functional link between nuclear RNA decay and transcriptional control mediated by the polycomb repressive complex 2
*These authors contributed equally
Contents

Paper I–III 3
Abstract i
Dansk Resumé iii
Publications included in this thesis v
Abbreviations ix

1 Introduction 1

1.1 Biogenesis of mRNA and ncRNA 2

1.1.1 mRNA 2

1.1.2 ncRNA 4

1.2 The RNA exosome 6

1.2.1 The structure of the RNA exosome complex 6

1.2.2 The nuclear RNA exosome degradation pathways 8

1.2.3 Substrates and mechanisms of nuclear RNA exosome degradation 9

1.2.4 Impact of the RNA exosome on cellular functions 11

1.3 Genome-wide techniques 12

1.3.1 Next-generation sequencing techniques 12

1.3.2 NGS data preparation 16

1.4 Machine learning techniques 16

1.4.1 General classification models 17

1.4.2 Random Forests 18

2 Present studies 21

2.1 Aim of the thesis 21
2.2 Paper I: The RNA exosome shapes the expression of key protein-coding genes ......................................................... 21

2.3 Paper II: Identifying determinants of nuclear RNA exosome degradation pathways using machine learning approaches ............................................ 23

2.4 Paper III: A functional link between nuclear RNA decay and transcriptional control mediated by the polycomb repressive complex ................................. 24

3 Conclusions and perspectives ........................................................................ 27

Acknowledgements ........................................................................................................ 29

References ......................................................................................................................... 31

Paper I–III .......................................................................................................................... 45
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5’ SS</td>
<td>5’ splice site</td>
</tr>
<tr>
<td>CAGE</td>
<td>cap analysis gene expression</td>
</tr>
<tr>
<td>CBC</td>
<td>cap binding complex</td>
</tr>
<tr>
<td>CBCA</td>
<td>CBC-ARS2</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CPA</td>
<td>cleavage and polyadenylation</td>
</tr>
<tr>
<td>CTD</td>
<td>carboxy-terminal domain</td>
</tr>
<tr>
<td>CTSS</td>
<td>CAGE tag start site</td>
</tr>
<tr>
<td>CUT</td>
<td>cryptic unstable transcripts</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>chromatin immunoprecipitation sequencing</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DSIF</td>
<td>DRB sensitivity-inducing factor</td>
</tr>
<tr>
<td>EB</td>
<td>embryoid body</td>
</tr>
<tr>
<td>eRNA</td>
<td>enhancer RNA</td>
</tr>
<tr>
<td>ESC</td>
<td>embryonic stem cell</td>
</tr>
<tr>
<td>FN</td>
<td>false negative</td>
</tr>
<tr>
<td>FP</td>
<td>false positive</td>
</tr>
<tr>
<td>GTF</td>
<td>general transcription factor</td>
</tr>
<tr>
<td>IncRNA</td>
<td>long non-coding RNA</td>
</tr>
<tr>
<td>m7G</td>
<td>7-methylguanosine</td>
</tr>
<tr>
<td>ML</td>
<td>machine learning</td>
</tr>
<tr>
<td>mNET-seq</td>
<td>mammalian NET-seq</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>ncRNA</td>
<td>non-coding RNA</td>
</tr>
<tr>
<td>NDR</td>
<td>nucleosome depleted region</td>
</tr>
<tr>
<td>NELF</td>
<td>negative elongation factor</td>
</tr>
<tr>
<td>NET-seq</td>
<td>Native Elongating Transcript sequencing</td>
</tr>
<tr>
<td>NEXT</td>
<td>nuclear exosome targeting</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>Description</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>NGS</td>
<td>next-generation sequencing</td>
</tr>
<tr>
<td>NNS</td>
<td>Nrd1-Nab3-Sen1</td>
</tr>
<tr>
<td>OOB</td>
<td>out-of-bag</td>
</tr>
<tr>
<td>P-TEFb</td>
<td>elongation factor-b</td>
</tr>
<tr>
<td>PABP</td>
<td>poly(A) tail binding protein</td>
</tr>
<tr>
<td>PAP</td>
<td>poly(A) polymerase</td>
</tr>
<tr>
<td>PAS</td>
<td>poly(A) site</td>
</tr>
<tr>
<td>PAXT</td>
<td>poly(A) exosome targeting</td>
</tr>
<tr>
<td>pc</td>
<td>protein-coding</td>
</tr>
<tr>
<td>PCPA</td>
<td>premature cleavage and polyadenylation</td>
</tr>
<tr>
<td>PIC</td>
<td>preinitiation complex</td>
</tr>
<tr>
<td>PPD</td>
<td>PABPN1 and PAP-mediated RNA decay</td>
</tr>
<tr>
<td>PRC2</td>
<td>polycomb repressive complex 2</td>
</tr>
<tr>
<td>PROMPT</td>
<td>PROMoter uPstream Transcript</td>
</tr>
<tr>
<td>PTT</td>
<td>premature transcription termination</td>
</tr>
<tr>
<td>Pol II</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>QC</td>
<td>quality control</td>
</tr>
<tr>
<td>RBP</td>
<td>RNA binding protein</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNP</td>
<td>ribonucleoprotein particle</td>
</tr>
<tr>
<td>RRM</td>
<td>RNA recognition motif</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>Saccharomyces cerevisiae</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>snRNP</td>
<td>small nuclear RNP</td>
</tr>
<tr>
<td>TES</td>
<td>transcript end site</td>
</tr>
<tr>
<td>TIF-seq</td>
<td>Transcript Isoform sequencing</td>
</tr>
<tr>
<td>TN</td>
<td>true negative</td>
</tr>
<tr>
<td>TP</td>
<td>true positive</td>
</tr>
<tr>
<td>TRAMP</td>
<td>Trf4/5-Air1/2-Mtr4 polyadenylation</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>TSS</td>
<td>transcription start site</td>
</tr>
<tr>
<td>TT-seq</td>
<td>transient transcriptome sequencing</td>
</tr>
<tr>
<td>uaRNA</td>
<td>upstream antisense RNA</td>
</tr>
<tr>
<td>WT</td>
<td>wild type</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The term transcriptome describes the ensemble of all types of ribonucleic acid (RNA) that are transcribed from deoxyribonucleic acid (DNA) by RNA polymerases. Based on several comprehensive transcriptome analyses across multiple cell lines using different techniques, about 75-90% of mammalian DNA is estimated to be transcribed, resulting in a large repertoire of RNAs [1–4]. Eukaryotic cells contain three types of RNA polymerases which produce distinct subsets of RNAs: RNA polymerase I synthesizes precursors of ribosomal RNA (pre-rRNA) except for 5S rRNAs, RNA polymerase II synthesizes precursors of messenger RNAs (pre-mRNAs), small nuclear RNAs and other less well-defined non-coding RNAs (ncRNAs), and RNA polymerase III synthesizes 5S rRNAs, transfer RNAs (tRNAs) and some other small RNAs [5, 6]. In this thesis, only RNA polymerase II (Pol II) transcribed RNAs are covered.

In a cell, the steady state RNA levels underlie the state of the cell and reflect its function. This steady state is determined by both transcription and degradation. As an essential cellular process for controlling RNA levels, the importance of RNA degradation is also reflected by the redundancy and complexity of the existing degradation pathways [7]. Here, we focus on a degradation machine – the RNA exosome, and set out to explore how the nuclear RNA exosome shapes transcriptome, establishes substrate specificity, and participates in regulation of cell differentiation using genomic and computational approaches.

In this chapter, I will first introduce the RNA species related to the thesis topic and their biogenesis, followed by discussing the RNA exosome including its structure, the nuclear RNA exosome pathways for selective degradation, the substrates and mechanisms of nuclear RNA exosome degradation, and the impact of the RNA exosome on cellular functions. Thereafter, I will summarize the genome-wide tech-
Chapter 1. Introduction

...techniques relevant to the thesis for profiling transcriptome and studying regulatory events. I will finish the chapter with an introduction to the machine learning techniques.

1.1 Biogenesis of mRNA and ncRNA

RNA is a single-stranded polymeric molecule made from ribose sugars, phosphates and four nucleobases, i.e., adenine (A), guanine (G), cytosine (C), and uracil (U) which replaces thymine (T) used in DNA. RNA is essential for all forms of life and exists in different types in cells. In this section, I will introduce two complementary RNA types: mRNA and ncRNA, with a focus on their biogenesis in the nucleus.

1.1.1 mRNA

mRNA is produced from protein-coding (pc) genes and, according to the central dogma, considered to be the intermediate between DNA and protein. In cells, the life cycle of mRNA undergoes several major events including: transcription, post-transcriptional processing, export, translation and degradation. In order to accurately transfer information encoded in the genome, these events are highly regulated and mRNA has to go through many processing steps important for its stability, subcellular location, and translation.

Transcription, as the first step, is the process by which pre-mRNA is synthesized from the DNA template of pc genes based on complementary base pairing. Transcription always starts from the 3’ end of the DNA template so that the RNA is synthesized in a 5’-3’ direction. There are three key stages in the transcription process, i.e., initiation, elongation and termination.

During transcription initiation, Pol II is recruited to a defined position on the DNA and initiates transcription. This position is called transcription start site (TSS), corresponding to the 5’ end of the subsequently transcribed RNA. The recruitment of the RNA polymerase II and the precise choice of TSS rely on binding of general transcription factors (GTFs) to a DNA sequence element called core promoter [8–11]. While the term promoter refers to an extended region upstream of TSS that contributes to transcription initiation, a core promoter refers to the region in the immediate proximity to the TSS required for GTF binding and initiaiion complex assembly, and is typically defined as a ±50 bp region around the TSS [10, 12]. Several types of core promoters are observed in Metazoan, e.g., core promoters with sharp initiation pattern and containing a TATA box and initiator
Chapter 1. Introduction

(INR) element, and core promoters with broad initiation and lacking a TATA box but with high G/C content or overlapping with CpG islands—genomic regions with long stretches rich in CpG dinucleotides, in contrast to other vertebrate genomic regions that are depleted of CpG dinucleotides. In the mammalian genome, the G/C-rich promoter is the more prevalent type [12–16].

The recruited Pol II forms a preinitiation complex (PIC) with GTFs and is activated for transcription [17]. After transcription starts, as soon as the 5’ end of the newly synthesized RNAs emerge from the Pol II exit channel, 7-methylguanosine (m7G) caps are added at their 5’ end [18]. At many genes, Pol II is subsequently observed to pause at regions 30-60 bp downstream of TSS, which is regulated by the core promoter element, negative elongation factor (NELF), DRB sensitivity-inducing factor (DSIF), and nucleosome [19–21]. To progress into productive elongation, activated positive transcription elongation factor-b (P-TEFb) is required to release paused Pol II [22].

Once Pol II is released from the pause site, it starts productive elongation. The newly transcribed pre-mRNA often does not exist on its own, instead it is packed into a ribonucleoprotein particle (RNP) by associating with RNA binding proteins (RBPs), which influence many important events in RNA life such as processing, nuclear export and translation [23]. The association of RBP and RNA is typically sequence-dependent and mediated through the recognition of short RNA sequence motifs [24]. During elongation, pre-mRNA often undergoes cotranscriptional splicing, a process in which introns are removed and exons are joined. The splicing reaction is catalyzed by a huge ribonucleoprotein complex called spliceosome, consisting of five small nuclear RNPs (snRNPs) and a large number of auxiliary proteins [25]. The assembly of the spliceosome on pre-mRNA relies on a set of consensus sequences that demarcate the intron, with a nearly invariant GU dinucleotide at the exon/intron boundary (5’ splice site) and AG at the 3’ end intron/exon boundary (3’ splice site) [26].

When approaching the end of the gene, transcription of Pol II needs to be terminated. For most pc genes in human cells, transcription termination is triggered by the recognition of poly(A) site (PAS) by Pol II and is tightly linked to the 3’ end processing of the nascent RNA. PAS is a hexamer with canonical sequence AAUAAA and a few minor variants [27, 28]. Once Pol II senses its passage through a functional PAS, the 3’ end cleavage and polyadenylation (CPA) complex is recruited to Pol II and cleaves the nascent RNA, which typically happens 10-30 bp downstream of the PAS [27, 29]. Following the cleavage, a poly(A) tail is added to the newly formed RNA 3’ end. The addition of the poly(A) tail releases the transcript from
Pol II [30]. Pol II will continue to transcribe further downstream before releasing the DNA template which, can be elicited by Pol II conformation change and/or Rat1/XRN2, a 5’-3’ exonuclease [29, 31]. The released transcript can undergo further processing events such as post-transcriptional splicing [32, 33] and matures into mRNA that will be exported into cytoplasm and translated into protein.

Degradation happens at many stages of the mRNA life cycle with the purpose of quality control (QC) or regulating its half-life. As discussed above, being a multi-step complex process, the biogenesis of mRNA is error-prone. QC is a critical step to ensure that mRNAs are accurately produced and correctly assembled. In nuclear RNA QC, RNA degradation pathways are often tightly coupled with transcription and RNA processing processes to remove aberrant transcripts; e.g., transcripts that fail to be capped after initiation are rapidly degraded by Rat1, and transcripts that are inefficient polyadenylated or poorly spliced are targeted by the RNA exosome [34–36]. While nuclear mRNA QC serves primarily the purpose of removing “wrong” transcripts, some QC processes could also be utilized to regulate gene expression, which will be discussed in detail in section 1.2.3. Processed mature mRNAs are usually promptly exported to the cytoplasm. Cytoplasmic RNA decay is often coupled with translation and, compared to nuclear RNA decay, it is considered to be the major contributor to regulating mRNA levels [7, 37].

1.1.2 ncRNA

As opposed to coding mRNA, the term ncRNA is generally used for RNA that does not translate into protein. While the biogenesis and function of some ncRNA species such as rRNA, tRNA, snRNA and microRNA have been well studied and characterized, those of a larger group of ncRNAs remain obscure. In the following, the term ncRNA is used to refer to the latter group of ncRNAs that are less well characterized and frequently discovered with the aid of high-throughput techniques.

Historically, many ncRNAs were first revealed by several large-scale transcriptome studies carried out by big consortiums like ENCODE and FANTOM. Pc genes were found to only comprise approximately 1% of the human genome, while the rest produces a large amount of ncRNAs [2, 4]. These studies used cell lines under normal growth conditions and only captured RNAs with a half-life long enough to have accumulated levels detectable by used techniques. Later, a group of highly unstable ncRNAs was identified following the development of nascent RNA sequencing techniques and the discovery that they are rapidly degraded by the nu-
clear RNA exosome; for instance, PROMoter uPstream Transcripts (PROMPTs)/up-stream antisense RNAs (uaRNAs), enhancer RNAs (eRNAs)[19, 38–41]. This also led to the realization that the eukaryotic genome is pervasively transcribed.

The classification/nomenclature of these ncRNA species is mostly descriptive based on observational discoveries and often does not reflect their biochemical roles in the cell. For example, long ncRNAs (lnRNAs) have been first defined as ncRNAs with lengths larger than 200 bp, a rather arbitrary cutoff. Another example are PROMPTs, which are named based on the location of their derived loci with respect to that of closely gene promoters. This adds difficulties for obtaining a unified model for ncRNA biogenesis. Here, shared and distinct features of the biogenesis of some ncRNAs compared to that of mRNA are listed in an exemplary manner.

The biogenesis of many ncRNAs share some similar features with that of mRNA during transcription initiation. For instance, they are both transcribed by Pol II, ncRNA also receives a 5’ cap, and Pol II also undergoes TSS proximal pausing [42]. Indeed, ncRNAs like PROMPTs/uaRNAs arise as results of the divergent transcription from mammalian promoters; as shown in Figure 1.1, the mRNA TSS and PROMPTs TSS are positioned in two independent core promoters flanking the edges of a nucleosome depleted region (NDR) that demarcates a gene promoter [43]. However, differences of genomic/sequence features around TSS are observed between ncRNA and mRNA. For example, chromatin modification environment around eRNA TSS show a high H3K4me1-to-H3K4me3 ratio and high H3K27ac levels while that around mRNA TSS have both high levels of H3K4me3 and H3K27ac [44–48]. In addition, the G/C content around the PROMPT TSS is in general lower than that around mRNA TSS [49].

Figure 1.1: Divergent transcription from a mammalian promoter. Figure is obtained from [43] and reprinted with permission from Elsevier and Copyright Clearance Center (License No. 5017171092761).
Chapter 1. Introduction

After initiation, the biogenesis of ncRNA often diverges significantly from that of mRNA, especially in processes like RNA processing and 3’ end formation. For example, ncRNA is generally not spliced and transcribed from loci with a depletion of 5’ splice sites (5’ SSs) and an enrichment of PASs downstream of the TSS [50, 51], which often leads to premature cleavage and polyadenylation. In addition, instead of PAS-dependent 3’ end processing and transcription termination, some ncRNAs were found to undergo different 3’ end processing and transcription termination, which is mediated by the Integrator (INT) complex. These ncRNAs are in general lacking poly(A) tails [52, 53]. Unlike mRNA, few ncRNAs are exported to the cytoplasm. Most of them either undergo rapid nuclear degradation or are retained in the nucleus to serve certain functions before being degraded.

The RNA exosome plays a prominent role in nuclear turnover of ncRNAs [37], which will be discussed in detail in section 1.2.3.

1.2 The RNA exosome

The RNA exosome is a highly conserved multi-subunit protein complex harboring 3’-5’ exonucleolytic and endonucleolytic activities; it is one of the major RNA degradation complexes in eukaryotes [54]. While the RNA exosome exists in both nucleus and cytoplasm, a majority of research has focused on its function in the nucleus. This is possibly due to its more essential roles in the nucleus compared to the cytoplasm, where 5’-3’ exonucleases carry out major tasks in RNA degradation [37]. In the nucleus, the RNA exosome serves dual roles: processing of some stable RNA species such as rRNA, snRNA, tRNA and complete turnover of a variety of RNA species including pervasive transcripts like PROMPTs, eRNAs and premature terminated transcripts from within pc genes [37, 55–57]. In this section, the structure of the RNA exosome, the pathways and the mechanisms of the nuclear RNA exosome degradation, and the impact of the RNA exosome on cellular functions will be discussed in detail.

1.2.1 The structure of the RNA exosome complex

The RNA exosome complex contains a barrel-shaped nine-protein subunit core, which does not harbor catalytic activities (see Figure 1.2a). It consists of a cap structure that includes three S1-KH-domain-containing proteins, Rrp4, Rrp40 and Csl4 and a ring structure that consists of six RNase PH-domain-containing proteins, Rrp41, Rrp42, Rrp43, Rrp45, Rrp46 and Mtr3. The nine-protein core can form a 10
subunit complex by associating with a distributive 3’-5’ exonuclease of the RNase D family Rrp6 on the top or a 3’-5’ exo- and endonuclease of the RNase II family Dis3 (also known as Rrp44, or homologue Dis3L) on the bottom; it can also associate Rrp6 on the top and Dis3 on the bottom at the same time to form a 11 subunit exosome [56, 57].

There are three known routes for the exosome degradation. The majority of RNAs are threaded through the central channel of the nine-protein subunit core in a 3’-5’ orientation and degraded by Dis3 on the bottom. In some cases, the RNA is degraded by Rrp6 on the top, where the RNA either does not go through the core (see Figure 1.2b, left) or enters between the cap and ring structure and bends upwards to reach Rrp6 on the top (see Figure 1.2b, right) [58–60].

In *Saccharomyces cerevisiae* (*S. cerevisiae*), the RNA exosome complex with Rrp6 can only be found in the nucleus while the forms with Dis3 are present in both nucleus and cytoplasm [61] (see Figure 1.2c, left). In human cells, Dis3 is excluded from the nucleolus while Dis3L is only present in the cytoplasmic exosome complexes (see Figure 1.2c, right) [56, 61].

**Figure 1.2**: Structure of the RNA exosome and its isoforms. a: a model of the RNA exosome complex associated with Rrp6 and Dis3. b: two possible pathways for an RNA substrate to the Rrp6 active site. c: Distribution of the RNA exosome complexes with different compositions within *S. cerevisiae* (left) and human cells (right). Figure is obtained from [56] and reprinted with permission from Springer Nature and Copyright Clearance Center (License No. 5017170741410).


1.2.2 The nuclear RNA exosome degradation pathways

In the nucleus, the RNA exosome participates in the 3’ end processing of many stable RNA species as part of their maturation process and at the same time targets a massive amount of various RNAs produced by the pervasive transcription of the genome. Fulfilling these roles requires both efficiency and specificity. While the RNA exosome complex itself is a highly processive and efficient degradation machine, it lacks substrate selectivity. The substrate specificity of the RNA exosome in vivo is achieved through its ability to be associated with other cofactor proteins. Here, I will only discuss degradation pathways of the RNA exosome, while the pathways for stable RNA processing is out of scope of the thesis.

In S. cerevisiae, the RNA exosome targets are co-transcriptionally bound by the Nrd1-Nab3-Sen1 (NNS) complex, which triggers transcription termination and simultaneously recruits the Trf4/5-Air1/2-Mtr4 polyadenylation (TRAMP) complex that is associated with the exosome complex. The nascent transcripts are then passed over to the RNA exosome through the NNS-TRAMP connection. In some cases, the NNS complex can also recruit the exosome directly [55, 62].

In mammalian cells, the TRAMP complex is restricted to the nucleoli and mainly involved in rRNA processing [63]. The exosome degradation specificity is achieved through associating with two major adaptors in the nucleoplasm - the nuclear exosome targeting (NEXT) complex and the related poly(A) exosome targeting (PAXT) connection (see Figure 1.3). NEXT consists of MTR4, the RNA recognition motif (RRM)-containing RBM7 and the Zn-knuckle protein ZCCHC8. MTR4 is also part of PAXT, which contains the Zn-finger protein ZFC3H1. Additionally, PAXT has a weaker connection with the nuclear poly(A) binding protein PABPN1 [63–68].

---

**Figure 1.3: Models of NEXT- and PAXT-dependent nuclear RNA degradation.** Figure is obtained from [66] and reprinted with permission from Elsevier and Copyright Clearance Center (License No. 5017170905062).

---

In human cells, MTR4 interacts with NEXT and PAXT in a mutually exclusive manner. Within NEXT, MTR4 forms a well-defined trimeric complex with RBM7 and ZCCHC8; whereas the PAXT connection is more complex; at its core MTR4
and ZFC3H1 form a stable dimer and make additional elusive contacts with the Zn-finger protein ZC3H3, one of the two RBM26/RBM27, and PABPN1. In addition, both NEXT and PAXT can mutually exclusively connect to the nuclear cap binding complex (CBC) through the zinc-finger protein, ZC3H18 and ARS2 [63, 66, 67, 69–71].

Present characterization of the substrates of NEXT and PAXT pathways in human cells shows that substrates of the NEXT pathway are in general short, un-spliced and non-adenylated, whereas PAXT pathway degrades RNAs that are often longer, spliced and polyadenylated [63, 65, 66, 72].

1.2.3 Substrates and mechanisms of nuclear RNA exosome degradation

Serving an important task in controlling the RNA output from the pervasive transcription of the eukaryotic genome, the nuclear RNA exosome targets a multitude of transcripts produced from non-coding loci for complete and rapid degradation. Such degradation activities are often intimately linked to transcription processes, especially transcription termination. In S. cerevisiae, most cryptic unstable transcripts (CUTs) are subject to transcription-termination coupled RNA decay. As mentioned in section 1.2.2, these transcripts are co-transcriptionally bound by the NNS complex, which mediates transcription termination and at the same time recruits the TRAMP complex to pass over the nascent transcripts to the RNA exosome. The RNA binding proteins Nrd1 and Nab3 recognize sequence motifs that are generally enriched in the genome but depleted in the coding regions, which allows the NNS-TRAMP-exosome pathway to target specifically those cryptic transcripts containing the motifs [62, 73]. In addition, Nrd1 also interacts with the carboxy-terminal domain (CTD) of Pol II, which is phosphorylated at Ser5 residues – a characteristic modification for early elongating Pol II enriched near TSS [74–76]. Therefore, the targeted transcripts are usually short.

Similarly in human cells, termination coupled degradation is employed to target ncRNAs like PROMPTs, which are produced from loci with depletion of 5' SSs and enrichment of PASs [50, 51]. Recognition of 5' SS by U1 snRNP is not only essential for spliceosome assembly and splicing, but also influences transcription output by keeping Pol II in an active transcribing state and suppresses downstream cryptic PAS-mediated premature cleavage and polyadenylation (PCPA) [77–79]. Due to lack of the 5' SS sequence feature, the PCPA events commonly occur in PROMPTs loci and the resulting prematurely terminated transcripts are targeted by
the RNA exosome (Figure 1.4a) [50, 51]. The reason why those PCPA-derived transcripts are generally subject to RNA exosome decay is speculated to be related to the inefficient polyadenylation or lack of poly(A) tail binding proteins [37]. While lack of 5’ SS or U1 snRNP binding facilitate the usage of downstream PAS for transcript cleavage; lack of the splicing process, on the other hand, might lead to subsequent inefficient polyadenylation as it has been observed that splicing factors could stimulate polyadenylation efficiency [80]. Additionally, given those transcripts are relatively short, their 3’ ends are close to the CBC-ARS2 (CBCA) complex, which might contribute to the inefficient polyadenylation and/or direct RNA degradation since the CBCA can connect with the RNA exosome through NEXT, which resembles the NNS-TRAMP-exosome degradation in S. cerevisiae [63, 66, 81, 82].

Figure 1.4: Nuclear degradation by the RNA exosome. a: PROMPTs/uaRNAs resulting from divergent transcription are degraded by the RNA exosome (shown as red ‘PacMan’) b: TSS-associated RNAs resulting from promoter-proximal pause, can be degraded by both XRN2 (shown as purple ‘PacMan’) and the RNA exosome. c: Prematurely terminated transcripts derived from pc genes are substrates of the RNA exosome. d: Aberrant or export-deficient mRNAs can be degraded by XRN2 or the RNA exosome. Figure is adapted from [37] and reprinted with permission from Springer Nature and Copyright Clearance Center (License No. 5017170612888).

In addition to target transcripts produced from non-coding loci, the RNA exosome also participates in degrading transcripts produced from within pc genes. It has been found to target transcripts arising as results from transcriptional regulation. Such regulatory events often lead to premature transcription termination (PTT), where transcription from pc genes is terminated before reaching the gene ends. PTT can first happen during Pol II pausing 30-60 bp downstream of TSS, regulated by NELF and DSIF. It has been found that some paused Pol II terminates and releases transcripts that can be degraded by both XRN2 and the RNA exosome (Figure 1.4b) [83, 84]. PTT also frequently happens in the first intron downstream
of the TSS-proximal pausing through PCPA, which is found to be associated with Pol II pausing at +1 stable nucleosome; the resulting prematurely terminated transcripts are targeted by the RNA exosome (Figure 1.4c) [78, 82, 85]. These PCPA events in the first intron are similar to those in PROMPTs loci discussed above, and can be suppressed by the U1 snRNP binding at upstream 5’ SS.

Moreover, the RNA exosome-mediated nuclear decay is tightly coupled with RNA processing processes like splicing to control mRNA quality. Pre-mRNA often undergoes efficient splicing co-transcriptionally or shortly after its release from the Pol II. It has been suggested that this process is closely monitored by the exosome engaged QC. The RNA exosome cofactors have been observed to interact with the spliceosome, disruption of which lead to increased pre-mRNA levels [86–88]. Transcripts with slow splicing or retained introns have been observed to be targeted by the RNA exosome in the nucleus [89–91]. Studies on the mechanistic link between inefficient splicing and RNA exosome-mediated decay suggest that it is correlated with deficient nuclear export and prolonged nuclear residing time of these transcripts, and likely to be regulated through the poly(A) tail binding proteins (PABPs) (Figure 1.4d) [92–98]. For example, in human cells, PABPN1 is found to participate in the degradation of nuclear-retained transcripts together with poly(A) polymerase (PAP) and the RNA exosome in a pathway called PABPN1 and PAP-mediated RNA decay (PPD) [98].

### 1.2.4 Impact of the RNA exosome on cellular functions

Given the diverse roles the RNA exosome plays in the cell, it is anticipated that disruption of the RNA exosome would lead to severe impairment of normal cell functions.

Through rapid and complete degradation of pervasive transcripts, the RNA exosome suppresses the RNA output from eukaryotic genomes [55]. Disruption of the RNA exosome leads to an increase of nuclear RNA levels. The resulting excessive amount of RNA poses a potential threat to the normal functions of the cell, e.g., the genome stability. Accumulation of nuclear RNAs might lead to increased RNA-DNA hybrids, which are also called R-loops. R-loops are regarded as a source of DNA damage and can induce genome instability [99, 100]. It has been reported that following exosome knockdown in mice, increased R-loops were observed in loci producing exosome-sensitive transcripts [101, 102]. Additionally, a study found that by disruption of the PAXT cofactors, MTR4 and ZFC3H1, their unstable substrates are stabilized and exported to the cytoplasm, which leads to a
global translational repression [67].

Besides, the nuclear RNA exosome is found to serve a regulatory role in controlling expression levels of some genes through mechanisms that are initially utilized in the exosome engaged QC; for example, regulating mRNA levels through intron retention [90, 103]. Interestingly, the intron retention-coupled exosome decay is utilized to autoregulate the PAXT cofactor PABPN1; PABPN1 can bind to the 3'-UTR of its own pre-mRNA and inhibit splicing of the 3'-terminal intron, which induces nuclear RNA exosome degradation [104]. This regulatory role of the RNA exosome is found to have important functional implications during cell differentiation. For instance, the nuclear RNA exosome was reported to regulate neuron-specific genes by targeting their intron-retained mRNAs in neuronal differentiation [105].

1.3 Genome-wide techniques

With the development of genome-wide high-throughput technologies and the corresponding bioinformatics tools, it becomes possible to interrogate the transcriptome and investigate related regulatory events globally. In this section, the genome-wide techniques used in the thesis will be introduced and general data preparation steps will be briefly discussed.

1.3.1 Next-generation sequencing techniques

At present, the most widely used genome-wide technology is next-generation sequencing (NGS), which directly determines the nucleotide sequence of a short fragment of DNA or Complementary DNA (cDNA) – transcribed from RNA – in a massively parallel manner. The NGS technology, at its core, relies on several key steps: template amplification, sequencing and imaging [106]. Different NGS platforms use different strategies for these key steps; here, the Illumina platform, which currently dominates the NGS market, is used as an example. In template preparation, libraries containing DNA and cDNA templates are clonally amplified by solid-phase amplification (Figure 1.5). Each amplified cluster is produced from a single-stranded DNA or cDNA template hybridized with two adaptors at each end. For each template, an immobilized primer first interacts with it to generate a complementary strand, the free template is washed away and the immobilized strand undergoes bridge amplification with closeby immobilized primers to form a cluster [106].
Chapter 1. Introduction

Figure 1.5: Solid-phase amplification by Illumina platform. Figure is adapted from [106] and reprinted with permission from Springer Nature and Copyright Clearance Center (License No. 5017170411926).

Figure 1.6: Four-color cyclic reversible termination by Illumina platform. **a:** Illumina/Solexa’s 3’-O-azidomethyl reversible terminator chemistry. **b:** The four-color images highlight the sequencing results from two clonally amplified templates. Figure is adapted from [106] and reprinted with permission from Springer Nature and Copyright Clearance Center (License No. 5017170411926).

After obtaining the clonally amplified clusters, the four-color cyclic reversible termination (CRT) method is used to sequence the template and capture the sequence information (Figure 1.6). In CRT, each cycle consists of three steps: incorporation, fluorescence imaging and cleavage. First, the DNA polymerase binds to
Chapter 1. Introduction

the template, adds only one fluorescently modified nucleotide and terminates the reaction; each type of nucleotide is labeled with a different color (Figure 1.6a). Then all free nucleotides are washed away and the incorporated nucleotides are determined by fluorescence imaging (Figure 1.6b). Last, the terminating group and fluorescent dye are cleaved and washed away, followed by the next cycle. The nucleobases are directly inferred from the fluorescence signal intensity during each cycle [106, 107].

As NGS uses cDNA/DNA libraries as input and does not require a priori knowledge about the genome or genomic features, it has become the most important tool to interrogate the transcriptome and study gene expression regulation in a genome-wide manner. In combination with techniques that capture specific molecular events, such as capping, protein-DNA/RNA binding, NGS can be customized to many more specific sequencing techniques, providing complementary information. The techniques used in the thesis will be briefly introduced in the following paragraphs, including Cap analysis gene expression (CAGE), RNA-seq and Native Elongating Transcript sequencing (NET-seq), Transcript Isoform sequencing (TIF-seq), chromatin immunoprecipitation sequencing (ChiP-seq) and RBP binding assays.

CAGE measures gene expression by only sequencing the first 20-30 bp of the 5' end of capped transcripts [108, 109]. In CAGE library preparation, complementary cDNA is first synthesized from the total RNA using random or oligo dT primers, and the cDNA that has reached the 5' end of the template RNA is selected by the cap-trapper methods. The RNA strand is then removed by RNase, and a biotinylated 5' linker containing endonuclease recognition site is added to the cDNA. A CAGE tag is subsequently generated by restriction enzyme cleavage and a 3' linker is added to the CAGE tag to generate a cDNA library. A unique feature of CAGE is that it provides TSS information at single nucleotide resolution. This makes CAGE a very powerful tool to identify alternative promoters, characterize promoter properties and study transcription regulation at TSS levels [15, 110–113]. With the discovery that active enhancers are bidirectionally transcribed, CAGE can be also used to find active enhancers, complementary to several other enhancer identification strategies [41]. While CAGE can also quantify gene expression, the complicated library preparation processes make it less favorable compared to RNA-seq.

RNA-seq is a technique used for profiling the transcriptome and quantifying RNA levels. The library preparation is relatively simple compared to CAGE: rRNAs are usually first removed, as they are the most abundant RNA molecules but of little interest. Total RNA or fractionated RNA (e.g., poly(A) tail selected) is frag-
mented; and the RNA fragments are then reversely transcribed to generate a cDNA sequencing library. RNA-seq is the most frequently used technique to quantify gene expression changes under different biological conditions. Additionally, it can also be used to identify and measure transcript isoforms, and study alternative splicing events [114]. As RNA-seq only measures steady state RNA levels, this limits its application to more stable RNA species but not those with very short half-times, like eRNAs or PROMPTs.

NET-seq tracks nascent RNA synthesis and maps transcribing Pol II density through measuring nascent RNA associated with actively transcribing Pol II. In current NET-seq protocols, two strategies have been used for isolating the Pol II complexes: by cell fractionation and by immunoprecipitation of Pol II. In the human NET-seq protocol developed by Mayer et al. [115, 116], the chromatin fraction is used upon cell fractionation, in which transcribing Pol II associated with nascent RNA is enriched. The chromatin associated nascent RNA is purified and the 3’ end of the purified nascent RNA is reversely transcribed to generate a cDNA library. In earlier yeast NET-seq and more recent mammalian NET-seq (mNET-seq) [117, 118], the active Pol II is first immunoprecipitated; the associated nascent RNA is then isolated and converted to a cDNA library. By using NET-seq, many details of Pol II transcription have been revealed. For example, antisense transcription has been observed downstream of promoter-proximal pausing at many promoters. In addition, detailed coordination between Pol II and co-transcriptional RNA processing has been uncovered by immunoprecipitation of Pol II with different CTD phosphorylation status [118, 119].

TIF-seq has been developed for profiling transcript isoforms defined by their 5’ and 3’ boundaries. It first selects RNAs with 5’ cap by RNA oligo capping and full length cDNAs are generated by priming the RNA poly(A) tail. The cDNA is annealed to 5’-biotinylated primer and 3’ primer and circularized following NotI digestion. The circular cDNA is subsequently fragmented. The cDNA fragments with the biotinylated 5’ and 3’ are isolated, multiplexing barcodes are then added to the 5’ and 3’ end to generate a cDNA sequencing library. An advantage of TIF-seq is that it provides direct evidence of transcripts isoforms. Although it lacks quantitative power in terms of isoform abundance, it can be used to measure the relative changes of the same isoform under different conditions [120, 121].

ChIP-seq maps protein binding sites or histone modifications on the genome. The key step in ChIP-seq is chromatin immunoprecipitation, in which antibody is used to select protein of interest or specific histone modifications. For protein binding, DNA is first cross-linked to proteins by formaldehyde and the chromatin
Chapter 1. Introduction

is fragmented; whereas for histone modification, the chromatin is directly fragmented without crosslinking. After fragmentation, an antibody specific to the protein of interest or a histone modification is used to immunoprecipitate the protein-DNA complex or modified nucleosome. The associated DNA is then purified and used for sequencing [122].

RBP binding assays are similar to ChIP-seq but capture RNA-protein interactions. While all RBP binding assays share the immunoprecipitation step, depending on whether proteins will be first cross linked to RNA, they can be classified into two types: RIP-seq and CLIP. RIP-seq does not require crosslinking, and RNA-protein complexes are directly immunoprecipitated with antibodies specific to the protein of interest. RNA protected by the protein is purified after RNase treatment and converted to a cDNA library. A major difference in CLIP is that RNA is first cross-linked to proteins before immunoprecipitation, which improves the binding strength of RBP to RNA. The simpler library preparation makes RIP-seq widely used for the study of general interactions between RBP and RNA, while CLIP is often used to obtain high-resolution RNA binding site information [123].

1.3.2 NGS data preparation

Data processing is an integral part in NGS applications. Processing of different types of NGS data have two initial steps in common, i.e., sequence reads quality control and reference genome alignment/de novo assembly. First, the quality of raw sequencing reads is assessed; the reads are trimmed and filtered to remove sequence artifacts like base calling errors and adaptor contamination. The filtered and trimmed reads are then aligned to the reference genome; this step adds genomic location information to each read. For some sequence techniques like RNA-seq, instead of genome alignment, de novo assembly is often used in order to identify unannotated transcripts and/or transcript isoforms. After the first two steps, the subsequent data processing varies depending on the specific technique and their applications. For example, for CAGE, following the genome alignment, the number of 5’ ends of mapped CAGE reads will be counted at each genomic position to give a unit of CAGE tag start site (CTSS) at single-base resolution [124].

1.4 Machine learning techniques

Generally, the term machine learning (ML) refers to the generation of knowledge from experience by artificial means, i.e., by computational methods. As such it is
a subfield of artificial intelligence, which is concerned with the automation and modelling of intelligent behavior. Specifically, ML refers to computer algorithms that can adapt their behavior and outcome by building an internal model depending on the data presented to them, a process termed training or learning. Training data consists of many training instances that are typically represented by vectors of numerical or categorical features.

Unsupervised and supervised learning are two main types of ML. In unsupervised learning, data is unlabeled, i.e., there is no specific outcome associated with each training instance. Clustering represents one of the most well-known unsupervised learning approaches. It is used to identify structural patterns in data by partitioning the data into groups of objects considered similar with respect to the chosen feature representation. In supervised learning, each training instance is associated with a label and the aim of a ML model becomes to predict the label of previously unknown instances. Depending on the type of label, either categorical or continuous, ML prediction becomes a classification or a regression problem [125]. In the thesis, we only utilized classification approaches to address the relevant questions.

1.4.1 General classification models

The basic approach to build ML classification models consists of two main steps, training and testing. During training, the model is built while the testing step evaluates its performance. In practice, training and test data are usually sampled as disjoint sets from the same pool of data. To evaluate the effect of random sampling, often training/test data splits are repeated and average and variance of performance are reported [125].

ML models can vary greatly in complexity, and the complexity determines how well the model can be fitted to the training data. While more complex models are able to reproduce the labels of training data well, they might fail to predict test data well – a problem known as overfitting. A simpler model will not be able to perform well on the training data but might be better able to predict unknown test instances, an effect known as the generalization capability of a model. Choosing the right model complexity is very important and many ML models allow the tuning of hyperparameters that control model complexity. Hyperparameters can affect the general performance and the generalization ability of a model and can be optimized during training. A common way of hyperparameter optimization utilizes cross validation. In cross validation, the training set is split multiple times
into a smaller training set and a validation set. Models are trained for different hyperparameters on the smaller training sets and their performance is evaluated on the validation sets. Hyperparameters with optimal performance on the validation sets are chosen for the final model and are expected to reduce overfitting and improve generalization.

In classification tasks, the quality of a model can be assessed based on the number of correct and wrong predictions for each class. For example, in binary classification, a model predicts either a positive or a negative label. A correctly predicted positive label is termed a true positive (TP), while a correctly predicted negative label is termed true negative (TN). Wrong predictions are termed false negative (FN) and false positive (FP), respectively. Accuracy is the overall fraction of correct predictions given by \((TP+TN)/(TP+TN+FN+FP)\). Recall is the fraction of correctly identified positive samples given by \(TP/(TP+FN)\) while precision is the ratio of correct predictions among all positive predictions and given by \(TP/(TP+FP)\). Model building in ML and hyperparameter optimization using cross validation use these or similar metrics to identify optimal parameters and models. However, these metrics are sensitive to the training set label composition. Highly unbalanced data set composition can influence the models and their performance. Depending on the application, e.g., when comparing different models using different data sets, it can be helpful to balance training and test sets in terms of positive and negative training examples, thus making these metrics comparable across different settings.

While the primary goal of ML classification models is to accurately predict a label, ML models can also be investigated with respect to the criteria, on which the model bases its decision. For instance, most standard ML methods require numerical, sometimes very high-dimensional, feature representations of the training and test instances; identifying individual or combinations of features that mostly drive the performance of ML models can aid the interpretation of the data and understanding of the underlying problem. Many popular ML methods like support vector machines or deep neural networks have so-called black box character, i.e., the rationale for a certain prediction cannot be elucidated; whereas other models like decision trees allow a direct assessment of feature contributions.

### 1.4.2 Random Forests

Random Forests (RFs) are an ensemble classification ML method [126]. The models are built by constructing a number of individual decision trees during training; RF assigns the label predicted by the majority of the trees to an instance. A deci-
Part of a document that reads:

Chapter 1. Introduction

A decision tree is a tree-structured model (see Figure 1.7), in which each internal node (shown as an ellipse in Figure 1.7) represents a test on a feature, each branch an outcome of the test, and each leaf node (shown as a triangle in Figure 1.7) the predicted class label. In a RF, each decision tree is built on a subset of training instances obtained by methods like bootstrap, i.e., iterative resampling with replacement, and a random subset of the features. Hyperparameters that influence model performance and generalization capability of RFs include the maximum depth of individual trees and the number of random features to consider at each split. As the decisions of RFs can be rationalized well, this facilitates the evaluation of the contribution of individual features to the performance of RF models. One commonly used measure of feature contribution is calculated from the out-of-bag error. When applying bootstrapping to generate subsets of training instances, the ones chosen for training are also called “in-the-bag” sets, while the ones not chosen are called “out-of-bag” (OOB) sets. To score the contribution of a given feature, for each tree built on the “in-the-bag” sets, the prediction error on OOB sets is recorded and the same is done after randomly permuting values of the feature; the difference between prediction errors before and after permutation are then averaged over all trees, and normalized by the standard deviation of the differences.

Figure 1.7: Schematic representation of a decision tree.
Chapter 2

Present studies

2.1 Aim of the thesis

This thesis centers around the RNA exosome, aiming at investigating regulation and functions of the nuclear RNA exosome in mammalian cells from three aspects: first, its impact on transcriptome derived from pc genes; second, contribution of molecular features to its substrate specificity; last, its role in embryonic stem cell (ESC) development.

The three papers included in the thesis each addresses one of the three problems. Paper I characterizes the nuclear RNA exosome substrates produced from protein-coding genes using genomic approaches. Paper II quantitatively identifies molecular features that distinguish the nuclear RNA exosome degradation pathways by machine learning approaches. Paper III investigates the role of the nuclear RNA exosome in mouse ESCs using both molecular biological and genomic approaches.

In the following, I will briefly introduce the background and summarize the results of the three papers.

2.2 Paper I: The RNA exosome shapes the expression of key protein-coding genes

As pointed out in section 1.1.1 and 1.2.3, although degradation of mRNA is carried out mainly in the cytoplasm instead of nucleus, it has been reported that transcripts produced from pc genes can be substrates of the nuclear RNA exosome; these substrates include both prematurely terminated transcripts and full-length
mRNAs. While previous studies have observed such transcripts, there has been no systematic characterization of exosome-sensitive transcripts originating from pc genes regarding transcript isoforms, whether they are produced from canonical/cryptic TSSs, and by which RNA exosome pathway they are targeted, etc. In this paper, we addressed these questions and comprehensively characterized the exosome-sensitive transcripts produced from within pc genes by using several complementary genomic techniques mentioned in section 1.3.1.

By using CAGE, we first captured around 2900 TSSs within protein-coding genes that produce exosome-sensitive transcripts in HeLa cells and termed them exoTCs. We found around 2000 pc genes harbor at least one exoTC, among which 380 pc genes use exoTC as the only annotated TSS. We then characterized transcripts produced from the exoTCs of the 380 pc genes using both TIF-seq and RNA-seq. Consistent with previous observations, there are two main types of exosome-sensitive transcripts – prematurely terminated transcripts mostly terminated in the first intron and full-length transcripts. Among the 380 pc genes, we found that genes with exoTC producing full-length exosome-sensitive transcripts are ubiquitously expressed across cells and tissues, many of which are immediate early genes and encode transcription factors, and the full-length exosome-sensitive transcripts are targeted by the PAXT pathway; while genes with exoTC mainly producing short exosome-sensitive prematurely terminated transcripts are often in head-to-head configurations with other genes, suggesting that they are likely transcribed due to strong transcription initiation of the genes on the upstream opposite strand; similar to PROMPTs, those short transcripts are targeted by both NEXT and PAXT pathways. In addition, we also found a group of genes harbor exoTCs that produce both short exosome-sensitive prematurely terminated transcripts and exosome-insensitive full-length transcripts; the prematurely terminated transcripts are primarily targeted by the PAXT pathway. This implies that the RNA exosome can participate in regulating gene expression through transcription attenuation.

We further investigated genes harboring multiple expressed TSSs, where at least one TSS is an exoTC. In these cases, exoTCs were found to have minor contributions to the overall gene expression. By characterizing transcripts produced from these exoTCs, we found the length of the exosome-sensitive transcripts is affected by the distance between the exoTCs and the other TSSs producing exosome-insensitive transcripts. This effect is most likely related to, first, whether the 5' SSs of the other TSSs are downstream of the exoTCs and second, the distance of the downstream 5' SSs to the exoTCs. When the 5' SSs are within 500 bp downstream of exoTCs, the exoTCs are observed to produce predominantly full-length tran-
scripts; interestingly, these full-length transcripts are not targeted by either PAXT or NEXT pathway; whereas if there are no downstream 5’ SSs or they are further away than 500 bp, the exoTCs produce mainly prematurely terminated transcripts targeted by the NEXT pathway.

In summary, this study reveals a complex landscape of sense transcription within pc genes shaped by the RNA exosome and suggests a role of the RNA exosome in regulating the mRNA expression levels of some genes.

2.3 Paper II: Identifying determinants of nuclear RNA exosome degradation pathways using machine learning approaches

The RNA exosome is highly efficient yet lacks substrate specificity by itself. In the nucleoplasm of mammalian cells, as discussed in 1.2.2, the RNA exosome connects to adaptors and selectively degrades RNAs through PAXT and NEXT pathways. While previous studies showed DNA/RNA features distinguishing the two pathways with respect to the features of loci producing the RNAs, the transcript splicing, and 3’ end processing patterns; there has been no comprehensive study to quantitatively evaluate the contribution of these features to different RNA exosome degradation pathways.

In this work, we set out to address this question and delineate the determinants for PAXT and NEXT pathways in HeLa cells using machine learning approaches. We first classified the targets of the RNA exosome pathways based on RNA-seq using the de novo HeLa-cell-specific transcriptome annotation from [53], which is generated from transient transcriptome sequencing (TT-seq) and RNA-seq data using hidden Markov Model approaches and refined by integrating CAGE and 3’ end-seq. We next systematically collected different types of relevant molecular features covering the relevant TSSs and transcript end sites (TESs), e.g., DNA/RNA sequence features, the chromatin environment and possible RBP-RNA interactions and organized them into four categories, i.e., TSS features, TSS-proximal RNA processing, TES-proximal RNA processing, TES features. Last, we developed predictive models using Random Forests to evaluate these features’ ability to distinguish RNA exosome targeting pathways; specifically, we performed three binary classifications: NEXT vs. non-exosome targets, PAXT vs. non-exosome targets and NEXT vs. PAXT targets. To measure the contribution of individual features to the prediction per-
formance, we defined the feature importance score as a z-score calculated from the change in performance by permuting the values of each feature. In order to obtain consistent and reproducible results, we performed iterative feature selection based on feature importance scores and obtained a stable set of consistently significant features, which was used for determining the most discriminative features.

Comparing the prediction results and discriminative features from the three binary classifications, we found the features relevant to TES-proximal RNA processing to be most predictive for distinguishing different exosome pathways. Specifically, the lack of canonical 3’end processing by the CPA machinery were found to be most characteristic for NEXT targets while other previously reported features relevant to exosome targeting such as TSS G/C content, TSS-proximal 5’SSs were only found to be distinct for NEXT targets but not able to distinguish PAXT from non-exosome targets.

2.4 Paper III: A functional link between nuclear RNA decay and transcriptional control mediated by the polycomb repressive complex 2

As briefly discussed in section 1.2.4, the degradation activities of the RNA exosome are crucial for maintaining normal cellular function. In this study, we investigated how the RNA exosome could impact cell differentiation using mouse ESCs. As I mainly participated in analyzing all the genomic data in this paper, in the following, I will focus on summarizing the results and conclusions drawn from this part of work.

ESCs are characterized by the ability to self-renew and their potential to differentiate into different cell types. The ESC pluripotency is maintained by pluripotency gene expression programs orchestrated by transcription factors (TFs) and reinforced by chromatin structure, whereas differentiation of pluripotent cells requires both activation of lineage-specific genes and chromatin-based silencing of pluripotency gene expression programs. While the importance of transcriptional regulation in ESCs is well established, less is known about the role of the post-transcriptional process. Some studies have suggested functional roles of long non-coding RNAs (lncRNAs) in regulating pluripotent states and differentiation of ESCs. The RNA exosome PAXT pathway is known to target polyadenylated (pA+) RNAs including a number of lncRNAs. In this work, by disrupting PAXT activity,
we explored the impact of excess nuclear pA\(^+\) RNA on mouse ESC pluripotency and differentiation.

The PAXT component ZFC3H1 was knocked out using CRISPR/Cas9 in mouse ESC; the Zfc3h1\(^{-/-}\) cells were viable under 2i+LIF growth conditions, which selects against cellular differentiation. However, after being transferred to serum-LIF media, which allow spontaneous development of embryoid bodies (EBs), Zfc3h1\(^{-/-}\) cells displayed defective differentiation. To compare the global gene expression difference between Zfc3h1\(^{-/-}\) and wild type (WT) cells, we performed differential analysis of total RNAseq data from cells cultivated under 2i+LIF condition (D0) and on serum-LIF media for 7 days (D7), respectively. Consistent with the observed differentiation defect, Zfc3h1\(^{-/-}\) D7 cells showed increased expression of pluripotency-associated TFs and decreased expression of lineage-specific factors, compared to WT D7 cells. For D0 samples, an increase of lineage-specific factors was observed in Zfc3h1\(^{-/-}\) vs. WT cells. The lineage-specific factors deregulated in D0 are normally subject to H3K27me3-mediated silencing in WT ESCs. As H3K27me3 is a histone modification deposited by polycomb repressive complex 2 (PRC2), we then compared the gene expression pattern of D0 Zfc3h1\(^{-/-}\) with PRC2 knockout ESCs; we found that about one-third of upregulated genes in Zfc3h1\(^{-/-}\) vs. WT cells overlap with those in PRC2 knockout vs. WT cells. Analyzing the intronic and exonic reads suggests these genes are transcriptionally upregulated. In summary, gene expression analysis using RNAseq showed a number of genes upregulated in Zfc3h1\(^{-/-}\) ESCs are from PRC2 target loci, and these genes are likely to be transcriptionally upregulated.

As many transcripts with increased expression in Zfc3h1\(^{-/-}\) ESCs are also upregulated in absence of PRC2, we investigated PRC2 status in D0 Zfc3h1\(^{-/-}\) cells. By using ChIPseq, we found that Zfc3h1\(^{-/-}\) cells displayed a global decrease in SUZ12 binding, a core component of the PRC2 complex, as well as reduced H3K27me3 levels. In addition, we found that in regions depleted of SUZ12 binding and H3K27me3 modification, the RNA expression is increased likely transcriptionally. These results suggest an upregulated expression of some genes in Zfc3h1\(^{-/-}\) cells resulted from loss of H3K27me3 due to reduced the recruitment of PRC2.

Furthermore, we observed disruption of the PRC2 complex in Zfc3h1\(^{-/-}\) cells. Given the RNA binding properties of PRC2, we hypothesized that RNA accumulation in absence of PAXT activities might contribute to PRC2 disruption. To investigate this possibility, we further analyzed RNA species bound to SUZ12 and EZH2, another core component of the PRC2 complex using RIPseq. However, we did not observe specific enrichment of PAXT targets. Instead, we found PRC2 disruption is
dependent on general increased RNA levels due to PAXT disruption.

In summary, all results underscore the importance of controlling nuclear RNA levels during ESC development and suggest a potential way to regulate transcription by bulk RNA.
Chapter 3

Conclusions and perspectives

The questions addressed in the three papers can be subsumed under the common theme of nuclear RNA turnover, what the substrates are, how it can be regulated and why it is important for cells. We focused on a main player in eukaryotic nuclear RNA turnover – the RNA exosome. With the aid of next generation sequencing techniques, we could address these relevant questions on a genomic scale. Computational techniques like machine learning helped us to gain new insights from the data.

By integrating different types of genomic data, we revealed a complex sense transcription landscape within pc genes shaped by the RNA exosome, and elucidated the role of the RNA exosome in regulating mRNA expression levels. Using machine learning approaches, we delineated the determinants of nuclear exosome degradation pathways and found molecular features of TES to be the most predictive for distinguishing PAXT and NEXT pathways, while TSS related features were only found to be distinct for NEXT substrates. By manipulating the PAXT activity in mouse ESCs, we established a functional link between exosome-mediated nuclear decay and PRC2-mediated transcriptional control and suggested a potential way to regulate transcription by bulk RNA.

Some limitations and unaddressed questions in these works might be interesting for future study. First, the loss-of-function approaches used in the thesis are all based on long-term stable depletion of the RNA exosome core subunit or cofactors, either by small interfering RNA (siRNA) or CRISPR/Cas9. Thus, our observations are outcomes of accumulated effects, many of which might be secondary, especially considering the diverse and important roles of nuclear decay activities in cells. For example, in paper I, following depletion of the RNA exosome, genes with upregulated full mRNA levels are often found to be immediate early genes.
Chapter 3. Conclusions and perspectives

It is known that expression of the early response genes can be induced by DNA damage [127, 128]; accumulation of RNA-DNA hybrids (R-loop) has also been observed upon depletion of the RNA exosome, which is considered a source of DNA damage [99, 100]. Therefore, it is possible that expression of these genes are up-regulated due to secondary effects like DNA damage, rather than being directly regulated by the RNA exosome. Also in paper III, we observed excess nuclear pA\(^+\) RNA could disrupt PRC2 function and impair cell differentiation; as the increased nuclear pA\(^+\) RNA levels are accumulated results, it is difficult to dissect the direct contribution of the RNA exosome and its substrates to such effect. To study the primary function of the RNA exosome, conditional knockdown/knockout techniques that allow for temporal alteration of the RNA exosome activities will be of great value.

Second, the genomic techniques used in the thesis are limited in the way they can directly identify the exosome-sensitive transcript isoforms. Although TIF-seq can accurately identify transcript isoforms, it lacks quantitative power, and only provides information about the start and end but not the splicing pattern of each transcript. While computational methods exist to infer and quantify transcript isoforms based on RNAseq, they often cannot accurately detect isoforms with low abundance [129], which typically are the cases with exosome-sensitive transcripts. To overcome these limitations, new techniques like long-read sequencing can be utilized. Related to this, currently only limited annotation is available for non-coding loci producing exosome-sensitive transcripts. With the aid of long-read sequencing, a complete and high resolution transcriptome annotation covering non-coding exosome-sensitive transcripts isoforms will be an important resource for addressing many questions related to the RNA exosome.

Furthermore, there are some interesting questions that are not addressed in the thesis. For example, what the determinants of pA\(^+\) RNA stability in the nucleus are; it might be interesting to explore the influence of differential polyadenylation status on pA\(^+\) RNA stability in a more quantitative manner. The application of techniques capturing poly(A) tail, e.g., direct full-length RNA sequencing, in combination with other complementary quantitative genome-wide techniques, promises to be able to identify more details on a genomic scale. In addition, direct full-length RNA sequencing techniques can provide new and more detailed information of RNA molecules and their processing, like splicing, poly(A) tail length, and RNA modification. It can facilitate further investigation on how those features and interactions of different processes contribute to pA\(^+\) RNA stability in the nucleus.
Acknowledgements

I would like to express my deepest gratitude to all the people that have helped and supported me in these three years.

First of all, my supervisor Albin Sandelin: thank you for all the guidance and encouragements throughout my PhD; you have always been patient with all kinds of doubts I had and willing to share your expertise and experience. My co-supervisor Torben Heick Jensen: thank you for always giving me sincere feedback, and pointing out the possible caveats of my work directly and constructively. Manfred Schmid: thank you for sharing your ideas and knowledge, and always patiently answering my constant repetitive questions. Robin Andersson: thank you for sharing your ideas, all the scientific advice, and discussions.

I would like to thank the Sandelin lab members. Yun Chen, Jette Bornholdt, Leonor Rib, Axel Thieffry, Laura Pikkuleura, Signe Schmidt Kjølner, Alfredo Rago, Arnaud Stigliani, Jiayi Yao, Christian Vaagenso for creating a friendly working environment. Kristoffer Vitting-Seerup, for helping me in getting familiar with the environment and work at the beginning of my PhD. Malte Thodberg, for sharing all the PhD tips, routinizing coffee break and gymnastics, and helping translate my thesis abstract into Dansk. Hans Ienasescu, for being a chill fellow, and fulflling my expectations for a Weichei.

The Andersson lab members: thank you to all PhD fellows for struggling all together; thank you to all non-PhD fellows for scientific discussions/philosophical reflections/beers.

Thank you to the THJ lab members for welcoming and helping me during my stay at Aarhus University. Special thanks to Will Garland, Marta Lloret-Llinares, and Effie Karadoulama, it has been great fun to work with you.

I would also like to thank all the people of BINF for many fond memories. Hanne Munkholm: thank you for being the greatest system administrator of all time.

Duc, Labo, Puja, Bony, Ele, Koa, Koco, Hedges, Moky and Toto, thank you for being supportive and always volunteering to share your academic/street wisdom.
Acknowledgements

To my family – to mom, dad, grandma, and grandpa, for all the love.
To Genís, for all the care and support; and finally to Martin, for everything.
References


References


References


References


References


References


Paper I–III
Paper I
The RNA exosome shapes the expression of key protein-coding genes

Mengjun Wu\textsuperscript{1}, Evdoxia Karadoulama\textsuperscript{1,2}, Marta Lloret-Llinares\textsuperscript{2,3}, Jerome Olivier Rouviere\textsuperscript{2}, Christian Skov Vaagenso\textsuperscript{1}, Martin Moravec\textsuperscript{2}, Bingnan Li\textsuperscript{4}, Jingwen Wang\textsuperscript{4}, Guifen Wu\textsuperscript{2}, Maria Gockert\textsuperscript{2}, Vicent Pelechano\textsuperscript{3}, Torben Heick Jensen\textsuperscript{2,*} and Albin Sandelin\textsuperscript{1,*}

\textsuperscript{1}The Bioinformatics Centre, Department of Biology and Biotech and Research Innovation Centre, University of Copenhagen, Ole Maaløes Vej 5, DK2200 Copenhagen N, Denmark, \textsuperscript{2}Department of Molecular Biology and Genetics, Aarhus University, C.F. Møllers Alle 3, Building 1130, Aarhus 8000, Denmark, \textsuperscript{3}European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and \textsuperscript{4}SciLifeLab, Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Solna 171 65, Sweden

Received April 13, 2020; Revised June 29, 2020; Editorial Decision July 01, 2020; Accepted July 03, 2020

ABSTRACT

The ribonucleolytic exosome complex is central for nuclear RNA degradation, primarily targeting non-coding RNAs. Still, the nuclear exosome could have protein-coding (pc) gene-specific regulatory activities. By depleting an exosome core component, or components of exosome adaptor complexes, we identify \textasciitilde2900 transcription start sites (TSSs) from within pc genes that produce exosome-sensitive transcripts. At least 1000 of these overlap with annotated mRNA TSSs and a considerable portion of their transcripts share the annotated mRNA 3′ end. We identify two types of pc-genes, both employing a single, annotated TSS across cells, but the first type primarily produces full-length, exosome-sensitive transcripts, whereas the second primarily produces prematurely terminated transcripts. Genes within the former type often belong to immediate early response transcription factors, while genes within the latter are likely transcribed as a consequence of their proximity to upstream TSSs on the opposite strand. Conversely, when genes have multiple active TSSs, alternative TSSs that produce exosome-sensitive transcripts typically do not contribute substantially to overall gene expression, and most such transcripts are prematurely terminated. Our results display a complex landscape of sense transcription within pc-genes and imply a direct role for nuclear RNA turnover in the regulation of a subset of pc-genes.

INTRODUCTION

RNA degradation is essential for maintaining transcript homeostasis in all cells. Together with transcription, it controls steady-state RNA expression levels, which underlie all major cellular transitions in development and disease. While RNA degradation in the cytoplasm is considered to be the main determinant for mRNA half-lives, the extent to which nuclear RNA decay is involved has been less clear. In the nucleus, transcript turnover is often coupled to transcription termination and/or processing of the nascent RNA (1–5). Moreover, it has been suggested that prolonged nuclear residence time correlates with the increased turnover of polyadenylated RNA species (5). Together, this serves to dampen the expression of a large amount of pervasively transcribed RNAs (2), thought to primarily include a multitude of long non-coding RNAs (lncRNAs), which as a group is prone to rapid nuclear degradation (6).

The highly conserved 3′-5′ exo- and endo-nucleolytic RNA exosome complex is a primary caretaker of the decay of capped RNAs in eukaryotic nuclei (3,4). In mammalian nuclei, the exosome is composed of a core unit with associated nucleolytic activities, which in the nucleoplasm may contact one of two exosome adaptor complexes in order to target RNAs for degradation (5). One such adaptor, the nuclear exosome targeting (NEXT) complex targets RNAs, that are primarily short, mono-exonic and non-adenylated RNAs (7–9). These can be lncRNAs, including subsets of enhancer RNAs (eRNAs) and promoter upstream transcripts (PROMPTs)/upstream antisense RNAs (uaRNAs) (10–13). The polyA exosome targeting (PAXT) connection targets similar RNA biotypes as the NEXT complex, but specifically those that are polyadenylated (9). Additionally, PAXT mediates the exosomal degradation of longer and
processed nuclear RNAs (14). Of interest, disruption of the nuclear exosome as well as of both the NEXT and PAXT pathways affect stem cell differentiation, suggesting a role for nuclear RNA decay in gene expression regulation (15–17).

Although mRNAs are generally not considered major targets of the nuclear exosome, early reports revealed that annotated mRNA TSSs may produce exosome-sensitive transcripts (12,18,19). Sequencing of capped RNA 5′ ends showed that a subset of alternative mRNA TSSs gives rise to exosome-sensitive RNAs although their exact nature was not established (6). These observations were rationalized in several recent papers, which established that premature termination of transcription (also referred to as ‘attenuation’) can affect the transcriptional output of full-length transcripts from pc-gene TSSs (reviewed in (20)). Such premature transcription termination can be mediated by nascent RNA cleavage by the canonical cleavage and polyadenylation (CPA) machinery or by the Integrator complex, and the resulting short transcripts were shown to be exosome sensitive (21–24). Interestingly, recent work also showed that a substantial number of full-length mRNAs might be nuclear exosome substrates (25). Collectively, these findings demonstrate that a portion of pc-genes emit transcripts that are affected by the nuclear exosome. These can either be prematurely terminated transcripts or full-length RNAs. However, a systematic analysis of such sensitivity, including the nature of the isoforms produced and whether they arise from major and/or cryptic TSSs commonly present in complex genomes (26), has been lacking. This is relevant to assess the impact of such transcription events on the overall output of pc-gene promoters.

Here, we confirm that a substantial number of pc-genes harbor TSSs, producing nuclear exosome-sensitive transcripts. Surprisingly, ∼360 such genes only employ one primary annotated TSS to produce full-length transcripts across diverse cells and tissues. These genes often encode transcription factors and immediate early response genes. Another set of pc-genes also employ a single annotated TSS, but primarily produce prematurely terminated transcripts. We show that this production is likely due to a bystander effect of strong and nearby mRNA initiation on the reverse strand. We also explore multi-TSS genes where at least one TSS produces exosome-sensitive RNAs, and find that such TSSs have a minor contribution to overall gene expression, where the length of exosome-sensitive RNAs produced is correlated to the distance to other TSSs producing exosome-insensitive RNAs. Overall, our work shows that the exosome shapes the expression of several pc genes, many of which are functionally important across cells.

**MATERIALS AND METHODS**

**HeLa cell culture and small interfering RNA (siRNA)-mediated knockdown**

HeLa Kyoto cells were cultured in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. siRNA transfections (for SLIC-CAGE and TIF-seq) were carried out using lipofectamine 2000 (Invitrogen) according to the manufacturer’s protocol. Cells were treated with 20 nM siRNA for 4 days, including a re-transfection 2 days after the initial transfection. siRNA sequences: siGFP: GA CGUAAACCGGCCACAGUdTdT; siRRP40: CAGCGCA CAGUACUAUGGUCdFdT; siZCCHC8: GGAAUGUA CCUCAGGAUAAdTdT; siZFC3H1: GAUAGAGUC CAUGAUAAdFdT. RNA was extracted using TRIZol (Invitrogen) and treated with TURBO DNase (Invitrogen) following the manufacturer’s instructions.

**Western blotting analysis**

Cells were lysed with lysis buffer (10 mM Tris–Cl pH 7.4, 100 mM NaCl, 2.5 mM MgCl2, 0.5% NP-40, 0.5% Triton X-100) on ice for 10 min, then centrifuged at 12 000 rpm for 20 min. The protein concentration in the supernatant was measured using Bradford solution (Bio-Rad). Equal amounts of proteins were loaded onto PAGE gels. After running, proteins were transferred to PVDF membranes, which were blocked with 5% skimmed milk/PBS-T for 1 h at room temperature (RT), and then incubated with primary antibodies diluted in PBS-T at 4°C overnight, followed by washing 3 × 10 min with PBS-T. Membranes were then incubated with HRP-conjugated secondary antibodies diluted in PBS-T for 1 h at RT, followed by washing 3 × 10 min with PBS-T. SuperSignal West Femto HRP substrate (ThermoFisher Scientific) was applied to the membranes and the signal was detected with X-ray film (Konica Minolta). Antibodies: RRP40: ProteinTech, 15062-1-AP, 1:1000; ZFC3H1: Sigma, HPA-007151, 1:1000; ZC-CHC8: Novus Biologicals, NB100-94995, 1:1000; Tubulin: Rockland, 200-301-880, 1:2500. Western blotting analysis are shown in Supplementary Figure S11.

**SLIC-CAGE library preparation, sequencing**

SLIC-CAGE preparation was performed as described in (27) with an input of 2000 ng of total RNA as starting material. Individually prepared SLIC-CAGE libraries with unique barcodes were pooled (8 per lane). The following 8 barcodes were used: # 1 (ACC), #2 (CAC), # 3 (AGT), # 4 (GCG), # 5 (ATG), # 6 (TAC), # 7 (ACG) and # 8 (GCT). All used primers and adaptors were purchased from Integrated DNA technologies (IDT). An Illumina NextSeq 500 instrument at the BRIC, University of Copenhagen, was used for sequencing.

**SLIC-CAGE data processing, quantification**

CAGE reads were trimmed to remove linker sequences at 5′ ends and incorrect ‘G’ calls at 3′ ends using cutadapt (version 1.14) (28) with parameters -u 5 -m 30 –nextseq-trim = 30 -l 70. Trimmed reads were filtered so that only reads with minimum sequence quality of 30 in at least 50% of the bases were kept. rRNAs were further removed using rRNAdust (http://fantom.gsc.riken.jp/5/suppl/rRNAdust/). Remaining reads were mapped to the human genome hg19 using bwa (version 0.7.16a-r1181) (29) with default settings. The number of 5′ ends of CAGE reads were counted at each genomic position to give a unit of...
CAGE tag start site (CTSS), at single-base resolution. The raw counts were normalized to tags per million mapped reads (TPM) for subsequent quantification.

Public data acquisition, processing and analysis
Public data used in this study were obtained from ENCODE and Gene Expression Omnibus (GEO); the accession numbers were as follows: DNase-seq (ENCODE, ENCSR959ZXU), HeLa S3 H3K4me3, H3K36me3 and H3K27ac ChIPseq (GEO, GSE29611), HeLa S3 nascent RNA-seq (GEO, GSE61332), nuclear RNA-seq of siRRP40 and siEGFP control (GEO, GSE108197), total RNA-seq of siRRP40, siZCCHC8, siZFC3H1 and siEGFP control (GEO, GSE84172), CAGE of siRRP40 and siEGFP control (GEO, GSE62047). For CAGE of siRRP40 and control, the triplicate HeLa siRRP40 and control CAGE libraries were computationally processed as described in (6). In brief, using the FASTX Toolkit (v0.0.13, https://hannonlab.cshl.edu/fastx_toolkit), reads were trimmed from the 5′ end to remove linker sequences, trimmed from the 3′ end to a length of 25 bp and subsequently filtered for a minimum sequencing quality of 30 in 50% of the bases. Trimmed and filtered reads were mapped to the human genome (hg19) using Bowtie (version 0.12.7) (30) with parameters –t –best –strata –v –k 10 –y –p 6 –phred33-quals –chunksmbs 512 –e 120 –q –un. The number of CAGE tag 5′ ends were counted in each genomic position and nearby 5′ ends on the same strand as in (31) to create tag clusters (TCs). The TCs read counts were normalized to tags per million mapped reads (TPM). The CAGE defined TCs were annotated using GENCODE v19 annotation (32) based on a hierarchical ranked classification, where in case of multiple classification overlaps the highest ranked was selected, the hierarchical model is shown in Figure 1C. The categories in priority order and their definitions were as follows: TCs within ±100 bp of the most upstream GENCODE annotated TSS of a gene—primary TSS; TCs within ±100 bp from all other GENCODE annotated TSSs of a gene—alternative TSS; TCs within 5′ UTRs of transcripts with annotated coding regions (CDS)—5′ UTR; TCs within CDS—CDS; TCs within 3′ UTRs of transcripts with annotated CDS—3′ UTR; TCs within exons of transcripts where no CDS is annotated—exon; TCs within introns—intron; TCs within a 10kb window upstream of the most upstream CENCODE annotated TSS of a gene—upstream. For non-CAGE data, replicates were pooled and signals were averaged over replicates for subsequent analysis. For gene level RNA-seq fold change (FC), strand-specific, uniquely mapped and properly paired reads across the GENCODE v19 gene models were counted using featureCounts from the R package Rsubread (1.3.2.1) (33), to minimize the expression differences between samples for genes with low read counts, a pseudocount of 7 was added when normalizing raw read counts to the library size. FC values were calculated between mean values of the normalized read counts from siRRP40, siZCCHC8 or siZFC3H1 and that from Ctrl libraries. For RNA-seq FC across gene bodies, FC values of siRRP40, siZCCHC8 or siZFC3H1 versus Ctrl were calculated using bigwigCompare from deepTools (34) over a 5 bp window. A pseudocount of 0.05 was added before FC calculation.

Sensitivity score calculation
A sensitivity score was designed to quantify the relative amount of expression increase or decrease after depletion of a given factor. It was calculated as:

\[
\text{Sensitivity} = \frac{(\text{Expression}_{\text{Depletion}} - \text{Expression}_{\text{Control}})}{\max(\text{Expression}_{\text{Depletion}}, \text{Expression}_{\text{Control}})} \in [-1, 1]
\]

where the Expression is the normalized strand-specific CAGE or RNA-seq expression for a given library.

Nascent RNA quantification and directionality calculation
Nascent RNA levels were quantified by data from (35). For quantifying nascent RNAs produced from exoTCs, the strand-specific genomic coverage in −100 to +500 bp regions was computed using computeMatrix reference-point from deepTools; for quantifying nascent RNAs produced from the upstream opposite strand of exoTCs, the strand-specific genomic coverage of NET-seq in -1 to −600 bp regions was computed the same way as exoTCs strand. A directionality score was designed to measure the biases of transcription or expression levels from opposite strands. It was calculated as follows:

\[
\text{Directionality} = \frac{(\text{Expression}_{\text{forward}})}{(\text{Expression}_{\text{forward}} + \text{Expression}_{\text{reverse}})} \in [0, 1]
\]

where Expression_{forward} is the transcription or expression levels of the TC on the forward or sense strand and Expression_{reverse} on the reverse or upstream opposite strand.

Definition of upstream opposite strand TCs
The upstream opposite strand TC of a given TC was defined as the closest CAGE TC, with TPM > 1 in siRRP40, that fell on the upstream opposite strand of the TC within 600 bp.

TIF-seq library preparation and sequencing
TIFseq2 library preparation was performed as described in (36) using 2500 ng of total RNA as starting material. In brief, 5′P RNA was dephosphorylated using calf intestinal alkaline phosphatase, purified and decapped using Cap-Clip. Newly exposed 5′P were ligated to chimic DNA/RNA oligos and reverse transcribed using barcoded oligo dT primers. Full-length cDNA was amplified by PCR and digested with NotI-HF to produce sticky ends. We then circularized the amplified cDNA, removed non-circular fragments and fragmented the purified circles by sonication. Fragments spanning the 5′ and 3′ cDNA ends and containing biotin were bound to streptavidin magnetic beads and then subjected to Illumina library preparation. Samples were sequenced using a NextSeq 500 instrument with the following options: read1 76 bp, read 2 76 bp, index1 6 bp and index2 6 bp.
TIF-seq data processing and analysis

Sequencing reads were converted by using bcf2fastq (v2.20.0) and demultiplexed according to the indices, allowing two mismatches in index 1 and one mismatch in index 2. TIF-seq2 sequencing primer (AGGTGACGCGAGGTT GT) and Illumina TruSeq adapter (AGATCGGAAAG) were converted to hg19 using UCSC liftOver tool ([PMID:28100584]) (v0.5.4) from the 5′ ends and extra adenine stretches in the 3′ ends were removed with cutadapt (v1.16) ([37]). Then, 8-bp unique molecular identifiers (UMIs) were extracted with UMI-tools ([PMID:91100584]) (v0.5.4) from the 5′ ends and extra adenine stretches in the 3′ ends were removed with cutadapt (v1.16). STAR (v2.5.3a) ([37]) were employed for aligning 5′ end reads and 3′-end reads separately to the human reference genome hg38, allowing maximum intron length as 1 Mb. A customized script adapted from UMI-tools was employed to remove PCR duplicates from uniquely mapped reads pairs on the same chromosome, allowing 1-bp shifting in the 5′ ends. The hg38 genome coordinates were converted to hg19 using UCSC liftOver tool ([38]). Paired 5′ end and 3′ end reads located on the same chromosome and opposite strand were used to form 5′-to-3′ end TIF transcripts. To avoid 3′ ends produced by spurious internal poly A priming by the oligo(dT) primer, the sequence immediately downstream of the 3′ end of each TIF transcript was further examined. If the downstream sequence started with five or more contiguous adenines, or had seven or more adenines in the first 10 bp, the corresponding TIF transcript was removed from this analysis. To remove artifically long TIF transcripts, GENCODE v19 genes were merged into transcription units using merge from bedtools (v2.23.0) ([39]) with parameter -s, TIF transcripts overlapped with more than one transcription units were removed. Replicates were pooled for subsequent analysis. To associate CAGE TCs with TIF transcripts, TIF transcripts whose 5′ ends fell within a ±100 bp window around TC peaks on the same strand were assigned to the corresponding TCs. If a CAGE TC was annotated as primary/alternative TSS of a protein-coding gene, the associated TIF transcripts were also assigned to the same gene. TIF-seq FC was calculated as the ratio between the library size normalized TIF transcript counts from pooled siRRP40- and Ctrl-libraries, a pseudocount of 1 was added. To annotate 3′ ends of TIF transcripts, a similar hierarchical approach as CAGE TC annotation was used; the 3′ end hierarchical model is shown in Supplementary Figure S1G. The transcription termination site (TES) was defined as the ±200 bp window region around the 3′ end of GENCODE v19 transcripts, the TSS was defined as the ±100 bp window region around the 5′ end of GENCODE v19 transcripts. 3′ UTR, 5′ UTR and CDS, exon regions were defined as in CAGE annotation. First intron was defined as the first intron of GENCODE v19 transcripts of all expressed genes shown in Figure 1D. Full-length TIF transcripts were defined as the TIF-seq reads with a 3′ end annotated as TES or 3′ UTR, premature terminated TIF transcripts were defined as TIF-seq reads with a 3′ end annotated as features within the gene body excluding 3′ UTR and TES.

Classification of exoTCs based on RNA-seq data

We devised a hierarchical decision tree to classify exoTCs that were associated with multi-exonic genes and their cognate transcripts into four classes (shown in Supplementary Figure S2A). This was based on (i) whether the exoTCs produced exosome-sensitive short transcripts, quantified by FC of siRRP40- versus Ctrl in the first intron 1 kb downstream of the first splice site, (ii) whether the TCs produce exosome-sensitive full-length transcripts, quantified by the same ratio but within all exons downstream of the first intron. The raw reads in the defined genomic regions were counted using featureCounts from the R package Rsubread ([132.1]) FC values were calculated between mean values of the library size normalized read counts from siRRP40- and Ctrl-libraries, a pseudocount of 7 was added. If the TCs did not produce exosome-sensitive full-length transcripts according to ii), they were further divided based on whether they produced any full-length transcripts, quantified by the RPKM normalized siRRP40 RNA-seq counts of all exons downstream of the first intron. Mono-exonic genes represented special cases since they have no introns: they were classified as Class 1 if FC of siRRP40 versus Ctrl exceeded the same threshold as in (ii).

Sequence analysis

Sequences were extracted from the reference genome (hg19) using getfasta from bedtools, G/C content was calculated using letterFrequencyInSlidingView function from the Biostrings R package over 5 bp window (version 2.50.2). For a site and 5′ splice site analysis, the motifs were obtained from ([40]), ASAP ([41]) was used to calculate motif prediction scores and a relative score cutoff of 0.9 was used for deciding the occurrence of the motifs.

Evolutionary conservation

Evolutionary conservation of a TC was calculated as the average phastCons score for a ±100 bp window region around the TC. The phastCons score for human genome (hg19) was calculated from multiple alignments with other 99 vertebrate species was used ([42]); data was downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/). As background, random intergenic and intronic regions of length 200 bp were extracted using shuffle from bedtools with default settings. The intergenic regions were randomly chosen from genomic regions that did not overlap with any GENCODE v19 genes. The intronic regions were randomly chosen from regions in the gene body that did not overlap with an exon from any GENCODE v19 annotated transcript isoforms.

FANTOM5 data processing and analysis

We used FANTOM5 CAGE TC expression data from primary cell groups and tissues, taken from SlideBase ([43]) processed data which in turn is based on CAGE data from ([44,45]).

Metagene plots

For metagene plots over gene bodies (Figures 2B–D and 4B, Supplementary Figures S1B–F, S4B), the transcript originating from a given TSS was used to represent the gene; for
TSSs that have multiple transcript isoforms, the most expressed transcript isoform was used. Salmon (v0.8.2) (46) was used for the isoform expression quantification and the lightweight-alignment (FMD-based) index was used. Strand-specific genomic coverage or log2 FC, was computed using computeMatrix scale-regions from deepTools, where all transcripts were stretched or shrunk to the same length. For metagene plots from a given genomic location, genomic coverage (ChIPseq) and strand-specific genomic coverage or log2 FC (RNA-seq, TIF-seq) was computed using computeMatrix reference-point from deepTools. In TIF-seq coverage plots (Figure 2B, Supplementary Figure S1B, Figure 6C, D), for each TC, the row-normalized relative coverage is calculated as the percentage of TIF-seq read counts at a given position relative to the total number of TIF-seq reads associated with the TC.

Data visualization and statistics
We used R (https://www.r-project.org/) and the ggplot2 R package (47) unless otherwise noted for visualizations.

RESULTS

Many TSSs within pc-genes produce exosome-sensitive transcripts

To assess the prevalence of TSSs within pc-genes, which produce exosome-sensitive RNAs, we measured capped RNA 5′ end abundances by Cap Analysis of Gene Expression (CAGE) data from (6) to compare RRP40/EXOSC3-depleted (siRRP40) HeLa cells with corresponding data from non-depleted control (Ctrl) cells, both in biological triplicates. We first merged nearby nucleotide positions with CAGE tags on the same strand into CAGE tag clusters (TCs) and calculated for each TC the average normalized expression (as TPM) in both the siRRP40- and Ctrl-libraries. For clarity, although many CAGE TCs overlap annotated TSSs, we will refer to them as ‘TCs’ and only use the term ‘TSS’ to indicate an annotated RNA 5′ end. TC expression values were then used to define an exosome sensitivity score, ranging from −1 to 1, where 0 corresponds to equal TPM values between the siRRP40- and Ctrl-libraries, while 1 and −1 correspond to exclusive expression in the siRRP40 and the Ctrl condition, respectively (see MATERIALS AND METHODS). TCs producing exosome-sensitive RNAs (sensitivity score > 0.5) were called ‘exoTCs’, while TCs with values in the range [−0.5, 0.5] were referred to as ‘non-exoTCs’. TCs with sensitivity values < −0.5 were excluded from this study. Because our focus was on transcription initiation events within pc-gene regions, we only analyzed CAGE TCs overlapping GENCODE v19 (32) pc-gene models defined as the gene body and a 10 kb upstream region on the coding strand, thereby omitting annotated antisense- and PROMPT-transcripts from the analysis.

Although the bulk of the analyzed TCs were not exosome-sensitive, regardless of the expression threshold applied, a substantial number of exoTCs could be detected, which declined with increasing expression level threshold (Figure 1A). Using a threshold of 2 TPM, nearly a third of exoTCs overlapped predominantly with annotated transcript 5′ ends (either primary or alternative TSSs, where the most upstream annotated TSS was defined as ‘primary’ and any other annotated TSSs as ‘alternative’); an additional ∼22% of exoTCs were located upstream of the primary TSS (Figure 1B, region definitions in Figure 1C). With increasing thresholds, higher fractions of exoTCs overlapped annotated TSSs. Conversely, lower expressed exoTCs were mostly found within introns. While eRNAs as a group are exosome sensitive, only 18% of these exoTCs overlapped previously defined intronic eRNA-producing loci (45). ExoTCs overlapping 5′- or 3′-UTRs, coding regions or other exons were generally rare, regardless of expression level.

Next, we asked how many pc-genes harbored exoTCs and found that while 59.8% of the 9803 expressed genes (TPM > 2 in either siRRP40- or Ctrl-samples, Supplementary Table S1) exhibited a single non-exoTC, 20.8% contained at least one exoTC (Figure 1D). Of these 2037 genes, 787 exclusively harbored exoTC(s), of which the majority (81.8%) were single exoTC cases. Finally, 12.8% of all expressed genes had combinations of exo- and non-exoTCs. Taking these observations together, we conclude that exoTCs occur within a substantial number of pc-genes. Moreover, it is noteworthy that for many genes having a single expressed TC, that TC was exosome sensitive, at least in HeLa cells. We therefore decided to first characterize such single exoTC cases (analyzed in Figures 2–5), and then later expand our analyses to more complex cases where multiple TCs are present in the same pc-gene (analyzed below in Figures 6–8).

Characterization of exoTCs from pc-genes with a single active TSS

As detailed above, we first focused our analysis on the 644 pc-genes harboring a single exoTC with an expression level > 2 TPM in either RRP40- or Ctrl-samples. The majority of these single exoTCs (59%, 380/644) overlapped GENCODE-annotated TSSs (Figure 2A, ‘primary and alternative TSS’), as compared to the higher fraction of single non-exoTCs (94%, 5514/5866, Supplementary Figure S1A). Moreover, 15% of single exoTCs were located upstream of the primary TSSs and 16% within introns. Because the majority of TSSs overlapped annotated TSSs, we focused our analysis on these 380 single exoTCs, using the set of 5514 genes having a single non-exoTC overlapping with annotated TSSs for comparison.

CAGE reads comprise only the first 30 nt of RNAs and therefore provide limited information about the nature of the RNAs produced from a given TC. We therefore prepared paired end transcript isoform sequencing (TIF-seq) (36,48) libraries from siRRP40- and Ctrl-cells, yielding reads which contain both the capped 5′- and the polyadenylated 3′-end of the same RNA, which can then be used to assess transcript length. We plotted the coverage and fold change (FC) of TIF-seq reads across the single exoTC pc-genes defined above, using a meta-gene heat map representation (Figure 2B and C) anchored at the positions of the TCs and the annotated gene 3′ ends (see MATERIALS AND METHODS). This revealed that many single exoTCs produced exosome-sensitive prematurely terminated RNAs (bottom panels in Figure 2B and C), even though...
these TCs by selection overlapped an annotated 5' end of a full-length transcript. Previously produced total RNA-seq data, from the same cell samples (14) showed similar results (Figure 2D, bottom panels of Supplementary Figure S1F), which prompted us to analyze the exact location of these premature 3' ends. Most (76.68%, Supplementary Figure S1H) were located in the first intron downstream of the exoTC and on average 1000 nt from the 5' splice site (Supplementary Figure S1I). This, and the exosome sensitivity of these transcripts, was further confirmed by RNA-seq FC in the first intron (Supplementary Figure S1J) and is consistent with previous results describing prematurely terminated exosome-sensitive transcripts (22,49). However, at least one third of the analyzed genes showed a substantial TIF-seq coverage across the whole gene (top panel in Figure 2B). Both TIF-seq and RNA-seq data confirmed that a substantial number of these cases represented full-length RNAs, displaying robust exosome sensitivity throughout the gene (Figure 2C and D, top panel). Others contained a mixture of shorter exosome-sensitive transcripts and longer transcripts covering the whole gene (middle panel in Figure 2B–D), where the longer transcripts in some cases were also exosome sensitive (Figure 2C and D, middle panel). In contrast, transcripts produced from single non-exoTCs were predominantly full-length and exosome insensitive (Supplementary Figure S1B–E).

Based on the above observations and in order to facilitate downstream analysis, we devised a hierarchical classification system of single exoTCs with decision rules based on the above properties (see decision tree in Supplementary Figure S2A, and MATERIALS AND METHODS). This comprised four classes with the following properties (visualized in Figure 2E, left; and with specific gene examples shown in Figure 2E, right): (i) Class 1 (N = 49) exoTCs almost exclusively producing full-length exosome-sensitive transcripts, (ii) Class 2 (N = 68) exoTCs producing both full-length and prematurely terminated transcripts, both of which were exosome sensitive, (iii) Class 3 (N = 99) exoTCs producing prematurely terminated exosome-sensitive transcripts, that also give rise to full-length exosome-insensitive transcripts (the exoTC captures only the 5' ends, and therefore the average sensitivity, of both transcript types) and (iv) Class 4 (N = 64) exoTCs almost exclusively producing prematurely terminated exosome-sensitive transcripts (Supplementary Table S2).

To investigate whether genes in these established classes might share specific functions, we performed Gene Ontology (GO) over-representation analysis. Class 1 and 2 genes were enriched for GO terms related to transcription factor and regulator activities, agreeing with previous results (25), and included well known immediate early response transcription factor genes such as JUN, KLF6, ATF3, MAFF...
Figure 2. Characterization of transcripts from single exoTC-containing pc-genes. (A) Overlap between single exoTCs and genic annotation. Y-axis shows the percentage of single exoTCs overlapping the respective gene annotation features (X-axis), visualized as in Figure 1B. Gray shading indicates the set of single exoTCs analyzed in Figures 2–5, referred to as single exoTCs. (B) Coverage of RNAs produced from single exoTCs by TIF-seq. Each row corresponds to one pc-gene body, with added flanking regions (1 kb in both directions), where the 'TSS' position corresponds to the exoTC and the 'TES' position corresponds to the GENCODE-annotated transcript 3' end. Blue bars show TIF-seq read coverage from siRRP40 samples, where the 5’- and 3’-end reads are connected by a blue line. Line color intensity shows the row-normalized relative TIF-seq coverage (see MATERIALS AND METHODS), and white color indicates the absence of TIF-seq reads. Blue lines crossing the TES positions are due to transcripts harboring multiple, distinct 3’ ends. Subpanels with callouts show cases where the majority of TIF-seq reads cover the whole gene (top), cases where most RNAs are prematurely terminated (bottom) and cases with a mixture of RNA lengths (middle). (C) TIF-seq-derived exosome sensitivity of RNAs produced from single exoTCs. Heat map representation following the same conventions as in B, but with color intensities showing siRRP40 versus Ctrl TIF-seq log_2 FC in 5 bp windows. Genes were sorted in the same order as in B. (D) RNA-seq-derived exosome sensitivity of RNAs produced from single exoTCs. Heat map representation following the same convention as in C, but using RNA-seq data to calculate FC and only analyzing exonic regions within each gene. Genes were sorted in the same order as in B. (E) Classification of single exoTC genes. Left sub-panel shows cartoons of features characterizing each class. Lines beneath gene models depict the RNAs produced. Dotted lines indicate exosome-sensitive RNAs, solid lines indicate exosome-insensitive RNAs. Right sub-panel shows genome-browser examples of each class with RNA-seq tracks from siRRP40- and Ctrl-libraries (average normalized signal per bp across triplicates) at each strand, where blue color indicates the same strand as the exoTCs, while the red color indicates the opposite strand. TSSs on each strand are indicated by arrows. RefSeq gene models (67) are shown on top. (F) Gene Ontology (GO) over-representation analysis of Class 1 and 2 genes. X-axis shows -log_{10}(FDR) of top 5 terms. Numbers on the right of each bar indicate the number of genes from the two classes annotated with the respective GO term.

and DDIT3 (Figure 2F). Immediate early response genes often encode short primary transcripts with few exons (50). Consistently, Class 1 and 2 genes produce shorter primary transcripts and with fewer exons than RNAs from genes with single non-exoTCs (Supplementary Figure S2B). Class 1 genes in particular were often mono-exonic and had longer first exons than other classes, and both classes had shorter first introns, consistent with the above (Supplementary Figure S2B). However, Class 1 and 2 genes did not exhibit substantially higher degrees of intron retention than other classes (Supplementary Figure S2B). Overall, this led us to conclude that the exosome likely participates in regulating mRNA levels of such early-response transcription factors. While we found no significantly enriched GO terms for Class 3 and 4 genes, it is interesting to note that Class 3 included the PCF11 gene, which was recently reported to autoregulate its expression levels by transcription attenuation (22).
ferent, perhaps exosome-related, mechanisms to regulate their expression. In Class 1, the exosome might control full-length mRNA expression, while the Class 3 genes are likely subjected to partial premature termination of transcription within their first introns, which may influence their final overall gene expression (reviewed in (20)). RNAs deriving from such an attenuation mechanism would then be substrates of the exosome. Class 2 transcripts can be viewed as a hybrid of classes 1 and 3, where both prematurely terminated and full-length transcripts are exosome-sensitive, while Class 4 constitutes cases where premature transcription termination dominates and full-length transcripts are rare.

Prematurely terminated transcripts are often reverse strand byproducts of nearby mRNA TSSs

Next, we investigated the transcription levels of exoTCs of each class using native elongating transcript sequencing (NET-seq) of nascent RNA from HeLa cells (35). Interestingly, exoTCs from all four gene classes exhibited similar transcription levels, which in turn were on average slightly lower than those of single non-exoTCs (Figure 3A, left panel). As discussed in the introduction, the large majority of human gene promoters are bidirectionally transcribed (51,52). Analysis of opposite strand transcription upstream of the respective TCS showed that Class 4 gene promoters were highly balanced with roughly equal amounts of transcription in the forward and reverse directions, while promoters from the other gene classes displayed a higher transcription on the strand from which the exoTC of interest was present (Figure 3A, left and right panels). Notably, for Class 4 exoTCs, the reverse strand TSSs were in 60% of cases overlapping an annotated pc-gene TSS within 600 bp (Figure 3B). In other words, more than half of Class 4 exoTCs were components of annotated mRNA-mRNA bidirectional promoters, which was roughly twice as much as that of any other class, despite the fact that Class 4 TCs were required to overlap annotated pc TSSs. We therefore reasoned that Class 4 exoTCs, and their predominantly prematurely terminated RNA products, might be consequences of highly transcribed mRNA TSSs on the other strand, similar to canonical mRNA-PROMPT pairs. Consistently, opposite strand TCS upstream of Class 4 promoters were typically non-exoTCs that are not exosome sensitive, as opposed to other corresponding opposite strand TCS upstream of classes 1-3 exoTCs (Figure 3C, selected examples are shown in Figure 3D). Moreover, the region downstream of such opposite-strand TCS was more evolutionarily conserved than the corresponding region downstream of Class 4 exoTCs, with similar conservation levels as regions downstream of non-exoTCs (Figure 3E).

Chromatin data from the HeLa cell ENCODE project (53) showed enrichment patterns consistent with the above observations; while exoTCs of all classes showed similar chromatin accessibility and levels of H3K27ac; H3K4me3 histone marks, implicating active transcription. These levels were higher upstream of Class 4 exoTCs. Notably, Class 4 exoTCs themselves also lacked a gene body enrichment of the H3K36me3 histone mark, consistent with their inefficient transcription elongation (Supplementary Figure S3).

Next, we asked whether the properties of each TC class might be related to its surrounding sequence content. As previously reported, polyadenylation (pa) sites and 5′ splice sites (5′ SSs) are over- and under-represented, respectively, downstream of 5′ ends of known exosome-sensitive transcripts (e.g. PROMPTs) compared to their forward strand mRNA counterparts (18,54). All four single exoTC classes fell between these two reference sets in terms of predicted pa site occurrence, where Class 1 exoTCs were the most similar to single non-exoTCs, while Class 4 exoTCs showed a similarly strong pa enrichment as PROMPTs from ~1200 bp downstream from the exoTC (Figure 3F, left panel). This is roughly consistent with the typical position of prematurely terminated transcript 3′ ends (median Class 4 transcript length by TIF-seq was 1391 bp). Class 2 and 3 genes displayed a similar enrichment of predicted 5′ SSs as non-exoTCs, while Class 4 genes had a similar 5′ SSs enrichment profile as PROMPT regions from ~1200 bp downstream from the exoTC (Figure 3F, right panel). Class 1 genes fell between these two, possibly due to the fact that many Class 1 transcripts were short and mono-exonic. Related to the above, we have previously shown that pa site depletion, downstream of the TSS of exosome-insensitive transcripts, often coincides with CpG-enriched regions, which in turn are often limited to the first 500 bp (55). Plotting G/C content up- and downstream of TCSs in each class, showed that exoTCs from classes 1–3, which all largely produce full-length transcripts, had a clear G/C enrichment around the exoTC peak, which often extended downstream (Figure 3G). In contrast, Class 4 exoTCs had a much higher G/C enrichment upstream the exoTC than downstream, which likely reflects a G/C enrichment around their commonly occurring upstream mRNA TSSs on the opposite strand, as discussed above.

In summary, several lines of inquiry - transcription initiation bidirectionality, evolutionary conservation and sequence motif enrichment/depletion indicate that Class 4 exoTCs and their associated transcripts are by-products of the initiation of canonical, exosome-insensitive mRNA up and on the opposite strand of the exoTCs. In this sense, Class 4 exoTC regions share properties with canonical PROMPT regions, although Class 4 transcripts are on average 2–3 times longer than typical PROMPTs (1391 vs. ~500 bp (18)). This difference in length is also reflected in sequence content: once the 3′ end is reached for Class 4 transcripts, the pa site and 5′ SS enrichment is similar to those of PROMPTs (Figure 3F).

Exosome-sensitive full-length mRNAs are PAXT targets

Exosome-directed decay of nuclear RNA is mediated by adaptor complexes (36). We therefore investigated which adaptor is implicated in the degradation of transcripts deriving from the four single exoTC classes, focusing on the PAXT connection and the NEXT complex, which primarily target longer polyadenylated and short non-adenylated transcripts, respectively (7,9,14). To enable such analysis, we prepared CAGE libraries from cells subjected to siRNA-depletion of ZFC3H1 (PAXT) or ZCCHC8 (NEXT), and plotted the distribution of CAGE exoTC sensitivity scores of siZFC3H1 versus Ctrl and siZCCHC8 versus Ctrl (Fig-
Figure 3. Class 4 exoTCs are by-products of strong upstream TSSs on the opposite strand. (A) Bidirectional transcription at single exoTC promoters. Left: Boxplots show NET-seq signal distributions (Y-axis) in −100 to +500 bp regions of single exoTCs from each exoTC class (opaque colors), and the −1 to −600 bp region on the opposite strand (pale colors). Genes with a single non-exoTC were analyzed for comparison (red). Right: Boxplots show distributions of corresponding bidirectionality scores for each TC class from the same data, ranging from −1 (only signal upstream of exoTC on the opposite strand) to +1 (only signal on the exoTC strand). (B) Gene annotation overlap of TCs located upstream of, and on the opposite strand of, exoTCs. Bar plots show the percentages of TCs that are upstream and on the opposite strand of single exoTCs and which overlap a given gene annotation feature (GENCODE v19), split by exoTC class as above. (C) Exosome sensitivity of TCs located upstream of, and on the opposite strand of, exoTCs. Boxplots show distributions of exosome sensitivity scores based on CAGE (calculated as in Figure 1A). TCs analyzed as in panel B. (D) Genome-browser examples of Class 4 single exoTCs. The tracks show average normalized RNA-seq signal per bp across triplicates from siRRP40- and Ctrl-libraries at each strand. Blue color indicates the same strand as the exoTCs, while red color indicates the opposite strand. TSSs on each strand are indicated by arrows. RefSeq gene models (67) are shown not top. (E) Evolutionary conservation of exoTCs and their upstream opposite strand TCS. Y-axis shows distributions of evolutionary conservation scores (phastCons 100 vertebrate species, where 0 corresponds to least conserved and 1 most conserved) in the ±100 bp regions around exoTCs (opaque colors) and upstream opposite strand TCs defined as above (pale colors). X-axis shows TC type. (F) Enrichment of predicted 5′ sites and 5′ SSs downstream of exoTCs. X-axis shows the distance in bp from exoTCs. Y-axis shows the cumulative fraction of regions having one or more predicted sites at a given bp, moving left to right. ExoTC regions were split by exoTC class as above, as indicated by line color. Regions downstream of PROMPT TSSs and single non-exoTCs are shown for comparison (red and black lines, respectively). (G) G/C content centered on exoTCs of different classes. G/C content per base is calculated as the fraction of C or G nucleotides in 5 bp sliding windows. Color intensity indicates average G/C content per base over a 10 bp window.

In general, all exoTC classes were to some degree sensitive to both nuclear RNA decay pathways (median sensitivity > 0), however, classes 1 and 2 showed significantly (P = 0.016 and 0.037, respectively; one-sided Mann–Whitney test) higher siZFC3H1 than siZCCHC8 sensitivity, while Class 4 displayed the opposite pattern (Figure 4A). Corresponding analyses using RNA-seq data from the same cell samples, summing all reads across the gene models, gave consistent results (Supplementary Figure S4A). Importantly, these analyses showed the average sensitivity of all transcript isoforms from loci starting at the respective exoTCs. To investigate changes in PAXT- and NEXT-sensitivity across gene bodies, we plotted the average FC of RNA-seq signals for factor depletions versus Ctrl as metagenic profiles (Figure 4B). This revealed that for classes 1 and 2, the whole gene body showed only moderately higher...
RNA-seq signal in ZCCHC8-depleted cells versus Ctrl cells, while RNA-seq signal was strongly increased in ZFC3H1-depleted cells. Conversely, for Class 4, the increase in RNA-seq signal in ZCCHC8- and ZFC3H1-depleted cells compared to Ctrl cells was only visible in the first ~20% of the gene body, consistent with the premature termination of these transcripts.

For Class 3 genes, no substantial RNA-seq signal increase was observed in ZCCHC8- or ZFC3H1-depleted cells except for a modest increase in ZFC3H1-depleted cells in the first ~30% of the gene body. Similar trends were observed when plotting log2 FC of TIF-seq data from RRP40 depleted cells (Supplementary Figure S4B). We interpret this pattern as a mixture between production of primarily PAXT-sensitive prematurely terminated RNAs and the production of longer, exosome-insensitive RNA isoforms.

While mRNAs are generally quickly exported to the cytoplasm and therefore are not usually targets of the nuclear exosome, studies have shown that some mRNAs, that are retained in the nucleus, undergo decay (25,57,58). Consistently, exosome-sensitive full-length transcripts produced from exoTCs of classes 1 and 2 were enriched in RNA-seq libraries from nuclear RNA versus total RNA samples from control HeLa cells, compared to full-length transcripts from Class 3 exoTCs or from single non-exoTCs (Figure 4C).

Overall, we conclude that the examined exosome-sensitive transcripts are substrates of both the PAXT and NEXT decay pathways. However, longer exosome-sensitive RNAs (from classes 1 and 2) are primarily targeted by PAXT, consistent with the similarity of these transcripts to canonical polyadenylated mRNAs. In line with NEXT primarily targeting short transcripts, this decay pathway plays a more prominent role for Class 4 genes.

Long exosome-sensitive RNAs from pc-genes are ubiquitously expressed across cells and tissues

An important question is to what extent exoTCs, and in particular those that primarily produce full-length exosome-sensitive transcripts, are used across normal cells and tissues, and if so, whether these TCs are the main expression contributors of their cognate genes. To investigate this, we employed CAGE data from the FANTOM5 consortia, covering most human primary cells and tissues (45). Heatmap visualization of the expression of Class 1–4 exoTCs showed that classes 1–3 were expressed roughly uniformly across primary FANTOM cell groups; classes 1 and 3 exoTCs were more highly expressed while Class 4 exoTCs, as expected, were lower expressed across all cell facets (Figure 5A, B). Corresponding analysis on FANTOM5 CAGE tissue samples showed similar patterns (Supplementary Figure S5).

We then wondered whether exoTCs might be the main drivers of gene expression across cell types, or if they merely correspond to secondary TSSs with more modest expression contributions in non-HeLa cells. To address this, we calculated, for each exoTC, an expression contribution score, representing the fraction of FANTOM5 primary cell types in which the TC had the highest CAGE expression of all FANTOM5 TCs within the specific gene (Figure 5C). This revealed that Class 1–3 exoTCs were the main contributors in ~95% of primary cell groups (median contribution score 0.94–0.99), while Class 4 exoTCs showed more variance and contributed less, albeit still with a high median contribution score of 0.9.

Taken together these data imply that single exoTCs identified in HeLa cells correspond to bona fide TSSs used across most cell types, and often corresponding to the most used TSS. Thus, the exoTCs identified here are likely physiologically relevant given that many of the genes with exoTCs are functionally important, e.g. JUN, KLF6, ATF3, MAFF and DDI73. Moreover, the exosome sensitivity of these transcripts suggest that their gene expression might be regulated post-transcriptionally at the level of nuclear RNA turnover.

The NEXT sensitivity of exoTCs is correlated with their proximity to non-exoTCs

In the above analyses, we have focused on ‘simple’ cases where a given pc-gene was utilizing a single TC. However, as many genes contain multiple active TCs, and thereby the potential to employ combinations of exo- and non-exoTCs (Figure 1D, also exemplified in Figure 6A), we set out to explore such relations in terms of genomic distance, sequence...
content, expression level and utilization across primary cells and tissues.

For this analysis, we focused on pc-genes harboring at least two TCs on the coding strand, using the same expression cut-offs as employed for our single TC analysis. We considered all adjacent pairs of TCs within 3000 bp of one another and within the same gene, including a 10 kb upstream region of its primary annotated TSS. These pairs were then stratified by whether their TCs were exosome-sensitive or not, which resulted in the four TC combinations shown in Figure 6B. TC pairs consisting of two non-exoTCs did, as expected, account for the large majority of cases (67%, 2611 pairs), while the remaining pairs, involving at least one exoTC, were roughly evenly divided between the remaining three possible pair types. The spacing between TC pairs was typically 300–500 bp, except for exoTC:non-exoTC pairs, which displayed highly varied spacing but on average were further apart (Supplementary Figure S6A, median 876 bp, \( P < 2.2\times 10^{-16} \), two-sided Mann-Whitney test). Intersection with gene annotations showed that non-exoTCs were primarily overlapping annotated TSSs, whereas exoTCs were not; in exoTC:non-exoTC pairs the exoTC was primarily located in the unannotated upstream region, while in non-exoTC:exoTC pairs the exoTC was typically located within introns (Supplementary Figure S6B).

We hypothesized that the lengths of transcripts deriving from exoTCs might be influenced by the distance to the closest up- or down-stream non-exoTC. Indeed, while TIF-seq reads from exoTCs were generally short (median 1281 nt in siRRP40), when an exoTC was close to a non-exoTC (<200 bp), transcripts initiating at the exoTC were longer (Figure 6C, D: heat map visualizations to the left, TIF-seq length distributions to the right, \( P = 8.159e^{-08} \) and \( 2.729e^{-11} \) for exoTC:non-exoTC and non-exoTC:exoTC pairs, respectively) regardless of whether the exoTC preceded the non-exoTC or vice versa. The 3' ends of these longer transcripts overlapped the annotated TES of the pc-gene in 80% of cases, similar to the TIF-seq reads originating from non-exoTCs (Supplementary Figure S7A).

To inquire whether these longer transcripts originating from exoTCs were exosome sensitive, we assessed the TIF-seq FC between siRRP40- and Ctrl-library data (while corresponding RNA-seq data was available, the overlapping transcripts originating from TC pairs made it difficult to assess the exosome targeting of individual transcripts using RNA-seq data). This revealed that both prematurely terminated as well as full-length transcripts, originating from exoTCs within the pair types analyzed above, were similarly exosome-sensitive (Figure 6E, F: heat map visualizations to the left, distributions of log₂FC to the right). This in turn suggested that the CAGE-based depletion sensitivities...
Figure 6. Characterization of exosome-sensitive transcripts from genes with multiple TCs. (A) Genome-browser examples of exoTC:non-exoTC and non-exoTC:exoTC pairs within genes. The CAGE tracks show average normalized signal per bp from siRRP40- and Ctrl-libraries at each strand. Blue color indicates the same strand as the gene with exoTC and non-exoTC pairs. Red color indicates the opposite strand. TSSs on each strand are indicated by arrows. RefSeq gene models (67) are shown on top. (B) Schematic overview of the analyzed combinations of exoTCs and non-exoTCs. The top schematic shows a fictive gene model with five TCs. The bottom schematic shows all pairs of adjacent TCs that were analyzed, and the number of such pairs found across all pc-genes. (C) Length of transcripts originating from non-exoTC:exoTC pairs. Each heat map row shows one non-exoTC:exoTC pair, centered
were reasonable approximations for whole transcript sensitivities.

We hypothesized that exoTCs close to non-exoTCs would give rise to predominantly PAXT-targeted transcripts similar to the classes 1–2 single exoTC cases analyzed above, while transcripts from distal exoTCs might give rise to predominantly NEXT-targeted transcripts similar to those from Class 4. To test this, we analyzed PAXT and NEXT sensitivities of exoTCs from the same TC pairs as above, using our CAGE siZFC3H1 and siZCCHC8 libraries. When averaging over all TC pairs, exoTCs were generally NEXT but not substantially PAXT-sensitive, regardless of pair type (Figure 6G, H, top panels, and Supplementary Figure S7B). Moreover, the NEXT sensitivity of exoTCs increased with the distance between exoTCs and non-exoTCs, where exoTCs close to non-exoTCs were neither substantially NEXT nor PAXT-sensitive despite being RRP40 sensitive (Figure 6G, H, bottom panels). While this increase in NEXT sensitivity with increased TC-TC distance was compatible with the initial hypothesis, the putative exosome adaptor responsible for the observed RRP40/EXOSC3 sensitivity of exoTCs close to non-exoTCs is presently unclear.

**ExoTC proximity to annotated splice sites is correlated with the generation of long, exosome-sensitive transcripts**

Next, we asked whether sequence content around TCs could explain the correlation between transcript length and TC pair distance. By displaying predicted pA sites, 5′SSs, and G/C content for each TC pair analyzed above as a heat map (Figure 7A–C, Supplementary Figure S8A–C), we observed drastic shifts in sequence content at the respective TC positions and its immediate downstream region in pairs with one exoTC and one non-exoTC (Figure 7A–C). This was not observed in other TC pairs (Supplementary Figure S8A–C).

In particular, non-exoTCs displayed a strong occurrence of predicted 5′SSs just after, but not before their TC peak, and a corresponding depletion of predicted pA sites. In the case of non-exoTC:exoTC pairs (Figure 7A–C, top row), the depletion of pA sites extended up until the exoTC location, while the 5′SSs accumulation was strongest just after the non-exoTC. The same pattern was also evident when assessing G/C content. Thus, such TC pairs delineated a G/C (and predicted 5′SSs)-rich boundary in between them. For exoTC:non-exoTC pairs, the properties were expectedly reversed: there was an increase in G/C content before the exoTC and after the non-exoTC, but a depletion in between. The same pattern was reflected in predicted pA sites, which were depleted downstream of the non-exoTC only, and 5′SSs, which were highly enriched directly downstream of the non-exoTC, but depleted between the TCs.

For both pair types, the exception to the above observations was when TCs were close (<200–500 bp), where the sequence properties of the non-exoTCs overtook that of the exoTC, most visible in terms of G/C content (Figure 7A–C, indicated with red arrows). These sequence properties of exoTCs close to non-exoTCs coincided with their production of full-length transcripts as analyzed in Figure 6C and D. Based on this observation, we hypothesized that the occurrence of longer exoTC transcripts, when TCs were close, was not due to the TC distance itself but rather because exoTC transcripts might co-opt the first 5′SS used by the non-exoTC transcript. We tested this hypothesis by focusing on the subset of TC pairs where the non-exoTC overlapped with an annotated TSS, or was within the 5′UTR of a GENCODE gene model (71% of analyzed pairs), so that analyzed non-exoTCs were always associated with an annotated first exon. Consistent with the hypothesis, in non-exoTC:exoTC pairs, exoTCs were rarely located within the first exon originating from the non-exoTC unless the distance between TCs was <200 bp (Figure 7D). ExoTCs that were located within the first exon initiated significantly longer transcripts than exoTCs located downstream of the first exon, as assessed by median TIF-seq lengths (P = 1.731e–09, one-sided Mann–Whitney test, Figure 7E).

For exoTC:non-exoTC pairs, the exoTC can per definition not be within the same annotated exon, originating from the non-exoTC, so the same analysis was not meaningful. However, we observed a clear correlation between the median TIF-seq lengths of transcripts originating from the exoTC and the distance to the next downstream annotated 5′SS: predominantly full-length transcripts were produced from exoTCs within ∼500 bp of a 5′SS, and the proportion of prematurely terminated short transcripts increased when exoTCs were further from annotated 5′SSs, especially when the distance was >1 kb (Figure 7F, G). This was consistent with our observations for non-exoTC:exoTC pairs discussed above; transcripts become long if the exoTC is proximal to a strong 5′SS, which either could be specific for on the non-exoTC position and sorted by increasing TC pair distance. TC positions are shown with vertical dotted lines. X-axis shows the distance from the non-exoTC in bp. Left heat map shows TIF-seq reads starting from the non-exoTC, where blue bars show TIF-seq coverage and the color intensity indicates the relative coverage as in Figure 2B. Horizontal dashed lines indicate specific distances between TCs (indicated on the Y-axis). The number of analyzed TCs per TC pair distance category are indicated on the left side. Right heat map follows the same conventions, but shows TIF-seq reads starting from the downstream exoTC. Schematics on top show the specific TC type analyzed in each pair, highlighted by a red box. Right violin-boxplots show the distribution of median length (log2-scaled) of TIF-seq transcripts from upstream non-exoTCs (turquoise) and downstream exoTCs (red), split by TC pair distances (Y-axis), summarizing TIF-seq data shown in the two heat maps to the left. (D) Length of transcripts originating from exoTC:non-exoTC pairs. Organized as in panel B, but analyzing exoTC:non-exoTC pairs. Heat maps were centered on the exoTC position. (E) Exosome sensitivity of transcripts originating from non-exoTC:exoTC pairs. Heat maps organized as in panel B, but with bar color showing TIF-seq siRRP40 versus Ctrl log2FC. Combined violin-box plots to the right show TIF-seq siRRP40 versus Ctrl log2FC. For each TC, the maximum value of log2FC of positions covered by the associated TIF-seq reads was used to represent the TIF-seq log2FC of that TC. (F) Exosome sensitivity of transcripts originating from exoTC:non-exoTC pairs. Organized as in D, but analyzing exoTC:non-exoTC pairs. Heat maps were centered on the exoTC position. (G) PAXT and NEXT sensitivities of non-exoTC:exoTC pairs. Schematics on top show the TC types analyzed. The upper violin-boxplot shows the overall distribution of PAXT and NEXT sensitivity scores of TCs based on CAGE data calculated as in Figure 4A. The lower violin-boxplot shows the distribution of PAXT and NEXT sensitivities of the exoTCs in the TC pair, split by the distance to the paired non-exoTCs in bp (X-axis). (H) PAXT and NEXT sensitivities of exoTC:non-exoTC pairs. Organized as in E, but analyzing exoTC:non-exoTC pairs.
Figure 7. Analysis of the relation between TC pair distance, sequence content and transcription outcome. For all plots, the type of TC pairs analyzed is shown as a schematic on top. (A) Heat map representation of predicted pA sites around TCs in non-exo:exoTC pairs (top panel) and exoTC:non-exoTC pairs (bottom panel). The heat maps were organized as in Figure 6C. Positions of TCs are indicated by black lines. Dots indicate predicted pA sites, where the color intensity indicates the motif prediction score (see MATERIALS AND METHODS). Number of analyzed TCs per distance category are indicated on the left side. Dashed lines indicate specific distances between TCs. Red dashed lines (at 200 and 300 bp distances, respectively) and red arrows refer to specific main text discussion points. (B) Heat map representation of predicted 5SSs around TCs in non-exoTC:exoTC pairs (top panel) and exoTC:non-exoTC pairs (bottom panel).
the exoTC or shared with the non-exoTC. Otherwise, they would prematurely terminate before the non-exoTC.

In summary, these analyses demonstrated that distances between exoTCs and non-exoTCs correlate strongly with the lengths of the transcripts produced from exoTCs, a property most likely linked to the availability of strong splice donor sites utilized by exosome-insensitive transcripts, that can only be co-opted if the exoTC is proximal to the splice site.

**The majority of exoTCs within multi-TC genes are secondary TSSs across cells and tissues**

An important parameter for establishing whether exoTCs within multi-TC genes are perhaps functionally relevant is their relative expression levels compared to their paired non-exoTCs. To answer this question, we interrogated FANTOM5 CAGE data across primary cells, as when analyzing single exoTCs above (Figure 5). In order to make results between these two analyses (single versus multi-TC genes) comparable, we focused on cases where both TCs in exoTC:non-exoTC and non-exoTC:exoTC pairs overlapped annotated TSSs (25%, 97/385 and 29%, 148/514 of pairs, respectively). In general, exoTCs paired with non-exoTC were substantially less expressed across FANTOM5 primary cells (median 0.9 TPM) than their paired non-exoTCs (median 3.9 TPM), and at the same level as exoTCs in exoTC:exoTC pairs (Figure 8A, Supplementary Figure S9). Moreover, the expression level of exoTCs of any pair type was significantly \( P = 6.321 \times 10^{-13} \), one-sided Mann–Whitney test) lower than those of single exoTCs for classes 1–3 analyzed above (see dotted lines in Figure 8A). Somewhat surprisingly, there was no substantial expression difference between exoTCs that produced full-length or prematurely terminated transcripts as defined in Figure 6 (Figure 8B). Similar trends were observed when analyzing expression data at the level of tissues (Supplementary Figure S10).

Taken together, these results indicated that most exoTCs are minor contributors to overall gene expression, provided other non-exoTCs are present in the same gene. This is regardless of their position and whether they produce prematurely terminated transcripts or not. In that sense these exoTCs are more similar to single exoTCs of the Class 4 genes analyzed above, and indeed have comparable expression levels (see dotted lines in Figure 8A). However, our observation that Class 4 single exoTCs were often linked to mRNA TSSs on the opposite strand was not mirrored for exoTCs within multi-TC genes (only 19 versus 58% in Class 4 single exoTCs). Overall, these observations indicate that most exoTCs, and their transcripts, within multi TC genes have limited physiological relevance, since they are lowly expressed and their lengths are most likely a side-effect of their proximities to non-exoTCs and their downstream splice sites (Figure 7).

**DISCUSSION**

In the present study, we have established that TSSs that produce exosome-sensitive transcripts occur within many pc-genes. In many cases, these TSSs, captured as exoTCs using CAGE, overlap annotated mRNA TSSs, established by full-length cDNA sequencing. We found that the properties of exosome-sensitive transcripts can be classified largely based on two parameters: (i) whether the exoTC is the only active TC in the gene region and (ii) whether the produced transcripts are prematurely terminated, or whether they share their 3’ ends with the annotated full-length mRNAs (Figure 9).

Genes that only use one exoTC and no non-exoTCs (Figure 9A) are arguably interesting, as the abundance of their expressed transcripts will depend on exosome availability. Interestingly, a subset of these exoTCs overlap an annotated TSS and mainly produce full-length exosome-sensitive transcripts, which are primarily PAXT targets (classes 1–2). Those TSSs remain active across many cells and tissue types and in most cases constitute the main TSSs for their respective host genes. Genes in this category include well-known transcription factors as JUN, KLF6, ATF3, MAFF, and DDIT3, many of which are so-called immediate early response genes. Single exoTCs of these classes share many properties with canonical non-exoTCs (53, 57). It is somewhat surprising that immediate early gene full-length transcripts would benefit from being exosome sensitive, since they by definition need to exhibit strong transcriptional/expression responses within minutes after cells are exposed to external stimuli. A high nuclear exosome sensitivity would have the effect of decreasing the overall expression amplitude, but could be important for the rapid removal of remaining RNA copies once the transcriptional burst has ended. Moreover, constitutive nuclear degradation of full-length transcripts could serve to allow these genes to be constitutively lowly transcribed instead of fully inactive when stimuli are not present, which would allow for a faster response upon stimulation. When cells are not stimulated, the exosome would dampen RNA copy numbers, while during induction, robust production of transcripts would saturate the PAXT/exosome pathway across cells and tissues.
Figure 8. Expression of exoTCs and non-exoTCs across primary cells. For all plots, the type of TC pairs analyzed are shown as a schematic on top or on the left side. (A) Expression of TCs across primary cells. Combined violin-boxplots show the distribution of CAGE expression of TCs as log_{10}TPM across the same FANTOM5 cell type groups as in Figure 5, split by TC pair type. Color indicates which TC is analyzed (turquoise for the most upstream in the pair, red for the most downstream). Y-axis shows the types of TC pairs and X-axis shows CAGE log_{10}TPM. Median expression values of single exoTCs of different classes and single non-exoTCs across the same FANTOM5 cell facets (from Figure 5B) are plotted as vertical dotted lines. (B) Expression of TCs across primary cells, split by the lengths of the produced transcripts. Combined violin-boxplots show the distribution of CAGE log_{10}TPM of TCs in non-exoTC:exoTC (left panel) and exoTC:non-exoTC (right panel) pairs across cell type groups as in A, but split by the ratio of full-length transcripts (ratio < 0.5: TC produces mostly prematurely terminated transcripts; ratio > 0.5: TC produces mostly full-length transcripts). The reference lines are the same as in A.

and thereby achieve sufficient expression. A similar model has been proposed for nutrient response in budding yeast (59). Alternatively, the exosome sensitivity of these transcripts may be a necessary tradeoff: gene features that facilitate early gene response (short gene lengths, few or no introns) are also correlated to high exosome sensitivity (6).

Classes 2–4 also produce short, prematurely terminated RNAs which are targeted by both the NEXT and PAXT pathways. ExoTCs initiating such transcription exhibit a higher and lower average occurrence of predicted TSS-proximal pA sites and 5′SS, respectively. This is quite similar to the sequence features present within PROMPT regions, although with enrichments/depletions of lower magnitude. Class 3 genes, which produce long, exosome-insensitive RNAs in combination with short prematurely terminated RNAs, are the most similar to canonical mRNA genes. We speculate that this class may be characterized by less efficient RNAPII elongation, meaning that higher transcription initiation is needed to attain a given mRNA copy number. Possibly related to this, reduced RNAPII elongation has been shown for genes that undergo premature termination by the Integrator complex at sites of paused RNAPII (23), resulting in transcription attenuation. Because the CAGE, TIF and RNA-seq techniques we employ only assess RNAs that are >100 nt, we will in most cases not capture attenuation events close to the TSS, but premature termination further downstream the gene is widespread in pc-genes and could also lead to attenuation (reviewed in (20)). The short prematurely terminated transcripts observed in classes 2–4 could therefore be results of transcription attenuation, as reported for PCF11 (22), a Class 3 gene in our analysis.

Surprisingly, our data demonstrate that some genes contained in Class 4 harbor single TCs that almost exclusively produce prematurely terminated and exosome-sensitive transcripts. This is despite the fact that the related exoTCs correspond to annotated mRNA TSSs, which in most cases are the dominant TSSs for these genes across most cell and tissue types. However, in many cases the expression of such TSSs is likely a bystander effect of strongly expressed upstream mRNA TSSs on the other strand. Thereby, these RNAs, which share properties with canonical PROMPTs, are under lower selective pressure.

An outstanding question is which features drive exosome targeting of the sensitive transcripts. A commonly accepted idea is that mRNAs that are slowly or inefficiently processed are subject to inefficient nuclear RNA export (60) and therefore targeted by the PAXT pathway. In line with this, Class 1 genes are often mono-exonic, and Class 1 and 2 transcripts were found to be more enriched in the nucleus than exosome-insensitive transcripts. Hence, this may, at least in part, explain the PAXT sensitivities of these transcripts (56, 61). Prematurely terminated transcripts in classes 2–4 may share many of these features, since their 3′ ends predominantly reside in the first intron and they are likely not spliced. However, a clear difference to the above is the high incidence of pA sites, reminiscent of the link between exosome targeting of PROMPTs and TSS-proximal pA sites (18, 54). As many of these prematurely terminated transcripts are enriched in both PAXT and NEXT depletions, these transcripts are targeted by the exosome through different mechanisms. As reported previously (9), the recruitment of PAXT could occur through recognition of pA sites and conventional 3′end processing by the CPA machinery. NEXT could be recruited to these transcripts through its interaction with the CBC (62–64). For both full length mRNAs and prematurely terminated transcripts, it is also possible that exosome targeting is further increased by additional RNA-bound proteins. However, since the transcripts are short-lived, it is challenging to comprehensively identify the mechanisms that target them for degradation.

While single exoTCs are probably largely physiologically relevant since they remain the main TSS across cells and tissues, exoTCs co-existing with non-exoTCs (Figure 9B) within the same gene are more common, but likely of lesser physiological importance. This is because they are more lowly expressed than non-exoTCs within the same gene, both in HeLa cells and across the FANTOM5 cells and tis-
Figure 9. Models for exoTCs positioning and expression within pc-genes. For each gene/TC cartoon model, typical transcripts and their most common targeting fates are shown. Solid lines represent exosome-insensitive RNAs, dotted lines represent exosome-sensitive RNAs, where the primary exosome adaptor responsible for degradation is indicated. Below: typical enrichment/depletion of pA sites, 5' SSs and G/C content are shown as solid density plots: the dotted lines show corresponding enrichments at non-exoTCs as a reference. Also see DISCUSSION. (A) Genes with a single TC. Cartoon models of single TC genes: non-exoTCs and exoTCs of classes 1–4 are shown from top to bottom. ExoTCs were divided into four classes based on the lengths of transcripts and their exosome sensitivities. Similar to non-exoTCs, exoTCs that are at least partially producing full-length transcripts, have canonical PROMPT regions on their reverse strand, and share similar downstream sequence patterns (pA sites, 5' SSs, G/C content) as non-exoTC (in classes 1–4, the black dashed line indicates sequence pattern of single non-exoTCs as a reference), although for classes 2–3, which also produce exosome-sensitive prematurely terminated transcripts, have a slightly higher occurrence of pA sites downstream. Class 4 exoTCs produce predominantly short, prematurely terminated transcripts and most are initiated head-to-head with the TSS of another pc-gene, thus appearing like 'PROMPTs' of the other gene with the sequence pattern more similar to PROMPTs, and produce short transcripts, that are redundantly targeted by the NEXT and PAXT pathways. (B) Genes with multiple TCs. A non-exoTC:exoTC pair cartoon is shown in the upper panel, and an exoTC:non-exoTC cartoon in the lower panel. In non-exoTC:exoTC pairs: when exoTCs are close to non-exoTCs, they are mostly within the same exon, and share downstream sequence properties with the non-exoTC due to their proximity. When such exoTCs are positioned further away from non-exoTCs, they are mostly located in introns with sequence patterns similar to PROMPT TSSs and producing short transcripts that are targeted by NEXT. In exoTC:non-exoTC pairs, there is a similar distance effect as non-exoTC:exoTC pairs: when exoTCs are close to non-exoTC, exoTCs are more likely to be annotated TSSs and produce long exosome-sensitive transcripts, that likely use the same splice sites as transcripts from the non-exoTC, due to that the sequence pattern of non-exoTC 'bleeds into' that of exoTC. More distal exoTCs are not subject to sequence constraints of the non-exoTC, and produceNEXT-sensitive short transcripts, and are similar to PROMPTs in terms of sequence patterns and products. For both configurations, long exosome-sensitive transcripts appear not to be targeted by NEXT or PAXT, suggesting the possible existence of additional adaptors.

Moreover, these exoTCs mostly produce short, prematurely terminated transcripts. Longer transcripts produced from such exoTCs are rare and likely consequences of proximity to stronger non-exoTCs on the same strand, where DNA sequence constraints, in particular splice sites, of the non-exoTC and its transcripts are also imposed on the exoTCs and their products. Conversely, when exoTCs are positioned further away from non-exoTCs, either in introns or in the region upstream of the non-exoTC, the sequence pattern downstream of them will be similar to that of PROMPTs and consequently produce exosome-sensitive short transcripts.

Our analysis leads to two important open questions regarding the metabolism of longer exosome-sensitive transcripts, originating from exoTCs from within multi-TC genes. First, their exosome sensitivity is surprising given that they share 3' ends with transcripts originating from non-exoTCs, that reside only a few hundred bp away. Hence, the DNA that encodes them is almost identical. This would suggest that either the local sequence downstream of each TC is highly informative for exosome targeting, or that the transcripts are biochemically different in some other way, e.g. with regards to processing and/or nuclear export efficiency. To investigate the latter in detail, it would be necessary to employ long read sequencing, as isoform convolution from standard RNA-seq does not have the resolution to distinguish different RNAs initiating from nearby TSSs. Second, although the longer transcripts ap-
pear clearly RRP40/EXOSC3 sensitive in both our CAGE and TIF-seq experiments, they are not enriched in neither NEXT nor PAXT-depleted cells. Perhaps, these transcripts are targeted by alternative exosome adaptors, which remain to be discovered.

Given our results discussed above, there may be several mechanisms underlying the production of exosome-sensitive transcripts within mRNA genes. Exosome-sensitive full-length mRNAs might have evolved to be co-regulated with exosome-sensitive transcripts. Exosome sensitivity may be a ‘necessary evil’ to accommodate other constraints, like mediating burst transcription of immediate early response genes. Such transcripts are typically PAXT sensitive and this might be a contributing reason to why depletion of a factor in the PAXT pathway impairs mouse embryonic stem cell differentiation (16). Other exosome-sensitive transcripts are prematurely terminated and may in many cases be bystander effects of the transcription initiation of other loci. Alternatively, such RNAs may be the results of processes that are not directly linked to the host gene, e.g. transcription of mRNAs or other non-coding RNAs. Close-by TSSs may also affect each other functionally by the act of transcription, even though they are not producing full-length transcripts, e.g. by transcriptional interference by RNAPII elongation through downstream TSSs (65).

More generally, sensitive 5′ sequencing-based methods like CAGE have the ability to discover a wealth of uncharacterized alternative TSSs, but not all of these will produce physiologically relevant RNAs (66,66). Exosome sensitivity and the lengths of the produced RNAs, as presented here, are therefore important features for predicting alternative TSS relevance within complex genomes.

DATA AVAILABILITY

All TIF-seq, SLIC-CAGE datasets are available at the Gene Expression Omnibus (GEO) under accession number: GSE147655. Code for TIF-seq processing and CAGE annotation is available at GitHub (https://github.com/PelechanoLab/TIFseq2 and https://github.com/MengjunWu/Exosome_sensitive_TSS). Supplementary Data are available at NAR online.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS


FUNDING

Work in the Sandelin laboratory was supported by grants from the Lundbeck Foundation; Danish Cancer Society; Novo Nordisk Foundation; Independent Research Fund Denmark; Carlsberg Foundation; Work in the Jensen laboratory was supported by the Danish Cancer Society; Lundbeck Foundation; Independent Research Fund Denmark; Work in the Pelechano laboratory was supported by the Swedish Research Council; Wallenberg Academy Fellowship; Swedish Foundations’ Starting Grant (Ragnar Söderberg Foundation); Karolinska Institutet. Funding for open access charge: Novo Nordisk Foundation.

Conflict of interest statement. None declared.

REFERENCES


Supplementary Information

Figure S1

A: Overlap between single non-exoTCs and genic features. Y-axis shows the percentage of single non-exoTCs overlapping the respective gene features (X-axis), visualized as in Fig. 1B. Grey shading indicates single non-exoTCs which were used for further analysis.

B: Coverage of RNAs produced from single non-exoTCs by TIF-seq. Heatmap follows the same conventions as in Fig. 2B, but each row corresponds to a gene with a single non-exoTC.
C, D: Exosome sensitivity of RNAs produced from single non-exoTCs as measured by TIF- (C) and RNA-seq (D). Heatmaps follow the same conventions as in Fig. 2C and 2D, respectively, but each row corresponds to a gene with a single non-exoTC. Genes were sorted in the same order as in B.

E: Gene coverage of RNAs produced from single non-exoTCs by RNA-seq. Heatmaps follow the same conventions as in B, but color intensities show the normalized (count per million) RNA-seq coverage. White color indicates the absence of RNA-seq coverage. Results from Ctrl- (left panel) and siRRP40-samples (right panel) are shown.

F: RNA coverage from single exoTCs by RNA-seq. Heatmaps and subpanels follow the same conventions as in Fig. 2B. Genes were sorted in the same order as in Fig. 2B.

G: Schematic representation showing gene features used to annotate 3’ ends of TIF-seq reads. For details, see Methods.

H: Fractions of prematurely terminated transcripts from single exoTCs and non-exoTCs. Left: the Y-axis shows the percentage of prematurely terminated TIF transcripts (see Methods), originating from exoTCs or non-exoTCs, indicated at the X-axis and split by sample siRRP40- and Ctrl-samples. Dotted box indicates the TIF-seq reads used for analyzing the genomic location of their 3’ ends (right). Right: Y-axis shows the percentage of 3’ ends of TIF-seq reads overlapping with a given genomic location (X-axis). ‘Other location’ corresponds to genomic features (as shown in G), that are within the gene body but excluding the first intron and 3’UTR.

I: Distribution of the distance from first 5’ SS to the 3’ end of prematurely terminated transcripts. Y-axis shows the length distribution of the distance between single exoTC TIF-seq 3’ ends to the first upstream 5’ SS. Only TIF-seq reads with the 3’ end located in the first intron are included in the analysis. X-axis shows the distance in bp to the first upstream 5’ SS. Colors indicate library.

J: Exosome sensitivity of RNAs falling within the first intron. Y-axis shows the average RNA-seq fold change (siRRP40 vs Ctrl) as a function of the distance to the annotated 5’ SS of the first (black) and second (grey) introns.
Figure S2: Classification and properties of single exoTC genes.

A: Classification of single exoTCs. Left sub-panel shows the decision tree utilized for classification of single exoTCs and their genes (see Methods). Right sub-panel shows cartoons of features characterizing each class: the same colors are used to distinguish classes in Figures S2-S5.

B: Properties of single exoTC gene classes. Boxplots show, from left to right, i) distributions of annotated transcript lengths ii) number of exons, iii) lengths of the first exon, iv) length of first intron and v) intron/exon total RNA seq RPKM ratio using Ctrl cells. All measurements are based on the location of the exoTC and its linked gene/transcript annotation. All X axes are log2 scaled. Colors indicate exoTC class as defined in A. Genes with a single non-exoTC were analyzed for comparison (red).
**Figure S3**: Chromatin ChIP-seq signals around Class 1-4 single exoTCs and single non-exoTCs. The profile plots show, from left to right, average i) DNase I hypersensitivity site (DHS-seq) signals, ii) H3K4me3 ChIP-seq signals, iii) H3K27ac ChIP-seq signals and iv) H3K36me3 ChIP-seq signals, centered around single non-exoTCs and exoTCs stratified by class and with an upstream and downstream 5 kb added.

**Figure S4**: Exosome sensitivity of single exoTCs based on RNA-seq and TIF-seq data.

**A**: Exosome decay pathway sensitivities for single exoTC genes based on RNA-seq data. Combined violin-boxplots showing the distribution of sensitivity scores calculated for siRRP40, siZFC3H1- and siZCCHC8-samples vs. their Ctrl samples based on RNA-seq data, and stratified by single exoTC classes (X-axis). Y-axis shows the sensitivity score calculated as in Fig. 1A. RNA-seq reads summed over the gene model of exoTC genes were used for the sensitivity score calculation.

**B**: Exosome sensitivity of single exoTC transcripts across gene bodies based on TIF-seq data. Metagene profile plot showing the average TIF-seq log₂FC of RRP40- vs Ctrl library, stratified by exoTC class as indicated by color. The log₂FC are plotted across TSSs and the annotated gene 3' ends with 1 kb added both up- and down-stream.
Figure S5: Expression of single exoTCs across tissues

Panels A-C were organized as those in Figures 5A-C, but based on CAGE data from FANTOM5 tissue groups instead of groups of primary cells.

Figure S6
Figure S6: Analysis of combinations of exoTCs and non-exoTCs within genes with multiple TCs

A. Distribution of distances between TCs in TC pairs. Combined violin-boxplots showing the distribution of distances in bp between TCs in pairs, split by pair type. X-axis shows the distance and Y-axis shows the type of TC pairs, also shown as cartoons.

B: Overlap between TCs in TC pairs and genic features. Matrices showing the percentage of TCs within the four types of TC pairs overlapping with a given feature. The type of TC pair is shown as a schematic on the top and the left side of the table. Rows show genic features overlapping with the downstream TC highlighted by a red box on the side schematic. Columns show genic features overlapping with the upstream TC highlighted by a red box on the top schematic. Matrix cells show the percentage of TC pairs having a given combination of overlaps.
Figure S7: Characterization of transcripts from genes with multiple TCs

A. Lengths of transcripts originating from non-exoTC:non-exoTC pairs. Heat maps were organized as in Fig. 6B, but analyzing non-exoTC:non-exoTC pairs. Heat maps were centered on the most 5’ non-exoTC.

B. PAXT and NEXT sensitivities of exoTC:exoTC pairs. Boxplot-violin plots were organized as in Fig. 6F, but analyzing exoTC:exoTC pairs.
Figure S8. Analysis of sequence content in TC pairs

Heat map representation of predicted pA sites (A), predicted 5' SSs (B) and G/C content (C) around TCs in exoTC:exoTC pairs (top panel) and non-exoTC:non-exoTC pairs (bottom panel). Heat maps were organized as in Fig. 7A, 7B, 7C, respectively, but analyzing exoTC:exoTC and non-exoTC:non-exoTC pairs.
**Figure S9:** Expression differences between paired TCs across primary cells. Combined violin-boxplots showing the distribution of expression differences of CAGE log$_2$TPM between paired TCs across the same primary cell type groups as in Fig. 8, split by TC pair type (Y-axis). X-axis shows the expression difference values calculated as log$_2$(upstream TC expression)-log$_2$(downstream TC expression), where 0 (no difference in expression between paired TCs) is indicated as a black dashed line.

**Figure S10**
**Figure S10. Expression of exoTCs and non-exoTCs across tissues**

For all plots, the type of TC pairs are shown as a schematic on the top or on the left side.

**A. Expression of TCs across tissues.** Organized as in Fig. 8A, but analyzing expression across the same FANTOM5 tissue facets as in Fig. S5.

**B. Expression difference between paired TCs across tissues.** Organized as in Fig. S9, but analyzing tissue facets as in A.

**C. Expression of TCs across tissues, split by length of produced transcripts.** Organized as in Fig. 8B, but analyzing tissue facets as in A.

---

**Figure S11. Western blot verification of siRNA depletions**
Paper II
Identifying determinants of nuclear RNA exosome degradation pathways using machine learning approaches

Mengjun Wu¹, Manfred Schmid², Torben Heick Jensen², Albin Sandelin¹

¹ The Bioinformatics Centre, Department of Biology and Biotech and Research Innovation Centre, University of Copenhagen, Ole Maaloes Vej 5, DK2200 Copenhagen N, Denmark
² Department of Molecular Biology and Genetics, Aarhus University, C.F. Møllers Alle 3, Building 1130, Aarhus 8000, Denmark

Abstract

The RNA exosome serves an important nuclear degradation task targeting diverse transcript species. In human cells, the substrate specificity of nucleoplasmic RNA exosome degradation is achieved through two major exosome adaptors – the nuclear exosome targeting (NEXT) complex and the poly(A) exosome targeting (PAXT) connection. Although previous studies have revealed specific DNA/RNA features distinguishing these decay pathways, no comprehensive study has evaluated their contributions quantitatively. Here, we systematically collect relevant sequence features and organize them into four categories, develop machine learning models using Random Forests that predict their RNA exosome targeting pathway, and select the most discriminative features by quantitatively assessing the contribution of each feature. We identify features of the transcript end site (TES) to be most predictive for distinguishing the NEXT and PAXT pathways; specifically, the lack of canonical 3’ end processing by the cleavage and polyadenylation (CPA) machinery was found to be most characteristic of NEXT targets. Previously well-known features relevant to exosome targeting such as GC content of transcription start site (TSS), TSS-proximal 5’ splice site (5’ SS) were only found to be distinct for NEXT targets but not able to distinguish PAXT from non-exosome targets.

Introduction

The pervasive transcription of mammalian genomes gives rise to an excessive amount of RNAs, many of which are targeted in the nucleus by the RNA exosome, a highly conserved 3’-5’ exo- and endonucleolytic multisubunit complex (Houseley & Tollervey, 2009; Jensen et al., 2013; Kilchert et al., 2016; Schmid & Jensen, 2008, 2018). Nuclear RNA exosome targets are composed of a variety of species, including prematurely terminated RNAs from within protein-coding loci, a number of long non-coding RNAs (lncRNAs), such as promoter upstream transcripts (PROMPTs)/upstream antisense RNAs (uaRNAs), enhancer RNAs (eRNAs), and several stable nuclear RNAs such as rRNAs, tRNAs, small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) (Core et al., 2008; Flynn et al., 2011; Preker et al., 2008; Schmid & Jensen, 2018; Seila et al., 2008).
While the RNA exosome is highly efficient in degrading the large amount of RNAs presented to it, the enzyme complex itself is incapable of specifying its individual substrates. This is of critical importance given that the exosome functions on both complete substrate decay as well as in the maturation of RNAs as part of their functionalization. In the nucleoplasm of human cells, substrate specificity relies on the ability of the exosome to connect to one of two exosome adaptors: the nuclear exosome targeting (NEXT) complex or the poly(A) exosome targeting (PAXT) connection (Lubas et al., 2011; Meola et al., 2016), which connect to the RNA exosome through the common RNA helicase hMTR4/Skiv2L2 in a mutually exclusive manner (Schneider & Tollervey, 2013; Schuch et al., 2014). Besides hMTR4, the NEXT complex consists of the Zn-knuckle protein ZCCHC8 and the RNA recognition motif (RRM)-containing RBM7 (Lubas et al., 2011), whereas the core moiety of PAXT consists of hMTR4 heterodimerizing with the Zn-finger protein ZFC3H1, making additional elusive contacts with the ZC3H3, RBM26/RBM27, and nuclear poly(A) binding (PABPN1) proteins (Meola et al., 2016; Silla et al., 2020).

Present characterization of the RNA substrates of the NEXT and PAXT pathways suggests that NEXT primarily targets short, unspliced and non-adenylated RNAs (Lubas et al., 2011, 2015), while PAXT mediates the exosomal degradation of RNAs that are often longer and polyadenylated. In addition, the PAXT pathway may also target spliced RNAs with long nuclear residence times (Meola et al., 2016; Ogami et al., 2017; M. Wu et al., 2020). Despite these reported distinguishing features, other genomic elements and/or biological processes may contribute to RNA targeting by these different pathways. For example, it is well established that nuclear exosome targets are generally transcribed from loci with high densities of transcription start site (TSS)-proximal poly(A) sites (PASs) and low densities of 5’ splice sites (5’ SSs) (Almada et al., 2013; Chen et al., 2016a; Ntini et al., 2013; M. Wu et al., 2020); however, it is not entirely clear whether these features are equally important for both NEXT and PAXT targeting. A previous study also showed some major differences in 3’ end processing of the two pathways (G. Wu et al., 2020); however, it is unclear to what extent such differences contribute to the substrate specificity. Finally, although previous case studies have uncovered that some RNA binding proteins (RBPs) are important for recruiting the RNA exosome to its target RNAs (Houseley & Tollervey, 2009; Schmid & Jensen, 2018), no systematic genome-wide investigation has been conducted.

Here, we used machine learning approaches to comprehensively evaluate the contribution of relevant genomic and RNA processing features to differentiate NEXT and PAXT pathways. We find the features relevant to TES-proximal RNA processing to be most predictive for distinguishing different exosome pathways; specifically, lack of canonical 3’ end processing by cleavage and polyadenylation (CPA) machinery were found to be most characteristic of NEXT targets. While other previously reported
features relevant to exosome targeting such as TSS GC content, TSS-proximal 5’ splice site (5’ SS) were only found to be distinct for NEXT targets but not able to distinguish PAXT from non-exosome targets.

Results

Exosome target classification, feature design and machine learning framework

To identify determinants for PAXT- and NEXT-mediated RNA decay pathways, we used the *de novo* HeLa transcriptome annotation from (Lykke-Andersen et al., 2021) as a framework for an RNA-seq based classification of transcription units (TUs) into PAXT-, NEXT- or ‘non-exosome’ targets (Meola et al., 2016) (see Methods), which yielded 1043 PAXT, 2552 NEXT and 2810 non-exosome targets, respectively. Consistent with previous observations, non-exosome targets were mostly full-length protein-coding RNAs, that are long and derive from multi-exonic genes; whereas NEXT targets were generally short and mono-exonic RNAs; while PAXT targets mainly contained mono-exonic protein-coding RNAs and IncRNAs with lengths between those of non-exosome and NEXT targets (Figure S1A-C).

We next considered different types of molecular features, covering the relevant TSSs and TESs; such as DNA/RNA sequence feature, chromatin modification environment and possible RBP-RNA interaction. These features were organized into four classes (Figure 1A and Table 1). For RBP binding measured by CLIP-seq, we initially collected 401 sets of CLIP-seq peaks called from 171 RBPs across different cell lines, using several peak callers. In addition, we designed the feature ‘GC spread’ to be a proxy of the width of region with G/C enrichment immediately downstream of TSSs (see Methods); as observed in our previous works, the G/C enriched regions downstream of non exosome-sensitive TSSs are often broader than that of exosome-sensitive TSSs (Chen et al., 2016a; M. Wu et al., 2020). Comparing the calculated GC spread between PROMPT TSSs obtained from (Chen et al., 2016a) with non exosome-sensitive TSSs from (M. Wu et al., 2020) showed the feature well reflected the width of G/C enriched regions immediately downstream of TSSs, and by design it is more sensitive to capture narrow G/C enriched regions compared to broad ones (Figure S2A). We then examined the correlation of all the features by taking the values in all three exosome target categories and calculating the pairwise Spearman correlation coefficients; the results showed only few features with high correlation while the whole feature space is not highly correlated (Figure S2B).

To predict the exosome target pathways, we employed Random Forests as the machine learning model, and performed three binary classifications independently: NEXT targets vs. non-exosome targets, PAXT targets vs. non-exosome targets and NEXT targets vs. PAXT targets. Binary classifications were chosen instead of multi-class classification as the former allow us to identify specific features
distinguishing the individual classes. We used z-scores calculated from the change in performance by permuting feature values to assess the individual feature importances. However, this feature importance assessment suffers from overemphasizing features with low information content, which might still have consistent but very small effects on performance, resulting in a high z-score due to scaling by very small variances in prediction error observed during feature value permutations (see Methods). As such features would not be considered to be generally discriminative, we used entropy to eliminate those low information features and retain features with entropy larger than 0.5 across three exosome target categories (Table 1, Figure S2C). Promoter features – TATA box and INR element – are filtered out, agreeing with the previous observations that most human promoters are TATA box and INR element independent (Andersson & Sandelin, 2020; Carninci et al., 2006). For CLIP-seq RBP binding sites, 31 sets of peaks of 24 RBPs for Class 2 and 58 sets of peaks of 35 RBPs for Class 3 are retained after filtering. In order to obtain consistent and reproducible results, we opted to choose a relatively low importance score of 0.5 as threshold and performed iterative feature selection by training a model, removing features below the threshold and then retraining the model on the remaining features (Figure 1B). Training and feature selection was repeated until the feature set did not change anymore. A stable feature set of consistently significant features was obtained after at most 10 iterations in each case (data not shown). Feature selection resulted in only very minor changes of prediction performance and the importance scores of the selected features of the final models were used for determining the most discriminative features.

Figure 1. Feature design and machine learning framework
Determinants for classifying NEXT and non-exosome targets

The iterative feature selection yielded a large reduction of the number of features in Class 2 and 3 while the number of features in Class 1 and 4 retained all features (see Figure S2C, Figure 2A). The feature selection resulted in little change in classification performance for all classes (Figure. 2B). All single-class and class-combination models showed good to very good performance (F1 score ranging from 0.75 to 0.9 and AUC from 0.83 to 0.95) (Figure 2B). Combination of features of different classes did
not affect the performance positively or negatively, indicating that there were no synergistic effects of features from different classes. The predictive models showed RNA processing proximal to TSS and TES (Class 2 and 3) and chromatin environment around TES (Class 4) to be much more predictive (with F1 scores around 0.90 and AUC around 0.95) than the Class1 TSS related features with an F1-score of 0.75 and AUC of 0.83.

Figure 2. Predictive model of NEXT vs. non-exosome targets. 

A. Number of features retained after iterative selection. Bar plot showing the number of features (X-axis) for each feature class (Y-axis) after iterative selection. B. Classification performance by Random Forest for different combination of features. Bar plots in the upper panel showing the average performance (F1 score on the left, AUC on the right) over 10 repetitions for initial feature set and consistent feature set after iterative selection.
error bar shows the standard deviation of the performance over 10 repetitions. The lower panel shows the features class (dark dots) or combination of features classes (dark dots connected by black solid line) used for classification. C. Feature importance of Class 1 features. Bar plot (left panel) showing the feature importance score (X-axis) of top ranked features (Y-axis); the rows are ordered by importance score. The distribution of selected features for the two exosome target categories are shown as density plots on the right panel. To obtain best visualibility, two types of density plots (histogram or smoothed histogram by kernel density estimation) and two scales (original and logarithmic) of feature values on X-axis are used. D. Feature importance of Class 2 features. Organized as in panel C, but analyzing Class 2 features. E. Gene Ontology (GO) over-representation analysis of RBPs in significant Class 2 features. X-axis shows -log_{10}(FDR) of up to 15 top terms. Numbers on the right of each bar indicate the number of RBPs annotated with the respective GO term. F. Feature importance of Class 3 features. Organized as in panel C, but analyzing Class 3 features. G. GO over-representation analysis of RBPs in significant Class 3 features. Organized as in panel E, but analyzing Class 3 features. H. Feature importance of Class 4 features. Organized as in panel C, but analyzing Class 4 features.

We then assessed the significant features in each feature class (z-score > 1, only up to 15 most significant features were shown in the following figures: Figure 2C-D,F,H). For Class 1, sequence features GC spread and G/C content were among the highly significant features, where GC spread, as a proxy for width of G/C enriched region downstream of TSS, was more discriminative than the accumulated GC content around the TSS (Figure 2C, left panel). Other top ranked features included histone modifications H3K4me1, H3K4me3, H3K4me2, transcription levels (NET-seq) and Pol II loading. In order to identify how differences of individual features between the two target categories affect classification, we plotted the distributions for selected top-ranked features (Figure 2C, right panel). This revealed that the GC spread was on average smaller with higher variance in NEXT targets compared to the non-exosome targets, consistent with previous observations (Chen et al., 2016a; M. Wu et al., 2020). For histone modifications, NEXT targets had higher average levels of H3K4me1 and lower average levels of H3K4me3, which indicates enrichment of enhancer chromatin marks in NEXT targets (Andersson & Sandelin, 2020; Heintzman et al., 2007; Robertson et al., 2008) and agreeing with many of NEXT targets being enhancer RNAs (Meola et al., 2016). The distribution of transcription levels showed trends consistent with reported observations, i.e., that exosome targets are more lowly transcribed compared to non-exosome targets (Andersson et al., 2014; Lloret-Llinares et al., 2018).

For Class 2 features, the performance relied to a large extent on a single feature - the binding of DDX3X, which belongs to the DEAD-box helicases family (Figure 2D). DDX3X is reported to function both in the nucleus and cytoplasm, and play a diverse role in regulating transcription, mRNA maturation, export and translation. Notably, other top features were mainly related to RNA splicing, including 5’ SS motifs, binding of RBPs, which function in splicing and assembling spliceosomes, such as PRPF8, EFTUD2, U2AF2 (Figure 2D, left panel). In addition, to further investigate what biological processes the significant RBPs might be involved in, we performed Gene Ontology (GO) over-representation analysis of RBPs from all significant CLIP-seq binding and RBP-binding motif features, the top GO terms showed that the significant RBPs are generally involved in RNA processing, especially splicing (Figure 2E). Plotting distributions of the top features showed a depletion of all these features in NEXT compared
to non-exosome targets (Figure 2D, right panel). This possibly implies that NEXT targets in general lack RNA processing like splicing, compared to non-exosome targets. Interestingly, while the enrichment of TSS-proximal PAS motifs in exosome targets has been previously observed, it was not ranked as an important feature for distinguishing NEXT and non-exosome targets.

The top features of Class 3 showed similar importances (Figure 2F), which implied several possibilities: either that the good predictive performance relied on a combination of these features, or that the top features were highly correlated. We examined the latter possibility by plotting the pairwise Spearman correlation of feature values across the two exosome target categories for the top 15 features. The correlation heatmap indeed shows that most RBP binding features are highly correlated while the features related to PAS motifs and strength are weakly to moderately correlated (Figure S3).

Examining the top features in combination with their distributions (Figure 2F), we found for TES-proximal RNA processing in NEXT targets lack features relevant to canonical 3’ end RNA processing by the cleavage and polyadenylation (CPA) machinery, e.g., a well-positioned PAS motif upstream of TES (cleaved PAS), PAS strength and binding of cleavage related RBPs such as CSTF2T, CPSF7, NUDT21 (CPSF5). GO analysis showed that in addition to splicing factors also observed before, the significant RBPs in class 3 are also over-represented in terms related to 3’ end processing like RNA cleavage for polyadenylation (Figure 2G). This is consistent with previous studies that many NEXT targets do not undergo canonical 3’ end processing (Lykke-Andersen et al., 2021; G. Wu et al., 2020).

For Class 4 features, the performance, like Class 2, relied mostly on a single standout feature - H3K36me3 (Figure 2H). Higher average levels of H3K36me3 around TES were observed in non-exosome vs. NEXT targets. As H3K36me3 is known to be enriched in the gene body during active transcription and associated with efficient transcription elongation (Wagner & Carpenter, 2012), additionally, H3K36me3 has been previously reported to be only deposited downstream of the first intron (Huff et al., 2010), lower levels of H3K36me3 around NEXT TES agree with the fact that NEXT targets are usually short, lack efficient elongation, and not spliced.

To compare the relative importances of features of all four classes in predicting NEXT and non-exosome targets, we further assessed the importances of the features in the full model (shown in Figure S4). Consistent with the prediction power (Figure 2B), the top-ranked features are mostly from Class 2-4, while the Class 1 features were lower-ranked. In accordance with the individual models, the two top-ranked features correspond to the two stand-out features of Class 2 and 4, followed by top-rank features of Class 3.
In summary, NEXT and non-exosome targets could be well classified from all four feature classes, although the TSS features were less predictive compared to RNA processing features at both 5’ and 3’ end, as well as TES features. We further measured the extent to which features of different classes contributed to the predictive model, and found the width of the G/C enriched region downstream of TSS, the H3K4 methylation status to be the most discriminative features in TSS. In TSS-proximal RNA processing, 5’ SS motif and RBPs related to splicing were the most discriminative features, while interestingly PAS was not considered as the important discriminative feature at 5’ end. PAS motifs and RBPs related to canonical 3’ end processing like RNA cleavage were the most discriminative in TES-proximal RNA processing. Finally, H3K36me3 was the most discriminative feature at the TES.

**Determinants for classifying PAXT and non-exosome targets**

We next applied the model for classifying PAXT and non-exosome targets. Iterative feature selection resulted in fewer features compared to the NEXT vs. non-exosome classification for all classes, particularly prominent in Class 2 and 3 (Figure 3A, Figure 2A). The classification performance for PAXT vs. non-exosome targets was not affected by feature selection, however it was in general worse for all four classes compared to NEXT vs. non-exosome (Figure 3B, Figure 2B), while F1 score and AUC for the best performing class were around 0.75 and 0.84, the worst performing Class 1 had values around 0.6 and 0.65, respectively. No synergistic effects were observed when combining different feature classes. The model with Class 3 features of TES-proximal RNA processing yielded the best performance in classifying PAXT and non-exosome targets, whereas Class 1 TSS related features had the worst performance.
Given the performance, Class 1 features showed limited predictability with respect to PAXT and non-exosome targets, but was better than random classification. Examining the significant features, we found, in accordance with the limited predictive power, the distribution of the three top features were largely similar between the two target categories but the NET-seq and GC spread showing higher variance in PAXT targets (Figure 3C).

Class 2 features gave good prediction performance. Assessing the top ranked features, we found all of them are RBP bindings and share large similarity with the significant features in NEXT vs. non-exosome targets, where DDX3X stands out as the most dominant discriminative feature, followed by
similar splicing related RBPs like EFTUD2, PRPF8 (Figure 3D, left panel). Distribution of these features showed PAXT targets to also have a general lack of binding of the RBPs (Figure 3D, right panel). GO analysis also showed the significant RBPs were over-represented in terms relevant to RNA processing like splicing (Figure 3E). Nevertheless, despite the similarities, the enrichment of TSS-proximal 5’ SS motif was not considered a discriminative feature for classifying PAXT and non-exosome targets, indicating that the lack of splicing in PAXT targets might be due to other motif-independent mechanisms.

Class 3 features are the most predictive feature class. The top ranked features, most of them RBP bindings, also showed similar importances (Figure 3F) and the majority having only moderate correlation between them (Figure S5). While sharing some similar RBPs with top Class 3 features in NEXT vs. non-exosome targets, they differed insofar that the 3’ end cleavage and polyadenylation related features were not among the most discriminative features for classifying PAXT vs. non-exosome targets. GO analysis showed significant RBPs over-represented in terms of general biological processes like regulation of gene expression, mRNA metabolic process (Figure 3G).

Class 4 features are moderately predictive, though less than Class 2 and 3 features, where the H3K36me3 is the single standout significant feature again, which has average lower levels around TES of PAXT targets compared to that of non-exosome targets (Figure 3H).

We further analyzed the relative feature importance of four classes in the full model. Similar to NEXT vs. non-exosome classification, the result to a large degree is consistent with prediction power and feature rank observed in the individual classes (Figure S6).

Taken together, PAXT vs. non-exosome target models were less predictive compared to NEXT vs. non-exosome target models. RNA processing related features had better predictive power, and the TES-proximal RNA processing features gave the best performance. While PAXT vs. non-exosome targets could not be classified well from TSS features; for other feature classes, which were more predictive, RBPs related to splicing but not the 5’ SS motif were the most discriminative features in TSS-proximal RNA processing. In TES-proximal RNA processing, different from NEXT vs. non-exosome classification, PAS motifs and CPA related RBP bindings were not ranked as discriminative features, while RBP binding related features were ranked as the most discriminative features, the relevant RBPs were only found to be enriched in some general rather than specific biological processes. Similar to NEXT vs. non-exosome classification, H3K36me3 was also the most discriminative features in TES.

**Determinants for classifying NEXT and PAXT targets**
Finally, we applied the Random Forest model to select features distinguishing NEXT from PAXT targets. After iterative feature selection, there were fewer features of all classes, in particular for Class 2 and 3, compared to the other two classifications above (Figure 4A). Feature selection only had a small effect on classification performance. The performance was in general worse than NEXT vs. non-exosome and slightly better than PAXT vs. non-exosome classification, with F1 score and AUC ranging from 0.65 to 0.75 and 0.7 to 0.84, respectively. Similar to PAXT vs. non-exosome targets, Class 3 features of TES-proximal RNA processing gave the best prediction performance (Figure 4B).

Figure 4. Predictive model of PAXT vs. NEXT targets. Panel A-H are organized as those in Figure 2A-H, but based on NEXT and PAXT targets.
While Class 1 features were not considered to be very discriminative given the prediction power, we found, similar to NEXT vs. non-exosome targets, three features – H3K4me1, H3K4me3 and GC spread – that showed considerably higher importance than others (Figure 2C, Figure 4C), indicating that for Class 1 features, PAXT and non-exosome targets are somewhat similar, which is also reflected by the poor prediction performance of Class 1 features for PAXT vs. non-exosome targets. By plotting the distribution of the three most prominent features, we found that NEXT targets have on average higher levels of H3K4me1 and lower levels of H3K4me3, implying that NEXT targets comprise more eRNAs than PAXT targets. The GC spread distribution indicates the width of the G/C enriched region downstream of TSS is smaller on average in NEXT compared to PAXT targets, which displayed a more well-defined width as evidenced by the smaller variance.

Class 2 features showed slightly improved performance compared to Class 1 features (Figure 4B). Examining the top ranked features and their distribution, we found DDX3X and 5’ SS motif to be the two most important features showing much higher importance than the other features (Figure 4D). In addition, NEXT targets in general are more lacking in RNA processing features compared to PAXT targets, e.g. 5’ SS, RBP motifs, and RBP bindings. GO analysis of significant RBPs showed they are over-represented mostly in terms relevant to RNA processing processes like RNA splicing (Figure 4E).

Combining Class 1 and 2 features showed some positive synergistic effects on the performance, indicating they provided complementary information for classifying NEXT with PAXT targets (Figure 4B). However, we did not observe such effects in the previous two comparisons (Figure 2-3B).

Class 3 features yielded the best performance (Figure 4B). Two features, PAS strength around TES and well-positioned PAS upstream of TES, stood out as the most discriminative features, and both had higher values in PAXT than NEXT targets (Figure 4F), consistent with previous studies that showed the 3’ end processing of PAXT targets are dependent on the CPA machinery while NEXT targets are not (G. Wu et al., 2020). Analyzing other top ranked RBPs, showed these RBPs had less binding in NEXT compared to PAXT targets. Consistently, GO analysis showed that the significant RBPs were over-represented in terms related to RNA 3’ end processing like RNA cleavage (Figure 4G).

Class 4 features gave slightly better prediction performance as Class 1 but worse than Class 2 features (Figure 4B). While H3K36me3 is still the most important feature for the classification performance, which has on average higher levels in PAXT targets than NEXT targets, other features correlated to active transcription initiation such as H3K27ac, H2A.Z with higher levels in NEXT than PAXT targets showed similar importance score (Figure 4H).
Assessing the relative feature importance of four classes in the full model showed consistent results that agree with prediction power and feature rank observed in the individual classes (Figure S7).

In summary, the predictive model of NEXT vs. PAXT targets showed a slightly better but similar trend of performance as PAXT vs. non-exosome target classification, and TES-proximal RNA processing is the most predictive among all classes. Class 3 features related to canonical 3’ end processing like PAS motifs, CPA related RBPs were the most discriminative.

Exosome target pathway specific features
Based on the previous three binary classifications, Class 2 and 3 features had overall good prediction performance while Class 3 had the best results. We then investigated whether we could identify exosome pathway specific features in these two classes. To determine NEXT specific features, we considered significant features that occurred in both NEXT vs. PAXT and NEXT vs. non-exosome target classification. Similarly, features that occurred in both NEXT vs. PAXT and PAXT vs. non-exosome target classification were considered PAXT specific; and features that occurred in both NEXT vs. non-exosome and PAXT vs. non-exosome target classification were considered as non-exosome specific features. Additionally, significant features shared in all comparisons were considered discriminative for all three target categories (Figure 5, left panel). Notably, we found no PAXT specific features in either Class 2 or 3. The NEXT specific features are generally depleted in NEXT for both Class 2 and 3 (indicated by the grey arrows in Figure 5, right panel). In Class 2, in addition to RBP-binding related features, the NEXT specific features also included 5’ SS motif, implying different from PAXT targets, lack of splicing in NEXT targets might be due to depletion of splice information in the sequence. In Class 3, NEXT specific features were mainly concerned with RNA cleavage, e.g. cleaved PAS motif, RBP binding of CSTF2, CPSF6, NUDT21. Although PAS strength around TES was among the shared features, the importance score of the feature was much higher in NEXT vs. PAXT and NEXT vs. non-exosome target classification (5.1 and 3.0, respectively) compared to PAXT vs. non-exosome targets (1.1); therefore, it could also be considered as a more NEXT specific feature. In addition, most shared and non-exosome specific features in Class 2 and 3 are RBP bindings. The corresponding RBPs were observed to be generally enriched in non-exosome targets compared to PAXT and NEXT targets. Inspecting functions of these RBPs showed they are involved in various processes related to RNA biogenesis and gene expression regulation (Figure 5, right panel).
Figure 5. Exosome target pathway specific features. Schematic Venn Diagram (left panel) showing the definition of pathway specific features. Right panel shows the corresponding Venn Diagram of Class 2 (upper box) and Class 3 (lower box), the pathway specific and shared features are listed in colored boxes. Arrows in front of NEXT and non-exosome specific features indicate whether the feature is depleted (arrow downward) or enriched (arrow upward) in NEXT or non-exosome targets compared to the other two targets.

We then asked how well features of the four classes were able to distinguish the exosome target pathways in a multi-class classification model. To this end, we used the significant features of the four classes across all three comparisons, then trained a multi-class Random Forest model. The result was consistent with the individual binary models (Figure 6). Class 1 TSS related significant features had the worst prediction performance with average accuracy around 0.5, still comparing favorably to a random three class model with an expected accuracy of 0.33. Significant Class 2 features of TSS-proximal RNA processing and Class 4 features of TES chromatin environment had similarly good accuracy above 0.6, while significant Class 3 features of TES-proximal RNA processing had the best accuracy around 0.7. We observed no obvious synergetic effects.
Figure 6. Multi-class classification. Bar plots (upper panel) showing the average accuracy over 10 repetitions for significant features of the four classes across all three comparisons and combinations of feature classes. An error bar shows the standard deviation of the performance over 10 repetitions. The lower panel shows the features class (dark dots) or combination of features classes (dark dots connected by black solid line) used for classification.

Discussion

NEXT and PAXT define two main nucleoplasmic degradation pathways of the RNA exosome. Previous observations have shown that RNPs targeted by these pathways harbor different characteristics with respect to features of loci producing the RNAs as well as the transcript splicing and 3’ end processing patterns (Almada et al., 2013; Chen et al., 2016b; Meola et al., 2016; Ntini et al., 2013; G. Wu et al., 2020). To further delineate features driving these two pathways, we here categorized examined features into four different classes (depending on the location, i.e., TSS or TES and type of the feature, i.e. RNA processing or TSS/TES related configuration). Based on the different feature classes, we trained Random Forest models to assess their abilities to distinguish targets and to identify relevant features within the classes that drive the prediction performance. In contrast to other machine learning models that often display a ‘black box’ character, decisions of Random Forest models can be rationalized well and the importance of individual features can be directly quantitatively assessed. Feature importances were therefore used to eliminate less relevant features in an iterative process to obtain a final model containing a minimal set of high-priority and discriminative features.

It was previously noticed that many of the Class 1 TSS-related features are different between exosome- and non-exosome targets; e.g., exosome targets tend to be lower expressed (Andersson et al., 2014; Lloret-Lliinares et al., 2018), and the corresponding TSSs differ in G/C content and histone modification patterns (Andersson & Sandelin, 2020; Chen et al., 2016a; M. Wu et al., 2020). Interestingly, however, we found that Class 1 features were only able to clearly distinguish NEXT- from non-exosome targets, of which the width of the G/C enriched regions immediately downstream of TSSs, and the H3K4
methylation status were the most discriminative features. Instead these had only limited predictable value with respect to PAXT vs. non-exosome targets. At the same time, Class 1 features were to some extent able to distinguish NEXT- from PAXT-targets, indicating that Class 1 features of PAXT targets share some similarities with non-exosome targets and some with NEXT targets.

Class 2 features are related to TSS-proximal RNA processing including 5’SS, PAS, and RBP motifs and RBP bindings. The predictive models of the three pairwise comparisons showed Class 2 features could well distinguish exosome pathways and the discriminative features are mostly splicing related. It is well characterized that many exosome target RNAs are transcribed from loci with depleted 5’ SS and enriched PAS motifs proximal to TSS. Interestingly, of all the top ranked discriminative features, we found a lack of 5’ SS is only specific for the NEXT but not PAXT pathway while enrichment of TSS-proximal PAS was not considered to be a discriminative feature for either NEXT or PAXT pathway. It has been reported that 5’ SS could serve as a regulation point to keep polII in active transcribed status through binding of U1 snRNP, this subsequently suppresses 3’end processing and premature termination (Chiu et al., 2018; Kaida et al., 2010; Zhang et al., 2021). However, this model is not entirely in accordance with previous studies showing that NEXT targets often do not undergo canonical 3’end processing mediated by CPA complexes (G. Wu et al., 2020). It still remains unclear what the mechanism link between lacking 5’ SS and NEXT targeting is.

Class 3 features of TES-proximal RNA processing consist of RBP motifs, RBP bindings and features more specific to 3’end processing like well-positioned PAS and PAS strength around TES. Class 3 features yielded the best prediction performance among all four feature classes to distinguish exosome pathways in both binary and multi-class models. Features related to the RNA cleavage process in canonical 3’end processing were found to be most discriminative for distinguishing NEXT from both PAXT and non-exosome targets. This agrees with previous studies that transcription termination of NEXT targets is mostly mediated by Integrator rather than CPA complexes (Lykke-Andersen et al., 2021). While PAXT targets were also best distinguished from non-exosome targets using Class 3 features and the most discriminative features being mostly RBP bindings, with the exact mechanisms that mediate PAXT targeting due to lacking of such RBPs remain unclear.

Additionally, we observed many discriminative features from Class 2 and 3 were RBP bindings detected by CLIP-seq, which were in general much less abundant in NEXT and PAXT than non-exosome targets. Inspecting functions of individual RBPs and GO analysis showed they fit well with previous observations. This is in line with the views proposed by (Bresson & Tollervey, 2018), in which nuclear decay is considered as a default fate for all transcripts that lack specific protective features. However, technical biases introduced by CLIP-seq could not be ruled out: the instability of exosome targets in the
presence of the exosome makes RBP bindings difficult to be captured by CLIP-seq, as most CLIP-seq data used in this study was performed in cells with exosome function unperturbed.

Class 4 features of the TES chromatin environment also gave good overall prediction performance for distinguishing exosome pathways, H3K36me3 stood out as the most predictive feature in all comparisons. H3K36me3 is known to be enriched in active transcribed gene body and an indicator of efficient elongation, this indicates NEXT and PAXT targets might both lack efficient elongation but to a different extent.

In summary, our systematic study was consistent with previous observations. By using predictive models together with feature importance assessment, we added quantitative evidence and validated recent insights in the differences between PAXT and NEXT targeting pathways.

Methods

1. Public data acquisition and processing
HeLa S3 NETseq were described in (Mayer et al., 2015) and obtained from GEO: GSE61332. The ChIPseq datasets used in this study were described in the ENCODE project (ENCODE Project Consortium, 2012), HeLa S3 H3K4me1, H3k4me2, H3K4me3, H3K9ac, H3K9me3, H3K27me3, H3K27ac, H4K20me1, H2A.Z, PolIIb were obtained from Gene Expression Omnibus (GEO): GSE29611 and DNase-seq was obtained from ENCODE: ENCSR959ZXU. The replicates were pooled and signals were averaged over replicates. The hg19 genome coordinates were converted to hg38 using the UCSC liftOver tool. For data with replicates, replicates were pooled and signals were averaged over replicates for subsequent analysis.

CLIP binding sites for RNA binding proteins were obtained from POSTAR2 (http://lulab.life.tsinghua.edu.cn/postar/index.php) (Zhu et al., 2019). Binding site datasets called by different peak calling methods from the same CLIP data were treated as independent datasets when extracting the features.

2. Transcriptome annotation
De novo HeLa transcriptome annotation from (Lykke-Andersen et al., 2021) was used in this study.

3. Classification of exosome targets by pathways
For classification, we used RNAseq counts derived from total RNA of siEGFP, siRBM7, siZCCHC8, siZFC3H1 -treated HeLa cells first described in Meola et al. (GSE84172) and similar data from siZC3H3 -treated cells first described in Silla et al. 2020 (GSE131255). Counts in exonic regions of major transcript isoforms in our in-house HeLa transcriptome annotations were collected using featureCounts tool from subread package (v2.0.0) (Liao et al., 2013), using parameters [-p -C -s 2 -t exon]. These counts were then subjected to differential expression analysis using DESeq2 (v1.22.2)
(Love et al., 2014) using default settings except that batch information (see Silla et al. 2020 for details) was included in the design. Transcripts significantly upregulated (log_2 FC > 0 and padj < 0.1) in siRBM7, siZCCHC8, siZFC3H1 and siZC3H3 were selected and used to defined the set of NEXT targets (sig. upregulated in siRBM7 and siZCCHC8, but not sig. upregulated in siZC3H3 and siZFC3H1), PAXT targets (sig. upregulated in siZC3H3 and siZFC3H1, but not sig. upregulated in siRBM7 and siZCCHC8) and non-exosome targets (log_2 FC > -0.5 and log_2 FC <= 0 in any of the 4 knock-downs).

4. Sequence analysis

Sequences were extracted from the reference genome (hg38) using getfasta from bedtools. The R package Biostrings (version 2.54.0) was used for the following sequence analysis. G/C content is defined as the percentage of DNA that is G or C, and computed using letterFrequencyInSlidingView function over a 10 bp window. Raw position frequency matrix (PFM) of the TATA box, initiator (INR) element, 5' splice site and pA site motifs were obtained from (Portales-Casamar et al., 2010), and converted to position weight matrix (PWM) using R function PWM. RNA binding protein (RBP) motif PWM were obtained from the CISBP-RNA database (Ray et al., 2013). countPWM function was used to scan for motif occurrences, a minimum score of 90% was used for counting a motif hit.

5. Extract signals from sequencing data.

Signals from ChIPseq and strand specific signals from NET-seq over a given window were extracted using R package rtracklayer (version 1.46.0).


Nats entropy was calculated for each feature by taking the values across three exosome target categories using the entropy function from R package entropy (version 1.2.1). Features with entropy smaller than 0.5 were removed in this study.

7. GC spread estimation

The GC spread metric was designed as a proxy to quantify the boundary of the G/C enriched region immediately downstream of TSS. It is calculated as the width of regions with 75% of the total G/C content in a defined region downstream of TSS. G/C content for each nucleotide is firstly computed as in 4 and normalized as max(0, G/C-0.5), the normalized G/C content is then multiplied by a dynamic scaling factor calculated using a Gaussian function as follows: \( s = 2^{-\frac{x^2}{a^2}} (0,1) \), here x is the distance of the nucleotide from TSS and a is set to half of the defined region width so that the scaling factor will be decreased to 0.5 at the middle point of the region. By using the scaling factor, we put more weight on nucleotides close to TSS and less weight on those more distant, thus minimizing the influence of the random G/C content fluctuation more distant from TSS. In addition, to avoid high GC spread from TSSs with generally low G/C content, we reduce the GC spread if the average normalized and unscaled G/C content (GC’avg) in the calculated GC spread is less than 0.1, by a factor of GC’avg/0.1.

8. PAS strengths estimation
PAS strengths were estimated as described in (G. Wu et al., 2020) using the deep neural network model APARENT (python package V0.1) (Bogard et al., 2019) and depicted as log odds of the prediction score.

9. Random forest model and prediction performance evaluation
The random forest model was built using R package caret (version 6.0.86). We first randomly selected a number of samples from the larger data class corresponding to the number of samples in the smaller data class, in order to avoid the biases caused by unbalanced data of different classes. We then used the resulting balanced dataset for classification, where 70% of the dataset was used for training and the remaining for testing. When training the model, we chose 500 trees and used 5-fold cross-validation with 5 repeats in order to tune the parameter “number of randomly selected features at each tree split”; the optimal model with the largest average accuracy value was selected to evaluate the performance on the test data. To measure the prediction performance, we used “area under the receiver operator characteristic” (AUC) and F1 score, i.e., the harmonic mean of precision and recall. To ensure the down-sampling process accurately reflects the larger data class and to minimize the biases of random splitting training and testing data, we repeated the above processes 10 times, and calculated the mean and standard deviation of AUC and F1 score.

10. Feature importance score calculation
The importance score of a feature in the random forest model was measured by calculating the average values of the difference in prediction error with and without permuting the values of the feature on the out-of-bag portion of data over all trees, normalized by the standard deviation of the difference yielding a z-score. We used the varImp function in Caret to compute the feature importance score.

11. Data Visualization
We used R and the ggplot2 R package (Wickham, 2009) unless otherwise noted for visualizations.

References


Ntini, E., Järvelin, A. I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A.,


Supplementary Information

**Figure S1. Characterization of nucleoplasmic exosome targets.**

**A.** Bar plots showing the percentage (%) (Y-axis) of specific RNA biotypes for each exosome target category (X-axis).

**B.** Combined violin-boxplots showing the distribution of the number of exons (Y-axis) for each pathway category (X-axis).

**C.** Combined violin-boxplots showing the distribution of TU length (Y-axis) for each exosome target category (X-axis).

**Figure S2. Feature properties.**
A. Heat map showing G/C content 1 kb downstream of PROMPTs (left) and non exosome-sensitive (right) TSSs, each row corresponds to one TSS. Color intensity indicated average G/C content per base calculated as the fraction of C or G nucleotides in 10 bp sliding windows. Red line indicates the GC spread computed for each TSS. B. Heat map showing the absolute pairwise Spearman correlation for all features of the four classes. The correlation is calculated by taking values of all three exosome target categories. C. Combined bar plots showing the number of filtered features (X-axis) in each feature class (Y-axis).

**Figure S3.** Evaluation of feature correlation in NEXT vs. non-exosome classification model. Heat map showing the absolute pairwise Spearman correlation for top 15 Class 3 features by taking values of both NEXT and non-exosome target categories.

**Figure S4.** Feature importance of the full model of NEXT vs. non-exosome targets. Bar plot showing the feature importance score (X-axis) of top ranked features (Y-axis) from the full model of NEXT vs. non-exosome using a combination of all four feature classes. The rows are ordered by importance score.
Figure S5. Evaluation of feature correlation in PAXT vs. non-exosome classification model. Heat map showing the absolute pairwise Spearman correlation for top 15 Class 3 features by taking values of both PAXT and non-exosome targets.

Figure S6. Feature importance of the full model of PAXT vs. non-exosome targets. Bar plot showing the feature importance score (X-axis) of top ranked features (Y-axis) from the full model of PAXT vs. non-exosome targets using a combination of all four feature classes. The rows are ordered by importance score.
Figure S7. Feature importance of the full model of NEXT vs. PAXT targets. Bar plot showing the feature importance score (X-axis) of top ranked features (Y-axis) from the full model of NEXT vs. PAXT targets using a combination of all four feature classes. The rows are ordered by importance score.
Paper III
A Functional Link between Nuclear RNA Decay and Transcriptional Control Mediated by the Polycomb Repressive Complex 2

**Highlights**

- Depletion of ZFC3H1 in mouse ESCs results in differentiation defects
- PRC2 target genes are deregulated in Zfc3h1<sup>-/-</sup> cells
- Chromatin binding of PRC2 and H3K27me3 is reduced in Zfc3h1<sup>-/-</sup> cells
- Increased binding of RNA impairs PRC2 complex stability

**Authors**

William Garland, Itys Comet, Mengjun Wu, ..., Albin Sandelin, Kristian Helin, Torben Heick Jensen

**Correspondence**

helink@mskcc.org (K.H.), thj@mbg.au.dk (T.H.J.)

**In Brief**

ZFC3H1 targets pA<sup>+</sup> RNA for decay by the nuclear RNA exosome. Garland et al. report a disruptive relationship between excess RNA and PRC2 upon depletion of ZFC3H1 in mouse ESCs. In such conditions, RNA is bound by PRC2 components, which show reduced binding to chromatin and fellow PRC2 proteins.
A Functional Link between Nuclear RNA Decay and Transcriptional Control Mediated by the Polycomb Repressive Complex 2

William Garland,1,2 Itys Comet,2,3,6 Mengjun Wu,2,4,6 Aliaksandra Radzisheuskaya,2,3,5,6 Leonor Rib,2,4,6 Kristoffer Vitting-Seerup,2,4 Marta Lloret-Llinares,1 Albin Sandelin,2,4 Kristian Helin,2,3,5,7* and Torben Heick Jensen1,7,*

1Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark
2Biotech Research and Innovation Centre (BRIC), Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
3The Novo Nordisk Foundation for Stem Cell Biology, Faculty of Health and Medical Science, University of Copenhagen, Copenhagen, Denmark
4The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
5Cell Biology Program and Center for Epigenetics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
6These authors contributed equally
7Lead Contact
*Correspondence: helink@mskcc.org (K.H.), thj@mbg.au.dk (T.H.J.)
https://doi.org/10.1016/j.celrep.2019.10.011

SUMMARY

Pluripotent embryonic stem cells (ESCs) constitute an essential cellular niche sustained by epigenomic and transcriptional regulation. Any role of post-transcriptional processes remains less explored. Here, we identify a link between nuclear RNA levels, regulated by the poly(A) RNA exosome targeting (PAXT) connection, and transcriptional control by the polycomb repressive complex 2 (PRC2). Knockout of the PAXT component ZFC3H1 impairs mouse ESC differentiation. In addition to the upregulation of bona fide PAXT substrates, Zfc3h1−/− cells abnormally express developmental genes usually repressed by PRC2. Such de-repression is paralleled by decreased PRC2 binding to chromatin and low PRC2-directed H3K27 methylation. PRC2 complex stability is compromised in Zfc3h1−/− cells with elevated levels of unspecific RNA bound to PRC2 components. We propose that excess RNA hampers PRC2 function through its sequestration from DNA. Our results highlight the importance of balancing nuclear RNA levels and demonstrate the capacity of bulk RNA to regulate chromatin-associated proteins.

INTRODUCTION

Embryonic stem cells (ESCs) are distinguished by their dual ability to self-renew and differentiate, both of which require tight regulatory control. ESC pluripotency is maintained by a complex molecular network centered around key transcription factors (TFs), including OCT4, SOX2, NANOG, and KLF2 (Morey et al., 2015; Takahashi and Yamanaka, 2006; Zhou et al., 2007). In addition, epigenetic mechanisms establish and maintain specialized chromatin through DNA methylation and histone modifications to allow the activation and repression of genes during development (Bibikova et al., 2008; Chen and Dent, 2014). In consequence, perturbation of DNA methyltransferases (DNMTs) or chromatin repressive complexes, for example, can disrupt mammalian development through the dysregulation of normal gene expression programs (Laugesen and Helin, 2014; Smith and Meissner, 2013).

The exit of ESCs from pluripotency requires concerted silencing of pluripotency factors and activation of lineage-specific genes (Loebel et al., 2003). A key player here is the polycomb repressive complex 2 (PRC2), which catalyzes the formation of facultative heterochromatin via trimethylation of lysine 27 on histone 3 (H3K27me3). PRC2 is dispensable for the maintenance of self-renewal in ESCs but functions to prevent inappropriate transcriptional activation of lineage-specific differentiation factors. These include highly conserved homeobox (HOX) factors, which are crucial for regulating axial patterning in development (Laugesen and Helin, 2014; Pearson et al., 2005), among other developmentally associated TFs that specify cell fate. In ESCs, these genes are repressed through a combination of transcriptional and epigenetic control requiring PRC2 and H3K27me3 (Mallo and Alonso, 2013). Furthermore, PRC2 is crucial for cell fate transitions during development from ESCs, where knockouts (KOs) of core complex components EZH2, SUZ12, and EED result in a block in differentiation (Chamberlain et al., 2008; Pasini et al., 2007; Shen et al., 2008).

Although research on ESC regulation has focused predominantly on transcriptional or epigenetic control, a role of post-transcriptional events, including their putative coupling to transcriptional control, has been less explored. A proper balance of RNA processing and decay ensures homeostasis, whereby different steps in gene regulation buffer one another to maintain a stable expression profile within cell types (for recent reviews see Schmid and Jensen, 2018; Timmers and Tora, 2018). Consequently, malfunction of nuclear RNA decay pathways are therefore linked to developmental disorders and human disease (Corbett, 2018). With improved sequencing technologies, the complexity of the non-protein-coding genome has been revealed (Carninci et al., 2005; Djebali et al., 2012). Functional roles...
Figure 1. Zfc3h1⁻/⁻ Cells Display Defects in EB Differentiation

(A) Schematic representation of the nuclear exosome complex (EXO13) and its PAXT connection. Question marks denote yet-to-be-defined PAXT components.

(B) Western blotting analysis of WT and three independent Zfc3h1⁻/⁻ cell lines (#1–#3). Blots were probed with the indicated PAXT-related antibodies and actin (ACTB) as a loading control.

(C) qRT-PCR analysis of the indicated PAXT targets from total RNA isolated from WT and Zfc3h1⁻/⁻ cell lines. Primers were designed to span exon-exon junctions in order to amplify spliced host gene transcripts. Results are shown relative to Rplp0 mRNA (RPO) and normalized to average WT values. Columns represent average values of technical triplicates per sample, with error bars denoting SD. Individual data values from replicates are indicated as points.

(D) Phase contrast microscopy images of WT and Zfc3h1⁻/⁻ (#1–#3) colonies after 8 days of EB induction. Scale bars denote 200 μm.

(legend continued on next page)
have emerged for long non-coding RNAs (lncRNAs), which form ribonucleoprotein (RNP) complexes capable of regulating various stages of gene expression (Geisler and Coller, 2013; Rinn and Chang, 2012), including the maintenance of ESC pluripotency and differentiation (Guttman et al., 2011; Luo et al., 2016). Levels of lncRNAs in eukaryotic nuclei are tightly regulated by RNA decay systems, which consequently must affect biological processes regulated by such transcripts. Furthermore, it has been shown that many key regulators of gene expression also have RNA binding activity (Hendrickson et al., 2016; Khail et al., 2009). PR2C, for example, binds promiscuously to RNA both in vitro and in vivo, with all core components (EZH2, SUZ12, and EED) contributing to varying degrees (Cifuentes-Rojas et al., 2014; Davidovich et al., 2013; Zhao et al., 2008, 2010).

The function of PR2C-RNA binding is not fully understood, but results have suggested roles of both transcript-mediated recruitment and eviction of PR2C to and from chromatin (Davidovich et al., 2013; Kaneko et al., 2013; Rinn et al., 2007; da Rocha et al., 2014).

The RNA exosome is an essential 3’-5’ ribonucleolytic complex involved in the regulation of the majority of nuclear transcripts (Kilchert et al., 2016; Mitchell et al., 1997; Schmid and Jensen, 2008). Assessing possible links between transcription regulation and the post-transcriptional balancing of RNA levels, the nuclear exosome stands out with its global activity in the processing of precursor RNAs and its ability to efficiently remove transcriptional by-products and otherwise nuclear retained RNA (Schmid and Jensen, 2018). To facilitate recognition and targeting of its plethora of transcript targets, the nuclear exosome associates with adaptor complexes; that is, two nucleoplasmic decay pathways are guided by the nuclear exosome targeting (NEXT) complex and the poly(A) exosome targeting (PAXT) connection, respectively (Lubas et al., 2011, 2015; Meola et al., 2016; Silla et al., 2018). NEXT and PAXT share a common subunit in the RNA helicase MTR4, which connects these adaptors to the exosome. PAXT is also composed of a large zinc finger protein, ZFC3H1, which bridges MTR4 to the nuclear poly(A) binding protein (PABPN1), aiding the targeting of polyadenylated (pA) nuclear RNAs for exosome-mediated decay (Beaujieu et al., 2012; Bresson and Corrad, 2013; Meola et al., 2016; Ogami et al., 2017).

At steady state, IncRNA levels in mammalian cells are generally low, with estimates suggesting that less than 1,000 IncRNAs are present in more than one copy per cell (Djebali et al., 2012; Seiler et al., 2017). Thus, most IncRNAs are stoichiometrically inferior to their putative protein effectors, often weakening the associated mechanistic models claiming function of individual IncRNAs. The abundance of IncRNAs is regulated through nuclear RNA decay pathways and is considerably enriched upon removal of exosome components (Lubas et al., 2011; Meola et al., 2016; Silla et al., 2018). Manipulating PAXT activity, through the depletion of ZFC3H1, therefore allows an approach to study the general effects of excess pA RNA in the nucleus.

Here, we establish a functional link between PAXT activity and transcriptional control mediated by PR2C. Zfc3h1−/− cells are unable to differentiate and exhibit phenotypes reminiscent of cells deficient for PR2C activity. Consistently, normal PR2C function is impaired in Zfc3h1−/− cells and we provide evidence that this is due to PR2C binding to stabilized RNAs. Our results highlight the importance of controlling nuclear RNA levels during key regulatory stages of ESC development and imply that modulation of bulk RNA levels is a potent way of IncRNA-mediated transcription regulation.

RESULTS

Zfc3h1−/− Cells Exhibit Defective Embryoid Body Differentiation

To assess the role of exosome-mediated decay of nuclear pA RNAs in ESC pluripotency and differentiation, we used CRISPR/Cas9 to generate homozygous KOs of the PAXT component ZFC3H1 (Figure 1A). Zfc3h1 was specifically targeted because MTR4 and PABPN1 also reside in alternative nuclear complexes. Three biologically independent Zfc3h1−/− ESC lines with disrupted Zfc3h1 ORFs were derived from single-cell KO clones (Figure S1A). In agreement with our previous observations in human cells, the expression of other known PAXT-related (Figure 1B) and exosome-related (Figure S1B) proteins was unaffected by ZFC3H1 depletion (Meola et al., 2016). Still, PAXT-mediated RNA decay was disrupted, which resulted in an approximately 2-fold accumulation of total nuclear pA RNA (Figure S1C), including spliced small nuclear RNA (snRNA) host gene (Shhg) IncRNAs (Meola et al., 2016; Figure 1C).

Zfc3h1−/− cells were viable under 2i-LIF growth conditions, which selects against cellular differentiation, and appeared morphologically similar to wild-type (WT) cells (Figure S1D). Furthermore, expression of the key pluripotency TFs NANOG, ESRRB, KLF2, SOX2, and OCT4 was not perturbed (Figure S1F). Zfc3h1−/− KO cells still retained an ESC-like morphology, with gene expression related to ESC pluripotency and differentiation (Figure S1F). Zfc3h1−/− KO cells retained an ESC-like morphology, with smaller rounded colonies and a lack of differentiated cell types (Figure S1D). The cells also showed a lack of cystic EBs, cavities that normally form during differentiation because of programmed cell death, but instead remained as solid aggregates, apparent as dark spots on microscopy images (Figure 1D). Finally, between 7 and 8 days of differentiation, WT cells normally develop...
Figure 2. PRC2 Target Genes Are Upregulated in Zfc3h1−/−/− ESCs
(A) Differential gene expression analysis of RNA-seq data from Zfc3h1−/−/− versus WT cells after 7 days of EB induction shown as an MA plot. The y axis shows Zfc3h1−/−/− versus WT RNA-seq log2 FC. The x axis shows average normalized expression as log2 counts per million (CPM) values across all biological replicates. Each dot indicates a gene, and blue/yellow color denotes significant differential expression (edgeR FDR < 0.05). Red lines denote log2 FC > 1, < 1. Upregulated genes involved in pluripotency and downregulated genes involved in differentiation are highlighted.
(B) MA plot as in (A) but for undifferentiated (D0) Zfc3h1−/−/− versus WT cells. Values are averages of two WT and three Zfc3h1−/−/− biological replicates. Upregulated genes highlighted indicate PRC2 target genes involved in developmental processes.
(C) Bar plot of upregulated (log2 FC > 0.5) gene types in D0 Zfc3h1−/−/− cells. GENCODE RNA biotypes are color coded as indicated, and numbers of affected genes are shown in parenthesis.

(legend continued on next page)
spontaneous beating colonies, indicative of cardiomyocyte formation (Doetschman et al., 1985), which was not observed up to 10 days of EB induction of Zfc3h1−/− cells.

We then collected samples across the EB differentiation time course to analyze the expression of cell type-specific markers (e.g., pluripotency TFs normally decrease rapidly when cells differentiate). In contrast, Zfc3h1−/− cells retained high expression of the pluripotency TFs OCT4, SOX2, ESRRB, and KLF2 at later stages of the time course as assessed by RNA and protein analyses (Figures 1E and 1F). qRT-PCR primers spanning exon-intron (ExIn) borders of Oct4 and Nr0b1 pre-mRNAs showed that intronic sequences were elevated in Zfc3h1−/− cells, with little or no expression of endodermal, mesodermal, and ectodermal TFs (Figure 1G). Taken together, we conclude that Zfc3h1−/− cells show EB differentiation phenotypes and retain an expression profile reminiscent of undifferentiated cells.

**Transcripts from PRC2-Targeted Genes Are Enriched in Zfc3h1−/− Cells**

To obtain a global impression of the disparities between WT and Zfc3h1−/− cells, sequencing of rRNA-depleted total RNA (RNA-seq) was performed on samples harvested after 7 days (D7) of EB induction. This time point was chosen because it showed large gene expression differences between WT and Zfc3h1−/− cells as measured by qRT-PCR (Figures 1E and 1G). In addition, total RNA from ESC (D0) samples was sequenced to assess the undifferentiated starting point of the experiment. Biological replicates of RNA-seq libraries were overall highly correlated (Figures S2A and S2B). Differential expression (DE) analysis of D7 samples mirrored our qRT-PCR analyses: Zfc3h1−/− cells displayed elevated expression of pluripotency-associated TFs and decreased expression of germ layer-specific transcripts (Figure 2A). This, in conjunction with the morphological phenotypes, suggested that Zfc3h1−/− cells are unable to exit from the embryonic stem (ES) state and activate the normal genes required to initiate differentiation. We surmised that this inability to exit from pluripotency might be due to gene expression aberrations in the D0 ES state disrupting normal progression into differentiation pathways. Consistent with this notion, DE analysis revealed higher levels of mRNAs encoding lineage-specific factors normally associated with differentiated cells in Zfc3h1−/− versus WT D0 samples (Figure 2B). A large number of these up-regulated lineage markers comprised TFs, including HOX genes, which are involved in early developmental processes (Pearson et al., 2005). At first glance, such an expression profile would seemingly contrast our observation that Zfc3h1−/− ESCs appear morphologically similar to WT ESCs. However, although exit from pluripotency requires the expression of developmental TFs, this must occur concomitantly with suppression of pluripotency TFs, which predominantly define the cellular state of ESCs (Young, 2011). Moreover, the selective pressure of the 2i/LIF culture condition maintains the pluripotent state by blocking MEK and GSK activity and activating the STAT3 pathway (Wray et al., 2010; Ying et al., 2008).

Nevertheless, abnormal expression of developmental genes in the Zfc3h1−/− ESCs indicated a general deregulation of differentiation-associated genes. In WT ESCs, these genes are normally repressed in an inactive chromatin environment highly decorated with H3K27me3 (Malo and Alonso, 2013). Furthermore, the repression of pluripotency genes during differentiation is regulated by H3K27me3 (Obier et al., 2015; Pasini et al., 2007). As this histone mark is solely catalyzed by the PRC2 complex, we compared the gene expression profile of Zfc3h1−/− ESCs with published PRC2 KO (Ezh1−/−/Ezh2−/−) RNA-seq data derived from the same parental mouse ESC line (Højfeldt et al., 2018).

Of the 1,804 upregulated transcripts in the Zfc3h1−/− D0 samples (log2 fold change [FC] > 0.5, false discovery rate [FDR] < 0.05, edgeR), approximately 25% could be designated as non-coding RNAs (ncRNAs) (Figure 2C), including known PAXT targets upregulated because of their diminished decay, while ~75% of cases were protein coding. Almost one-third of the latter transcripts showed a significant (p < 1.1e-18, hypergeometric test) overlap with transcripts upregulated in PRC2 KO cells (Figure 2D). For these shared transcripts, exonic and intronic reads were evenly upregulated in the Zfc3h1−/− D0 data (Figure 2E), suggesting that increased mRNA levels were based on increased transcription. This was validated by qRT-PCR analysis of upregulated HOX transcripts using ExIn-specific primers on chromatin-associated RNA to enrich for pre-mRNA (Figure 2F). We conclude that Zfc3h1−/− ESCs have higher levels of transcripts derived from a subset of PRC2 target loci, and these genes appear to be more transcriptionally active in the absence of PAXT.

**SUZ12 Chromatin Occupancy and H3K27me3 Levels Are Decreased in Zfc3h1−/− Cells**

As Zfc3h1−/− cells displayed increased expression of transcripts also upregulated in the absence of PRC2, we investigated the status of the PRC2 complex in the Zfc3h1−/− background. Expression of its core components, EZH2, SUZ12, and EED, was unaffected by the absence of ZFC3H1 (Figure S3A). Therefore, we decided to conduct chromatin immunoprecipitation sequencing (ChIP-seq) on the three Zfc3h1−/− cell lines, along...
Figure 3. SUZ12 DNA Occupancy and H3K27me3 Levels Are Decreased in Zfc3h1−/− Cells

(A) MA plots showing global changes in H3K27me3 and SUZ12 ChIP-seq densities in Zfc3h1−/− versus WT cell lines. Dots indicate sliding genome windows, where color intensity indicates the density of overlapping points. Y axes show log2 FC Zfc3h1−/− versus WT normalized ChIP signal, and x axes show average log2 ChIP signal, all from three independent biological replicates.

(B) Boxplot distributions of log2 FC Zfc3h1−/− versus WT normalized H3K27me3 and SUZ12 ChIP signals on regions centered on 3678 SUZ12 peaks identified in WT cells.

(C) Average SUZ12 ChIP-seq signals for Zfc3h1−/− and WT cell lines in regions centered on the SUZ12 peaks as in (B). Zfc3h1−/− ChIP data are the average of three biological replicates.

(D) As in (C) but for H3K27me3 signal from regions centered on the SUZ12 peaks from (B).

(E) Heatmap representation of the relation between SUZ12 ChIP-seq, H3K27me3 ChIP-seq, and RNA-seq log2 FC in Zfc3h1−/− versus WT cells. Columns represent SUZ12 peaks identified in WT cells (N = 1,485), sorted after SUZ12 Zfc3h1−/− versus WT log2 FC. Rows correspond to SUZ12 ChIP-seq, H3K27me3 ChIP-seq, and RNA-seq Zfc3h1−/− versus WT log2 FC values.
with the parental WT cell line, using antibodies specific for H3K27me3 and SUZ12. Zfc3h1/C0 cells displayed reduced H3K27me3 levels and SUZ12 DNA occupancy in comparison with WT cells (Figures 3A, 3B, and 3C). These reductions were generally found in regions with high average signal, and the effect on H3K27me3 levels was substantially stronger than the effect on SUZ12 binding (Figure 3B). Still, while H3K27me3 levels were clearly decreased in Zfc3h1/C0 cells, they were not abolished (average reduction to ~75% versus WT, Figure S3D; as opposed to previously reported KOs of different PRC2 components in which H3K27 methylation is absent, Højfeldt et al., 2018; Pasini et al., 2007; Shen et al., 2008; see Discussion).

PRC2-bound regions, defined by SUZ12 ChIP-seq peaks in WT cells, had significantly (p < 2.2e-16, one-sided Mann-Whitney test) lower average SUZ12 and H3K27me3 occupancy in Zfc3h1/C0 conditions (Figures 3B–3D). Moreover, regions depleted for PRC2 and H3K27me3 in Zfc3h1/C0 cells showed a concomitant increase in RNA expression (Figure 3E), which was equally elevated whether exonic or intronic reads were interrogated (Figure 3F).

ChIP-seq analyses of active and primed chromatin modifications (H3K4me1, H3K4me3, H3K27ac) were also carried out in WT and Zfc3h1/C0 cells. Genes depleted for SUZ12 and H3K27me3 ChIP signals had a marked increase in H3K4me1 and H3K4me3 levels (Figures 3G and 3H). We also obtained low-sequence depth RNA polymerase II (RNAPII) ChIP-seq data, which showed increased occupancy at select PRC2 target genes in Zfc3h1/C0 cells (Figure S3H). Altogether, this fits the notion that PRC2 loci are more transcriptionally active in the absence of normal polycomb-mediated repression via H3K27me3. We conclude that recruitment of SUZ12 to PRC2 target genes is reduced in Zfc3h1/C0 cells, resulting in loss of H3K27me3 at these regions and abnormal RNA expression due to increased transcription.

**Decreased PRC2 Complex Integrity in Zfc3h1/C0 Cells**

The discovered correlation between PAXT-mediated RNA decay and PRC2-mediated transcriptional repression is unprecedented. We therefore sought to address how impaired PRC2 function in Zfc3h1/C0 cells relates to the primary phenotype of stabilized nuclear pA- RNAs. As mentioned above, steady-state levels of PRC2 components remain unchanged in Zfc3h1/C0 cells (Figures S3A, 4A, and 4B, “Inputs”). However, co-immunoprecipitation (co-IP) analyses of SUZ12 (Figures 4A and S4A) or EZH2 (Figures 4B and S4B) from whole-cell lysates revealed their reduced binding to the remaining PRC2 core. Similar effects were observed when conducting IPs from nuclear extracts (Figures S4C and S4D). Such weakened PRC2 complex formation in Zfc3h1/C0 cells was further supported by analyzing the sedimentation of PRC2 components through glycerol gradients: Although the overall protein distribution and sedimentation of loading controls remained unchanged (Figures S4E and S4F), SUZ12 and EZH2 showed a marked shift from high-molecular weight fractions toward lower molecular weight fractions of the gradient derived from Zfc3h1/C0 lysates (Figures 4C and 4D). The distribution of EED was more dispersed throughout the gradient from WT lysates but still showed a shift toward lower fractions in Zfc3h1/C0 samples (Figure S1F). Despite these shifts, a fraction of PRC2 sedimented normally in Zfc3h1/C0 samples and consistently remained bound in co-IPs (Figures 4A–4D and S4F). Together, these analyses suggested that the PRC2 complex is partially compromised in Zfc3h1/C0 cells, which could be further elaborated by subcellular fractionation of WT and Zfc3h1/C0 cells, giving rise to a slight, but significant, shift in the distribution of SUZ12, EZH2 and EED from chromatin-bound to nucleoplasmic fractions (Figures S4G and S4H). Previous reports demonstrated that the complete loss of SUZ12 or EED results in reduced levels of the remaining PRC2 core components (Højfeldt et al., 2018; Montgomery et al., 2005; Pasini et al., 2007), suggesting that PRC2 complex stability depends on the interactions of all core proteins. In contrast, here we found lower levels of the PRC2 complex in Zfc3h1/C0 cells despite normal expression of SUZ12, EZH2, and EED.

Given the RNA accumulation observed in the absence of PAXT (Meola et al., 2016; Figures 1C, 1C, and S1C) and the RNA-binding properties of PRC2, we hypothesized that accumulation of PAXT targets could contribute to PRC2 disruption, linking the primary Zfc3h1/C0 phenotype with a secondary PRC2 phenotype. Hence, we carried out native IPs of SUZ12 and EZH2, using lysates prepared from WT and Zfc3h1/C0 ESCs, and measured the amount of co-isolated RNA. From WT cells, EZH2 IPs showed a greater enrichment of RNA over SUZ12 IPs and the IgG control (Figure 4E), which is in agreement with EZH2 being the PRC2 component with the strongest capacity for RNA binding (Cifuentes-Rojas et al., 2014). However, elevated RNA levels were isolated from both SUZ12 and EZH2 IPs of Zfc3h1/C0 versus WT cell lysates. Thus, concomitant with complex disruption, more RNA was bound to PRC2 components upon PAXT depletion. RNA isolated from SUZ12 and EZH2 IPs was sequenced (RIP-seq) to interrogate any differences in PRC2-bound transcripts from WT versus Zfc3h1/C0 ESCs. Consistent with previous native RIP-seq analysis of PRC2 components (Davidovich et al., 2013; Khalil et al., 2009; Zhao et al., 2010), we found that EZH2 and SUZ12 showed promiscuous binding to both protein-coding RNA and ncRNA. In Zfc3h1/C0 samples, both SUZ12 and EZH2 IPs revealed PAXT targets, including pA- nuclear lncRNAs (Figure S5A), along with transcripts from de-repressed PRC2 target genes (Figure S5B). However, for transcripts upregulated in Zfc3h1/C0 ESCs, there was no significant enrichment of coding RNAs or ncRNAs over the input samples (Figure S5C). Furthermore, there was no specific

---

(F) Correlation of Zfc3h1/C0 versus WT log2, FC RNA-seq signal in intronic versus exonic regions. Points displayed are intron-containing transcripts with log2 FC > 0 in Zfc3h1/C0 versus WT overlapping with SUZ12 peaks that show log2 FC > 0 in Zfc3h1/C0 ChIP-seq data (N = 365). Axes show Zfc3h1/C0 versus WT log2 FC RNA-seq in exonic (y) and intronic (x) regions of these transcripts. (G) Genome Browser views of four PRC2 target genes (Hoxd8, Sox21, Gsc, and Nefm). Displayed tracks include H3K4me1, H3K4me3, H3K27ac, H3K27me3, and SUZ12 ChIP-seq data as well as RNA-seq data from WT and Zfc3h1/C0 (#1) cell lines. RNA-seq tracks on both strands are shown (+ and −, respectively). Gene models are based on RefSeq. Genomic coordinates (mm10) are indicated for each panel.
enrichment of RNAs upregulated in Zfc3h1/C0/C0 ESCs only, in both Zfc3h1/C0/C0 and Ezh1/C0/C0 Ezh2/C0/C0 ESCs (Højfeldt et al., 2018) or for transcripts upregulated at SUZ12-depleted loci (Figure S5D). This suggests that increased nuclear RNA levels in Zfc3h1/C0/C0 ESCs leads to more RNA binding to PRC2 components with no general specificity.

To address whether the disruption of PRC2 in Zfc3h1/C0/C0 cells was indeed dependent on increased RNA levels, we repeated the SUZ12 coIP analyses using cell extracts that were either mock or RNaseA treated. In the presence of RNase, SUZ12 IPs from Zfc3h1/C0/C0 extracts recapitulated the coIP efficiency of equivalent WT extracts (Figures 4F and S5E). More bait protein was generally pulled down in the RNase-treated samples, which suggested that RNA already interfered with the IP in WT extracts, but to a greater extent in Zfc3h1/C0/C0 extracts. Taken together, we propose that increased RNA binding weakens PRC2 complex formation, thereby decreasing its normal function, stability, and recruitment to chromatin (Figure 4G; see Discussion).

**DISCUSSION**

The molecular decisions that govern correct progression through cellular differentiation require a complex agreement of checks and balances. Here, the transcriptional and epigenetic profiles at the ESC stage are important starting points, whose dysregulation may be deleterious for development. In the present study, we discovered a link between excess nuclear RNA and PRC2-mediated transcriptional control in ESCs, suggesting an essential role of nuclear RNA turnover in cellular commitment to differentiation.

Deletion of the nuclear exosome adaptor ZFC3H1 to some extent phenocopies ESCs depleted for PRC2 components; cells
self-renew and appear morphologically similar to WT cells but show a loss of H3K27me3, deregulation of PRC2 target genes, and difficulty initiating differentiation (Boyer et al., 2006; Chamberlain et al., 2008; Pasini et al., 2007; Shen et al., 2008). In contrast to PRC2-depleted cells, Zfc3h1−/− ESCs retain a degree of SUZ12 binding and H3K27me3 modification, which presumably explains the less severe deregulation of PRC2 target genes. Still, the defect appears to be sufficient to restrict the progression into differentiation. In agreement with previous studies, we find that deregulation of PRC2 target genes results in abnormal transcription of developmental genes (Boyer et al., 2006; Lee et al., 2006). It would seem counterintuitive for ESCs expressing developmental genes to retain self-renewal ability and to lose the ability to differentiate upon induction of EB formation. However, it has been suggested that the expression of pluripotency TFs is sufficient for self-renewal and overrides any abnormal expression of lineage markers (Chamberlain et al., 2008). As cells exit pluripotency, PRC2 functions to silence key maintenance factors such as Oct4, Sox2, and Nanog by depositing H3K27me3 at their loci (Ozier et al., 2015). With PRC2 function decreased in Zfc3h1−/− cells, this presumably explains their retained transcriptional activity of pluripotency TFs and block in differentiation (Figures 1E, 1F, and S1G). Together this reiterates that PRC2 and, in turn, ZFC3H1 are dispensable for self-renewal but are essential for the progression into differentiation (Chamberlain et al., 2008; Pasini et al., 2007; Shen et al., 2008).

An RNA-binding ability of the PRC2 complex has been widely documented, with suggested models for transcript-mediated recruitment or eviction of PRC2 to or from DNA (Davidovich et al., 2013; Kaneko et al., 2013; Rinn et al., 2007; da Rocha et al., 2014). A primary phenotype of ZFC3H1 depletion is the stabilization of nuclear pA+ RNAs (Meola et al., 2016; Ogami et al., 2017; this study), which are, by nature, unstable and typically present only in trace amounts under normal conditions. Through removing the targeting mechanisms for decay, this increases the concentration of pA+ RNAs (Silla et al., 2018). Taken together with increased binding of RNA to EZH2 and SUZ12 in Zfc3h1−/− cells, we therefore propose that increased transcript levels negatively affect PRC2 function through its increased RNA binding (Figure 4G). Previous studies initially suggested that excess RNA can inhibit the methyltransferase activity of EZH2 in vitro (Cifuentes-Rojas et al., 2014; Kaneko et al., 2014), which was further elaborated to suggest that decreased catalytic activity was due to RNA titrating PRC2 off nucleosomes (Wang et al., 2017). This was supported by observations that DNA- and RNA-binding capabilities of PRC2 are mutually exclusive in vitro (Beltran et al., 2016; Wang et al., 2017). More recently, an RNA-binding region was identified at an allosteric regulatory region of PRC2 in close proximity to the methyltransferase region of EZH2, which is subsequently inhibited by RNA binding (Zhang et al., 2019). It is therefore plausible that increased nuclear RNA levels dually affect PRC2 function by decreasing its catalytic activity as well as its DNA-binding capacity. We also find that the interaction between PRC2 subunits is compromised in Zfc3h1−/− cells with reduced binding between core subunits in coIP and glycerol gradient assays (Figures 4A–4D). Such decreased interaction between SUZ12 and EZH2 may contribute to an explanation of the stronger loss of H3K27me3 compared with SUZ12 ChIP-seq signal in Zfc3h1−/− cells (Figure 3B); that is, residual SUZ12, uncoupled from EZH2, may still bind DNA. In line with this, our ChIP and subcellular fractionation data showed that a fraction of PRC2 components are still associated with chromatin. This partial phenotype allows us only to speculate at this point and will require further investigation to understand the status of PRC2 complex proteins that are still able to bind DNA. However, in support of this possibility, the N-terminal region of SUZ12 recapitulates SUZ12-binding patterns but lacks EZH2 interaction and thereby does not rescue H3K27me3 activity (Højfeldt et al., 2018). The use of such mutants, in combination with Zfc3h1−/− cells and extended ChIP datasets, might allow unpacking of the molecular basis behind these observations.

Finally, we show that the disruption between core PRC2 subunits in Zfc3h1−/− cell extracts can be rescued upon RNAse treatment (Figure 4F). These results echo previous data showing that RNAse treatment increases chromatin association of PRC2 in cells and, reciprocally, recombiant PRC2 can be titrated off nucleosomes by increasing nuclear RNA levels (Beltran et al., 2019). We propose that this antagonism is a general effect of increased nuclear RNA, as we do not see specific enrichment of particular transcripts in RIP-seq experiments: RNAs that are upregulated in the Zfc3h1−/− inputs are also upregulated in the RIP data. This appears in line with previous findings, that the RNA-binding capability of PRC2 is non-specific and promiscuous in nature.

Although PRC2 has garnered considerable attention, RNA-binding capacities have also been reported for other chromatin regulators, including DNMTs, histone deacetylases (HDAC1), chromatin remodeling proteins (ATRX), DNA demethylases (TET1/2), and other histone methyltransferases (G9a) (Castellanos-Rubio et al., 2016; He et al., 2016; Hendrickson et al., 2016; Holz-Schietinger and Reich, 2012; Li et al., 2018; Di Ruscio et al., 2013; Sarma et al., 2014). Interestingly, RNA has been suggested to have a regulatory function in a number of models, either locally at specific loci or by more global mechanisms; that is, similar to PRC2, RNA binding has been proposed to sequester DNMT1 from transcriptionally active regions as a regulatory mechanism to prevent DNA methylation at these loci (Di Ruscio et al., 2013). Indeed, RIP experiments demonstrate that SUZ12 and DNMT1 both tend to associate with the 5’ ends of RNA (Hendrickson et al., 2016), and both proteins have a greater affinity for RNA over DNA (Di Ruscio et al., 2013; Wang et al., 2017). RNA sequestering is also suggested in some disease models in which IncRNAs are overexpressed and affect chromatin modifiers through their abnormal titration (Gupta et al., 2010; Li et al., 2018; Merry et al., 2015; Prensner et al., 2013). Taken together with our results, this highlights the importance of maintaining a stable nuclear transcriptome through active RNA decay to prevent off-target effects as a result of RNA accumulation. Moreover, our results demonstrate the capacity of changed bulk RNA levels to affect cellular transcription programs. Although dysregulation of PRC2 on a global level is highlighted here, it is equally feasible that an unbalanced transcriptome might affect the function of other bivalent chromatin/RNA-binding proteins either locally or globally.
STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **LEAD CONTACT AND MATERIALS AVAILABILITY**
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - mES cell culture and differentiation
- **METHOD DETAILS**
  - CRISPR/Cas9 KOs
  - RNA isolation
  - pA RNA purification
  - qRT-PCR analysis
  - RNA-seq library preparation
  - Western blotting analysis
  - IP experiments
  - Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) control
  - Subcellular fractionation
  - RIP experiments
  - RIP-seq
  - ChIP experiments
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - Processing and analysis of RNA-seq data
  - Processing and analysis of RIP-seq data
  - Processing and analysis of ChIP-seq data
- **DATA AND CODE AVAILABILITY**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2019.10.011.

ACKNOWLEDGMENTS

We thank Nadia Laura Schmidt, Dorthe Risshøj, and Anna Fossum for excellent technical assistance; Manfred Schmid and Peter Refsing Andersen for critical comments; and members of the T.H.J., K.H., and A.S. groups for advice and discussion. Work in the T.H.J. laboratory was supported by the European Research Council (ERC) (grant 339953), the Lundbeck Foundation, and the Novo Nordisk Foundation (NNF). A.R. was funded by the European Union’s Horizon 2020 research and innovation program under Marie Skłodowska-Curie Actions grant agreement 749362, through a center grant from the NNF to the NNF Center for Stem Cell Biology (NNF17CC0027852) and through the Memorial Sloan Kettering Cancer Center Support Grant (NIH grant P30 CA008748). Work in the A.S. laboratory was funded by the Lundbeck Foundation 55-2015-8825 and through a center grant from the NNF (NNF16OC0023234), through a center grant from the European Research Council (ERC) (grant 339953), the Lundbeck Foundation, and the Novo Nordisk Foundation (NNF). Work in the K.H. laboratory was supported by the European Union’s Horizon 2020 research and innovation program under Marie Skłodowska-Curie Actions grant agreement 749362, through a center grant from the NNF to the NNF Center for Stem Cell Biology (NNF17CC0027852) and through the Memorial Sloan Kettering Cancer Center Support Grant (NIH grant P30 CA008748). Work in the A.S. laboratory was funded by the Lundbeck Foundation.

AUTHOR CONTRIBUTIONS

W.G., A.S., K.H., and T.H.J. conceived the project. W.G. designed and performed the majority of experiments. I.C. performed the ChIP experiments. M.W., L.R., and K.V.S. carried out the bioinformatics analysis. A.R. and M.L.-L. contributed to the experimental design and cell line generation. T.H.J., K.H., and A.S. supervised the project. W.G. and T.H.J. wrote the manuscript with input from all co-authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES


is implicated in the initial Xist-induced targeting of PRC2 to the inactive X chromosome. Mol. Cell 53, 301–316.


# STAR★METHODS

## KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antibodies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSRC2 (ZFC3H1) antibody</td>
<td>Novus Biologicals</td>
<td>Cat# NB100-68267; RRID:AB_2218116</td>
</tr>
<tr>
<td>ZFC3H1 antibody</td>
<td>Sigma-Aldrich</td>
<td>Cat# HPA-007151; RRID:AB_1846133</td>
</tr>
<tr>
<td>MTR4 (SKIV2L2) antibody</td>
<td>Abcam</td>
<td>Cat# ab70551; RRID:AB_1270701</td>
</tr>
<tr>
<td>PABPN1 antibody</td>
<td>Abcam</td>
<td>Cat# ab75855; RRID:AB_1310538</td>
</tr>
<tr>
<td>EXOSC4 (RRP41) antibody</td>
<td>In house (Pruijn lab)</td>
<td>Cat# Brouwer et al., 2001</td>
</tr>
<tr>
<td>DIS3 antibody</td>
<td>Sigma-Aldrich</td>
<td>Cat# HPA039281; RRID:AB_10795583</td>
</tr>
<tr>
<td>ZCCHC8 antibody</td>
<td>Abcam</td>
<td>Cat# ab88739; RRID:AB_1271512</td>
</tr>
<tr>
<td>ACTB (β-ACTIN) antibody</td>
<td>Sigma-Aldrich</td>
<td>Cat# A2226; RRID:AB_476697</td>
</tr>
<tr>
<td>NANO2 antibody</td>
<td>Cell Signaling</td>
<td>Cat# 8822; RRID:AB_11217637</td>
</tr>
<tr>
<td>ESRRB antibody</td>
<td>R+D Systems</td>
<td>Cat# PP-H6705-00; RRID:AB_2100412</td>
</tr>
<tr>
<td>KL2F antibody</td>
<td>Millipore</td>
<td>Cat# 09-820; RRID:AB_10607287</td>
</tr>
<tr>
<td>SOX2 antibody</td>
<td>Santa Cruz</td>
<td>Cat# sc-17320; RRID:AB_2286684</td>
</tr>
<tr>
<td>OCT4 antibody</td>
<td>Abcam</td>
<td>Cat# ab19857; RRID:AB_445175</td>
</tr>
<tr>
<td>VCL (VINCULIN) antibody</td>
<td>Sigma-Aldrich</td>
<td>Cat# V9131; RRID:AB_477629</td>
</tr>
<tr>
<td>SUZ12 antibody</td>
<td>Cell Signaling</td>
<td>Cat# 3737; RRID:AB_2196850</td>
</tr>
<tr>
<td>EZH2 antibody</td>
<td>Cell Signaling</td>
<td>Cat# 5246; RRID:AB_10694683</td>
</tr>
<tr>
<td>EZH2 antibody</td>
<td>In house (Helin lab)</td>
<td>Cat# AC22</td>
</tr>
<tr>
<td>EZH2 antibody</td>
<td>Abcam</td>
<td>Cat# ab195409</td>
</tr>
<tr>
<td>EED antibody</td>
<td>In house (Helin lab)</td>
<td>Cat# Bracken et al., 2003</td>
</tr>
<tr>
<td>SFPQ antibody</td>
<td>Abcam</td>
<td>Cat# ab50935; RRID:AB_882523</td>
</tr>
<tr>
<td>TUBA1B (α-TUBULIN) antibody</td>
<td>Rockland</td>
<td>Cat# 600-401-880; RRID:AB_2137000</td>
</tr>
<tr>
<td>RPB1 NTD antibody</td>
<td>Cell Signaling</td>
<td>Cat# 14958; RRID:AB_2687876</td>
</tr>
<tr>
<td>H3K27me3 antibody</td>
<td>Cell Signaling</td>
<td>Cat# 9733; RRID:AB_2616029</td>
</tr>
<tr>
<td>H3K27ac antibody</td>
<td>Cell Signaling</td>
<td>Cat# 8173; RRID:AB_10949503</td>
</tr>
<tr>
<td>H3K4me1 antibody</td>
<td>Cell Signaling</td>
<td>Cat# 5326; RRID:AB_10695148</td>
</tr>
<tr>
<td>H3K4me3 antibody</td>
<td>Cell Signaling</td>
<td>Cat# 9751; RRID:AB_2616028</td>
</tr>
<tr>
<td>H3 antibody</td>
<td>Abcam</td>
<td>Cat# ab1791; RRID:AB_302613</td>
</tr>
<tr>
<td>H4 antibody</td>
<td>Millipore</td>
<td>Cat# 05858; RRID:AB_390138</td>
</tr>
<tr>
<td>IgG</td>
<td>Millipore</td>
<td>Cat# PP64; RRID:AB_97852</td>
</tr>
</tbody>
</table>

Bacterial and Virus Strains

| DH5x Chemically Competent Cells | Prepared in lab | N/A |

Chemicals, Peptides, and Recombinant Proteins

| GSK3 inhibitor (CHIR99021) | Sigma-Aldrich | Cat# SML1046 |
| MEK1/2 inhibitor (PD0325901) | Sigma-Aldrich | Cat# PZ0162 |
| N-2 Supplement | Thermo Fisher Scientific | Cat# 17502048 |
| B-27 Supplement | Thermo-Fisher Scientific | Cat# 17504044 |
| TRIZOL Reagent | Thermo-Fisher Scientific | Cat# 15598018 |
| Lipofectamine 2000 Transfection Reagent | Thermo-Fisher Scientific | Cat# 11680019 |
| Protein A Dynabeads | Thermo-Fisher Scientific | Cat# 10008D |
| Agencourt AMPure XP Beads | Thermo-Fisher Scientific | Cat# 10136224 |
| Benzonase nuclease | Millipore | Cat# 70746 |
| RNaseA | Thermo-Fisher Scientific | EN0531 |

(Continued on next page)
Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Torben Heick Jensen (thj@mbg.au.dk).

All unique/stable reagents generated in this study are available from the lead contact upon request.

**EXPERIMENTAL MODEL AND SUBJECT DETAILS**

**mES cell culture and differentiation**

E14TG2a mouse ESCs (male genotype, XY) were cultured on 0.2% gelatin coated plates in 2i/LIF containing medium (1:1 mix of Neurobasal (GIBCO) and DMEM/F-12 (GIBCO) supplemented with 1x Pen-Strep (GIBCO), 2 mM Glutamax (GIBCO), 50 μM...
β-mercaptoethanol (GIBCO), 0.1 mM nonessential amino acids (GIBCO), 1 mM sodium pyruvate (GIBCO), 0.5x N2 supplement (GIBCO), 0.5x B27 supplement (GIBCO), 3 μM GSK3i (CHIR99021), 1 μM MEKi (PD0325901) and Leukemia Inhibitory Factor (LIF; produced in house). Cells were passaged every 2-3 days by aspirating medium, dissociating cells with 0.25% trypsin-EDTA (GIBCO) briefly at 37°C before the addition of an equal volume of 1x trypsin inhibitor (Sigma) and gentle disruption by pipetting. Cells were pelleted by centrifugation, washed in 2i-LIF to remove excess trypsin and pelleted again before resuspending and plating ~1x10⁶ cells/10 cm plate.

For differentiation into EBs, dissociated cells were washed 2x in Serum-LIF (GMEM (Sigma) supplemented with 10% heat-inactivated fetal bovine serum, 1x Pen-Strep (GIBCO), 2 mM Glutamax (GIBCO), 0.1 mM nonessential amino acids (GIBCO), 1 mM sodium pyruvate (GIBCO), 50 μM (β-mercaptoethanol (GIBCO)) before seeding 1.5x10⁶ cells into 10 cm Petri-dishes containing Serum-LIF media. Media was changed at days 2, 3, 5 and 7. At day 7, EBs were transferred to 0.2% gelatin coated plates and grown for a further 3 days.

Phase contrast microscopy images were captured using an Olympus IX73 inverted microscope using the cellSens Entry software (Olympus).

**METHOD DETAILS**

**CRISPR/Cas9 KOs**

KO cell lines were generated by CRISPR/Cas9 targeting of Zfc3h1 in WT ESC. Single guide (sg) RNAs (Table S1) were cloned into the pSPCas9(BB)-2A-GFP vector (pX458, Addgene plasmid ID: 48138) as previously described (Ran et al., 2013) and transfected into ES cells using Lipofectamine 2000 (Thermo). Single cell clones were isolated by GFP sorting using FACS into 0.2% gelatin coated plates and grown for a further 3 days. KO clones were screened by western blotting analysis and validated by Sanger sequencing of amplified genomic DNA around the cut site. Three independent Zfc3h1−/− cell lines were derived from expanded single cell clones.

**RNA isolation**

Total RNA was isolated using the RNeasy Mini Kit (QIAGEN) according to the manufacturer’s instructions or by Trizol extraction (Thermo) using the standard protocol. For chromatin associated RNA, samples were prepared as previous described (Conrad and Ørom, 2017).

**pA+ RNA purification**

pA+ RNA was isolated from nuclear RNA samples using the Dynabeads mRNA Purification Kit (Thermo). For isolation of nuclei, 2x10⁶ cells were resuspended in nuclear isolation buffer (NIB) (10 mM Tris pH 7.4, 150 mM NaCl, 0.15% Igepal CA-630) supplemented with protease inhibitors and lysed at 4°C on a rotating wheel for 5 minutes. Lysates were overlaid onto 1 mL Sucrose buffer (10 mM Tris pH 7.4, 150 mM NaCl, 24% sucrose) in a DNA LoBind tube (Eppendorf) and nuclei were pelleted for 10 minutes at 2000 x g. Nuclei were resuspended in 1 mL Trizol (Thermo) and RNA was extracted using the standard protocol. 50 μg of nuclear RNA extracts were heated to 65°C and cooled on ice before incubating with oligo dT(25) Dynabeads (Thermo). Bead complexes were washed twice before elution in 10 mM Tris pH 7.5 and recovered RNA were assessed using a NanoDrop Lite Spectrophotometer (Thermo).

**qRT-PCR analysis**

cDNA was prepared from 500 ng of total RNA with TaqMan Reverse Transcription reagents (Thermo) using random hexamers. qRT-PCR was performed using the LightCycler 480 SYBR Green I (Roche) in technical triplicates. Primers used in qRT-PCR are listed in Table S2.

**RNA-seq library preparation**

RNA-seq libraries were prepared from 1 μg of total RNA using the TruSeq Stranded Total RNA library prep kit with RiboZero Gold (Illumina) according to the manufacturer’s instructions. Three biological replicates from each sample were prepared. RNA integrity and library quality were assessed on a Bioanalyzer 2000 using RNA Nano and DNA 1000 chips (Agilent), respectively. Libraries were quantified and normalized for multiplexing using the KAPA library quantification Kit for Illumina (KAPA Biosystems) and sequenced on an Illumina NextSeq 550 (75-bp, paired-end).

**Western blotting analysis**

Protein lysates were prepared using TOPEX+ buffer (Rising et al., 2014) (300 mM NaCl, 50 mM Tris-HCl pH 7.5, 0.5% Triton X-100, 1% SDS) freshly supplemented with protease inhibitors, 1 mM DTT and 33.3 U/ml Benzonase (Novagen). SDS-PAGE and western blotting analysis were carried out according to standard protocols with the antibodies listed in the Key Resources Table and HRP conjugated secondary antibodies (Vector Laboratories and Agilent). Bands were visualized by Super Signal West Pico chemiluminescent ECL (Thermo) and exposed either on Amersham Hyperfilm ECL films (GE Healthcare) and developed (Ferrania Imagine Technologies) or digitally captured using an Amersham Imager 600 (GE Healthcare). Images were processed and quantified using ImageJ (Schneider et al., 2012).
**IP experiments**

For whole cell IPs, $1 \times 10^7$ cells/IP were resuspended in HT150 extraction buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 0.5% Triton X-100) supplemented with protease inhibitors and sheared mechanically using 22 G and 27 G needles sequentially with 6 strokes each. For nuclear IPs, $1 \times 10^7$ cells/IP were resuspended in nuclear isolation buffer (NIB) (10 mM Tris pH 7.4, 150 mM NaCl, 0.15% Igepal CA-630) supplemented with protease inhibitors and lysed at 4°C on a rotating wheel for 5 minutes. Lysates were overlaid onto 1 mL Sucrose buffer (10 mM Tris pH 7.4, 150 mM NaCl, 24% sucrose) in a DNA LoBind tube (Eppendorf) and nuclei were pelleted for 10 minutes at 2000 x g. Nuclei were resuspended in 250 μl/IP RNA isolation buffer (RIB) (25 mM Tris pH 7.4, 150 mM KCl, 0.5 mM DTT, 0.5% Igepal CA-630) supplemented with protease inhibitors and sheared mechanically using 22 G and 27 G needles sequentially with 6 strokes each.

Claried lysates were treated with DNasel (Thermo) for 20 minutes at 37°C before pre-clearing with rabbit IgG (Millipore) and Protein-A Dynabeads (Thermo) for 2 hours at 4°C. Supernatants were incubated with either IgG, SUZ12 (Cell Signaling) or EZH2 (made in-house) antibodies overnight at 4°C with Protein-A Dynabeads. Beads were washed 3 times with the respective extraction buffer, transferring beads to a fresh tube on the final wash. Proteins were eluted by boiling in 1X NuPAGE loading buffer (Invitrogen). 10X reducing agent (Invitrogen) was added to the supernatants before denaturing for 10 minutes at 95°C and proceeding with western blotting analysis.

**Glycerol gradient sedimentation analysis**

The glycerol gradient sedimentation analysis was performed as previously described (Chu et al., 2014), with minor modifications. Briefly, whole cell extracts from $\sim 2 \times 10^7$ cells were resuspended in BC100 buffer (5 mM HEPES pH 7.5, 100 mM NaCL, 1 mM MgCl$_2$, 0.5 mM EGTA, 0.1 mM EDTA, 10% v/v glycerol, 1 mM DTT) supplemented with protease inhibitors, lysed by sonication (3 x 5 s, amplitude 2) and centrifuged at 14,000 rpm for 20 minutes. Clarified lysates were loaded on 10%–50% (v/v) glycerol gradients prepared in BC100 buffer and centrifuged at 35,000 rpm for 28 hours using a SW41 rotor (Beckman). Gradients were separated into 18 fractions and protein content was assessed using the Bradford assay or using the Blue Silver modified Neuhoff’s colloidal Coomassie Blue G-250 stain (Candiano et al., 2004).

**RNaseA treatment**

Cells were treated with RNaseA as previously described (Beltran et al., 2016). Cells were trypsinised and permeabilised with 0.05% Tween-20 (Sigma) in PBS for 10 minutes on ice. Cells were washed once, resuspended in PBS and either mock-treated or treated with 1 mg/ml RNaseA (Thermo) for 30 minutes at RT with gentle agitation. Cells were washed twice with PBS before proceeding with lysis and IP.

**Subcellular fractionation**

Cells were separated into cytoplasmic, nuclear and chromatin fractions using the Subcellular Protein Fractionation Kit (Pierce) according to the manufacturer’s instructions. Cells were harvested in 1x10$^7$ aliquots, split into 2 samples for either whole cell extraction using TOPEX+ buffer or fractionation. Equivalent lysate volumes were separated by SDS-PAGE and analyzed by western blotting.

**RIP experiments**

Whole cell IPs were performed as described above with alterations. All buffers were additionally supplemented with 100 U/ml RiboLock RNase Inhibitor (Thermo). Following overnight IP with IgG, SUZ12 or EZH2 antibodies, beads were washed 3 times with HT150 buffer, transferring beads to a fresh tube on the final wash. RNA was isolated from the IPs by the addition of 1 mL Trizol, homogenization and incubating for 1 hour on ice. Beads were removed magnetically and RNA was isolated from Trizol using the standard protocol.

**RIP-seq**

Strand specific libraries were prepared from $\sim$200 ng RNA isolated in SUZ12 and EZH2 RIP experiments by BGI Tech Solutions (Europe) according to their IncRNA-seq library preparation protocol. Two biological replicates were prepared from each sample. RNA integrity was assessed using a BioAnalyzer 2000 (Agilent) using RNA Nano chips. Samples were ribodepleted using Ribop Zero (Illumina) and libraries prepared using the TruSeq Stranded mRNA library prep kit (Illumina). Libraries were sequenced on a BGISEQ-500 (100 bp, paired end).

**ChIP experiments**

ChIP experiments were carried out according to standard protocols. Briefly, ES cells were cross-linked by the addition of 1% form- aldehyde (Sigma) in the dish for 10 minutes at RT before quenching with glycine. DNA was sheared to $\sim$200 bp fragments by sonication using a Biorupter (Diagenode) and validated by agarose gel electrophoresis. ChIPs were carried out using 200 μg of chromatin and 2-3 μg of the indicated antibodies (Key Resources Table). Libraries for ChIP-seq were prepared using NEBNext Ultra II DNA Library prep kit (NEB) using AmpureXP beads (Beckman) for size selection. Libraries were assessed on a Bioanalyzer 2000 (Agilent).
using High Sensitivity DNA chips and quantified using the Qubit dsDNA HS assay kit (Thermo). Libraries were sequencing on an Illumina NextSeq 550 (75 bp, single end).

QUANTIFICATION AND STATISTICAL ANALYSIS

Processing and analysis of RNA-seq data
Quality control of sequence reads was done using FastQC v0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Illumina adaptors (as provided with the FastQC tool), low quality bases, the first 12 bases and reads shorter than 25 nt were removed with Trimmomatic v0.32, using settings "ILLUMINACLIP:<TrueSeq3_PE_2>:-2:30:10 HEADCROP:12 LEADING:22 SLIDING WINDOW:4:22 MINLEN:25" (Bolger et al., 2014). Both paired and unpaired (due to the trimming) reads were mapped using HISAT v0.1.6.beta (Kim et al., 2015), against the mouse genome (mm10), where a list of GENCODE M12 (Frankish et al., 2019) annotated splice sites was also provided. HISAT was furthermore run with maximum fragment length set to 1000 and the `-rf` parameter (for the upstream/downstream mate orientation), otherwise default settings was used.

To generate visualization, properly paired reads mapping onto unique genomic locations were selected. Genomcov from bedtools v2.23.0 (Quinlan and Hall, 2010) was used to calculate strand-specific per-base genome coverage in bedgraph format. Bedgraph files were converted into bigwig format for using the UCSC Genome Browser Utility ‘bedGraphToBigWig’ (Kent et al., 2002). Finally, the per base coverage was CPM normalized. Genome browser images are generated from IGV (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

Paired and uniquely mapped exonic reads for GENCODE M12 genes were counted using featureCounts from the R package Rsubread (1.32.1) for individual RNA-seq libraries. RNA-seq read counts for PRC2 KO (Ezh1−/−/Ezh2−/−) and corresponding WT samples were obtained from the Gene Expression Omnibus (GEO), accession number: GSE103685 (Hajfeldt et al., 2018). Differential expression analysis was performed using the R package edgeR (version 3.24.1) with default parameters (McCarthy et al., 2012; Robinson et al., 2010). Intronic regions of a gene were defined as regions in the gene body that do not overlap with an exon from any GENCODE M12 annotated transcript isoform. Intronic reads were counted and differential expression analysis was performed using the same method as for exonic reads.

Processing and analysis of RIP-seq data
Quality control of sequence reads was done using FastQC v0.11.2. Illumina adaptors, low-quality bases with Phred score lower than 20, and reads shorter than 25 bp were removed using Trim Galore (version: 0.4.4, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/); reads were further trimmed using parameters–clip_R1 13–clip_R2 13–three_prime_clip_R1 1–three_prime_clip_R2 1. Trimmed reads were mapped as described for RNA-seq data processing. BAM alignment files were further processed to select uniquely mapped and properly paired reads using SAMtools (version 1.6.1) (Li et al., 2009). Duplicated reads were removed using MarkDuplicates (version: 2.8.1) from GATK (McKenna et al., 2010) with default settings and processed BAM files were used for downstream analysis. The strand specific genomic coverage was calculated using bamCoverage from deepTools (Ramirez et al., 2014). Raw reads for GENCODE M12 genes were counted using the same method as for RNA-seq, and differential analysis was performed using edgeR. To calculate the enrichment of transcripts pulled down by SUZ12 and EZH2 IPs in Zfcrh1/C0 compared to WT, contrast in the differential analysis was made as (Zfcrh1/C0 IP – Zfcrh1/C0 input) – (WT IP – WT input) and the enrichment score was defined as log2 fold change calculated from the contrast.

Processing and analysis of ChIP-seq data
Reads were mapped onto the mouse (mm10) genome with Bowtie (Langmead and Salzberg, 2012), selecting only hits with the best stratum with up to two mismatches in the seed and reporting up to four good alignments per read. Two only copies of identical alignments were kept. The fragment sizes could be deduced using the ChiP-Cor tool (Ambrosini et al., 2016). Reads were shifted to the center of their fragment by half of the deduced fragment size. Genomcov from bedtools v2.23.0 (Quinlan and Hall, 2010) was used to calculate strand specific per base genome coverage in bedgraph format. Bedgraph files were converted into bigwig format for using the UCSC Genome Browser Utility ‘bedGraphToBigWig’ (Kent et al., 2010). Finally, the per base coverage was CPM normalized.

For global MA plots, ChIP signals from all libraries were quantified on genomic windows of 2 kb sliding by 500 bp using the UCSC Genome Browser Utility ‘bigWigAverageOverBed’ (Kent et al., 2010). SUZ12 peaks were called using MACS2 (Zhang et al., 2008) (version 2.1.1.20160309) with parameters–value 0.05–broad–broad-cutoff 0.3, ENCODE blacklisted peaks (ENCODE Project Consortium, 2012) and low quality peaks (-log10(value) ≤ 1) were filtered out. Consistent SUZ12 peaks between Zfcrh1/C0 replicates were defined as peaks overlapping in at least two replicates. A single SUZ12 reference peak set for WT and Zfcrh1/C0 was obtained by pooling SUZ12 peaks of WT and consistent SUZ12 peaks of Zfcrh1/C0 and merging overlapping peak regions into a single region using mergeBed from bedtools (v2.23.0). Raw reads of H3K27me3 and SUZ12, for SUZ12 peaks in the reference peak set, were obtained using featureCounts from the R package Rsubread (1.32.1), a pseudocount of 1 was added when normalizing raw read counts to the library size. log2FC values were calculated between mean values of the normalized read counts from Zfcrh1/C0 replicates and the normalized read counts from WT. For mean signal plots, genomic coverage from bigWig files was calculated using computeMatrix from deepTools (version 2.5.3) (Ramirez et al., 2014). Regions without read coverage in bigWig files were treated as 0, mean values from the replicates were calculated and mean values at each position were plotted. For the heatmap, a list of genes
overlapping with at least one SUZ12 peak from the reference peak set was obtained, H3K27me3 and SUZ12 were quantified in the
gene bodies, log₂FC values of H3K27me3 and SUZ12 between Zfc3h1−/− and WT were calculated using the method described
above, and log₂FC values of gene exonic expression were computed from the differential expression analysis of RNA-seq described
above.

DATA AND CODE AVAILABILITY

All high-throughput RNA-seq, ChIP-seq and RIP-seq datasets generated during this study are available at the Gene Expression
Omnibus (GEO) under accession code GSE137491.
Supplemental Information

A Functional Link between Nuclear RNA Decay and Transcriptional Control Mediated by the Polycomb Repressive Complex 2

FIGURE S1, related to Figure 1

A

WT | gRNA
---|---
ACGGGAAAGGCCGCGCCCGCCGGGTCCGTT AGGACAGAGGGCGGAGGAG

Zfc3h1⁻/⁻

#1

ACGGGAAAGGCCGCGCCCGCCGGGTCCGTT TAGGGAGGAG

#2

ACGGGAAAGGCCGCGCCCGCCGGGTCCGTT AGGACAGAGGGCGGAGGAG

#3

ACGGGAAAGGCCGCGCCCGCCGGGTCCGTT TAGGGAGGAG

∆13

∆7

∆14

∆7

∆1

∆1

B

kDa

<table>
<thead>
<tr>
<th>WT</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
</tr>
</thead>
<tbody>
<tr>
<td>35 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>130 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>100 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>40 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

C

Nuclear pA⁺ RNA

% pA⁺ RNA from total nuclear RNA

WT

Zfc3h1⁻/⁻

D

WT

Zfc3h1⁻/⁻ (#1)

E

kDa

<table>
<thead>
<tr>
<th>Zfc3h1⁻/⁻</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
</tr>
</thead>
<tbody>
<tr>
<td>35 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>40 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>40 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>100 -</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

F

Samples taken at days 0, 3, 5, 7 and 10

Beating cells observed ~ day 8/9

EB's transferred to gelatinised plate

1.5x10⁶ cells seeded in low attachment plates in serum - LIF media

Day 0

Day 3

Day 5

Day 7

Day 10

Differentiation

G

Oct4 ExIn

Nr0b1 ExIn

Time after EB induction (days)
Supplemental Figure 1, related to Figure 1

(A) Genomic validation of CRISPR/Cas9-engineered Zfc3h1<sup>−/−</sup> cell lines. PCR was carried out on genomic DNA isolated from three Zfc3h1<sup>−/−</sup> cell lines derived from single cell clones. To distinguish individual alleles, amplicons were cloned into the pCR4 vector and sequenced. The gRNA sequence is highlighted and the PAM motif is shown in bold. Base pair deletion sizes are indicated per allele (Δn).

(B) Western blotting analysis using the same lysates as in Figure 1B. Blots were probed with the indicated antibodies and actin (ACTB) was used as a loading control.

(C) Quantification of pA<sup>+</sup> RNA purified from WT and Zfc3h1<sup>−/−</sup> nuclei. Results are shown as percentage of total nuclear RNA. Columns represent the average value of biological triplicates with error bars denoting the SD. Individual data values from technical triplicates are indicated as points.

(D) Phase contrast microscopy images of undifferentiated (D0) WT and Zfc3h1<sup>−/−</sup>(#1) cells. Scale bars denote 200 µm.

(E) Western blotting analysis of extracts from WT and Zfc3h1<sup>−/−</sup> cell lines to assess the expression of pluripotency markers. Blots were probed with the indicated antibodies and vinculin was used as a loading control.

(F) Overview of the EB differentiation assay. ES cells, growing in 2i/LIF media, were transferred to serum-LIF media and seeded in low attachment plates. Samples for RNA and protein analyses were taken at the indicated days (0, 3, 5, 7, 10). At day 7, EBs were transferred to gelatinised plates and cultured for an additional 3 days.

(G) qRT-PCR analysis of total RNA from Figure 1E but using exon-intron (ExIn) primers to amplify pre-mRNA. Results are shown relative to Rplp0 (RPO) mRNA as in Figure 1E.
FIGURE S2, related to Figure 2

A

WT #2
(Day 0)

S: 0.99
P: 0.99

WT #3
(Day 0)

Zfc3h1-/- #1
(Day 0)

Zfc3h1-/- #2
(Day 0)

Zfc3h1-/- #3
(Day 0)

B

WT #1
(Day 7)

S: 0.96
P: 0.98

WT #2
(Day 7)

S: 0.97
P: 0.99

WT #3
(Day 7)

Zfc3h1-/- #1
(Day 7)

S: 0.98
P: 0.97

Zfc3h1-/- #2
(Day 7)

S: 0.78
P: 0.84

Zfc3h1-/- #3
(Day 7)

S: 0.72
P: 0.82

FIGURE S2, related to Figure 2
Supplemental Figure 2, related to Figure 2

(A) Pairwise comparisons of expression from replicate RNAseq samples from WT and Zfc3h1-/- cell lines after 7 days of EB differentiation. All genes from GENCODE were included and expression were measured as TPM. Spearman and Pearson correlation coefficients, (S) and (P) respectively, are shown for each plot. (B) As in (A) but with samples taken from WT and Zfc3h1-/- undifferentiated (D0) cells.
FIGURE S3, related to Figure 3

A. Zfc3h1−/−

B. H3K27me3 ChIP

C. SUZ12 ChIP

D. WT Zfc3h1−/−

E. SUZ12 ChIP

F. H3K4me1 ChIP

G. H3K4me3 ChIP

H. H3K27 ac ChIP

Distance from SUZ12 peak (bp)
Supplemental Figure 3, related to Figure 3

(A) Western blotting analysis of PRC2 components using samples prepared as in Figure 1B. Blots were probed with the indicated antibodies and Tubulin (TUBA1B) was used as a loading control. (B-C) MA plots of H3K27me3 (B) and SUZ12 (C) ChIPseq profiles in Zfcs3h1−/− vs. WT cell lines as in Figure 3A, but showing all 3 replicate experiments. (D) Top panel: Western blotting analysis of global H3K27me3 levels from three biological replicates of WT and Zfcs3h1−/− ES cells. Blots were probed with H3K27me3-specific antibody and H4 was used as a loading control. Bottom panel: Quantification of western results from top panel. H3K27me3 values were normalised to H4. Individual points indicate biological replicates. (E) Mean SUZ12 and H3K27me3 ChIPseq signals as in Figure 3C-D, but showing all 3 replicate experiments (F) Average H3K4me1, H3K4me3 and H3K27ac ChIPseq signals from WT and Zfcs3h1−/− cell lines in regions centred on SUZ12 ChIP peaks as in Figure 3C-D. (G) Genome browser views of 4 PRC2 target genes as in Figure 3G, but showing all biological replicate experiments. (H) Genome browser views of 5 PRC2 target genes. Displayed tracks include H3K4me1, H3K4me3, H3K27ac, H3K27me3, SUZ12 and RNAPII ChIPseq data as well as RNAseq data from WT and Zfcs3h1−/− (#1) cell lines. RNAseq tracks on both strands are shown (+ and – respectively). Gene models are based on RefSeq. Genome coordinates (mm10) are indicated for each panel.
FIGURE S4, related to Figure 4

**A**
![Graph showing relative co-IP protein levels](image)

**B**
![Graph showing relative co-IP protein levels](image)

**C**
![Image of nuclear inputs](image)

**D**
![Graph showing protein distribution](image)

**E**
![Image of nuclear inputs](image)

**F**
![Image of nuclear inputs](image)

**G**
![Image of nuclear inputs](image)

**H**
![Graph showing relative protein amounts](image)
Supplemental Figure 4, related to Figure 4

(A) Quantification of western blotting analysis from Figure 4A, and including 2 additional replicates (data not shown). EZH2 and EED values from IP samples were normalised to SUZ12 values from IP samples. Individual points indicate replicates. (B) As in (A) but for results from Figure 4B, and including 2 additional replicates (data not shown). SUZ12 and EED values from IP samples were normalised to EZH2 values from IP samples. (C) Western blotting analysis of SUZ12 IPs as in Figure 4A, but using lysates prepared from isolated nuclei. Actin (ACTB) was used as a loading control for nuclear inputs. (D) As in (A) but for results from Figure S4C. EZH2 and EED values from nuclear IP samples were normalised to SUZ12 values from nuclear IP samples. (E) Total protein content in glycerol gradient sedimentation assays from Figure 4C and 4D. Left panel: Input and 2% of indicated fractions were separated on SDS-PAGE gels and stained with Coomassie Blue 250G. Right panel: Protein concentration from each fraction was measured by Bradford assay. Axes show the OD_{595} reading (Y) in each fraction (X) for WT (blue) and Zfc3h1^{-/-} (green) samples. Values are the average of two replicates and error bars show the SD. A smoothed curve was fit to show the distribution of protein throughout the gradient. (F) Left panels: As in Figure 4C but showing the sedimentation of EED, Vinculin (VCL) and Tubulin (TUBA1B) in glycerol gradients from WT and Zfc3h1^{-/-} cell extracts. Right panels: Quantification of western blots as in Figure 4C. (G) Western blotting analysis of equivalent lysates isolated from whole cell extract (WCE), cytoplasmic (CYT), nucleoplasmic (NUC) and chromatin-associated (CHR) fractions of WT and Zfc3h1^{-/-} cells. Tubulin (TUBA1B), SFPQ and H3 were used as cytoplasmic, nucleoplasmic and chromatin-associated protein markers, respectively. (H) Quantification of western blots from (G). Columns show the average values from biological triplicates and error bars denote the SD. Individual points indicate replicates.
FIGURE S5, related to Figure 4

A

B

C

D

E

SUZ12 RIP

EZH2 RIP

SUZ12 RIP

EZH2 RIP

SUZ12 IP

Relative co-IP protein (Relative to SUZ12 IP levels)

EZH2

EED

WT (- RNaseA)

Zfc3h1-/- (-RNaseA)

WT (+ RNaseA)

Zfc3h1+/+ (+RNaseA)
Supplemental Figure 5, related to Figure 4

(A) Genome browser views of 3 PAXT target genes (Hotairm1, Fam120aos and Snhg10). Displayed tracks include RNAseq data from input samples along with SUZ12 RIPseq and EZH2 RIPseq data from WT and Zfc3h1⁻/⁻ cell lines. SUZ12 and EZH2 RIPseq tracks were coloured red and blue respectively. RNAseq tracks on both strands are shown (+ and – respectively) and include biological duplicates (#1 and #2). Gene models are based on RefSeq. Genome coordinates (mm10) are indicated for each panel.

(B) As in (A) but displaying 3 PRC2 target genes (Hoxd8, Hoxd9 and Pitx2).

(C) Box plot distribution showing relative enrichment of RNAs in SUZ12 and EZH2 RIPseq samples from WT and Zfc3h1⁻/⁻ cells as indicated. Transcripts shown are protein coding and ncRNAs, that were upregulated in Zfc3h1⁻/⁻ RNAseq data (log₂FC > 0.5, FDR < 0.05). Enrichment scores were defined as log₂FC ((Zfc3h1⁻/⁻ IP – Zfc3h1⁻/⁻ input) – (WT IP – WT input)).

(D) Quantification of western blots from Figure 4F, and including two additional replicates (data not shown). Values were plotted as in Figure S4A.
### SUPPLEMENTAL TABLES

#### Table S1, related to STAR methods

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zfc3h1_CRISPR_2_F</td>
<td>CACCGGCCGGCGCCGCGGTCGTT</td>
</tr>
<tr>
<td>Zfc3h1_CRISPR_2_R</td>
<td>AAACAACGGACCCCGGGCCGCGGCC</td>
</tr>
</tbody>
</table>

#### Table S2, related to STAR methods

<table>
<thead>
<tr>
<th>Name</th>
<th>Forward</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct4</td>
<td>CAGCAGATCACATCACATCGCCA</td>
<td>GCCTCATACTCTTCTCGTTGGG</td>
</tr>
<tr>
<td>Sox2</td>
<td>AACGGCAGCATACAGCATGATGC</td>
<td>CGAGCTGCTATGAGGTATAC</td>
</tr>
<tr>
<td>Klf2</td>
<td>CACCTAAAGGCATCTGCCTGA</td>
<td>GTGACCTGTTGCTTTCCGTTAG</td>
</tr>
<tr>
<td>Esrrb</td>
<td>GACATTGCTCTGCTACCAC</td>
<td>ACTTGCGCTCCGTTTGGTGAT</td>
</tr>
<tr>
<td>Gata4</td>
<td>GCCCTCTACACAAGATGAAGGC</td>
<td>TACAGGCTCACCTCGGCAATT</td>
</tr>
<tr>
<td>Tbx1</td>
<td>ATCCACCGAAGACTCGCCTGAAATT</td>
<td>CTCTCAGATGTGAATGCCAGG</td>
</tr>
<tr>
<td>Nkx2.5</td>
<td>CTCCGATCCATCCCCACTTTA</td>
<td>AGTGTGAAATCCGTCGAAGG</td>
</tr>
<tr>
<td>Oct4 ExIn</td>
<td>GGGCTTTGGCCAAGTGCAATAC</td>
<td>CCAAGGTCTCCGATTTGCATAT</td>
</tr>
<tr>
<td>Nr0b1 ExIn</td>
<td>GGACCGTGCTCTTTACACCAG</td>
<td>CAAGTGGGCTCCAAACCC</td>
</tr>
<tr>
<td>Hoxd8 ExIn</td>
<td>GCCGAGGACCACAGACCAC</td>
<td>GCAGATTCTCCCTTTCTCCCC</td>
</tr>
<tr>
<td>Hoxd9 ExIn</td>
<td>TCCGCTCTCGCTCGCTAAAC</td>
<td>TCCGCTCTGTGCTGCAAAC</td>
</tr>
<tr>
<td>Snhg10</td>
<td>GCCCTTCATGCTCCACCG</td>
<td>GAATCAGAGGATCTGGCAGA</td>
</tr>
<tr>
<td>Snhg20</td>
<td>CCGAAGGCGAATCTGGGAT</td>
<td>CAACACAGAAAGCGCCAGT</td>
</tr>
<tr>
<td>Snhg13 (Dancr)</td>
<td>GAAACCGTGACTGAATGGC</td>
<td>GAAACTCACATGGGCCCT</td>
</tr>
<tr>
<td>Hottirm1</td>
<td>GCACGAGGACAAAGGGAG</td>
<td>GCACGAGGACAAAGGGAG</td>
</tr>
<tr>
<td>Rplp0 (RPO)</td>
<td>TCCATGGGGAGCAAGCAC</td>
<td>CAGCGGTGTCCCAGAGC</td>
</tr>
<tr>
<td>GAPDH</td>
<td>TTGATGGCAACAATCTCCAC</td>
<td>CGTCCGTCGAGCAAAATGGT</td>
</tr>
</tbody>
</table>