Title
Chromatin Association of mRNAs Regulates Expression of Genes Important in Mouse Embryonic Stem Cell Biology

Author
Lim, Han Young

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Chromatin Association of mRNAs Regulates Expression of Genes Important in Mouse Embryonic Stem Cell Biology

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Molecular Biology

by

Han Young Lim

2022

ABSTRACT OF THE DISSERTATION

Chromatin Association of mRNAs Regulates Expression of Genes Important in Mouse Embryonic Stem Cell Biology

by

Han Young Lim

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2022

Professor Douglas L. Black, Chair

Gene expression involves multiple layers of regulation to change the amount of proteins produced. The complexity of this phenomenon starts in the nucleus, where various cis-regulatory elements near genes can engage in transcriptional regulation by association with transcription factors that often function in diverse combinations to enhance or silence transcription. In addition, chemical modifications of DNA such as methylation can have varying impacts on the rate of transcription. Furthermore, molecular events that occur to a maturing pre-mRNA synthesized by RNA polymerase II, notably 5’ capping, splicing and 3’ end processing, all contribute to the intricacies of gene regulation. Whereas these processes need to successfully occur to a maturing pre-mRNA to make it competent for export, malfunction in any of these steps can compromise the RNA’s nuclear export, often leading to its sequestration in the nucleus. Despite ongoing research in the nuclear retention of various RNAs, including mRNAs, mechanisms and tissue-specific consequences of this surveillance mechanism are

poorly understood. This work aims to better our understanding of the molecular mechanisms leading to chromatin association of RNAs, particularly mRNAs, by RNA binding proteins PTBP1 and NXF1. Our work demonstrates that these proteins play key roles in altering gene expression in mouse embryonic stem cells (mESCs) through manipulation of splicing and subcellular localization of their target RNAs. We discovered that these mechanisms can have direct consequences on the biology of the mESCs, affecting their pluripotency and differentiation.

The dissertation of Han Young Lim is approved.

Feng Guo

Kathrin Plath

Yi Xing

Douglas L. Black, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to first acknowledge my advisor Douglas Black, whose guidance, support and patience throughout my graduate school career have shaped me into a better scientist. I would also like to thank all members of the Black lab, past and present, for mentoring and supporting me. I wish to specifically thank Xiaojun Wang, Wen Xiao, Nivedita Damodaren, Xinyuan Chen, Kyu-Hyeon Yeom, Julia Nikolic, Andrey Damianov, Nazim Mohammad and Chia-Ho Lin, in no particular order, for always being willing to help me with techniques that were critical for research and continuous support and guidance, both academic and emotional. Furthermore, I would like to thank Kathrin Plath, Feng Guo and Yi Xing for scientific guidance as committee members. Without these wonderful people, I would not have been able to get to this point.

Next, I would like to acknowledge all the friends that I have made since the beginning of graduate school. Anthony Chau, Lisa Truong and Nhan Phan, who I met in the first week of graduate school and are friends ever since, have not only made my graduate school experience more enjoyable, but also provided continuous support and guidance for matters related to being a PhD student and life in general – I thank all of them for that. I also would like to thank friends outside of graduate school, especially Dakota Im for being a great friend and supporter of my work in the program despite having limited knowledge in biology. Finally, I would like to thank my partner, Ada Dai, for continuous support and being the best partner in life.

Lastly, I would like to acknowledge my family. Despite being over 5,000 miles away across the Pacific Ocean, my older brother, Han Soo Lim, and my parents, Yong

Han Lim and Seon Hee Han, have continuously supported me no matter what happened. I would like to dedicate this thesis to them.

Funding sources: NIH R35 GM136426 to DLB, the Broad Stem Cell Research Center to DLB, and the Keck Foundation to DLB, and the UCLA Molecular Biology Institute to HYL.

VITA

EDUCATION

2015 Bachelor of Science, Microbiology, Immunology and Molecular Genetics
University of California, Los Angeles | Los Angeles, CA

PUBLICATIONS

Yeom KH, Pan Z, Lin CH, **Lim HY**, Xiao W, Xing Y, Black DL. Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring. Genome Res. 2021 Jun;31(6):1106-1119. doi: 10.1101/gr.273904.120. Epub 2021 Apr 8. PMID: 33832989; PMCID: PMC8168582.

Dai X, Gong D, **Lim H**, Jih J, Wu TT, Sun R, Zhou ZH. Structure and mutagenesis reveal essential capsid protein interactions for KSHV replication. Nature. 2018 Jan 25;553(7689):521-525. doi: 10.1038/nature25438. Epub 2018 Jan 17. PMID: 29342139; PMCID: PMC6039102.

Yang K, Wills E, **Lim HY**, Zhou ZH, Baines JD. Association of herpes simplex virus pUL31 with capsid vertices and components of the capsid vertex-specific complex. J Virol. 2014 Apr;88(7):3815-25. doi: 10.1128/JVI.03175-13. Epub 2014 Jan 22. PMID: 24453362; PMCID: PMC3993549.

Chapter 1: Introduction

Gene expression is a highly complex process that begins with transcription of nascent pre-mRNA on the chromatin and ends with translation of the mature mRNA in the cytoplasm. In the nucleus, numerous processes exist to modulate processing of a pre-mRNA to make it suitable for nuclear export and translation (Bentley 2014). 5' capping, which is the first step in mRNA processing and occurs as early as after synthesis of 20 nucleotides, has been implicated to occur in a co-transcriptional manner due to association of capping proteins with the C-terminal domain of RNA polymerase II (Ho and Shuman 1999; Rasmussen and Lis 1993). Splicing, another crucial process in pre-mRNA processing, has been long known to occur by recruitment of spliceosome factors to a growing strand of mRNA during transcription (Görnemann et al. 2005). 3' end cleavage and polyadenylation, which is considered one of the final processing steps as full-length transcription of the pre-mRNA completes, most often follows completion or near-completion of splicing (Schmidt et al. 2011). Since each major processing event involves recruitment and dissociation of multiple RNA binding proteins on and from a pre-mRNA, with interactions often occurring between proteins from more than one event, it is not surprising that misregulation and appropriate quality check measures exist to prevent export of incompletely processed transcripts (Fasken and Corbett 2009).

While the vast majority of RNAs exist in the cytoplasm as mature transcripts for translation, incompletely-spliced transcripts are often found in the nucleus at specific sites that show as puncta representing their sites of transcription (Vargas et al. 2011). When splicing isn't complete following transcription, mRNAs can be retained in the nucleus, specifically nuclear speckles, until splicing finishes in a post-transcriptional

manner and nuclear export ensues (Girard et al. 2012). Some RNAs, including mRNAs with detained introns and long non-coding RNAs, have been shown to be sequestered in the nucleus and have varying consequences related to gene expression and chromatin structure (Boutz, Bhutkar, and Sharp 2015; Quinn and Chang 2016). An example of this was described in a previous work where abundant full-length yet incompletely spliced transcripts accumulate on and associate with the chromatin in lipid A-stimulated macrophages (Bhatt et al. 2012; Pandya-Jones et al. 2013). These transcripts are mostly likely productive mRNAs as opposed to nonproductive “dead-end” transcripts because their accumulation on chromatin was observed to follow accumulation in the nucleoplasm and cytoplasm with a defined temporal delay (Bhatt et al. 2012; Pandya-Jones et al. 2013). Similar mechanism was observed in neurons where polyadenylated yet incompletely spliced transcripts retained in the nucleus are fully spliced and loaded onto ribosomes in response to stimulation (Mauger, Lemoine, and Scheiffele 2016). These findings reveal an intricate mechanism for regulating gene expression in a spatial and temporal manner involving delayed splicing and nuclear/chromatin sequestration of mRNAs encoding regulatory proteins. Despite ongoing research such as ones mentioned, modulation of gene expression through nuclear and intron retention of regulatory transcripts is still largely not understood. Specifically, polyadenylated yet incompletely spliced transcripts, which have been frequently observed in the nucleus, require further investigation since their subnuclear localization and mechanism of retention remain mostly unknown.

In **Chapter 2**, we present a detailed work of understanding subcellular localization and intron retention of transcripts in mouse embryonic stem cells (mESCs),

neuronal progenitor cells and postmitotic neurons. Using biochemical fractionation and RNA sequencing of the individual fractions, we find defined RNA partitioning between cellular compartments of genes that undergo co-transcriptional and sometimes post-transcriptional splicing of select introns. We also categorize polyadenylated transcripts into four groups based on their level of intron retention across the three subcellular fractions, and we discover that many of the introns are differentially spliced across the three cell types. Finally, we discuss polypyrimidine tract binding protein (PTBP1) as an important regulator of intron retention in neuronal development and present transcripts of the Gamma-Aminobutyric Acid Type B Receptor Subunit 1 (GABBR1) gene as chromatin-associated polyadenylated mRNAs that are developmentally regulated in expression through splicing and localization. This work has been published as:

Yeom KH, Pan Z, Lin CH, Lim HY, Xiao W, Xing Y, Black DL. Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring. Genome Res. 2021 Jun;31(6):1106-1119. doi: 10.1101/gr.273904.120. Epub 2021 Apr 8. PMID: 33832989; PMCID: PMC8168582.

In addition to perturbation in splicing and release of RNA from chromatin, nuclear export of transcripts is a crucial step in controlling gene expression. Joining of two exons after splicing deposits a protein complex called exon junction complex (EJC) that lies ~24 nucleotides upstream of the exon-exon junction. Several reports showed that the EJC serves as a landing pad for export factors and thus a primary driver of mRNA nuclear export (Le Hir et al. 2001; Singh et al. 2012). However, reports that showed inconsistent results pointed to the cap binding complex as the main recruiter of export

factors including the TREX complex (Cheng et al. 2006). A recent work seemed to reconcile these two mechanisms through an *in vivo* study that revealed binding of export adapters such as NXF1 occurring throughout the RNA during 5' capping and splicing but before 3' end processing (Viphakone et al. 2019). Despite the varying observations on the mechanism of RNA export, NXF1 and its cofactor NXT1 have always been recognized as the prominent heterodimeric export factor that shuttles mature mRNA through the nuclear pore into the cytoplasm. However, recent transcriptomic analyses using cell fractionation and RNAi knockdown revealed that NXF1 and another export factor TPR are selective in their export targets (Lee et al. 2020; Zuckerman et al. 2020). In two separate studies conducted using human breast cancer MCF7 or osteosarcoma U2OS cells, researchers observed that transcripts of short and intron-poor genes are most strongly affected in export by NXF1 or TPR (Lee et al. 2020; Zuckerman et al. 2020). Nevertheless, exploration of NXF1's role in RNA export has largely not been done in multiple cell lines, and the mechanism and biological consequence of its selectivity remain poorly understood.

In **Chapter 3**, we present a comprehensive approach to investigating NXF1's role in RNA export of polyadenylated mESC transcripts. We show through subcellular fraction and RNA sequencing that a defined set of short and intron-poor RNAs encoding immediately early response proteins and other transcription factors are most strongly regulated by NXF1. In addition, we present an *in vivo* binding data of NXF1 showing that the export factor's association with RNAs occurs throughout the transcript. Lastly, we present evidence of NXF1 targeting mRNAs critical in pluripotency for nuclear export, and its depletion leads to differentiation.

References

Bentley, David L. 2014. “Coupling MRNA Processing with Transcription in Time and Space.” *Nature Reviews Genetics* 15 (3): 163–75. https://doi.org/10.1038/nrg3662.

Bhatt, Dev M., Amy Pandya-Jones, Ann-Jay Tong, Iros Barozzi, Michelle M. Lissner, Gioacchino Natoli, Douglas L. Black, and Stephen T. Smale. 2012. “Transcript Dynamics of Proinflammatory Genes Revealed by Sequence Analysis of Subcellular RNA Fractions.” *Cell* 150 (2): 279–90. https://doi.org/10.1016/j.cell.2012.05.043.

Boutz, Paul L., Arjun Bhutkar, and Phillip A. Sharp. 2015. “Detained Introns Are a Novel, Widespread Class of Post-Transcriptionally Spliced Introns.” *Genes & Development* 29 (1): 63–80. https://doi.org/10.1101/gad.247361.114.

Cheng, Hong, Kobina Dufu, Chung-Sheng Lee, Jeanne L. Hsu, Anusha Dias, and Robin Reed. 2006. “Human MRNA Export Machinery Recruited to the 5’ End of MRNA.” *Cell* 127 (7): 1389–1400. https://doi.org/10.1016/j.cell.2006.10.044.

Fasken, Milo B., and Anita H. Corbett. 2009. “Mechanisms of Nuclear MRNA Quality Control.” *RNA Biology* 6 (3): 237–41. https://doi.org/10.4161/rna.6.3.8330.

Girard, Cyrille, Cindy L. Will, Jianhe Peng, Evgeny M. Makarov, Berthold Kastner, Ira Lemm, Henning Urlaub, Klaus Hartmuth, and Reinhard Lührmann. 2012. “Post-Transcriptional Spliceosomes Are Retained in Nuclear Speckles until Splicing Completion.” *Nature Communications* 3: 994. https://doi.org/10.1038/ncomms1998.

Görnemann, Janina, Kimberly M. Kotovic, Katja Hujer, and Karla M. Neugebauer. 2005. "Cotranscriptional Spliceosome Assembly Occurs in a Stepwise Fashion and Requires the Cap Binding Complex." *Molecular Cell* 19 (1): 53–63. https://doi.org/10.1016/j.molcel.2005.05.007.

Ho, C. Kiong, and Stewart Shuman. 1999. "Distinct Roles for CTD Ser-2 and Ser-5 Phosphorylation in the Recruitment and Allosteric Activation of Mammalian MRNA Capping Enzyme." *Molecular Cell* 3 (3): 405–11. https://doi.org/10.1016/S1097-2765(00)80468-2.

Le Hir, H., D. Gatfield, E. Izaurralde, and M. J. Moore. 2001. "The Exon-Exon Junction Complex Provides a Binding Platform for Factors Involved in MRNA Export and Nonsense-Mediated MRNA Decay." *The EMBO Journal* 20 (17): 4987–97. https://doi.org/10.1093/emboj/20.17.4987.

Lee, Eliza S., Eric J. Wolf, Sean S. J. Ihn, Harrison W. Smith, Andrew Emili, and Alexander F. Palazzo. 2020. "TPR Is Required for the Efficient Nuclear Export of MRNAs and LncRNAs from Short and Intron-Poor Genes." *Nucleic Acids Research* 48 (20): 11645–63. https://doi.org/10.1093/nar/gkaa919.

Mauger, Oriane, Frédéric Lemoine, and Peter Scheiffele. 2016. "Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity." *Neuron* 92 (6): 1266–78. https://doi.org/10.1016/j.neuron.2016.11.032.

Pandya-Jones, Amy, Dev M. Bhatt, Chia-Ho Lin, Ann-Jay Tong, Stephen T. Smale, and Douglas L. Black. 2013. "Splicing Kinetics and Transcript Release from the Chromatin Compartment Limit the Rate of Lipid A-Induced Gene Expression." *RNA (New York, N.Y.)* 19 (6): 811–27. https://doi.org/10.1261/rna.039081.113.

Quinn, Jeffrey J., and Howard Y. Chang. 2016. “Unique Features of Long Non-Coding RNA Biogenesis and Function.” *Nature Reviews. Genetics* 17 (1): 47–62. https://doi.org/10.1038/nrg.2015.10.

Rasmussen, E. B., and J. T. Lis. 1993. “In Vivo Transcriptional Pausing and Cap Formation on Three Drosophila Heat Shock Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (17): 7923–27. https://doi.org/10.1073/pnas.90.17.7923.

Schmidt, Ute, Eugenia Basyuk, Marie-Cécile Robert, Minoru Yoshida, Jean-Philippe Villemin, Didier Auboeuf, Stuart Aitken, and Edouard Bertrand. 2011. “Real-Time Imaging of Cotranscriptional Splicing Reveals a Kinetic Model That Reduces Noise: Implications for Alternative Splicing Regulation.” *The Journal of Cell Biology* 193 (5): 819–29. https://doi.org/10.1083/jcb.201009012.

Singh, Guramrit, Alper Kucukural, Can Cenik, John D. Leszyk, Scott A. Shaffer, Zhiping Weng, and Melissa J. Moore. 2012. “The Cellular EJC Interactome Reveals Higher-Order MRNP Structure and an EJC-SR Protein Nexus.” *Cell* 151 (4): 750–64. https://doi.org/10.1016/j.cell.2012.10.007.

Vargas, Diana Y., Khyati Shah, Mona Batish, Michael Levandoski, Sourav Sinha, Salvatore A. E. Marras, Paul Schedl, and Sanjay Tyagi. 2011. “Single-Molecule Imaging of Transcriptionally Coupled and Uncoupled Splicing.” *Cell* 147 (5): 1054–65. https://doi.org/10.1016/j.cell.2011.10.024.

Viphakone, Nicolas, Ian Sudbery, Llywelyn Griffith, Catherine G. Heath, David Sims, and Stuart A. Wilson. 2019. “Co-Transcriptional Loading of RNA Export Factors

Shapes the Human Transcriptome.” *Molecular Cell* 75 (2): 310-323.e8. https://doi.org/10.1016/j.molcel.2019.04.034.

Zuckerman, Binyamin, Maya Ron, Martin Mikl, Eran Segal, and Igor Ulitsky. 2020. “Gene Architecture and Sequence Composition Underpin Selective Dependency of Nuclear Export of Long RNAs on NXF1 and the TREX Complex.” *Molecular Cell* 79 (2): 251-267.e6. https://doi.org/10.1016/j.molcel.2020.05.013.

Chapter 2: Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring

Resource

Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring

Kyu-Hyeon Yeom,[1,6] Zhicheng Pan,[2,3,6] Chia-Ho Lin,[1] Han Young Lim,[1,4] Wen Xiao,[1] Yi Xing,[3,5] and Douglas L. Black[1]

[1]*Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, California 90095, USA;* [2]*Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, Los Angeles, California 90095, USA;* [3]*Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA;* [4]*Molecular Biology Interdepartmental Doctoral Program, University of California, Los Angeles, Los Angeles, California 90095, USA;* [5]*Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

Steps of mRNA maturation are important gene regulatory events that occur in distinct cellular locations. However, transcriptomic analyses often lose information on the subcellular distribution of processed and unprocessed transcripts. We generated extensive RNA-seq data sets to track mRNA maturation across subcellular locations in mouse embryonic stem cells, neuronal progenitor cells, and postmitotic neurons. We find disparate patterns of RNA enrichment between the cytoplasmic, nucleoplasmic, and chromatin fractions, with some genes maintaining more polyadenylated RNA in chromatin than in the cytoplasm. We bioinformatically defined four regulatory groups for intron retention, including complete cotranscriptional splicing, complete intron retention in the cytoplasmic RNA, and two intron groups present in nuclear and chromatin transcripts but fully excised in cytoplasm. We found that introns switch their regulatory group between cell types, including neuronally excised introns repressed by polypyrimidine track binding protein 1 (PTBP1). Transcripts for the neuronal gamma-aminobutyric acid (GABA) B receptor, 1 (*Gabbr1*) are highly expressed in mESCs but are absent from the cytoplasm. Instead, incompletely spliced *Gabbr1* RNA remains sequestered on chromatin, where it is bound by PTBP1, similar to certain long noncoding RNAs. Upon neuronal differentiation, *Gabbr1* RNA becomes fully processed and exported for translation. Thus, splicing repression and chromatin anchoring of RNA combine to allow posttranscriptional regulation of *Gabbr1* over development. For this and other genes, polyadenylated RNA abundance does not indicate functional gene expression. Our data sets provide a rich resource for analyzing many other aspects of mRNA maturation in subcellular locations and across development.

[Supplemental material is available for this article.]

After transcription initiation, the maturation of pre-messenger RNA (pre-mRNA) requires splicing, polyadenylation, and release of the RNA from the chromatin template before export to the cytoplasm for translation. For many genes, the bulk of expressed RNA exists in the cytoplasm as mature mRNA, whereas nascent, intron-containing transcripts are limited to small nuclear puncta at the sites of transcription (Vargas et al. 2011; Coulon et al. 2014). For other genes, unspliced introns may remain after transcript completion but are ultimately excised to allow export (Girard et al. 2012; Popp and Maquat 2013; Stewart 2019). These nuclear transcripts are not necessarily found at their gene loci, but some polyadenylated transcripts, including many noncoding RNAs, are tightly associated with chromatin (Quinn and Chang 2016). Although proteins affecting processes such as DNA template release, RNA export, and nuclear RNA decay have been identified (Schmid and Jensen 2018; Stewart 2019), the global distribution of RNA transcripts between subcellular compartments and the alteration of their maturation and location with development have not been well studied.

In earlier studies, we examined the kinetics of transcription, splicing, and nuclear export for macrophage transcripts induced by inflammatory stimuli (Bhatt et al. 2012; Pandya-Jones et al. 2013). By following inflammatory gene transcripts, we found that partially spliced but polyadenylated transcripts in the chromatin fraction completed splicing over time and were released to the soluble nucleoplasmic fraction before appearing in the cytoplasm as functional mRNAs (Bhatt et al. 2012; Pandya-Jones et al. 2013). These studies focused on introns whose slow splicing impacted the rate of inflammatory gene expression. However, polyadenylated, partially spliced RNA has been long been observed in nuclei, where its interactions and localization are largely unknown.

[6]**These authors contributed equally to this work.**
Corresponding authors: xingyi@email.chop.edu, dougb@microbio.ucla.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.273904.120.

The above analyses used a fractionation procedure to enrich for nucleoplasmic or chromatin-associated RNA (Wuarin and Schibler 1994; Pawlicki and Steitz 2008; Pandya-Jones and Black 2009; Khodor et al. 2012; Herzel and Neugebauer 2015; Yeom and Damianov 2017). Nucleoplasmic and chromatin compartments are operationally defined as the supernatant and pellet fractions, respectively, after nuclear lysis in a stringent buffer containing NP-40, urea, and NaCl. This solubilizes many components such as the U1 snRNP, while leaving other molecules associated with the high-molecular-weight chromatin pellet (Wuarin and Schibler 1994). The cytoplasmic fraction is enriched for mature mRNA, whereas the nucleoplasmic fraction contains recently matured transcripts released from the chromatin that have not yet reached the cytoplasm (Bhatt et al. 2012; Pandya-Jones et al. 2013), as well as some mature mRNAs associated with ER and mitochondria (Yeom and Damianov 2017). The chromatin pellet is enriched for nascent RNA bound by elongating RNA Pol II but also contains substantial polyadenylated RNA, including the *Xist* noncoding RNA tightly bound to chromatin (Pandya-Jones et al. 2020) and the *Malat1* noncoding RNA, which is enriched in nuclear speckles that are adjacent to chromatin but only partially in contact with it (Hutchinson et al. 2007; Fei et al. 2017).

The consequences of intron retention (IR) are diverse and complex to dissect. Splice sites and binding of spliceosomal components can prevent nuclear RNA export (Hautbergue 2017; Stewart 2019; Garland and Jensen 2020). Nevertheless, some intron-containing transcripts are exported to the cytoplasm as alternative mRNA isoforms that either encode an alternative protein or are subject to altered translation and decay (Jacob and Smith 2017; Wegener and Müller-McNicoll 2018). Other introns slow to be excised relative to transcription are ultimately removed and their transcripts exported as fully spliced mRNAs (Ninomiya et al. 2011; Bhatt et al. 2012; Hao and Baltimore 2013; Pandya-Jones et al. 2013; Frankiw et al. 2019a). Such transcripts can create a nuclear pool of partially spliced RNA, which acts as a reservoir to feed the cytoplasmic mRNA pool upon splicing. A group of these introns found in genes affecting growth control and cell division was named "detained introns" (DIs) to distinguish them from classical "retained introns" found in cytoplasmic mRNA (Boutz et al. 2015; Braun et al. 2017). A similar pool of incompletely spliced transcripts affecting synaptic function is found in neurons, where cell stimulation induces their processing to allow transcription-independent changes in mRNA pools (Mauger et al. 2016). The term "retained intron" thus encompasses a wide range of molecular behaviors.

Retained introns are more difficult to characterize than other patterns of alternative splicing in whole-transcriptome RNA-seq data. Overlapping patterns of alternative processing can be miscalled as IR by sequence analysis tools (Wang and Rio 2018; Broseus and Ritchie 2020). Many RNA-seq studies have identified conditions leading to higher levels of unspliced introns across the transcriptome (Wong et al. 2013; Braunschweig et al. 2014; Edwards et al. 2016; Pimentel et al. 2016; Jacob and Smith 2017; Naro et al. 2017; Schmitz et al. 2017; Parra et al. 2018). These studies have not always distinguished between nuclear and cytoplasmic RNA or examined the fate of the partially spliced transcripts, information that is essential to understanding the biological role of these regulatory mechanisms.

Here we undertook a broad examination of how RNAs are distributed between subcellular compartments and how this compartmentalization changes with development. Our goals were to distinguish transcripts in the nucleoplasmic and chromatin-associated RNA pools from cytoplasmic mRNAs and assess how their processing and localization to chromatin tracked with expression of mature cytoplasmic mRNA.

Results

Both coding and noncoding RNAs show defined partitioning between cellular compartments

To broadly categorize RNAs enriched in different cellular locations and to gain insight into how this compartmentalization might be regulated across cell types, we generated deep RNA-seq data from mouse embryonic stem cells (mESCs), a neuronal progenitor cell line derived from embryonic mouse brain (mNPC), and explanted mouse cortical neurons cultured in vitro for 5 d (mCtx) (Fig. 1A). RNA was isolated from three fractions of each cell: cytoplasm, soluble nucleoplasm, and chromatin pellet as previously described (Wuarin and Schibler 1994; Pandya-Jones and Black 2009; Bhatt et al. 2012; Yeom and Damianov 2017). The quality of subcellular fractionation was assessed by immunoblot for GAPDH and tubulin, alpha 1A (TUBA1A) proteins as cytoplasmic markers, SNRNP70 fractionating with the soluble nucleoplasm, and Histone H3.1 as a chromatin marker (Supplemental Fig. S1A; Supplemental Table S1B).

To provide information on the maturation of transcripts in each cell type and location, RNA was isolated as two separate pools. A total RNA pool depleted of ribosomal RNA [total] will include nascent incomplete transcripts. A polyadenylated pool [poly(A)$^+$] includes RNAs whose transcription and 3′ processing are complete. Each RNA pool from each fraction was isolated from three separate cultures of each cell type to yield biological triplicates of each experimental condition. The RNA pools were converted to cDNA libraries, sequenced on the Illumina platform to yield 100-nt paired end reads, and aligned to the genome (Supplemental Table S2). Gene expression markers for each of the three cell types confirmed the expected patterns of ESCs, NPCs, or immature neurons (Supplemental Fig. S1B). Clustering of gene expression values across all the data sets showed the expected segregation by cell type, fraction, and replicate, for both the poly(A)$^+$ and total RNA libraries (Supplemental Fig. S1C). The resulting 54 data sets constitute an extensive resource for examining multiple aspects of RNA maturation and its modulation during development (see Data access [GSE159919 for poly(A)$^+$ RNA and GSE159944 for total RNA]). In addition to the libraries used in this study, we also generated libraries of small RNAs (<200 nt) from all samples. As previously described, these can be used to assess miRNA maturation and other processes (Yeom et al. 2018). These 27 data sets are also available from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) (GSE159971).

Examining read distributions in the different RNA pools and fractions, we found that the housekeeping gene *Gapdh* (Fig. 1B) yields similar patterns of reads from either the poly(A)$^+$ or the total RNA populations, with the RNA being most abundant in the cytoplasm. The total *Gapdh* RNA on chromatin contains intron reads from the nascent transcripts (Fig. 1B, bottom). Although more abundant in the soluble nucleoplasm and especially in the cytoplasm, polyadenylated *Gapdh* transcripts are also found in the chromatin fraction but, in contrast to the total RNA, lack intron reads. We also examined the long noncoding RNA *Xist*, which condenses on the inactive X Chromosome in female cells (Fig. 1C). The mNPCs were isolated from female mice, and *Xist* is seen to partition almost completely to chromatin in these cells. The

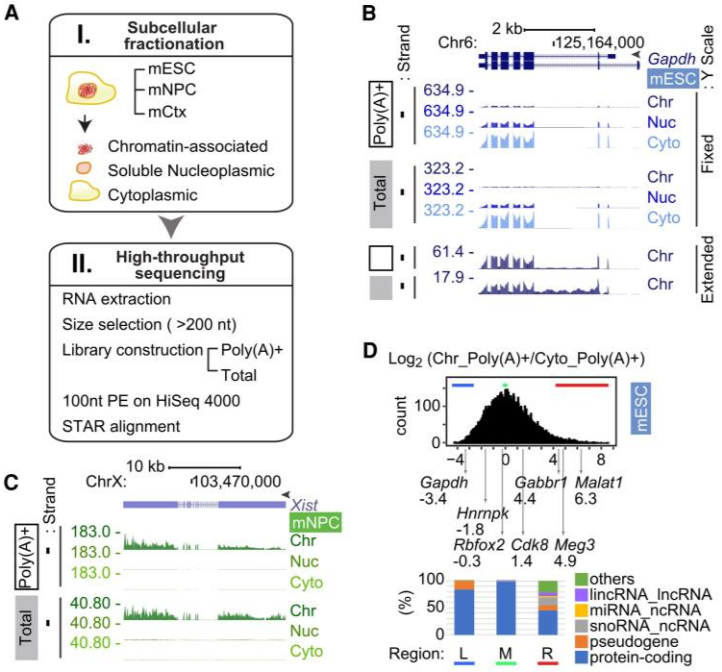


Figure 1. RNA partitioning between subcellular compartments. (*A*) Workflow used in this study. (*B*) Genome browser tracks of the *Gapdh* locus in mESCs. GENCODE annotated isoforms (M11) are diagrammed at the *top*. Poly(A)$^+$ RNA (open box), total RNA (gray box), and peak RPM are noted on the *left*. RNA from chromatin (Chr), nucleoplasmic (Nuc), and cytoplasmic (Cyto) fractions are labeled at the *right*. The fixed Y-scale (RPM) shows the strong enrichment of *Gapdh* RNA in the cytoplasm. The *bottom* tracks show chromatin RNA with an extended Y-scale to observe the intron reads. (*C*) Genome browser tracks of the *Xist/Tsix* locus in female mNPCs show strong chromatin enrichment of *Xist* RNA. (*D*) Distribution of chromatin partition indices. The chromatin/cytoplasm ratio [Chr_Poly(A)$^+$/Cyto_Poly(A)$^+$] of the averaged read counts of each gene are plotted as a distribution along the $\log_2$ scale, with partition indices of representative genes indicated *below*. Biotypes of the 400 genes from bottom (*left* [L]; blue bar), peak (*middle* [M]; green bar), and top (*right* [R]; red bar) of the distribution are presented in the bar graph *below*.

poly(A)$^+$ and the total RNA samples yielded very similar patterns of *Xist* reads, indicating that this RNA is largely spliced and polyadenylated (Brockdorff et al. 1992). Other noncoding RNAs yielded more complex patterns of subcellular partitioning that changed with cell type. The paraspeckle lncRNA *Neat1* is more highly expressed in mESCs than mNPCs or neurons (Supplemental Fig. S2A). The short polyadenylated form (*Neat1_1*) predominates in ESCs and is found mostly with chromatin but also in the nucleoplasm. The longer nonpolyadenylated *Neat1* RNA (*Neat1_2*) is seen in the total RNA samples and is also chromatin enriched. Whether this is a stable long isoform or nascent RNA is not clear. This longer RNA contributes a larger portion of the *Neat1* transcripts in mNPCs and neurons, consistent with observations that *Neat1* cleavage and polyadenylation may be modulated (Naganuma et al. 2012). Overall, we find that gene transcripts can show diverse patterns of enrichment and processing across the different fractions and cell types.

Because the relative transcript numbers and overall library complexity will differ between fractions, reads per million (RPM) values or other read number normalizations of individual genes cannot be directly compared between different subcellular fractions. By using qRT-PCR in mESCs to directly quantify individual transcripts in different fractions, we found that for cytoplasmic enriched transcripts in both the poly(A)$^+$ and the total RNA libraries, RPM values undercounted the RNA abundance in the cytoplasmic fraction relative to the chromatin and nucleoplasm (Supplemental Table S3). On the other hand, for RNAs that are primarily chromatin associated, qRT-PCR quantification yielded cytoplasmic-to-chromatin ratios that were similar to relative RPM numbers (Supplemental Table S3). Although the absolute transcript levels were not quantifiable by RPM, the ratios of these RPM values did reflect their relative enrichment in each fraction across a variety of genes. As an index for how RNAs partition between the chromatin and cytoplasmic pools, we used DESeq2 (Anders and Huber 2010) to measure the fold change in reads for each gene between the chromatin and cytoplasmic poly(A)$^+$ RNA. This returns the ratio of the averaged read counts for each gene between fractions. For genes that had a transcripts per million (TPM) value in chromatin over the median and that had read counts greater than zero in the cytoplasm (13,036 genes), this chromatin partition index was distributed over a 100-fold range centered on one ($\text{Log}_2 = 0$). Thus, a typical gene showed equal normalized read counts in chromatin and cytoplasm (Fig. 1D). By examining the Ensembl annotations (V.91) for genes in the left, middle, and right side of this distribution (400 genes each), we found that genes with predominately cytoplasmic reads as well as genes with roughly equal read numbers in cytoplasm and chromatin were annotated almost entirely as protein-coding genes. For example, on the left edge (Fig. 1D), *Gapdh* RNAs partition much more strongly to the cytoplasm than is typical. In the middle of the distribution, *Rbfox2* RNAs show slightly fewer reads on chromatin than in the cytoplasm, whereas *Cdk8* shows two- to threefold more chromatin reads (Fig. 1D). Thus, although the transcripts from protein-coding genes are usually most abundant in the cytoplasm, a substantial fraction of a gene's RNA product is often nuclear and chromatin associated. By comparing qRT-PCR quantification for select genes to their chromatin partition indices, we found that RNAs from genes showing a partition index above 3.6 were actually more abundant in chromatin than the cytoplasm. This included ~3% of protein-coding genes. At the right edge of the curve, the 400 most chromatin enriched transcripts included the expected noncoding RNAs, such as primiRNAs, snoRNAs, and lncRNAs, but also many protein-coding genes, including *Clcn2*, *Ankrd16*, and *Gpc2* (Supplemental Figs. S2C,D, S6B), and *Gabbr1*, which is analyzed further below. For these protein-coding genes, the majority of the polyadenylated product RNA is chromatin associated, where it is presumably inactive for protein expression (Supplemental Figs. S2C,D, S6B).

Examination of individual genes whose poly(A)$^+$ transcripts remain sequestered with chromatin showed that their splicing

was modulated across cell types. The chromatin-associated *Meg3* noncoding RNA is well expressed in mESCs and neurons but not in mNPCs (Supplemental Fig. S2B). *Meg3* is the host transcript for the miRNAs MiR-770 and MiR-1906-1. Mature MiR-770, processed from the last *Meg3* intron, is weakly expressed in neurons but absent from mESCs (Supplemental Table S4). This intron is absent from the RNA in mESCs, where it is apparently efficiently spliced. In contrast in neurons, this intron is abundant in the chromatin fraction of polyadenylated RNA, where its reduced excision might allow more efficient processing of *mir-770* (Supplemental Fig. S2B). This is consistent with observations that perturbations causing a host transcript to be released from chromatin reduce DROSHA cleavage and miRNA expression (Pawlicki and Steitz 2008; Liu et al. 2016a). The mESC small RNA data were previously used to examine expression of primary *mir-124a-1* in mESCs whose processing is blocked by PTBP1 in the chromatin fraction (Yeom et al. 2018). For *Meg3*, the processing of *mir-770* may be modulated by the excision rate of its host intron. The upstream portion of *Meg3* that includes *mir-1906-1* undergoes complex processing and shows more splicing in neurons than in mESCs. Thus, an additional product from the gene, possibly *mir-1906-1*, may also be differentially regulated between mESCs and neurons. These introns present in the polyadenylated RNA are not more abundant in the total RNA than adjacent exon sequences, indicating an absence of excised intron, which could also give rise to the miRNAs. Overall, the data indicate that splicing of the *Meg3* transcript is regulated on chromatin to allow differential expression of its mature products.

Chromatin-associated transcripts can be spliced either cotranscriptionally or posttranscriptionally

It is expected that most introns will be transient species within the chromatin RNA, with many introns excised before transcript completion, whereas some introns with slow kinetics will be removed later. Various studies estimate that 45%–84% of introns are cotranscriptionally excised in mammals (Ameur et al. 2011; Bhatt et al. 2012; Girard et al. 2012; Khodor et al. 2012; Tilgner et al. 2012; Windhager et al. 2012). Several approaches compare read numbers for spliced (exon–exon [EE]) and unspliced (exon–intron [EI] or intron–exon [IE]) junctions in nascent RNA to those in total RNA to measure cotranscriptional excision (Tilgner et al. 2012; Windhager et al. 2012; Herzel and Neugebauer 2015). To ensure that measurements are of the nascent RNA, this requires removal of polyadenylated RNA from the chromatin fraction and prevents parallel analysis of posttranscriptional events. Other studies identified sawtooth patterns of RNA read abundance in total cellular RNA, where reads peak in exons and then decline to the next exon or recursive splice site. Such a pattern is thought to indicate that the time needed to excise an intron is small relative to the time for RNA synthesis through the next intron downstream (Ameur et al. 2011; Duff et al. 2015; Sibley et al. 2015). Although sawtooth read densities can be observed on certain introns in the total chromatin RNA pools (Supplemental Fig. S3), these patterns were infrequent and lost on introns <50 kb, many of which are expected to be cotranscriptionally excised (Ameur et al. 2011).

As an alternative for defining cotranscriptional and posttranscriptional intron excision, we compared the total RNA from chromatin to the poly(A)$^+$ RNA from the same fraction. Introns remaining in polyadenylated RNA must be excised after transcription or be dead-end products. For example, in the *Sorbs1* gene (Fig. 2A), reads are observed across all the introns in the total RNA from

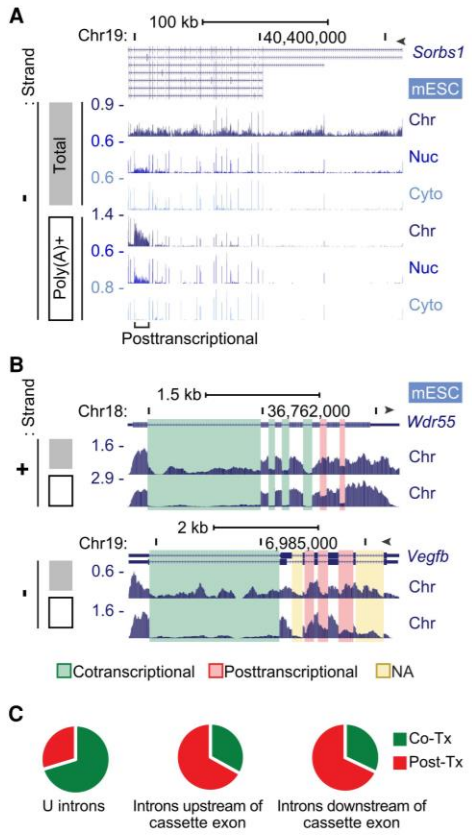


Figure 2. Cotranscriptional and posttranscriptional intron excision. (*A*) Genome browser tracks of the *Sorbs1* locus in mESCs. Total chromatin RNA (gray box) shows intron reads, but the poly(A)$^+$ RNA (open box) shows primarily exon reads except one posttranscriptional intron. (*B*) Genome browser tracks of chromatin RNA at the *Wdr55* and *Vegfb* loci in mESCs. Total (gray box) and poly(A)$^+$ (open box) are shown, with cotranscriptionally and posttranscriptionally spliced introns highlighted in green and red, respectively. Yellow highlighted introns were not analyzable owing to multiple processing patterns. (*C*) Proportions of co- and posttranscriptional splicing for 49,692 U introns in mESCs, using criteria described in Supplemental Figure S4, C through E. Introns upstream of (2779) and downstream from (2744) simple cassette exons were similarly analyzed.

chromatin, indicating the presence of unspliced introns in the nascent transcripts. In the polyadenylated RNA on chromatin, reads are largely absent from introns, indicating that by the time of polyadenylation or shortly after, these introns have been spliced out. However, one intron in *Sorbs*1 shows substantial read numbers in poly(A)$^+$ RNA on chromatin that are reduced in RNA from the nucleoplasm and absent from the cytoplasm (Fig. 2A). This intron is presumably excised after cleavage/polyadenylation. Although most introns are absent from the polyadenylated RNA and are likely spliced cotranscriptionally, there are many transcripts with one or more introns that are highly retained in the polyadenylated chromatin-associated RNA (Fig. 2A,B). The comparison of intron levels in total and poly(A)$^+$ RNA on chromatin provides a simple bioinformatic metric for distinguishing co- versus posttranscriptional excision.

To compare intron levels in the total and poly(A)$^+$ RNA pools, we determined fractional inclusion (FI) values (Supplemental Fig.

S4A) by counting reads across EI, IE, and EE junctions. Assessing IR by FI value can be confounded by alternative splicing, polyadenylation, or transcription initiation events occurring within the intron being measured (Supplemental Fig. S4B; Wang and Rio 2018; Broseus and Ritchie 2020). To avoid errors in IR measurements arising from other processes, we defined a set of introns showing a unique Ensembl v91 annotation without alternative processing events (Supplemental Fig. S4B). This set of 149,333 "unique" introns (U introns) across 28,733 genes was used for subsequent analysis. By focusing on the mESC RNA, we determined the FI values of all U introns in the total RNA and the poly(A)$^+$ RNA for genes above the median expression level as measured by kallisto (Bray et al. 2016). We included only introns excised by the major spliceosome with GU/AG splice junctions. Reads from poly(A)$^+$ RNA containing long unspliced introns can be biased toward the 3′ ends. To avoid undercounting in the poly(A)$^+$ samples, we removed genes in which reads per nucleotide length from the second exon were less than half that of the second to last exon. To filter out introns that were not measurable owing to anomalies in the generation of particular junction reads, we removed introns yielding a FI value below 0.1 in the total RNA, and introns with a zero value for one or more of the junction read counts. In mESCs, these criteria returned 49,629 U introns within 7672 genes for analysis.

Of the 49,629 U introns being measured, 34,939 introns (within 6952 genes) showed low FI values in the poly(A)$^+$ RNA (FI < 0.1) and are presumably spliced before transcript completion. Conversely, 14,753 introns within 5550 genes showed a FI value ≥0.1 in the poly(A)$^+$ RNA. These introns (29.7%) appear to be excised posttranscriptionally, with many highly unspliced in the chromatin poly(A)$^+$ RNA despite being fully spliced in other fractions. By this analysis, at least 70.3% of introns within our analysis set are excised cotranscriptionally, similar to estimates made by other methods (Fig. 2C; Supplemental Fig. S4C–F; Supplemental Table S5). On the other hand, the majority of genes (5550 out of 7672) have at least one posttranscriptionally spliced intron. If the analysis is restricted to the top quartile of expressed genes rather than the top half, the fractions of co- and posttranscriptional splicing change only slightly (70.7% cotranscriptional). The fraction of cotranscriptionally spliced introns is also essentially the same if the analysis is restricted to the first introns in each transcript or to internal introns. For introns that are the last intron transcribed before the polyadenylation site, a slightly higher fraction is classified as posttranscriptional, presumably because they are polyadenylated more rapidly after intron synthesis (Supplemental Fig. S4F). Thus, posttranscriptional splicing does not appear to be associated with higher or lower gene expression or with the position of an intron along the gene. Examples of introns defined as co- or posttranscriptional by these measures are shown in Figure 2B. Although in the minority, posttranscriptionally spliced introns are found across a wide range of genes and often show high FI values in the chromatin fraction, even though the cytoplasmic RNA is completely spliced.

In addition to the U introns analyzed above, we also analyzed a set of introns flanking simple cassette exons that could also be unambiguously measured for FI. By using the same parameters to define co- versus posttranscriptional splicing, we found a reversal in the percentages. Of these introns flanking alternative exons, ~67% show high read numbers (FI > 0.1) in the poly(A)$^+$ RNA and thus appear to be excised posttranscriptionally (Fig. 2C; Supplemental Fig. S4E). This was seen for introns both upstream of and downstream from the cassette exon. These data indicate that the majority of regulated splicing events occur with slower kinetics than the excision of typical constitutive introns.

Retained introns can be classified by their enrichment in the chromatin, nucleoplasmic, and cytoplasmic compartments

A variety of fates are possible for transcripts that retain introns after polyadenylation. Intron-containing transcripts can be sequestered in the nucleus until they are spliced or can undergo nuclear decay. Other intron-containing mRNAs are exported unspliced to the cytoplasm, where they can be translated or undergo nonsense-mediated mRNA decay (NMD). To categorize introns based on both their retention levels and location, FI values for the unique intron set in the polyadenylated RNA of all cells and fractions were subjected to X-means cluster analysis (Fig. 3A; Supplemental Table S6; Pelleg and Moore 2000). Consistently, in all three cell types, the clustering algorithm defined four groups of introns. The largest cluster Group A, containing 49,981 introns in mESCs, was almost entirely spliced in all three fractions. Introns in Group B (7529) showed measurable retention in the poly(A)$^+$ RNA from chromatin but showed nearly complete splicing in the nucleoplasm and cytoplasm (Fig. 3A). Group C introns (1351), including introns in *Zfp598* and *Neil3* (Fig. 3B), showed higher FI values in the chromatin and nucleoplasm than did Group B but were almost completely excised from the cytoplasmic RNA. The smallest cluster of only 247 introns in mESCs, Group D, was almost entirely retained in all three fractions. Each of the other two cell types also generated four clusters with similar splicing levels and similar numbers of introns in each group (Fig. 3A).

Group B and C introns that do not leave the nucleus can be seen to have different properties from Group D introns that also have high retention levels in the cytoplasm. A larger percentage of Group D introns are found in 5′ and 3′ UTR sequences, where they will not disrupt the primary reading frame but will likely affect translation and decay (Supplemental Table S7B). Group D introns were also found to be depleted of in-frame premature termination codons (PTC) compared with Groups A, B, and C (Fig. 3C), presumably owing to selection to prevent NMD in the cytoplasm. These observations indicate that the different intron clusters arise from selection for different functions in the intron-containing RNAs.

We found that among transcripts in which all introns were annotated as unique introns (Supplemental Fig. S4G), RNAs containing at least one Group C intron have a higher average chromatin partition index than transcripts with no Group C intron (Supplemental Fig. S4H). Previous work defined nuclear transcripts in mESCs containing what are called detained introns (DIs), whose splicing is modulated in cancer and growth control pathways (Boutz et al. 2015; Braun et al. 2017). Of 3150 DIs, 1021 were on our U intron list (Supplemental Tables S7A, S7B). Of these, 1000 introns passed the filters for FI measurement and are seen to fall predominantly into Groups B and C, in agreement with the earlier studies (Fig. 3D). However, the 1021 DIs were only a subset of the nearly 9000 retained introns identified in Groups B and C (Supplemental Table S8B). Similar to the DIs affecting growth control, as well as inflammatory and neuronal gene introns also identified previously (Bhatt et al. 2012; Hao and Baltimore 2013; Pandya-Jones et al. 2013; Mauger et al. 2016; Frankiw et al. 2019b), these new retained introns could affect cellular function by altering the movement of material through the gene expression pathway.

Predicting retained introns

To examine whether introns in different groups could be identified by their sequence features alone, we developed a deep learning model for predicting intron behavior. We extracted 1387 sequence features from the first and last 300 nucleotides (nt) of each intron and from the two flanking exons. For introns <300 nt, the intron interval includes some adjacent exon sequence. Analyzed features included short motif frequencies, predicted RBP binding elements, propensity to form local secondary structure, splice site strength scores, conservation scores, and nucleosome positioning scores (Supplemental Table S9A). This feature information was used to train a three-layer deep neural network (DNN) tasked with predicting whether an intron belonged in Group A, B, C, or D (Fig. 4A).

The performance of the model was assessed using receiver operating characteristic (ROC) curves plotting the false- and true-positive rates (Fig. 4B). The model was highly predictive in distinguishing Group D introns from A, yielding an area under the curve (AUC) of 0.94 (AUC = probability that any true positive will rank higher than any true negative). Group D introns could also be distinguished from Group B and C (AUC = 0.9 and 0.84, respectively), whereas Group B and C introns were distinguished from Group A with reduced accuracy (AUC = 0.68 and 0.76, respectively). Thus, the Group D introns are most different from the introns of other groups.

To assess the features of Group C and D introns that distinguish them from each other and from Group A, we isolated the top 15 features predictive of IR or its absence and used a t-distributed stochastic neighbor embedding algorithm (t-SNE) to project them onto two dimensions (Fig. 4C; for top 50 features, see Supplemental Table S9B). As previously observed, high splice site strength scores were predictive of Groups A and C over D, as well as Group A over C (Sakabe and de Souza 2007; Braunschweig et al. 2014). Other features redundant with splice site strength scores were also predictive of Groups A or C, including GTAAG count in the 5′ portion of the intron and the conservation of the splice site sequences. Translatability of the flanking exons and their spliced product was predictive of Groups A and C over D. This may reflect a greater percentage of Group D introns in 5′ and 3′ UTR sequences (Supplemental Table S7B). Conversely, the translatability of the exon–intron–exon unit containing the retained intron was predictive of Group D over Group C, in agreement with the Group D introns being depleted of in-frame termination codons (Fig. 3C) and adding a coding segment to the mRNA. Overall, the data indicate that IR is controlled by many factors each having relatively small effect.

We examined whether particular sequence elements correlated with the intron group assignments, indicative of regulatory protein binding sites. The model did not clearly identify known elements affecting nuclear localization or IR such as constitutive transport elements or decoy exons (Li et al. 2006; Parra et al. 2018). However, the sequence conservation score of the 5′ portion of the intron was predictive of Group D over Groups C or A, and conservation of both ends of the intron was predictive of C over A (Supplemental Table S9B). Particular triplet motif frequencies within introns or their flanking exons were also predictive of intron behavior. For example, CGA triplets in the 3′ portion of the intron were predictive of Group D over C, whereas TTG and GTT triplets in the 5′ intron segment were predictive of Group C over D. The predictive power of intron sequence conservation and of multiple triplets indicate that particular RNA/protein interactions likely determine the retention properties of these groups.

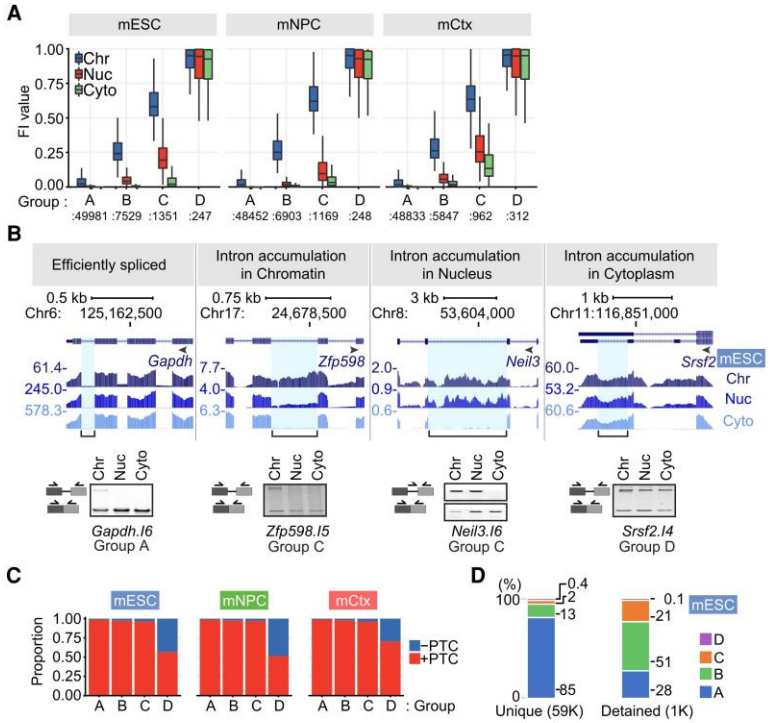


Figure 3. Intron groups defined by their retention level and fractionation behavior. (*A*) X-means clustering was applied to intron FI values and fraction enrichment in mESCs, mNPCs, and mCtx neurons. The FI distribution for introns in each subcellular fraction and group is shown. (*B*) Genome browser tracks (*top*) and RT-PCR validation (*bottom*) of representative transcripts in mESCs. Validated introns are indicated by a blue highlight and a bracket *below*. Gel images are one of three biological replicates. (*C*) The proportion of introns containing a PTC in frame with the upstream sequence is shown for each cluster and cell type. (*D*) Percentage of introns in each group for U introns from mESCs and for detained introns within the U intron set (Boutz et al. 2015).

IR and chromatin association are regulated with neuronal development

Because the X-means analysis yielded four intron clusters in each cell type, these cluster definitions allow bioinformatic analysis of IR regulation between cell types. Although many introns maintain their classification between cell types (Fig. 5A, left), some introns

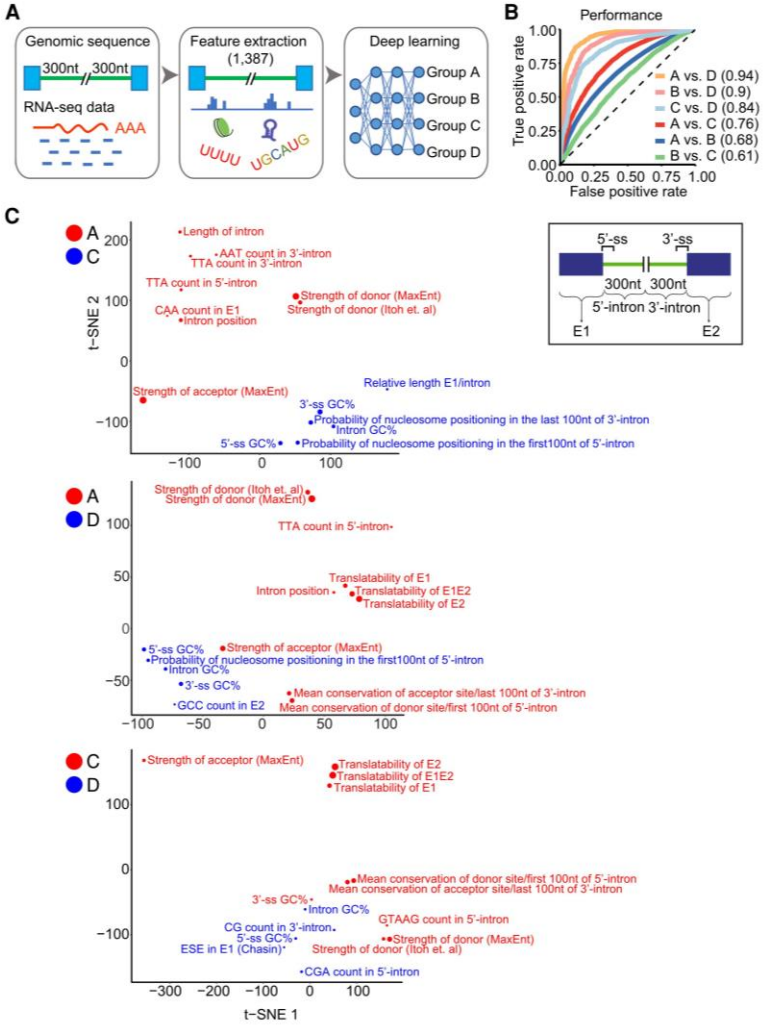


Figure 4. Deep learning analysis of intron groups. (*A*) Flow diagram for training the deep neural network. (*B*) Performance of the model in distinguishing introns of different groups. ROC curves were plotted for individual pairwise comparisons with AUC values shown in parentheses. (*C*) t-SNE plots of the 15 genomic features most predictive for distinguishing intron groups. Features distinguishing Group A from Groups C and D are shown *above* and those distinguishing Group C from Group D *below*. Features colored blue or red indicate the group for which they are positively correlated.

switched their group (Fig. 5A, right). One example is *Med22* (Fig. 5B), which contains a highly retained intron 3 (I3) in all three fractions of mESCs (Group D). This intron became more spliced in mNPCs and was classified as Group C and then became almost fully spliced as a Group B intron in neurons. The nearby intron 1 (I1) was maintained as a Group A intron in all three cell types. *Med22* encodes a subunit of the transcriptional mediator complex. The retention or splicing of *Med22* I3 creates MED22 proteins with different C-terminal peptides that likely alter mediator function in the two cell types. The group-switching introns are presumably part of the extensive alternative splicing programs modulated during neuronal development. By examining their Gene Ontology (GO) functions, we found that the 231 genes containing introns highly spliced in mESCs but unspliced in neurons (switching from Group A or B to Group C or D) were enriched in processes such as ribosome biogenesis, organelle assembly, and metabolism. These functional categories may reflect the different proliferation rates and metabolic status of the two cells. In contrast, 413 genes whose introns were unspliced in mESCs and became more spliced in neurons (switching from Group C or D to Group A or B) were enriched in GO biological processes of glutamatergic synaptic transmission and organelle localization by membrane tethering, in keeping with gene expression and cell morphology changes in the early neuronal state (Supplemental Fig. S5).

The changes in splicing between mESCs, mNPCs, and neurons are driven by changes in the expression of multiple protein regulators. In previous work, we and others characterized alternative splicing programs controlled by the polypyrimidine tract binding proteins PTBP1 and PTBP2 (Keppetipola et al. 2012; Vuong et al. 2016). In ESCs and other cells, PTBP1 maintains alternative splicing patterns characteristic of nonneuronal cells, and PTBP1 down-regulation is a key step in neuronal differentiation. Although the cultured NPCs are not true lineage precursors to the immature cortical neurons used here, the depletion of PTBP1 is common to many neuronal lineages. We previously reported neuronal cassette exons regulated by PTBP1 in ESCs (Linares et al. 2015), and PTBP1-regulated retained introns, including the *Med22* intron, have been described in a neuronal cell line (Yap et al. 2012). We next examined whether additional PTBP1 targets could be identified in the chromatin compartment of mESCs.

To assess PTBP1 regulation, we fractionated cells after *Ptbp1* knockdown and measured the splicing of polyadenylated RNA in the different compartments by RNA-seq. This confirmed the PTBP1 dependence of *Med22* I3, which shifted from Group D to Group C with *Ptbp1* depletion (Fig. 5B, right). By examining all the retained introns, we found that many more splicing changes could be observed in the chromatin-associated RNA than in the nucleoplasmic and cytoplasmic fractions (Fig. 5C). As shown previously with cassette exons, these PTBP1-dependent introns in ESCs also change with neuronal differentiation as PTBP1 levels drop (Fig. 5C). These include introns identified previously (Yap et al. 2012) as well as new introns. Other introns whose splicing changes with neuronal development but are not sensitive to PTBP1 are presumably regulated by other factors.

By examining the chromatin-associated RNA, our analysis identified substantially more PTBP1-regulated introns than

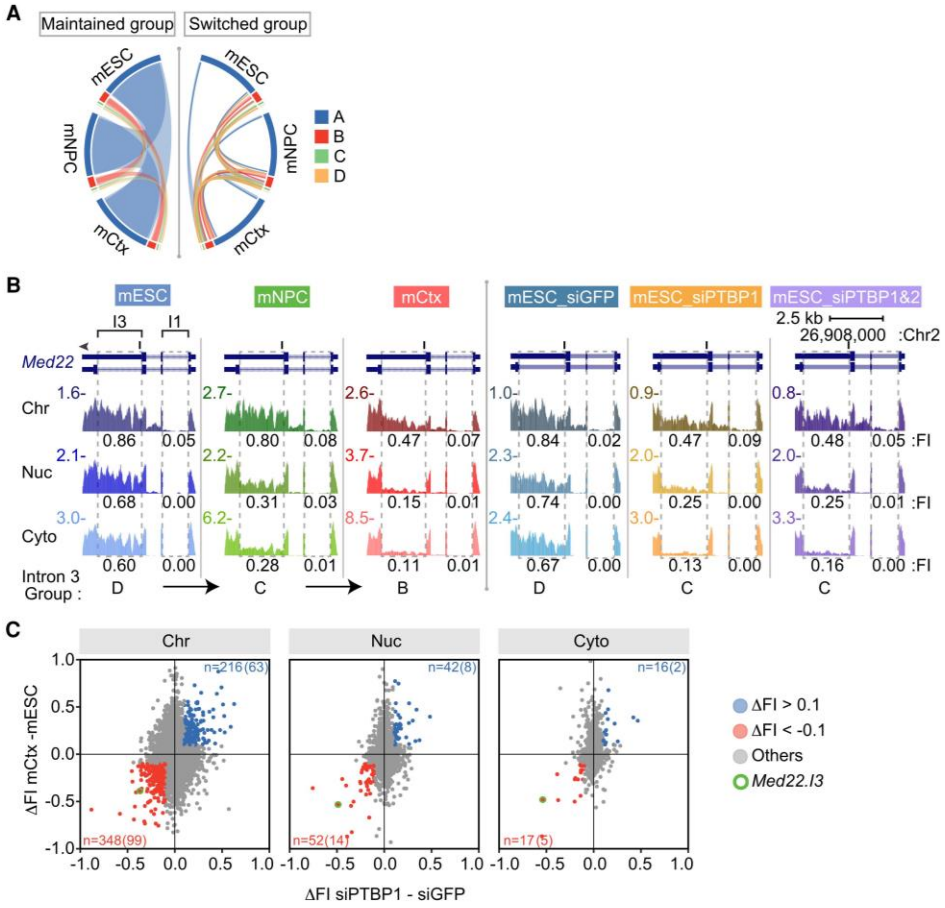


Figure 5. Regulation of intron retention and chromatin association during neuronal development. (*A*) Circos plot (Krzywinski et al. 2009; Gu et al. 2014) of intron group changes between cell types (mESCs, mNPCs, and mCtx neurons). Introns not changing groups are on the *left*. Introns switching groups between cell types on the *right*. (*B*) Genome browser tracks of *Med22* during neuronal differentiation (*left* three panels) and after *Ptbp* knockdown in mESCs (*right* three panels). Dashed boxes indicate U introns with measured FI values (introns 1 and 3) *under* each track. Group classification of intron 3 is at the *bottom*. (*C*) Scatter plots of FI change between mESCs and neurons (mCtx) plotted for each fraction against FI change after *Ptbp1* knockdown in mESCs. Introns with Δ FI < −0.1 in both conditions are in red and with Δ FI > 0.1 in blue. The number of introns showing these changes with the number carrying PTBP1 iCLIP tags in parentheses, is *above* and *below* (Linares et al. 2015). Intron 3 of *Med22* is circled in green.

previously recognized. The transcripts containing these introns may remain in the nucleus, similar to DIs, or may be exported to the cytoplasm and then lost to NMD. To assess this, we used data from a study of unfractionated polyadenylated RNA after *Upf1* knockdown that globally identified NMD targets in mESCs (Hurt et al. 2013). A majority of Group A, B, and C introns is predicted to induce NMD if their parent transcripts were exported to the cytoplasm (Fig. 3C). However, we find that of 871 genes containing PTBP1-dependent retained introns in the chromatin fraction, only 87 showed >10% transcript up-regulation after *Upf1* depletion (Supplemental Table S8C). Thus, the majority of the PTBP1-dependent retained intron transcripts likely stay in the nucleus and will be eliminated by nuclear RNA decay pathways.

By looking more broadly at whether NMD might create the apparent nuclear enrichment of some transcripts, we found that protein-coding genes with high chromatin partition indices were actually less likely to show increases after *Upf1* depletion than other genes across the distribution (Supplemental Table S8D). For the genes in the L, M, and R regions in Figure 1D, NMD targets constituted 4.2%, 7.2%, and 1.1%, respectively. Rather than NMD causing the observed nuclear enrichment by depleting the cytoplasmic RNA, the nuclear enrichment may buffer the effect of NMD on the level of total RNA. It would be interesting to assess this by examining the effect of *Upf1* knockdown specifically on the levels of cytoplasmic mRNA.

Posttranscriptional repression of *Gabbr1* expression

For the most part, transcripts enriched in the chromatin fraction of mESCs (Fig. 1D) were only mildly or unaffected by *Upf1* depletion (Supplemental Table S8D). Rather than cytoplasmic degradation, other processes prevent mRNA expression from these genes. A notable example is *Gabbr1*, which encodes GABBR1, an inhibitory neurotransmitter receptor whose cytoplasmic mRNAs are highly expressed in neurons, moderately expressed in mNPCs, but nearly absent in mESCs (Fig. 6A). By immunoblot, GABBR1 protein is only

observed in neurons (Fig. 6C). In the chromatin fraction of mESCs, the *Gabbr1* precursor RNA is present at high levels that nearly match those seen in mNPCs and neurons (Fig. 6A). This *Gabbr1* RNA is polyadenylated and most introns are excised, but introns 4 and 5, which show a complex pattern of alternative processing in neurons, are largely unprocessed in mESCs (Fig. 6A). *Gabbr1* mRNA expression is apparently blocked by a combined process of splicing inhibition and sequestration on chromatin. Upon differentiation into neurons, the chromatin partition index of *Gabbr1* RNA shifts from 4.43 to −0.69, as the RNA becomes fully processed and released from chromatin to appear in the cytoplasm as mature mRNA (Fig. 6A). Other protein-coding transcripts, including *Gpc2*, were found to behave similarly to *Gabbr1* with RNA abundant in mESC chromatin but low in cytoplasm. In neurons, this pattern was reversed with the *Gpc2* partition index shifting from 4.60 in mESCs to 1.09 in neurons (Supplemental Fig. S6B).

PTBP1 was previously found to regulate *Gabbr1* exon 15 in a neuronal cell line (Makeyev et al. 2007). To assess introns 4 and 5, we examined iCLIP maps of PTBP1 binding in mESCs (Linares et al. 2015), which showed prominent PTBP1 binding peaks in the intron 4–5 region, as well as confirming PTBP1 binding upstream of exon 15 and to the 3′ UTR (Fig. 6A,B). By examining the fractionated RNA-seq data, we found that *Ptbp1* knockdown led to processing of the *Gabbr1* RNA into the neuronal isoforms, including activation of exon 15 and activation of the exon 5 microexon encoding a 6-amino-acid linker of *Gabbr1a* (Fig. 6A, B). Some processed *Gabbr1* mRNA was present in the cytoplasm after *Ptbp1* knockdown, but more of this spliced RNA was in the soluble nuclear fraction. Even after *Ptbp1* depletion, a majority of the *Gabbr1* RNA was still in the chromatin fraction and still unprocessed in the intron 4–5 region, despite exon 15 being strongly activated for splicing in this fraction (Supplemental Fig. S6C). GABBR1 protein was also not observed in mESCs after *Ptbp1* knockdown (Supplemental Fig. S6D). Thus, although PTBP1 strongly affected the processing of *Gabbr1*, its depletion did not yield the predominantly cytoplasmic RNA seen in neurons. There must be additional factors preventing release of the RNA from chromatin in mESCs. *Gabbr1* is highly transcribed in mESCs, but its mRNA expression is blocked by a combination of splicing repression, NMD of transcripts that enter the cytoplasm, and sequestration of the unprocessed RNA on chromatin, with the latter mechanism having the largest effect.

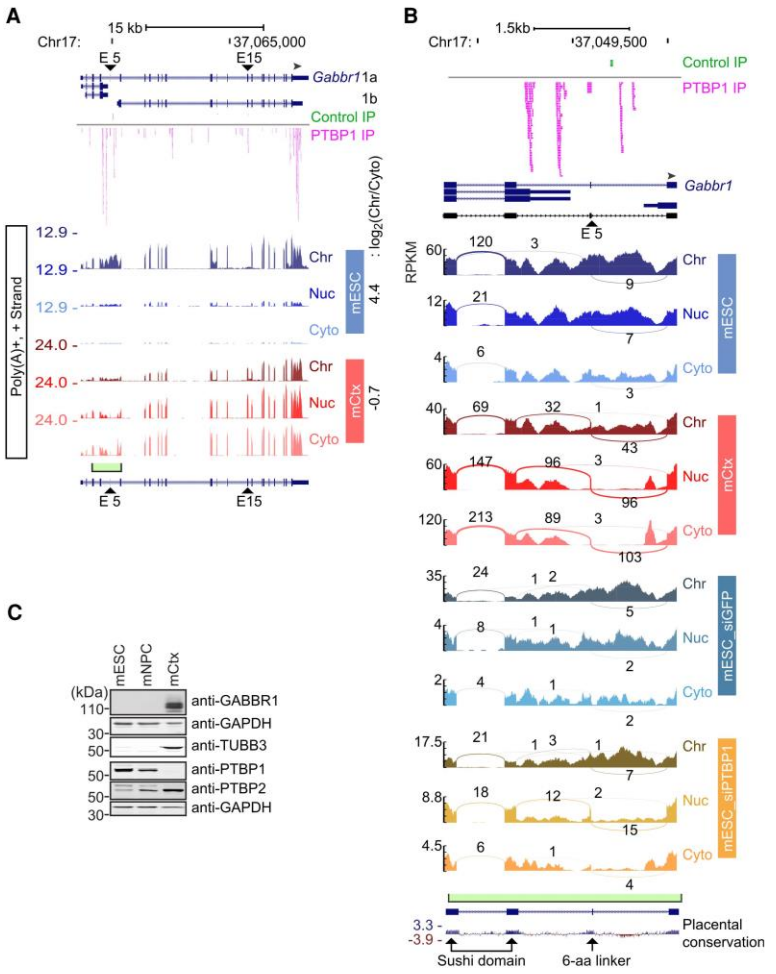


Figure 6. Chromatin enrichment and PTBP1 regulation of *Gabbr1* transcripts. (*A*) Genome browser tracks of *Gabbr1* in mESCs and mCtx neurons. PTBP1 iCLIP tags in mESCs are plotted *above* in pink. The *y*-axis indicates the maximum RPM in each cell type. The green box and bracket mark the intron 4–5 region expanded in panel *B*. PTBP1-responsive exons 5 and 15 are marked with arrowheads. (*B*) Sashimi plots of the *Gabbr1* intron 4–5 region in mESCs, in mCtx neurons, and after *Ptbp* knockdown in mESCs. RPKM is plotted on the *y*-axis. PTBP1-responsive exon 5 is marked with an arrowhead. Exons encoding the two sushi domains and the 6-aa linker are marked on the conservation track *below*. (*C*) Immunoblot showing expression of GABBR1 protein relative to other proteins in mESCs and cortical neurons.

Discussion

A resource for the analysis of RNA-level gene regulation

We developed extensive data sets to examine RNA maturation events across cellular location and developmental state. By applying these data to analyze IR, we compare total and polyadenylated RNA across subcellular fractions and cell types to define classes of introns showing different regulatory behaviors, and we uncover a novel form of gene regulation acting on chromatin-associated RNA. We find that a substantial fraction of the polyadenylated RNA product of some genes is incompletely spliced and still associated with chromatin. This points to a limitation for whole-transcriptome measurements of gene expression that assess total cellular

polyadenylated RNA. The RNA being measured in these studies is not all cytoplasmic mRNA. The presence of nuclear polyadenylated RNA may thus contribute to the observed lack of correlation between RNA and protein levels in global gene expression measurements (Edfors et al. 2016; Liu et al. 2016b).

The isolation of chromatin-associated RNA has frequently been used to enrich for nascent pre-mRNAs and other short-lived species (Pandya-Jones and Black 2009; Davidson et al. 2012; Herzel et al. 2017). We find that many introns are only observed in the total RNA of this fraction, whereas others are also present in the polyadenylated RNA. By quantifying this difference, we estimate that 70% of introns within our analysis set are spliced before the RNA has been completely transcribed. Although this roughly agrees with other studies, we believe it is a lower-bound estimate in our system because the criteria for counting cotranscriptionally excised introns required a measurable presence of the intron in the total RNA. In contrast, we find that introns flanking alternatively spliced cassette exons are mostly spliced posttranscriptionally, showing significant IR levels in the polyadenylated RNA. These introns may be spliced more slowly than typical constitutive introns because of the complex regulatory RNP structures that must assemble onto the sequences flanking alternative exons. By creating a pool of unspliced RNA for these genes, the delayed splicing may allow additional controls over the isoform choice. It will be interesting to examine whether the subset of exons whose inclusion is affected by transcription elongation rates and perturbations of RNA Pol II is among the 30% that appear to be cotranscriptionally excised (Herzel and Neugebauer 2015; Naftelberg et al. 2015; Saldi et al. 2016).

Our data provide a rich resource for examining other questions of RNA metabolism and its regulation over development. Besides introns, transient species one could observe in chromatin-associated RNA include upstream antisense RNAs and extended transcripts downstream from polyadenylation sites (Seila et al. 2008; Flynn et al. 2011; Vilborg and Steitz 2017). These data could also allow more sensitive detection of recursive or back-splicing and could inform studies of regulated RNA export. We have also examined regulated miRNA processing using parallel data from short RNA libraries (GSE159971) (Supplemental Table S4; Yeom et al. 2018).

Behaviors of retained introns

To characterize incompletely spliced transcripts, we assessed introns based on their retention levels across fractions and cell types. Unsupervised X-means clustering yielded four intron groups in each cell type. The largest cluster (Group A) were completely spliced in the poly(A)$^+$ RNA, including in the chromatin fraction, and are presumably excised before transcription termination. The smallest cluster (Group D) behaved like classical retained introns in being exported to the cytoplasm within the otherwise fully spliced mRNA. Two intermediate clusters of introns (Groups B and C) were fully spliced in the cytoplasm while showing different levels of retention on chromatin and, to some extent, the nucleoplasm. A DNN trained using a well-defined set of introns and a wide range of genomic features was able to distinguish introns in Group D from those in A or C with high accuracy. Group C introns were also distinguished from Group A with moderate accuracy (Fig. 4B). These data indicate that Groups D and C are functionally distinct and that the features that define them should give clues to their regulation. These features include those previously associated with retained introns, such as weak splice sites, conservation, and coding capacity (Sakabe and de Souza 2007; Jaillon et al. 2008; Braunschweig et al. 2014; Dvinge and Bradley 2015; Mauger et al. 2016; Parra et al. 2018). We found that introns of the different groups were defined by enrichment of particular short sequence motifs in their terminal regions and adjacent exons. We have not yet identified proteins whose binding sites might underlie the enrichment of these motifs. This may be because the recognition elements assigned to individual proteins are not sufficiently specific. Introns also may be regulated by so many different proteins that no single binding motif is strongly predictive. Proteins including PTBP1 and others are known to regulate particular retained introns (Yap et al. 2012; Horan et al. 2015; Pendleton et al. 2017; Frankiw et al. 2019b), but there may be many such factors, each regulating a subset of introns in a group. The extension of our approach to larger data sets will allow correlation of changes in intron group assignment with the expression of particular RNA-binding proteins.

Groups B and C include several previously described sets of interesting retained introns. DIs were defined as partially spliced introns in transcripts affecting growth control, whose excision can be modulated by cellular stimuli (Ninomiya et al. 2011; Boutz et al. 2015; Braun et al. 2017). These DIs are a subset of the Group B and, particularly, Group C introns we defined in mESCs. Another group of retained introns were shown to be regulated by PTB proteins in a neuronal cell line (Yap et al. 2012). Our analytical strategy identified many new PTBP1-dependent introns that remain as chromatin-associated transcripts in mESCs. In the total cellular polyadenylated RNA of mature primary neuronal cultures (Mauger et al. 2016), retained introns were characterized as transient or stable according to their splicing after transcription inhibition. In our data from less mature neurons, we found that the largest portion of transient introns were in Group C (40%). In contrast, of the stable introns that we could assay in our cultures, ~40% were in Group D (Supplemental Table S8E), consistent with the stable introns remaining in cytoplasmic mRNA after transcriptional shutoff. Mauger et al. (2016) found that similar to DIs, synaptic activation could change the splicing level of some retained introns. It will be interesting to examine whether these introns are associated with chromatin, but this will require improved isolation of nuclei from mature neuronal cultures.

Developmental regulation by splicing inhibition and chromatin sequestration

In previous studies, we showed how the neuronal-specific expression of certain genes is determined by the coupling of a PTBP1-dependent splicing event to NMD. RNAs for the neuronal PTBP2 and DLG4 (also known as PSD-95) proteins are expressed in ESCs and other nonneuronal cells, but through the action of PTBP1 are spliced as isoforms that are subject to NMD (Boutz et al. 2007; Makeyev et al. 2007; Spellman et al. 2007; Zheng et al. 2012; Linares et al. 2015). A similar mechanism affects *Gabbr1* through regulation of exon 15 by PTBP1 (Makeyev et al. 2007), but the change in RNA with loss of NMD is small (Hurt et al. 2013). Most protein-coding transcripts showing chromatin enrichment were not seen to be up-regulated by *Upf1* depletion, whereas some were modestly affected similar to *Gabbr1*. The nuclear pools of these RNAs may reduce the observed efficiency of NMD on total RNA levels, where transcripts show only partial depletion by the decay pathway even though near complete loss of protein is observed. Here we uncover another mechanism controlling the developmental-specific expression of a neuronal protein. The

Gabbr1 RNA is abundant in mESCs, but its splicing is incomplete, and its transcript remains in the chromatin compartment.

Gabbr1 is expressed as multiple isoforms (Kaupmann et al. 1997). The long *Gabbr1a* isoform comes from a promoter active in all three cell types studied here. *Gabbr1b*, which lacks N-terminal sushi domains, arises from an alternative promoter within intron 5 active in neurons (Vigot et al. 2006). There is also a short transcript derived from an alternative polyadenylation site in intron 4. A microexon 5 between these two introns adds a linker into the 1a isoform (Vigot et al. 2006). This complex intron 4–5 region is largely unprocessed in mESCs and becomes processed in neurons with the production of cytoplasmic mRNA including exon 5. The depletion of *Ptbp1* from mESCs leads to multiple changes in *Gabbr1* splicing, including activation of microexon 5 and downstream exon 15. This leads to some expression of neuronal mRNA isoforms but very limited protein expression. Much of the RNA remains nuclear, indicating that additional factors prevent its mobilization. Instead of regulation at the level of transcription or mRNA stability, incomplete *Gabbr1* splicing and sequestration of its RNA on chromatin are modulated to control gene output over development.

The *Gabbr1* transcript is extensively bound by PTBP1. Studies have shown that when binding RNA at high stoichiometry, PTBP1 can cause the condensation of RNA/protein liquid droplets in vitro (Lin et al. 2015). Extensive PTBP1 binding to the long noncoding RNA *Xist* is required for *Xist* condensation onto the X Chromosome during X inactivation (Pandya-Jones et al. 2020). PTBP1 also drives the condensation of the long noncoding RNA *PNCTR* in the perinucleolar compartment, and a similar mechanism may be involved in its interaction with LINE RNAs (Attig et al. 2018; Yap et al. 2018). It will be interesting to examine whether PTBP1 might create a nuclear condensate of *Gabbr1* RNA. Although *Ptbp1* knockdown led to increased splicing and increased mRNA in the nucleoplasm and cytoplasm, it did not eliminate the enrichment of the unspliced RNA in the chromatin. This may be because of the partial depletion of *Ptbp1* by RNAi, but it seems likely that other proteins will also contribute to the sequestration of *Gabbr1* RNA, as is seen with *Xist*. If the chromatin enrichment of protein-coding transcripts like *Gabbr1* involve similar mechanisms to those controlling lncRNA function, they may also have similar effects on chromatin condensation and gene expression.

Methods

Subcellular fractionation, RNA isolation, and library construction

Total RNA was isolated from mESCs, mNPCs, and mCtx neurons that were fractionated into cytoplasmic, soluble nuclear, and chromatin pellet compartments as described previously (Pandya-Jones and Black 2009; Wuarin and Schibler 1994; Yeom and Damianov 2017; Yeom et al. 2018). After checking RNA quantity and integrity, RNAs >200 nt (long RNA) and <200 nt (short RNA) were separated using RNeasy MinElute cleanup kit (Qiagen). Long RNAs were used for total and poly(A)$^+$ libraries, and short RNAs were used for small RNA library construction. See also the Supplemental Material.

Calculation of chromatin partition indices and biotype analysis

To analyze differential compartmentalization of RNAs, genes were selected that had chromatin expression greater or equal to the median TPM reported by kallisto (2.13 TPM) and had read counts greater than zero in the cytoplasmic fractions as measured by FeatureCount. This returned 13,036 genes for analysis. DESeq2 was used to measure fold change in read counts between the chromatin-associated and the cytoplasmic poly(A)$^+$ RNA by calculating the average read count among replicates of the chromatin fraction divided by the average read counts of the cytoplasmic fraction. The chromatin partition index was defined as the $\log_2$ of this ratio (Fig. 1D).

Biotypes were retrieved from Ensembl annotation (V.91). Of the 13,036 genes, 400 genes (3.1%) were analyzed in each of three ranges of the distribution. Partition indices were from −4.2 to −2.6 for region L, −0.1 to 0.1 for region M, and 4.1 to 8.6 for region R.

Measurement of IR

We developed systematic investigation of retained introns (SIRI), a tool to stringently quantify unspliced introns by deep sequencing (https://github.com/Xinglab/siri). In this tool, we first retrieved all introns from Ensembl gene transfer format (GTF) version 91 for the mouse mm10 genome (Hunt et al. 2018). The numbers of reads mapping to each EE, EI, and IE junctions were counted to determine the FI value of each intron. We selected only introns with a unique intron annotation (U introns) that are not involved in other alternative processing events (Supplemental Fig. S4B). Introns subjected to FI measurement were also required have an intron length ≥60 and have a sum of EE + EI + IE reads be ≥20 (Supplemental Table S6). From this set, IR events with EE reads no fewer than two in at least one cell compartment in one cell type were then kept for downstream analysis.

X-means clustering of IR events

X-means clustering was performed using the PyClustering tool (Novikov 2019) applied to the FI values determined in all three compartments of each cell type (Fig. 3A), with the maximum number of clusters set at six. The distance matrix for X-means clustering is based on the dynamic time warping (DTW) algorithm (Berndt and Clifford 1994) for the purpose of investigating directional changes of FI values from chromatin to nucleoplasm to cytoplasm. The Circos plot (Krzywinski et al. 2009) showing the intron group changes from one cell type to another cell type was produced using R (R Core Team 2020) package circlize (version 0.4.4) (Fig. 5A; Gu et al. 2014).

Predicting IR patterns by deep learning

To apply deep learning to IR group prediction, we constructed a compendium of 1387 intron features of five types: sequence motifs, transcript features, RNA secondary structure, nucleosome positioning, and conservation (Supplemental Table S9). Sequence motif features included splice site consensus sequences, position-specific matrices of RNA-binding proteins, and dinucleotide and trinucleotide frequencies of introns and flanking exons. Transcript features included the lengths of upstream exon (E1), downstream exon (E2), and intron (I) and intron number in the host gene. The translatability of E1, E2, E1 + E2, I and E1 + I + E2 were defined by confirming the absence of a stop codon in one of the three reading frames. To predict RNA secondary structure, RNA sequences from the regions from −20 to +20 nt relative to each splice site were examined. Sequence intervals from 1–70 nt, 70–140 nt, 140–210 nt from the 5′ portion of the intron and from −210 to −140 nt, −140 to −70 nt, and −70 to −1 nt from the 3′ portion of the intron were also examined. We computed the free energy of folding for each region with RNAfold (2.2.10) (Kerpedjiev et al. 2015) and used the free energy of unfolding for each region as features for the deep learning. The nucleosome

positioning was predicted by NuPoP (version 1.0, set to the mouse model) (Xi et al. 2010) on the last 50 nt of the upstream exon, the first 100 nt of 5′ intron region, the last 100 nt of 3′ intron region, and the first 50 nt of downstream exon. The training data set included introns that had grouping information in at least two cell types and excluded U11/U12 introns and other introns lacking GT or AG splice sites. We trained a DNN (LeCun et al. 2015) with these 1387 features to predict whether introns belong to Group A, B, C, and D for each cell type (Fig. 3A). The training was performed with fivefold cross-validation with area under the ROC curves on data held-out during training reported for performance evaluation (Pounraja et al. 2019). To evaluate the strengths of individual features, we assessed the decrease of AUC on held-out data when the values of each feature were substituted by its median.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE159944 for total RNA, GSE159919 for poly(A)$^+$ RNA, GSE159971 for small RNA, and GSE159993 for poly(A)$^+$ RNA in *Ptbp* knockdown experiments in Figures 5 and 6. Links to the data displayed on the UCSC Genome Browser are as follows: https://genome.ucsc.edu/s/Chiaho/Kay_fraction_total_hub_10202020 for total RNA and https://genome.ucsc.edu/s/Chiaho/Kay_fraction_polyA%2B_hub_10202020 for poly(A)$^+$ RNA. The source code of data analysis is available at GitHub (https://github.com/Xinglab/intron-retentionpaper), as well as in Supplemental Code files. The data resources used to reproduce the analysis are available at Zenodo (https://zenodo.org/record/4540589#.YJVgEC1h2v4).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Grigori Enikolopov (Cold Spring Harbor Laboratory) for the Nestin-GFP mouse line and Celine Vuong, Amy Pandya-Jones, Kathrin Plath, and members of the Black laboratory for help, discussions, and comments on the manuscript. Financial support was provided by W.M. Keck Foundation and National Institutes of Health (NIH) R01 MH109166 grants to D.L.B. and Y.X., NIH R01 GM088342 to Y.X., and R35 GM136426, R01 GM049662, and funding from the David Geffen School of Medicine and Division of Life Sciences at UCLA to D.L.B. K.-H.Y. received a postdoctoral fellowship from the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA. H.Y.L. was supported by Whitcome Warsaw Family fellowships from UCLA.

Author contributions: K-H.Y., Z.P., Y.X., and D.L.B. conceived and designed the study. K-H.Y. performed all the experiments and generated the RNA-seq data, except the *Ptbp* knockdown experiments and microscopy, which were performed by H.Y.L. and W.X. Z.P., C-H.L., and K-H.Y. processed and analyzed the sequence data. Z.P. defined the unique intron set and carried out the clustering and the deep learning analyses. K-H.Y. and D.L.B. wrote the manuscript with input from Z.P. and Y.X. D.L.B. and Y.X. supervised the project and provided funding.

References

Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L. 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18:** 1435–1440. doi:10.1038/nsmb.2143

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106. doi:10.1186/gb-2010-11-10-r106

Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, Zagalak JA, Emmett W, Smith CWJ, Luscombe NM, et al. 2018. Heteromeric RNP assembly at LINEs controls lineage-specific RNA processing. *Cell* **174:** 1067–1081.e17. doi:10.1016/j.cell.2018.07.001

Berndt DJ, Clifford J. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAIWS'94*, pp. 359–370. AAAI Press, Seattle.

Bhatt DM, Pandya-Jones A, Tong A-J, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST. 2012. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150:** 279–290. doi:10.1016/j.cell.2012.05.043

Boutz PL, Stoilov P, Li Q, Lin C-H, Chawla G, Ostrow K, Shiue L, Ares M, Black DL. 2007. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev* **21:** 1636–1652. doi:10.1101/gad.1558107

Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev* **29:** 63–80. doi:10.1101/gad.247361.114

Braun CJ, Stanciu M, Boutz PL, Patterson JC, Calligaris D, Higuchi F, Neupane R, Fenoglio S, Cahill DP, Wakimoto H, et al. 2017. Coordinated splicing of regulatory detained introns within oncogenic transcripts creates an exploitable vulnerability in malignant glioma. *Cancer Cell* **32:** 411–426.e11. doi:10.1016/j.ccell.2017.08.018

Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* **24:** 1774–1786. doi:10.1101/gr.177790.114

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34:** 525–527. doi:10.1038/nbt.3519

Brockdorff N, McCabe M, Norris P, Cooper J, Swift S, Kay F. 1992. The product of the mouse *Xist* gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71:** 515–526. doi:10.1016/0092-8674(92)90519-i

Broseus L, Ritchie W. 2020. Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput Struct Biotechnol J* **18:** 501–508. doi:10.1016/j.csbj.2020.02.010

Coulon A, Ferguson ML, de Turris V, Palangat M, Chow CC, Larson DR. 2014. Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife* **3:** e03939. doi:10.7554/eLife.03939

Davidson L, Kerr A, West S. 2012. Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. *EMBO J* **31:** 2566–2578. doi:10.1038/emboj.2012.101

Duff MO, Olson S, Wei X, Garrett SC, Osman A, Bolisetty M, Plocik A, Celniker SE, Graveley BR. 2015. Genome-wide identification of zero nucleotide recursive splicing in *Drosophila*. *Nature* **521:** 376–379. doi:10.1038/nature14475

Dvinge H, Bradley RK. 2015. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med* **7:** 45. doi:10.1186/s13073-015-0168-9

Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, Forsström B, Uhlén M. 2016. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* **12:** 883. doi:10.15252/msb.20167144

Edwards CR, Ritchie W, Wong JJ-L, Schmitz U, Middleton R, An X, Mohandas N, Rasko JEJ, Blobel GA. 2016. A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood* **127:** e24–e34. doi:10.1182/blood-2016-01-692764

Fei J, Jadaliha M, Harmon TS, Li ITS, Hua B, Hao Q, Holehouse AS, Reyer M, Sun Q, Freier SM, et al. 2017. Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J Cell Sci* **130:** 4180–4192. doi:10.1242/jcs.206854

Flynn RA, Almada AE, Zamudio JR, Sharp PA. 2011. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc Natl Acad Sci* **108:** 10460–10465. doi:10.1073/pnas.1106630108

Frankiw L, Baltimore D, Li G. 2019a. Alternative mRNA splicing in cancer immunotherapy. *Nat Rev Immunol* **19:** 675–687. doi:10.1038/s41577-019-0195-7

Frankiw L, Majumdar D, Burns C, Vlach L, Moradian A, Sweredoski MJ, Baltimore D. 2019b. BUD13 promotes a type I interferon response by countering intron retention in *Irf7*. *Mol Cell* **73:** 803–814.e6. doi:10.1016/j.molcel.2018.11.038

Garland W, Jensen TH. 2020. Nuclear sorting of RNA. *WIREs RNA* **11:** e1572. doi:10.1002/wrna.1572

Girard C, Will CL, Peng J, Makarov EM, Kastner B, Lemm I, Urlaub H, Hartmuth K, Lührmann R. 2012. Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat Commun* **3:** 994. doi:10.1038/ncomms1998

Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. *circlize* implements and enhances circular visualization in R. *Bioinformatics* **30:** 2811–2812. doi:10.1093/bioinformatics/btu393

Hao S, Baltimore D. 2013. RNA splicing regulates the temporal order of TNF-induced gene expression. *Proc Natl Acad Sci* **110:** 11934–11939. doi:10.1073/pnas.1309990110

Hautbergue GM. 2017. RNA nuclear export: from neurological disorders to cancer. *Adv Exp Med Biol* **1007:** 89–109. doi:10.1007/978-3-319-60733-7_6

Herzel L, Neugebauer KM. 2015. Quantification of co-transcriptional splicing from RNA-Seq data. *Methods* **85:** 36–43. doi:10.1016/j.ymeth.2015.04.024

Herzel L, Ottoz DSM, Alpert T, Neugebauer KM. 2017. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat Rev Mol Cell Biol* **18:** 637–650. doi:10.1038/nrm.2017.63

Horan L, Yasuhara JC, Kohlstaedt LA, Rio DC. 2015. Biochemical identification of new proteins involved in splicing repression at the *Drosophila* P-element exonic splicing silencer. *Genes Dev* **29:** 2298–2311. doi:10.1101/gad.268847.115

Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, et al. 2018. Ensembl variation resources. *Database (Oxford)* **2018:** bay119. doi:10.1093/database/bay119

Hurt JA, Robertson AD, Burge CB. 2013. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res* **23:** 1636–1650. doi:10.1101/gr.157354.113

Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, Chess A. 2007. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8:** 39. doi:10.1186/1471-2164-8-39

Jacob AG, Smith CWJ. 2017. Intron retention as a component of regulated gene expression programs. *Hum Genet* **136:** 1043–1057. doi:10.1007/s00439-017-1791-x

Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Ségurens B, et al. 2008. Translational control of intron splicing in eukaryotes. *Nature* **451:** 359–362. doi:10.1038/nature06495

Kaupmann K, Huggel K, Heid J, Flor PJ, Bischoff S, Mickel SJ, McMaster G, Angst C, Bittiger H, Froestl W, et al. 1997. Expression cloning of $GABA_B$ receptors uncovers similarity to metabotropic glutamate receptors. *Nature* **386:** 239–246. doi:10.1038/386239a0

Keppetipola N, Sharma S, Li Q, Black DL. 2012. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. *Crit Rev Biochem Mol Biol* **47:** 360–378. doi:10.3109/10409238.2012.691456

Kerpedjiev P, Hammer S, Hofacker IL. 2015. Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31:** 3377–3379. doi:10.1093/bioinformatics/btv372

Khodor YL, Menet JS, Tolan M, Rosbash M. 2012. Cotranscriptional splicing efficiency differs dramatically between *Drosophila* and mouse. *RNA* **18:** 2174–2186. doi:10.1261/rna.034090.112

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645. doi:10.1101/gr.092759.109

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521:** 436–444. doi:10.1038/nature14539

Li Y, Bor Y-C, Misawa Y, Xue Y, Rekosh D, Hammarskjöld M-L. 2006. An intron with a constitutive transport element is retained in a *Tap* messenger RNA. *Nature* **443:** 234–237. doi:10.1038/nature05107

Lin Y, Protter DSW, Rosen MK, Parker R. 2015. Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. *Mol Cell* **60:** 208–219. doi:10.1016/j.molcel.2015.08.018

Linares AJ, Lin C-H, Damianov A, Adams KL, Novitch BG, Black DL. 2015. The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. *eLife* **4:** e09268. doi:10.7554/eLife.09268

Liu H, Liang C, Kollipara RK, Matsui M, Ke X, Jeong B-C, Wang Z, Yoo KS, Yadav GP, Kinch LN, et al. 2016a. HP1BP3, a chromatin retention factor for co-transcriptional microRNA processing. *Mol Cell* **63:** 420–432. doi:10.1016/j.molcel.2016.06.014

Liu Y, Beyer A, Aebersold R. 2016b. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165:** 535–550. doi:10.1016/j.cell.2016.03.014

Makeyev EV, Zhang J, Carrasco MA, Maniatis T. 2007. The microRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27:** 435–448. doi:10.1016/j.molcel.2007.07.015

Mauger O, Lemoine F, Scheiffele P. 2016. Targeted intron retention and excision for rapid gene regulation in response to neuronal activity. *Neuron* **92:** 1266–1278. doi:10.1016/j.neuron.2016.11.032

Naftelberg S, Schor IE, Ast G, Kornblihtt AR. 2015. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu Rev Biochem* **84:** 165–198. doi:10.1146/annurev-biochem-060614-034242

Naganuma T, Nakagawa S, Tanigawa A, Sasaki YF, Goshima N, Hirose T. 2012. Alternative 3′-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J* **31:** 4020–4034. doi:10.1038/emboj.2012.251

Naro C, Jolly A, Di Persio S, Bielli P, Setterblad N, Alberdi AJ, Vicini E, Geremia R, De la Grange P, Sette C. 2017. An orchestrated intron retention program in meiosis controls timely usage of transcripts during germ cell differentiation. *Dev Cell* **41:** 82–93.e4. doi:10.1016/j.devcel.2017.03.003

Ninomiya K, Kataoka N, Hagiwara M. 2011. Stress-responsive maturation of Clk1/4 pre-mRNAs promotes phosphorylation of SR splicing factor. *J Cell Biol* **195:** 27–40. doi:10.1083/jcb.201107093

Novikov AV. 2019. PyClustering: data mining library. *J Open Source Softw* **4:** 1230. doi:10.21105/joss.01230

Pandya-Jones A, Black DL. 2009. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15:** 1896–1908. doi:10.1261/rna.1714509

Pandya-Jones A, Bhatt DM, Lin C-H, Tong A-J, Smale ST, Black DL. 2013. Splicing kinetics and transcript release from the chromatin compartment limit the rate of lipid A-induced gene expression. *RNA* **19:** 811–827. doi:10.1261/rna.039081.113

Pandya-Jones A, Markaki Y, Serizay J, Chitiashvili T, Mancia Leon WR, Damianov A, Chronis C, Papp B, Chen C-K, McKee R, et al. 2020. A protein assembly mediates *Xist* localization and gene silencing. *Nature* **587:** 145–151. doi:10.1038/s41586-020-2703-0

Parra M, Booth BW, Weiszmann R, Yee B, Yeo GW, Brown JB, Celniker SE, Conboy JG. 2018. An important class of intron retention events in human erythroblasts is regulated by cryptic exons proposed to function as splicing decoys. *RNA* **24:** 1255–1265. doi:10.1261/rna.066951.118

Pawlicki JM, Steitz JA. 2008. Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production. *J Cell Biol* **182:** 61–76. doi:10.1083/jcb.200803111

Pelleg D, Moore A. 2000. *X*-means: extended *K*-means with efficient estimation of the number of clusters. In *17th International Conference on Machine Learning, Stanford, CA*, pp. 727–734. Morgan Kaufmann, Stanford, CA.

Pendleton KE, Chen B, Liu K, Hunter OV, Xie Y, Tu BP, Conrad NK. 2017. The U6 snRNA m^6A methyltransferase METTL16 regulates SAM synthetase intron retention. *Cell* **169:** 824–835.e14. doi:10.1016/j.cell.2017.05.003

Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. 2016. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res* **44:** 838–851. doi:10.1093/nar/gkv1168

Popp MW-L, Maquat LE. 2013. Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu Rev Genet* **47:** 139–165. doi:10.1146/annurev-genet-111212-133424

Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. 2019. A machine-learning approach for accurate detection of copy-number variants from exome sequencing. *Genome Res* **29:** 1134–1143. doi:10.1101/gr.245928.118

Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17:** 47–62. doi:10.1038/nrg.2015.10

R Core Team. 2020. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Sakabe NJ, de Souza SJ. 2007. Sequence features responsible for intron retention in human. *BMC Genomics* **8:** 59. doi:10.1186/1471-2164-8-59

Saldi T, Cortazar MA, Sheridan RM, Bentley DL. 2016. Coupling of RNA polymerase II transcription elongation with pre-mRNA splicing. *J Mol Biol* **428:** 2623–2635. doi:10.1016/j.jmb.2016.04.017

Schmid M, Jensen TH. 2018. Controlling nuclear RNA levels. *Nat Rev Genet* **19:** 518–529. doi:10.1038/s41576-018-0013-2

Schmitz U, Pinello N, Jia F, Alasmari S, Ritchie W, Keightley M-C, Shini S, Lieschke GJ, Wong JJ-L, Rasko JEJ. 2017. Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biol* **18:** 216. doi:10.1186/s13059-017-1339-3

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322:** 1849–1851. doi:10.1126/science.1162253

Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, Trabzuni D, Ryten M, Weale ME, Hardy J, et al. 2015. Recursive splicing in long vertebrate genes. *Nature* **521:** 371–375. doi:10.1038/nature14466

Spellman R, Llorian M, Smith CWJ. 2007. Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol Cell* **27:** 420–434. doi:10.1016/j.molcel.2007.06.016

Stewart M. 2019. Polyadenylation and nuclear export of mRNAs. *J Biol Chem* **294:** 2977–2987. doi:10.1074/jbc.REV118.005594

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22:** 1616–1625. doi:10.1101/gr.134445.111

Vargas DY, Shah K, Batish M, Levandoski M, Sinha S, Marras SAE, Schedl P, Tyagi S. 2011. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell* **147:** 1054–1065. doi:10.1016/j.cell.2011.10.024

Vigot R, Barbieri S, Bräuner-Osborne H, Turecek R, Shigemoto R, Zhang Y-P, Luján R, Jacobson LH, Biermann B, Fritschy J-M, et al. 2006. Differential compartmentalization and distinct functions of $GABA_B$ receptor variants. *Neuron* **50:** 589–601. doi:10.1016/j.neuron.2006.04.014

Vilborg A, Steitz JA. 2017. Readthrough transcription: how are DoGs made and what do they do? *RNA Biol* **14:** 632–636. doi:10.1080/15476286.2016.1149680

Vuong CK, Black DL, Zheng S. 2016. The neurogenetics of alternative splicing. *Nat Rev Neurosci* **17:** 265–281. doi:10.1038/nrn.2016.27

Wang Q, Rio DC. 2018. JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proc Natl Acad Sci* **115:** E8181–E8190. doi:10.1073/pnas.1806018115

Wegener M, Müller-McNicoll M. 2018. Nuclear retention of mRNAs: quality control, gene regulation and human disease. *Semin Cell Dev Biol* **79:** 131–142. doi:10.1016/j.semcdb.2017.11.001

Windhager L, Bonfert T, Burger K, Ruzsics Z, Krebs S, Kaufmann S, Malterer G, L'Hernault A, Schilhabel M, Schreiber S, et al. 2012. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res* **22:** 2031–2042. doi:10.1101/gr.131847.111

Wong JJ-L, Ritchie W, Ebner OA, Selbach M, Wong JWH, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, et al. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154:** 583–595. doi:10.1016/j.cell.2013.06.052

Wuarin J, Schibler U. 1994. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* **14:** 7219–7225. doi:10.1128/MCB.14.11.7219

Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang J-P. 2010. Predicting nucleosome positioning using a duration hidden Markov model. *BMC Bioinformatics* **11:** 346. doi:10.1186/1471-2105-11-346

Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV. 2012. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes Dev* **26:** 1209–1223. doi:10.1101/gad.188037.112

Yap K, Mukhina S, Zhang G, Tan JSC, Ong HS, Makeyev EV. 2018. A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. *Mol Cell* **72:** 525–540.e13. doi:10.1016/j.molcel.2018.08.041

Yeom K-H, Damianov A. 2017. Methods for extraction of RNA, proteins, or protein complexes from subcellular compartments of eukaryotic cells. *Methods Mol Biol* **72:** 155–167. doi:10.1007/978-1-4939-7204-3_12

Yeom K-H, Mitchell S, Linares AJ, Zheng S, Lin C-H, Wang X-J, Hoffmann A, Black DL. 2018. *Polypyrimidine tract-binding protein blocks miRNA-124* biogenesis to enforce its neuronal-specific expression in the mouse. *Proc Natl Acad Sci* **115:** E11061–E11070. doi:10.1073/pnas.1809609115

Zheng S, Gray EE, Chawla G, Porse BT, O'Dell TJ, Black DL. 2012. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. *Nat Neurosci* **15:** 381–388. doi:10.1038/nn.3026

Received November 11, 2020; accepted in revised form April 1, 2021.

Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring

Kyu-Hyeon Yeom, Zhicheng Pan, Chia-Ho Lin, et al.

Genome Res. 2021 31: 1106-1119 originally published online April 8, 2021
Access the most recent version at doi:10.1101/gr.273904.120

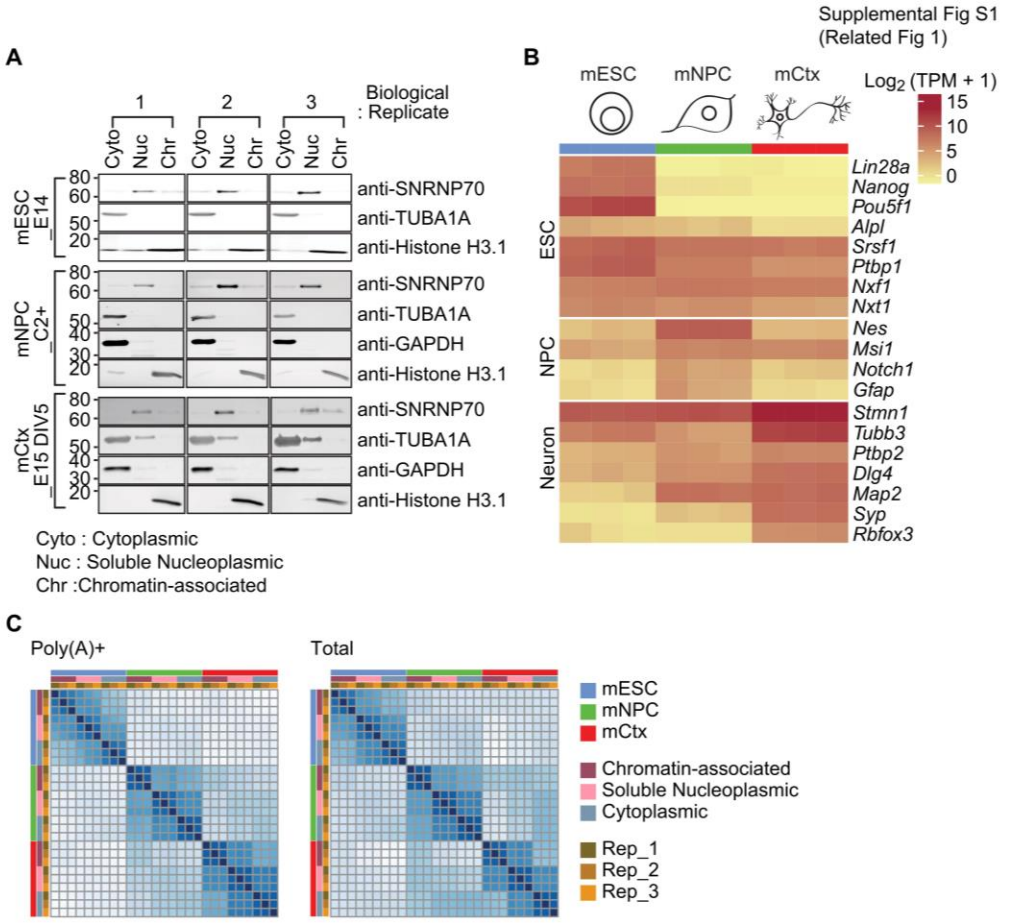
Supplementary Materials for

Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring.

Kyu-Hyeon Yeom[1,†], Zhicheng Pan[2,3,†], Chia-Ho Lin[1], Han Young Lim[1,4], Wen Xiao[1], Yi Xing[3,5,*], Douglas L. Black[1,*].

Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095
Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, CA 90095
Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104
Molecular Biology Interdepartmental Doctoral Program, University of California, Los Angeles, Los Angeles, CA 90095
Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104

This PDF file includes:



Supplemental Figure S1. (Related to Figure 1)

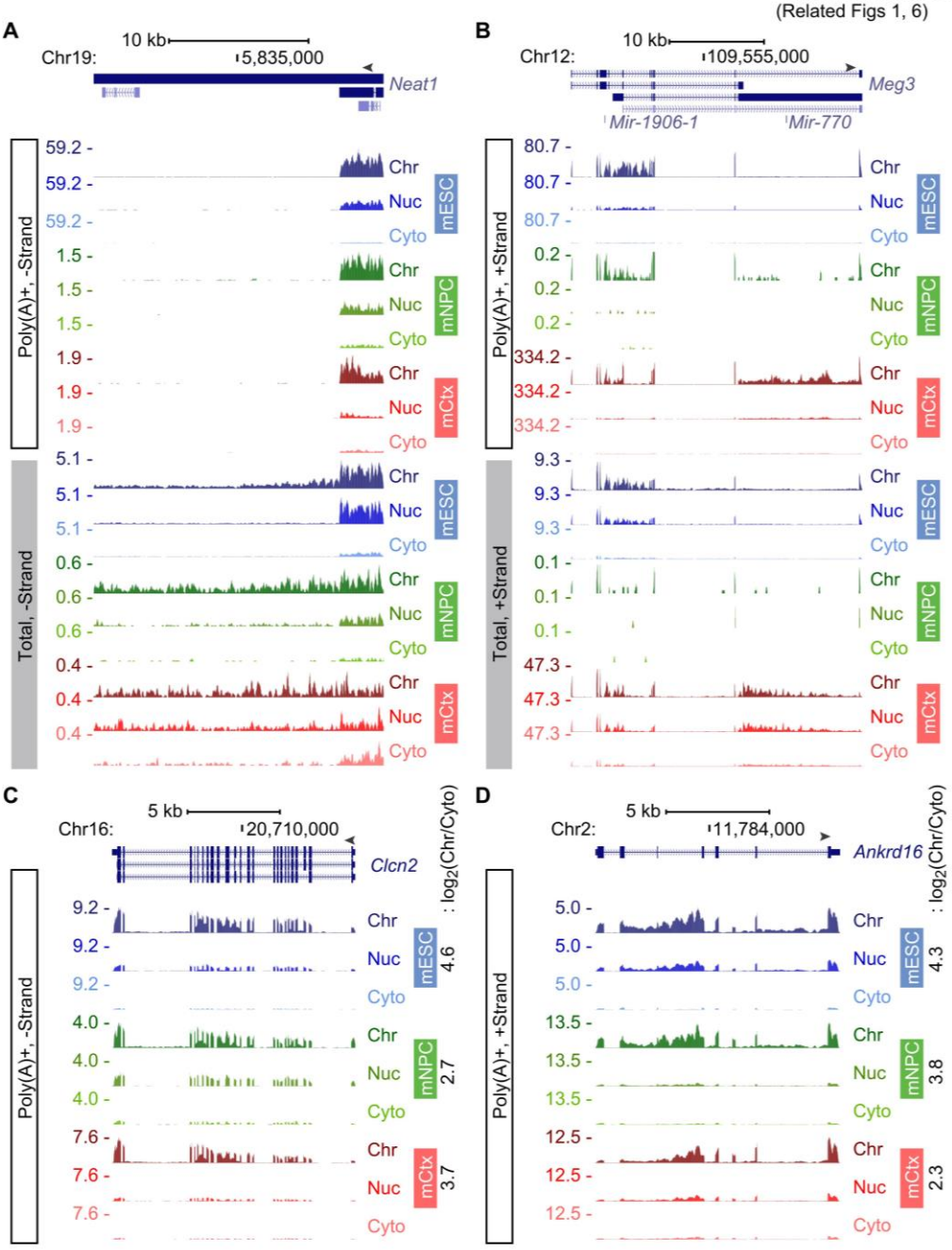
Validation of subcellular fractionation, cell type gene expression, and library consistency.

A. Confirmation of subcellular fractionation. Immunoblot analysis of diagnostic proteins in subcellular fractions. SNRNP70 for soluble nucleoplasm (Nuc), TUBA1A and GAPDH for cytoplasm (Cyto), and Histone H3.1 for chromatin pellet (Chr). Gel images include 3 biological replicates of mouse embryonic stem cells (line E14), mouse neuronal progenitor cell line C2+, and mouse cortical neurons after 5 days *in vitro* culture (E15DIV5; mCtx). Note that the immunoblot results of the third replicate of mESC_E14 are reprinted from Yeom et al. (Yeom et al. 2018).

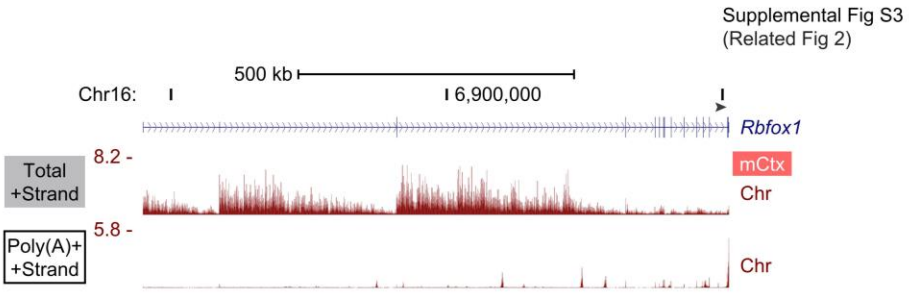
B. Confirmation of cell type specific gene expression. Heatmap presents the cytoplasmic expression as measured by kallisto for the indicated mRNAs in each cell type and replicate.

C. Confirmation of library similarity. Heatmap displays similarity of gene expression between pairwise comparisons of all cell types, fractions, and replicates. Color codes are indicated on right.

Supplemental Fig S2
(Related Figs 1, 6)

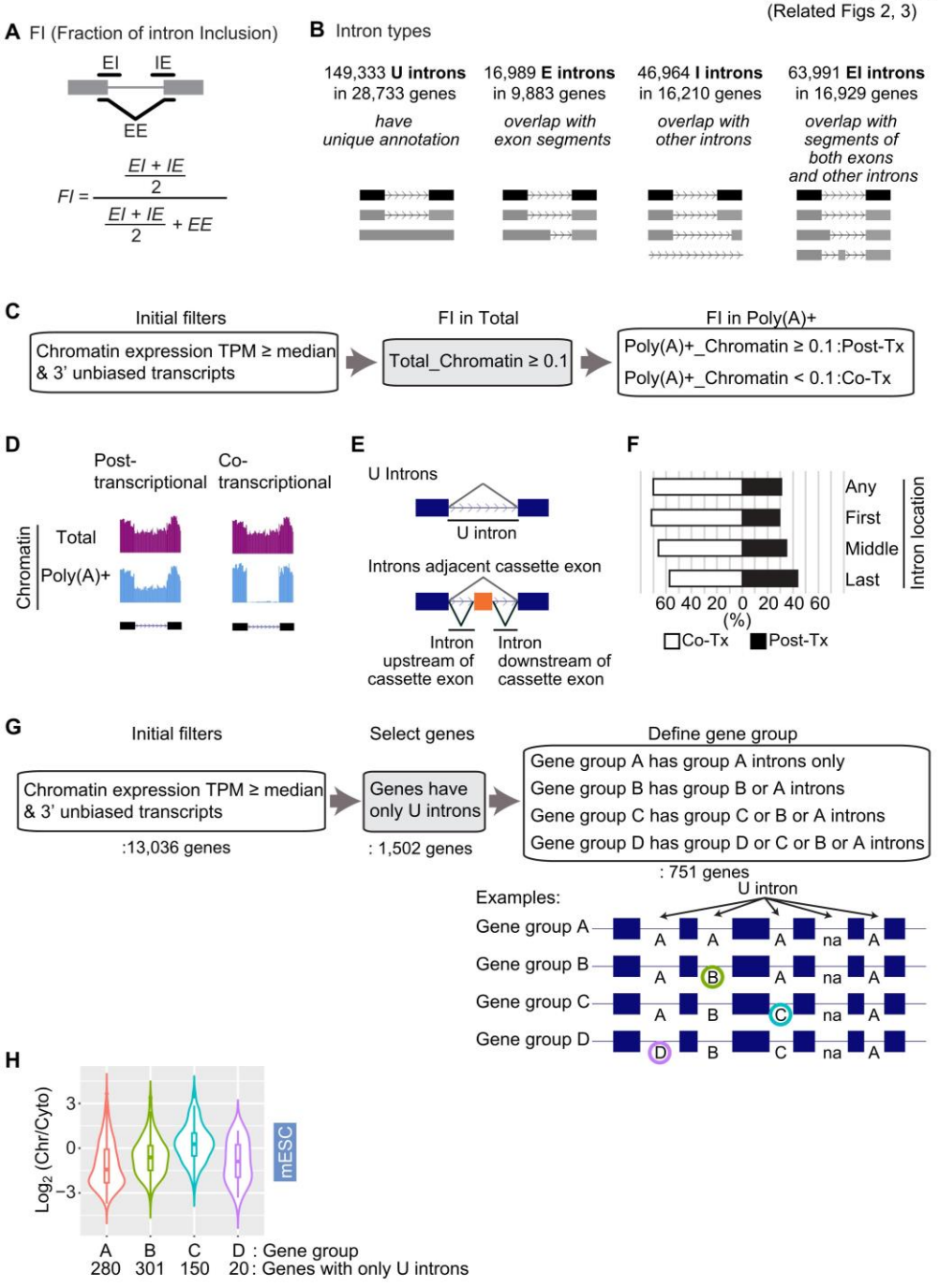


Supplemental Figure S2. (Related to Figures 1 and 6)
Example genome browser tracks of non-coding and coding RNAs.
A. *Neat1* expression in mESC, mNPC, and mCtx. Genome browser tracks of the *Neat1* locus for poly(A)+ and total libraries. Y axis shows RPM scaled to the highest value in the Chromatin-associated fraction.
B. *Meg3* expression in mESC, mNPC, and mCtx. Genome browser tracks of the *Meg3* locus in poly(A)+ and total libraries.
C. Genome browser tracks of the *Clcn2* locus in mESC, mNPC, and mCtx. Transcripts are enriched in the chromatin fraction and exhibit unspliced introns in poly(A)+ RNA. The partition index of *Clcn2* in each cell type is indicated on the right.
D. Genome browser tracks of the *Ankrd16* locus in mESC, mNPC, and mCtx. Transcripts are enriched in the chromatin fraction and exhibit unspliced introns in poly(A)+ RNA. Partition index of *Ankrd16* in each cell type is indicated on the right.



Supplemental Figure S3. (Related to Figure 2)
Very long introns exhibit declining reads 5' to 3' to create a sawtooth pattern.
Genome browser tracks of the *Rbfox1* locus for poly(A)+ and total libraries. Y axis shows RPM in each library.

Supplemental Fig S4
(Related Figs 2, 3)



Supplemental Figure S4. (Related to Figures 2 and 3)

Computational definition of introns and splicing.

A. Determination of FI value using read numbers for the exon-intron junction (EI), intron-exon junction (IE), and exon-exon junction (EE).

B. Introns were categorized as one of four types based on their Ensembl v91 annotation. Introns that are not partly overlapped with either exons or other introns are classified as U type introns. Introns that partly overlap with exons but not with other introns are classified E type introns. Introns that overlap with other annotated introns but not exons are called I type introns. EI type introns overlap with both exons and introns of other annotated isoforms.

C. Determination of cotranscriptional and posttranscriptional splicing. FI values were determined for all U introns from total and poly(A)+ chromatin associated RNA. Genes with overall expression above the median (2.13 TPM) were analyzed. Genes showing a bias for reads in the 3' end in the poly(A)+ RNA, and introns exhibiting FI values in total RNA below 0.1 were removed. A posttranscriptional splicing event was then defined as an intron having an FI value in poly(A)+ RNA greater than or equal to 0.1 (Post-tx). Cotranscriptional splicing of an intron generates an FI of less than 0.1 in the poly(A)+ RNA (Co-tx).

D. Illustration of post and cotranscriptional splicing. Introns with high read numbers on chromatin in both the total and poly(A)+ libraries were defined as posttranscriptionally spliced. Cotranscriptional splicing events exhibited reads in the total but not the poly(A)+ RNA.

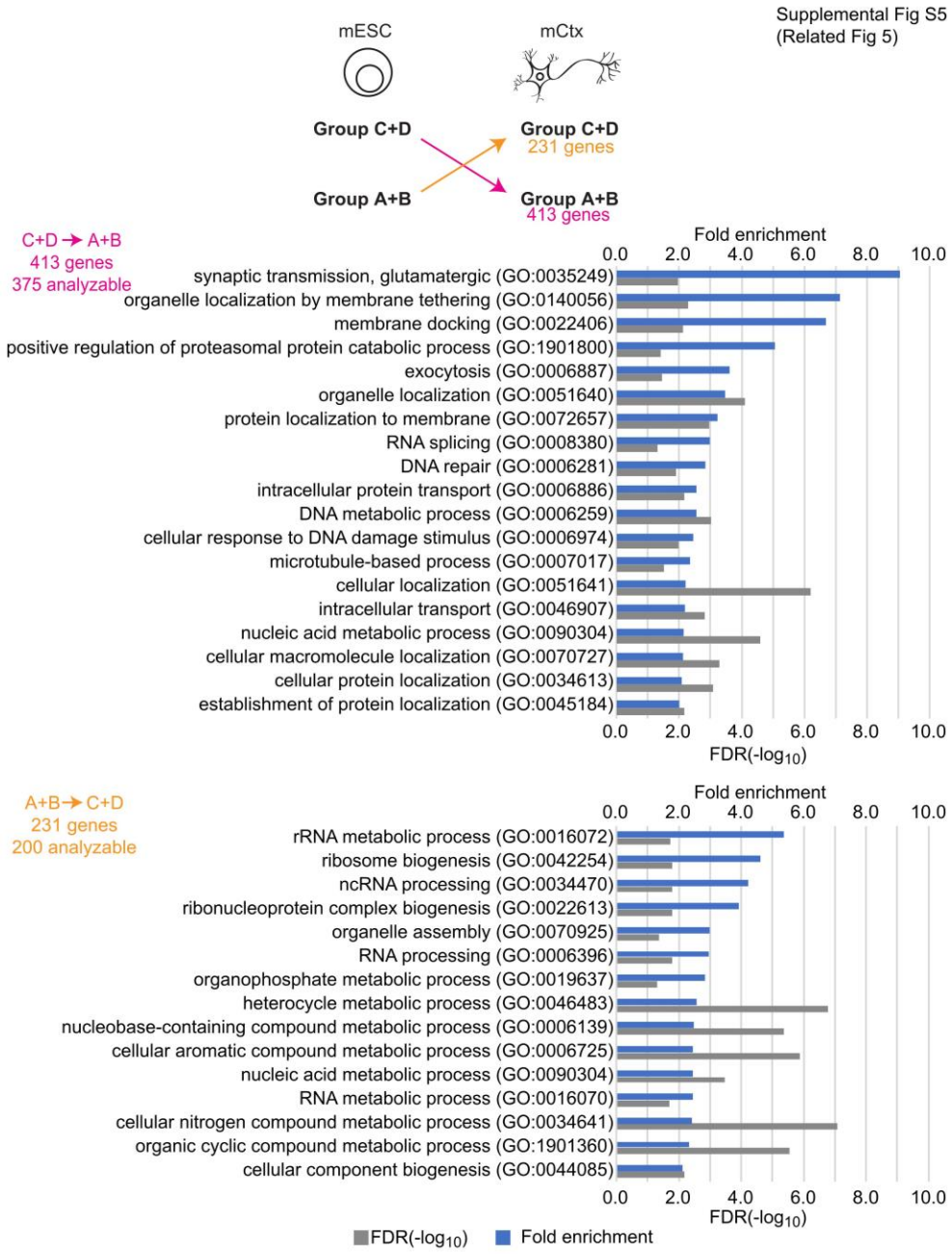
E. Diagrams of constitutive U introns and I introns adjacent to simple cassette exons that were assessed for co- and posttranscription splicing as presented in Figure 2C.

F. The proportions of co- and posttranscriptional splicing for all U introns and for first, middle and last introns in a transcript.

G. Transcripts with unspliced introns are enriched in the chromatin fraction.

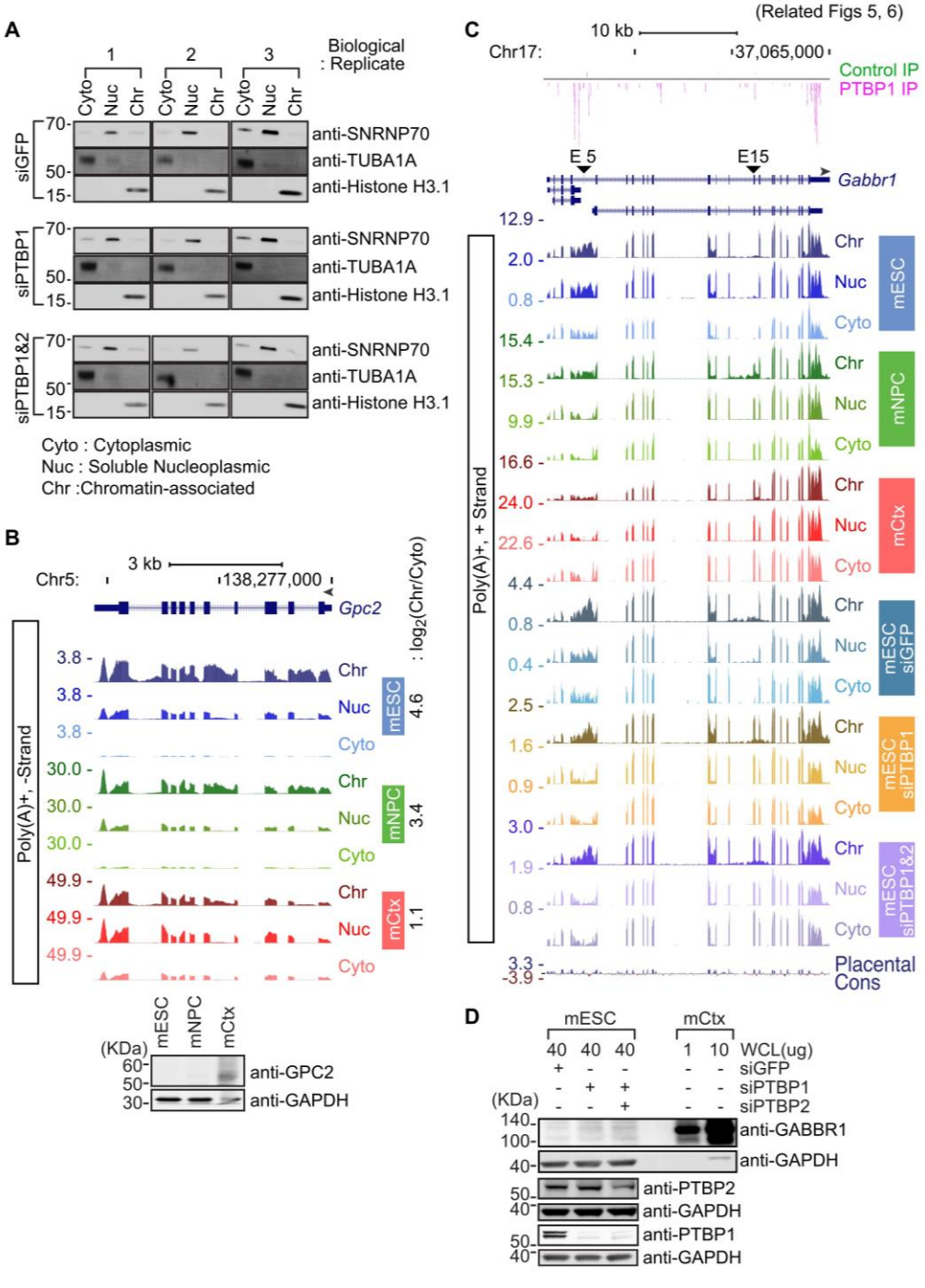
Genes having only U introns were selected from those whose overall expression was above the median (2.13 TPM). The gene group was then defined by the highest intron group within the gene (751 genes), where D > C > B > A. Introns marked 'na' indicate they were filtered by SIRI during X-means clustering.

H. Violin plots showing the distribution of chromatin partition indices (Log_2(Chr/Cyto)) of transcripts from different gene groups defined above. The number of genes in each gene group is indicated at the bottom.



Supplemental Figure S5. (Related to Figure 5)

GO analysis of genes containing introns that switch intron group during neuronal differentiation. Number of genes containing introns that changed group between mESC and mCtx is indicated at the top in yellow and pink. GO biological process enrichment these gene sets are listed at the bottom. Fold enrichment and FDR ($-\log_{10}$) shown in blue and grey bars, respectively.


A
Biological : Replicate
1
2
3
Cyto
Nuc
Chr
siGFP
siPTBP1
siPTBP1&2
70-
50-
15-
anti-SNRNP70
anti-TUBA1A
anti-Histone H3.1
Cyto : Cytoplasmic
Nuc : Soluble Nucleoplasmic
Chr :Chromatin-associated
B
3 kb
Chr5:
138,277,000
Gpc2
Poly(A)+, -Strand
: log2(Chr/Cyto)
3.8
30.0
49.9
Chr
Nuc
Cyto
mESC
mNPC
mCtx
4.6
3.4
1.1
(KDa)
60-
50-
30-
anti-GPC2
anti-GAPDH
C
10 kb
Chr17:
37,065,000
Control IP
PTBP1 IP
E5
E15
Gabbr1
Poly(A)+, + Strand
12.9
2.0
0.8
15.4
15.3
9.9
16.6
24.0
22.6
4.4
0.8
0.4
2.5
1.6
0.9
3.0
1.9
0.8
mESC siGFP
mESC siPTBP1
mESC siPTBP1&2
3.3
-3.9
Placental Cons
D
mESC
mCtx
40 40 40
1 10
WCL(ug)
siGFP
siPTBP1
siPTBP2
(KDa)
140-
100-
40-
50-
anti-GABBR1
anti-GAPDH
anti-PTBP2
anti-PTBP1

Supplemental Figure S6. (Related to Figures 5 and 6)
Validation of subcellular fractionation after *Ptbp* knockdown in mESC and genome browser tracks of *Gabbr1*.
A. Confirmation of subcellular fractionation. Immunoblot analysis of diagnostic proteins in sub cellular fractions. SNRNP70 for soluble nucleoplasm (Nuc), TUBA1A and GAPDH for cytoplasm (Cyto), and Histone H3.1 for chromatin pellet (Chr). Gel images include 3 biological replicates of mouse embryonic stem cells (line E14).
B. (Upper Panel) Genome browser tracks of the *Gpc2* locus in mESC, mNPC, and mCtx. Transcripts are enriched in the chromatin fraction and exhibit unspliced introns in poly(A)+ RNA. The partition index of *Gpc2* in each cell type is indicated on the right. (Lower Panel) Immunoblot measuring expression of GPC2 protein relative to GAPDH control in mESC, mNPC and cortical neurons (mCtx). Gel image is one of 3 biological replicates.
C. Complete genome browser tracks of the *Gabbr1* locus in mESC, mNPC, and mCtx, and for *Ptbp1* knockdown and *Ptbp1/2* double knockdown in mESC. PTBP1 iCLIP tags in mESC are shown at the top (Linares et al. 2015). Intron 4-5 region is shown with a bracket, and exons 5 and 15 are marked with arrowheads.
D. Immunoblot measuring expression of GABBR1 protein relative to GAPDH in *Ptbp1 and Ptbp1/2* double knockdown samples in mESC and in mCtx as positive control. 40ug of whole cell lysate (WCL) were loaded on the gel for mESC, and 1 and 10 ug of WCL for mCtx. Gel image is one of 3 biological replicates.

List of Supplemental Tables

Supplemental Table S8.

A. List of U introns reported as detained introns (DIs) in Boutz et al. (Boutz et al. 2015).

B. Proportion of introns assigned to groups.

C. List of genes containing PTBP1 dependent retained introns in chromatin fraction (871 genes). 87 genes out of 871 were reported upregulated by 10% after *Upf1* knockdown in data from Hurt et al. (Hurt et al. 2013).

D. Number of genes with different partition indices in mESC that were seen to be upregulated by 10% after *Upf1* knockdown. Data from Hurt et al. (Hurt et al. 2013).

E. Number of U introns and proportion of intron groups in mCtx that overlap with intron groups from Mauger et al. (Mauger et al. 2016).

Supplemental Table S9.

A. List of sequence features used for deep leaning model (1,387).

B. Predictive value of genomic features for distinguishing introns of different groups.

Supplemental Methods

Cell lines and tissue culture

The mouse embryonic stem cells (mESC) line, E14 (Hooper et al. 1987) was cultured on 0.1 % gelatin-coated dishes with mitotically inactivated mouse embryonic fibroblasts (MEF) (CF1, Applied StemCell, Inc.) in mESC media at least 2 passages from the initial thawing. Then, the mESCs were transferred onto 100 mm cell culture plates that contained only 0.1 % gelatin in preparation for RNAi and cell fractionation experiments. mESC media consisted of DMEM (Fisher Scientific) supplemented with 15 % ESC-qualified fetal bovine serum (Thermo Fisher Scientific), 1x non-essential amino acids (Thermo Fisher Scientific), 1x GlutaMAX (Thermo Fisher Scientific), 1x ESC-qualified nucleosides (EMD Millipore), 0.1 mM β-Mercaptoethanol (Sigma-Aldrich), and 10^3 units/ml ESGRO leukemia inhibitor factor (LIF) (EMD Millipore). Mouse primary cortical neuron cultures were prepared from gestational day 15 C57BL/6 embryos (E15) (Charles River Laboratories), as described previously (Zheng et al. 2010). Briefly, cortices were dissected out into ice cold HBSS and dissociated after a 10 min digestion in Trypsin (Thermo Fisher Scientific), then plated with plating media consisted of 70 % Neurobasal (v/v), 20 % horse serum (v/v), 25 mM sucrose, and 0.25x GlutaMAX plated at 5 million cells per 78.5 cm^2 on 0.1 mg/ml poly-L-lysine coated dishes. The half media was replaced with feeding media consisted of 98 % Neurobasal (v/v), 1x B27 (with vitamin A, Thermo Fisher Scientific), and 0.25x GlutaMAX at day 1 and 3. The neuronal culture maintained for 5 days *in vitro* (DIV5). Neurons represented 70 – 80 % of the cells in the culture. A mouse neuronal progenitor cell line (mNPC) was established from cortical cells of gestational day 15 embryos generated by crossing homozygous Nestin-GFP transgenic mice (Mignone et al. 2004) to wild type C57BL/6. GFP positive cells were collected by FACS and plated on uncoated culture dishes. These NPCs were

grown in DMEM/F12 supplemented with B27 (without vitamin A, Thermo Fisher Scientific), 1x GlutaMAX and antibiotics. EGF and FGF (PeproTech) were added every day at 10 ng/ml concentration. All experiments were approved by the UCLA Institutional Animal Care and Use Committee (ARC# 1998-155-53).

Knockdown of *Ptbp* in mESC.

siRNAs that target EGFP (Silencer Select AM4626), PTBP1 (Silencer Select s72337), and PTBP2 (Silencer Select s80149) were transfected into mESCs twice using Lipofectamine RNAiMAX (Invitrogen). The first transfection was performed while the cells were in suspension, and the second transfection was performed 24 hours after when the cells had already attached to the surface of the cell culture dish. The cells were harvested and fractionated 24 hours after the second transfection.

Subcellular fractionation, RNA isolation, and library construction in detail

mESCs, mNPC, and cortical neurons (mCtx) were fractionated into cytoplasmic, soluble nuclear, and chromatin pellet as described previously (Wuarin and Schibler 1994; Pandya-Jones and Black 2009; Yeom and Damianov 2017; Yeom et al. 2018). Briefly, cells were washed twice with washing buffer (1X PBS/1 mM EDTA [pH 8.0]) and gently trypsinized to collect the cell pellet by centrifugation. 1 X 10^7 cells from the pellet were retrieved in a 2.0 mL low adhesion microcentrifuge tube (USA Scientific) and washed twice with washing buffer to remove cell debris from prematurely-lysed cells. For neurons, the initial plating number on the day of dissection and culture was used for quantification, and cells were collected by scraping from the plate without trypsinization.

The cells were then incubated first in ice-cold hypotonic buffer (10 mM Tris–HCl [pH 7.5], 15 mM KCl, 1 mM EDTA [pH 8.0], 0.15 mM Spermine, 0.5 mM Spermidine, 0.5 mM DTT, and 1× Protease inhibitor, and 15 mM KCl (50 mM NaCl for mNPC and mCtx))

for 5 min and lysed in ice-cold lysis buffer (10 mM Tris–HCl [pH 7.5], 15 mM KCl, 1 mM EDTA [pH 8.0], 0.15 mM Spermine, 0.5 mM Spermidine, 0.5 mM DTT, and 1× Protease inhibitor, and 0.0375 %, 0.15 %, and 0.1 % of Igepal CA-630 for mESC, mNPC, and mCtx, respectively) for 1 min. The resulting lysate was immediately layered on top of a chilled 24 % sucrose solution (hypotonic buffer in 24 % (w/v) sucrose without detergent) and centrifuged for 10 min, 4 °C, 6000 x g. 10 % of the supernatant (cytoplasmic fraction) was used for immunoblot to check for contamination from nuclear materials, and the rest was mixed with TRIzol LS (Invitrogen) to extract cytoplasmic RNA. The nuclear pellet was washed once with washing buffer before being resuspended in 100 µl of chilled glycerol buffer (20mM Tris-HCl [pH 7.9], 75 mM NaCl, 0.5 mM EDTA [pH 8.0], 50 % glycerol (v/v), 0.85 mM DTT, and 0.125 mM PMSF). 100 µl of cold nuclear lysis buffer (20 mM HEPES [pH 7.6], 7.5 mM MgCl2, 0.2 mM EDTA [pH 8.0], 300 mM NaCl, 1M urea, 1 % Igepal CA-630, 0.3 mM Spermine, 1.0 mM Spermidine, 1× Protease inhibitor) was then added and the mixture incubated on ice for 2 min. After the 2 min incubation period, 1/8 of the nuclear lysate was transferred to a separate 1.5 mL microcentrifuge tube for each sample. After centrifugation for 2 min, 4 °C, 6000 x g, the supernatant (soluble nuclear fraction) from both tubes were pooled. 10 % of the pooled supernatant was used to check its purity by western blot and the rest was mixed with TRIzol LS to extract nucleoplasmic RNA. The resulting two insoluble nuclear pellets (one large, one small) were washed once with washing buffer before incubation of the large pellet in TRIzol at 50 °C until the pellet was completely solubilized to extract chromatin-associated RNA (from 7/8 nuclear lysate). After washing the small pellet (from 1/8 nuclear lysate) was resuspended in 5 % SDS sample buffer (5 % SDS, 30 % glycerol (v/v) and 150 mM Tris–HCl [pH 8.0]) at 55 °C for 15 min to extract proteins from the chromatin fraction. Each fractionation experiment was performed in triplicate.

RNA was treated DNase I (Takara) followed by phenol extraction. Before RNA quality check and the size selection steps, RNA was quantified on a NanoDrop spectrophotomer (Supplemental Table S2A), and RNA integrity was checked by either Bioanalyzer or RNA screen tape (Agilent) (Supplemental Table S2A). Note that the sum of the RNA from the chromatin, nucleus, and cytoplasmic fractions does not represent total cellular RNA due to extensive washing steps during fractionation.

5ug of RNA from each subcellular fraction was processed with the RNeasy MinElute Cleanup Kit (Qiagen). RNAs longer than 200 nt were eluted from the membrane to make long RNA libraries, while the flow through (< 200 nt) was also collected and ethanol precipitated to make short RNA libraries from each fraction. Half of the long RNA eluate was used to generate poly(A)+ libraries using the TruSeq Stranded mRNA Library Prep Kit (Illumina). The other half the long RNA eluate was used for total RNA library construction, using the ribosomal RNA removal kit (Ribo-Zero™ rRNA Removal Kits (Human/Mouse/Rat), MRZH116, Epicentre Biotechnologies) followed by library construction using TruSeq Stranded mRNA Library Prep Kit (Illumina) without the oligo-T bead purification step. Short RNA was used for small RNA library construction, using the ribosomal RNA removal kit (Ribo-Zero™ rRNA Removal Kits (Human/Mouse/Rat), MRZH116, Epicentre Biotechnologies) followed by library construction using the TruSeq Small RNA sample Prep Kit (Illumina). The small RNA library was further purified on a TBE gel by excising bands between 120 - 160 nt to generate small non-coding RNA reads that include mature-miRNAs (~ 22 nt) and piwi-interacting RNAs (piRNAs, ~ 30 nt).

For *Ptbp* knockdown experiments in mESC, poly(A)+ libraries were created using the TruSeq Stranded mRNA Library Prep Kit (Illumina).

RNA sequencing and alignment

The libraries of all fractions and cell types, were subjected to 100 nt paired-end sequencing at the UCLA Broad Stem Cell Center core facility on an Illumina HiSeq 4000 to generate about 21 million mapped reads per sample. The libraries of *Ptbp* KD samples in mESC were subjected to 75 nt paired-end sequencing at the UCLA Neuroscience Genomics Core on an Illumina HiSeq 4000 to generate about 25 million mapped reads per sample. Small RNA libraries were sequenced on an Illumina MiSeq machine using the MiSeq Reagent Kit v3 (Illumina). RNA-seq data were aligned using STAR (version 020210) (Dobin et al. 2013) on mm10.

Sequencing reads and mapping results are summarized in Supplemental Table S2B – S2D. Reads per million read values (RPM) were calculated and displayed in UCSC Genome Browser sessions including strand information. TPM values obtained from kallisto (version 0.43.0) were used in gene expression analyses (Bray et al. 2016). To examine the similarity of replicate samples, gene expression distances between samples were calculated using DESeq2 (Anders and Huber 2010), and plotted in heatmaps (Supplemental Fig S1) using ggplot2. For small RNA, sequencing reads and mapping results are summarized in Supplemental Table S2E. Small RNA expression was analyzed using the online tool SPAR (https://www.lisanwanglab.org/SPAR) (Kuksa et al. 2018).

RT-PCR and RT-qPCR validation

RNA was collected and extracted from subcellular fractions as described above. For the RT reaction, 0.8 - 1 µg of total RNA, 100 ng of random hexamers (or 250 ng of oligo(dT)), and 0.5 mM of dNTP mix (0.5 mM each) were incubated in a 7 µl reaction volume at 65 °C for 5 min. After 1min incubation on ice, 2 µl of 5x reaction buffer, 5 mM of DTT, and 100 units of SuperScript III RT (Thermo Fisher Scientific) were added. This

10 µl mixture was incubated at 25 °C for 10 min, 50 °C for 60 min, and 70 °C for 15 min. PCR was performed using Phusion DNA polymerase (Fisher Scientific). After the initial denaturing step at 98 °C for 1 min, amplification continued for 19 - 28 cycles of 98 °C for 1 min, 58 - 62 °C for 30 sec (annealing), and 72 °C for 20 - 25 sec (elongation). The reaction was completed with a final elongation at 72 °C for 10 min. Annealing temperature, elongation time, cycle numbers, and amplicon size are listed in Supplemental Table S1A. RT-PCR products were run on either agarose gels with EtBr staining or PAGE gels with SYBR Gold staining (Thermo Fisher Scientific), and visualized on a Typhoon imager (GE Healthcare) using the 492 nm excitation laser and 510 nm emission filters. For RT-qPCR the SensiFAST SYBR Lo-ROX Kit (Bioline) was used, and reactions contained 0.4ul of diluted cDNA (1:5 dilution in water) with 250nM each of forward and reverse primers in 6ul of RT-qPCR reactions. The mixtures were run on a QuantStudio 6 Real-Time PCR System (Thermo Fisher Scientific) with combined annealing and elongation cycles (Step1: 95 °C for 2 min, followed by 40 cycles of Step2: 95 °C for 5 sec and Step3: 55 or 60 °C for 30 sec. A list of primer sequences and PCR conditions is presented in Supplemental Table S1A.

Co- and posttranscriptional splicing analysis

For cotranscriptional splicing analysis, the U intron list was filtered for genes whose kallisto TPM in chromatin poly(A)+ RNA was over the median. Genes exhibiting 3' bias were identified by comparing the reads per nucleotide length of the 2nd exon to that of the 2nd to last exon for the longest transcript from the gene. If this ratio was less than 0.5 for all transcripts, the gene was excluded (Supplemental Fig. S4C). The same filters were applied to a list of simple cassette exons (SE) extracted from rMATS 4.0.1 (Shen et al. 2014). Cassette exons were selected that were located between I introns

when included, and become EI introns when skipped, as defined in Supplemental Fig. S4E.

To examine splicing at different intron locations, we compiled a list of 936 genes containing U introns as first, middle, and last introns. The middle intron was defined as intron number X. For genes with an odd number of total introns, X= (total intron number + 1) / 2. For genes with an even number of total introns, then X= total intron number / 2.

Immunoblotting

Protein samples for checking subcellular fractionation were prepared as described above. Whole-cell lysates (WCL) for checking GPC2, GABBR1, TUBB3, PTBP1 and PTBP2 were extracted in RIPA, quantified with the bicinchoninic acid (BCA) method (Smith et al. 1985), and loaded on 10 % SDS-PAGE gel. Images were taken on a Typhoon Imager (GE Healthcare) using fluorescent (Cy3 or Cy5) secondary antibodies. For GABBR1 detection HRP-linked secondary antibodies and a LAS-3000 Imaging System (Fuji) were used. All the experiments were repeated in three times. Antibodies used in this study were listed in Supplemental Table S1B.

GO analysis

Gene Ontology (GO) enrichments were determined using PANTHER (Protein ANalysis THrough Evolutionary Relationships) (Ashburner et al. 2000; Gene Ontology Consortium 2019). Molecular function terms were assessed for overrepresentation in genes containing introns that changed during neuronal differentiation from Groups C or D to Groups A or B, or from Groups A or B to Groups C or D compared to all mouse genes.

Identification of NMD target genes

A list NMD targets was extracted from a previous study in mESC (Hurt et al. 2013). We filtered for genes upregulated by greater than 10 % in two independent siRNA knockdowns of *Upf1* (siUpf1.1 and siUpf1.2). If a gene had multiple mRNA isoforms and one isoform was upregulated it was included as an NMD target on our list.

Identifying Premature Termination Codons (PTC) in retained introns

For each U intron, we retrieved all transcripts containing this intron from GTF. The transcripts without a clear open reading frame annotation or with termination codons after the intron were not considered. We also removed transcripts in which the intron was the last intron, because last intron retention should not cause Nonsense Mediated Decay (NMD). For each transcript that contained the intron, we inserted the intron sequence into the CDS of the transcript at junction of the intron. If the U intron was not the second last intron of the transcript, the intron was marked as PTC-containing if generated a stop codon inside the intron. If the intron was the second last intron of the transcript, the intron was defined as PTC-containing if an in frame stop codon within the intron was at least 50 nt upstream from the last exon-exon junction of the transcript. After checking all transcripts containing the intron, we defined the intron as a PTC-containing intron if it generated PTC in at least one transcript.

References for Supplemental Materials

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25–29.

Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev* **29**: 63–80.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Gene Ontology Consortium T. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**: D330–D338.

Hooper M, Hardy K, Handyside A, Hunter S, Monk M. 1987. HPRT-deficient (Lesch–Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326**: 292–295.

Hurt JA, Robertson AD, Burge CB. 2013. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res* **23**: 1636–1650.

Kuksa PP, Amlie-Wolf A, Katanić Ž, Valladares O, Wang L-S, Leung YY. 2018. SPAR: small RNA-seq portal for analysis of sequencing experiments. *Nucleic Acids Res* **46**: W36–W42.

Linares AJ, Lin C-H, Damianov A, Adams KL, Novitch BG, Black DL. 2015. The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation ed. B.J. Blencowe. *eLife* **4**: e09268.

Mauger O, Lemoine F, Scheiffele P. 2016. Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity. *Neuron* **92**: 1266–1278.

Mignone JL, Kukekov V, Chiang A-S, Steindler D, Enikolopov G. 2004. Neural stem and progenitor cells in nestin-GFP transgenic mice. *J Comp Neurol* **469**: 311–324.

Pandya-Jones A, Black DL. 2009. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15**: 1896–1908.

Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. 2014. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci* **111**: E5593–E5601.

Smith PK, Krohn RI, Hermanson GT, Mallia AK, Gartner FH, Provenzano MD, Fujimoto EK, Goeke NM, Olson BJ, Klenk DC. 1985. Measurement of protein using bicinchoninic acid. *Anal Biochem* **150**: 76–85.

Wuarin J, Schibler U. 1994. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* **14**: 7219–7225.

Yeom K-H, Damianov A. 2017. Methods for Extraction of RNA, Proteins, or Protein Complexes from Subcellular Compartments of Eukaryotic Cells. In *mRNA Processing: Methods and Protocols* (ed. Y. Shi), *Methods in Molecular Biology*, pp. 155–167, Springer, New York, NY https://doi.org/10.1007/978-1-4939-7204-3_12 (Accessed August 21, 2020).

Yeom K-H, Mitchell S, Linares AJ, Zheng S, Lin C-H, Wang X-J, Hoffmann A, Black DL. 2018. Polypyrimidine tract-binding protein blocks miRNA-124 biogenesis to enforce its neuronal-specific expression in the mouse. *Proc Natl Acad Sci* **115**: E11061–E11070.

Zheng S, Eacker SM, Hong SJ, Gronostajski RM, Dawson TM, Dawson VL. 2010. NMDA-induced neuronal survival is mediated through nuclear factor I-A in mice. *J Clin Invest* **120**: 2446–2456.

Chapter 3: Nuclear export of mRNAs encoding transcription factors by NXF1 influences mESC pluripotency

Nuclear export of mRNAs encoding transcription factors by NXF1 influences mESC pluripotency

Han Young Lim[1,2], Chia-Ho Lin[1], Kathrin Plath[3], Douglas L. Black[1]

[1]UCLA Department of Microbiology, Immunology, and Molecular Genetics

[2]Molecular Biology Interdepartmental Graduate Program at UCLA

[3]Department of Biological Chemistry, University of California, Los Angeles

Abstract

After synthesis and processing in the nucleus, messenger RNAs must be recognized as fully mature and ready for export to the cytoplasm for translation. The nuclear RNA export factor 1 (NXF1) protein acts as a shuttle for the nucleocytoplasmic transport of mRNAs, however the mechanisms of NXF1 recruitment are heterogeneous and data indicate that different classes of mRNA differ in the factors needed for their export. To broadly examine effects of NXF1 on the nuclear export of mRNAs, we depleted the protein from mouse embryonic stem cells (mESCs) before fractionating the cells into cytoplasmic, soluble nuclear, and chromatin-associated compartments. We extracted and sequenced the polyA+ RNA from each subcellular fraction, and assessed how loss of Nxf1 affected their nucleocytoplasmic distribution. We find that many mRNAs are relatively insensitive to Nxf1 depletion and Nxf1 dependence is variable between individual genes. Confirming the results of others, we find that transcripts that are short and have a low number of introns are most strongly affected. We further find that these genes strongly affected by NXF1 knockdown include many that encode transcription factors and other transcripts of immediate early genes. It appears that the rapid transcription and maturation of these RNAs is accompanied by rapid NXF1 binding and export. To examine Nxf1 recruitment to RNAs expressed in response to cellular stimuli, we performed individual-nucleotide resolution cross-linking and immunoprecipitation (iCLIP) analysis of FLAG-tagged NXF1 before and after serum stimulation. This identified new NXF1-sensitive mRNAs bound by FLAG-NXF1. Among

the mRNAs most dependent on Nxf1 for export is that encoding Sox2, a key regulator of pluripotency and differentiation. We find that loss of Nxf1 leads to a loss of Sox2 and other pluripotency factors in ESC, including Nanog, Oct4, and Rex1. Our findings indicate that mESC pluripotency requires the maintenance of cytoplasmic Sox2 and other mRNAs through their specific dependency on Nxf1.

INTRODUCTION

The subcellular compartmentalization of eukaryotic cells enables multiple posttranscriptional gene regulatory mechanisms. In the nucleus, after transcription and complex processing events a mature mRNA is exported to the cytoplasm for translation via the action of the Nxf1/Nxt1 heterodimer that shuttles through the nuclear pore complex (Hocine, Singer, and Grünwald 2010; Katahira 2015; Okamura, Inose, and Masuda 2015). The recognition of an mRNA as fully mature requires the TREX/THO complex that recruits Nxf1, and this recruitment can occur co-transcriptionally to the whole mRNA starting at the 5' end (Viphakone et al. 2019). In addition to TREX/THO, other proteins can act as adapters to recruit Nxf1, including SR proteins binding to exons (Moore and Proudfoot 2009; Müller-McNicoll et al. 2016; Huang and Steitz 2001; Reed and Cheng 2005). How TREX/THO can exclude incompletely processed mRNAs from the export pathway or sometimes allow export of alternative transcripts containing introns is not well understood. Some RNAs containing retained introns can be exported through the action of proteins that recruit exporters directly to the intron sequence. The unspliced viral RNAs of HIV are bound by the viral REV protein that binds to an intronic element in the HIV RNA and acts as an adapter to recruit the Crm1 exportin protein. Similarly, viral RNAs of the Mason Pfizer monkey virus (MPMV) contain a structure called the constitutive transport element (CTE) that directly recruits the host NXF1 for export of incompletely-spliced transcripts (Pasquinelli et al. 1997). Interestingly, the NXF1 transcript itself contains a CTE within intron 10. This causes some export of the NXF1 mRNA retaining intron 10, giving rise to a truncated NXF1 sNXF1 protein whose expression varies with cell type (Li et al. 2006; 2016). These findings indicate that mRNAs can employ multiple mechanisms for export than can be either NXF1 dependent or independent.

NXF1 is a 70 kDa protein with an N-terminal arginine-rich non-sequence specific RNA binding domain and a C-terminal domain that allows interaction with the FG-repeat nucleoporins during export (Viphakone et al. 2012). Several recent studies have indicated that NXF1 is

selective in targeting RNAs for export, presumably mediated by its different interactions with adapter proteins (Zuckerman et al. 2020; Lee et al. 2020; Aksenova et al. 2020). These findings imply that specialized factors may mediate export of different groups of RNAs.

Here we examine NXF1's role in the export of RNAs in mouse embryonic stem cells (mESCs). Using siRNA knockdown and subcellular fractionation, we find that depletion of NXF1 from mESCs primarily alters the nuclear/cytoplasmic accumulation of mRNAs encoding transcription factors and early response genes. NXF1 iCLIP-seq identified additional transcription factors and other targets of NXF1. Finally, we find that through its effect on the export of Sox2 and other transcription factors NXF1 plays an essential role in maintaining pluripotency in mESC.

RESULTS

Depletion of NXF1 from mESCs alters RNA partitioning between cellular compartments.

To broadly examine NXF1's role in the subcellular localization of RNAs, we depleted the protein from mESC by RNAi. SiRNAs targeting NXF1 transcripts reduced NXF1 protein levels by 70% (Figure 1A and 1B). This partial depletion led to a loss of cell proliferation, with increased cell numbers in the G2/M phase of the cell cycle (Figure 1C and 1D). These observations are consistent with results from NXF1 depletion by degron tagging, and from CRISPR screens indicating an essential role for NXF1 in cell growth, and indicate that full depletion of the protein may not be achievable without significant cell death (Aksenova et al. 2020; data from DepMap Portal [https://depmap.org/portal]). After siRNA treatment, we fractionated the cells into three compartments: chromatin, nucleoplasm and cytoplasm. The nucleoplasmic fraction contains nuclear material that is soluble in the nuclear lysis buffer containing urea and non-ionic detergent. The chromatin compartment contains insoluble buffer-resistant material that pellets with chromatin, including ternary RNA polymerase II complexes and nascent RNA, as well as molecules such as the nuclear speckle marker MALAT1 RNA (Figure 1E). Using α-tubulin, U1-70K and histone H3 as diagnostic protein markers for the cytoplasm, nucleoplasm and chromatin fractions, respectively, we confirmed the isolation of compartment-specific RNAs with minimal cross-contamination (Figure 1F). RNA was extracted from each fraction and polyA+ RNA was isolated for sequencing. RNA was converted into standard Illumina libraries and sequenced to identify RNAs whose subcellular localization is dependent on NXF1.

Transcripts encoding immediate early genes and transcription factors are among the most strongly dependent on NXF1 for nuclear export.

To examine the global effect of NXF1 depletion on RNA localization, we examined the ratios of read numbers between the chromatin, nuclear, or cytoplasmic compartments. We

defined the transcripts per million reads (TPM) value for each gene in each subcellular compartment, and calculated the ratios of the chromatin (CH) over cytoplasm (CY) read numbers. Since the complexity of the libraries differs between compartments, this is not an absolute measure of RNA ratio between the two compartments (Yeom et al. 2021). However, changes in this ratio provides a useful metric for the redistribution of RNAs between compartments. Comparing CH/CY values between control cells (siGFP) and after knockdown (siNXF1), we found that most transcripts showed less than two-fold changes. Transcripts of 60 genes were more than two-fold more enriched in the cytoplasm after NXF1 depletion, whereas 73 transcripts were more enriched in the chromatin and these tended to show larger changes than those shifted to the cytoplasm (Figure 2A). We also compared read numbers in the chromatin fraction to nucleoplasm fraction (CH/NP), as well as between nucleoplasm and cytoplasm (NP/CY) (Supplemental Fig. S1A). These comparisons yielded fewer transcripts shifting compartments after NXF1 depletion (Supplemental Fig. S1A). At this level of depletion, NXF1 is relatively selective in its effects on nucleocytoplasmic transport. It is likely that if NXF1 were depleted to lower levels, the effect mRNA export would become more profound, although we note that even at this level of depletion, cellular growth is strongly impacted. It is also possible that some mRNAs are exported through additional non-NXF1-dependent pathways, although many of them overlap with targets of other export factors TPR and GANP (Aksenova et al. 2020; Wickramasinghe et al. 2014).

We next examined whether the transcripts whose export was reduced by NXF1 depletion exhibited a sequence architecture that distinguishes them from other RNAs. Binning the genes into groups based on their fold change in CH/CY ratio after NXF1 depletion, we found transcripts showing a greater than 2 fold increase in this ratio arise from genes that are much shorter than average and have significantly fewer introns (Figure 2B). A similar enrichment of short genes with few introns was seen for transcripts that increase in CH/NP or NP/CY ratios after NXF1 depletion (Supplemental Fig. S1B), although these groups had many fewer genes.

Scatterplots displaying each transcript according to its change in CH/CY ratio and length or intron number are shown in Figure 2C. The majority of RNAs with a greater than two-fold change in CH/CY have three or fewer introns, with many having zero. These results are in agreement with previous findings, both for depletion of NXF1 and of other export factors such as TPR and GANP (Zuckerman et al. 2020; Lee et al. 2020; Wickramasinghe et al. 2014).

A change in CH/CY ratio after NXF1 depletion could be due to selective degradation of the RNA in the cytoplasm rather than a change in export. To assess this, we compared the total transcript abundance (TPM) in each subcellular compartment after NXF1 depletion compared to the control siRNA treatment. We found that the decrease in the cytoplasm TPM values is accompanied by an increase in TPM in the chromatin fraction (Supplemental Fig. S1C). Thus, the change in CH/CY ratio after NXF1 depletion is likely due to a loss of movement of RNA between fractions, although we can't rule out some contribution of RNA degradation in the cytoplasm (Wickramasinghe et al. 2014).

Short overall gene length and few introns are common characteristics of immediate early genes that are optimized for rapid expression in response to a stimulus (Herschman 1991; Bahrami and Drabløs 2016). To assess the types of genes most sensitive to loss of NXF1, we determined the gene ontology (GO) terms enriched in genes whose log 2 change in CH/CY was greater than or equal to 1 compared with all genes expressed in mESC (TPM > 1). Indeed, we found that the top GO term enriched in the NXF1 sensitive gene set was "DNA binding transcription factor activity" for RNA Pol II. The majority of the other enriched GO terms were also related to transcription (Figure 3A). Looking at specific genes within these groups, we found that polyA+ transcripts of Egr1, an immediate early response gene, strongly shifted to the chromatin fraction after NXF1 knockdown (Figure 3B). Transcripts of Sox2, a gene important for pluripotency in mESCs, also became more enriched in the chromatin (Figure 3B). Other transcription factor genes whose transcripts shifted from cytoplasm to chromatin after NXF1 depletion included immediate early genes Jun and Dusp6 (Supplemental Fig. S2A). Thus, we

find that mRNAs encoding transcription factors and particularly immediate early genes are among those most sensitive to the loss of NXF1.

iCLIP-seq to identify direct binding sites of NXF1 in the mESC transcriptome.

NXF1 can be directly recruited to mRNAs for export by specific RNA binding or be indirectly recruited via a variety of adapter proteins (Hocine, Singer, and Grünwald 2010; Müller-McNicoll et al. 2016). Binding to these adapters is thought to expose an RNA binding surface of the NXF1 protein that then binds non-sequence specifically to the RNA. This mode of binding may or may not allow efficient RNA/protein crosslinking by UV light. Nevertheless, there have been CLIP studies that have successfully identified NXF1 binding sites in mRNAs (Müller-McNicoll et al. 2016; Viphakone et al. 2019). To examine NXF1 binding to the NXF1 dependent transcripts identified in mESC, we performed iCLIP-seq in cells stably expressing a FLAG-tagged NXF1 gene introduced into a Flip-In locus at the COL1A1 gene (Figure 4A; Supplemental Fig. S3A). Some of the NXF1 dependent transcripts, such as Sox2, are well expressed in mESCs. Others such as Jun and Egr1 are not highly expressed unless the cells are stimulated to induce the immediate early response. To assay both sets of genes, we cultured the mESCs in the absence of serum and then performed iCLIP at 0 minutes and 60 minutes after serum addition, which stimulated expression of Jun, Egr1 and other genes (Figure 4A). Triplicate cultures of these cells were crosslinked and then subjected to a modified iCLIP protocol described previously (Damianov et al. 2016). RNA fragments interacting with NXF1 were recovered with FLAG antibodies and compared to fragments from isolated cells not expressing FLAG-tagged NXF1. RNAs were converted to Illumina sequencing libraries, sequenced, and after processing aligned to the mouse genome. This generated an average of 700,000 aligned reads per replicate from the no serum condition and 300,000 reads per replicate from serum stimulated cells. The RNA isolated on FLAG antibodies from cells not expressing FLAG-NXF1 yielded very low read numbers. This minimal background from

antibodies not binding to FLAG-NXF1, and from the gel isolation of RNA fragments indicates that virtually all the reads isolated from the FLAG-NXF1 expressing cells arise from RNA interactions of the FLAG-NXF1 protein.

Since NXF1 is thought to largely be recruited to RNA by adapter proteins and not direct RNA interactions, its crosslinking is likely to vary in efficiency between different targets. Indeed, the NXF1 bound fragments sometimes distributed across mRNAs rather than coalescing into a limited number of peaks. This may be indicative of more distributed, less sequence specific, binding to its targets. We found that the peaks of recovered RNA fragments were variable between conditions and replicates indicating that the recovery of binding sites was not saturating. To identify a set of confident binding sites for use in subsequent analyses, we compiled iCLIP peaks that were present in at least two replicates. The most consistent iCLIP peak was within the NXF1 transcript itself at the constitutive transport element, a known direct binding site within intron 10 (Li et al. 2006; Figure 4B). NXF1 binding sites were abundant in 5' and 3' UTRs and in coding regions, but relatively fewer were found in introns, consistent with the protein binding to mature mRNAs prior to export (Figure 4C). Peaks in the serum stimulated datasets were much sparser, but largely overlapped with the more numerous peaks from the unstimulated cells (Figure 4D).

An earlier study performed iCLIP analysis of GFP-NXF1 in mouse P19 embryonal carcinoma cells (Müller-McNicoll et al. 2016). A majority of the target genes in our analyses overlap with the targets identified in P19 cells, indicating good agreement between these two analyses (Supplemental Fig. S3B). GO analysis of our combined NXF1 targets yielded enriched terms related to translation and tRNA maturation, as well as terms for pluripotency and developmental processes (Supplemental Fig. S3C). These enrichments indicate that NXF1 mediated export may play a special role in maintaining mESC pluripotency.

Among transcripts affected by NXF1 depletion, Sox2 exhibited multiple NXF1 binding sites. These isolated fragments were distributed broadly across the Sox2 3' UTR, with additional

fragments in the coding region. The control IP produced virtually no apparent background binding. NXF1 bound fragments of SOX2 mRNA were also observed in the serum stimulated samples, but again were sparser (Figure 5A).

The Jun and Egr1 mRNAs are induced by serum. These showed essentially no reads in the unstimulated samples but NXF1 binding could be observed after stimulation that like Sox2 was distributed across the RNA (Supplemental Fig. S4A). Again samples from the cells not expressing FLAG tagged NXF1 generated no reads, so these reads appear to reflect true binding events. However, the limited numbers of reads in the Flag-NXF1 sample did not produce much overlap between replicates (Supplemental Fig. S4A). To confirm NXF1 binding to these RNAs, we repeated UV crosslinking and immunoprecipitation of the FLAG-NXF1-expressing mESCs and then performed RT-PCR to assay for Jun and Egr1 transcripts within the immunoprecipitated RNA. Krt17 RNA, which was not observed to bind NXF1 binding by iCLIP, served as a negative control (Figure 5B). These data confirmed that NXF1 binds to the Jun and Egr1 mRNAs after serum stimulation *in vivo* (Figure 5B). Despite the sparse coverage, the iCLIP-seq performed after serum stimulation allowed discovery of new NXF1 targets.

NXF1 controls expression of pluripotency factors in mESCs.

We found that the NXF1 targeted transcripts identified by iCLIP were enriched in factors affecting pluripotency, including Sox2 (Figure 5A). Sox2 is a key factor in the maintenance of pluripotency. Relatively small changes in Sox2 concentration determine whether the pluripotent state will be sustained or the cells will exit from it. We found that after NXF1 depletion, Sox2, Oct4 and Nanog all decreased relative to housekeeping factors such as GAPDH and U1-70K (Figure 6A; Supplemental Fig. S5A). These pluripotency factors were thus specifically sensitive to the loss of NXF1 (Figure 6A).

Another protein marker for naïve pluripotency in mESC is Rex1, which is repressed very early in differentiation and the exit from the pluripotent state. To examine the effect of NXF1

depletion on Rex1 expression, we used an engineered mESC line carrying a Rex1::GFPd2 (RGd2) reporter. In this cell line, one allele of Zfp42 (coding for Rex1) is replaced with a gene encoding a destabilized GFP (Kalkan et al. 2017). When pluripotency is compromised in response to cues such as removal of leukemia inhibitory factor (LIF), the loss of Rex1 can be monitored by GFP flow cytometry. We treated these RGd2 ESCs with NXF1 siRNA, and after 72 hours analyzed their GFP fluorescence by flow cytometry (Figure 6B). In agreement with the loss of other pluripotency factors, NXF1 knockdown led to dramatic loss GFP signal from the Rex1 reporter indicating an exit from naïve pluripotency (Figure 6B).

DISCUSSION

As part of a larger study assessing the role of RNA binding proteins in maintaining the chromatin association of nuclear RNA, we tested NXF1 as a factor that should affect the cytoplasmic concentration of mRNAs. NXF1 is widely studied as the major export factor for mRNAs, yet we found that upon 70% depletion of NXF1 from mESC most transcripts maintained their cytoplasmic concentrations. Despite a profound reduction in cell proliferation, only a subset of mRNAs were shifted to the nucleus and chromatin fractions by the loss of NXF1. It is likely that the unaffected RNAs are able to make use of the remaining NXF1 to sustain their nuclear export, and the proliferation defect indicates that it will be difficult to generate cells showing more complete depletion. The transcripts that are observed shifting to the chromatin fraction are presumably those most sensitive to the concentration of NXF1.

We find that in mESCs, transcripts of genes that are short and have few introns are most strongly regulated by NXF1. These properties have been previously in transcripts most affected by NXF1 depletion, as well as by depletion of export cofactors such as TPR and GANP (Zuckerman et al. 2020; Lee et al. 2020; Wickramasinghe et al. 2014). We note that this short gene architecture is commonly found in immediate early genes that are needed to rapidly respond to cellular stimuli. Indeed we find that immediate early genes and other transcription factor genes are enriched in the set most strongly affected by NXF1 depletion (Figure 2B; Figure 3A). These genes with short length and minimal splicing, along with induced transcription allow them to mediate rapid changes in gene expression in response to stimuli. Their unusual dependence on NXF1 levels indicates that rapid nuclear export also likely also contributes to their response to extracellular stimuli.

To confirm that NXF1 was mediating the export of immediate early genes and other transcription factors, we performed iCLIP-seq to identify NXF1 binding sites. Consistent with previous findings, NXF1 predominantly binds within exons (Müller-McNicoll et al. 2016; Viphakone et al. 2019). From our analysis, we find that binding primarily occurs in the 3’ UTR,

which often serves as a landing pad for diverse array of RNA binding proteins to regulate transcript expression at the post-transcriptional level. Moreover, GO analysis of NXF1 targets from 0m and 60m combined resulted in terms that are similar to some of the terms identified in the earlier studies of NXF1 targets. The differences in these lists likely result from different numbers of genes identified and different cell types analyzed. While showing reproduceable binding at the intron 10 of the NXF1 gene by all biological replicates, iCLIP reads were in general sparse. This indicates that NXF1 most likely binds to target RNAs in a non-specific manner, and its activity seems to be mostly from recruitment by export adapters that bind first and “hand-over” the mRNA to NXF1-NXT1 for exit through the nuclear pore complex (Hautbergue et al. 2008). It would be interesting to see whether NXF1 and other export factors such as TPR bind at the same sites of a transcript or co-regulate its export by CLIP strategy. If such sites are discovered, discovery of a common binding motif would reveal a conserved mechanism of RNA export.

Our flow cytometry results describe NXF1 as a major regulator of mESC pluripotency, with its depletion leading to substantial loss of Rex1 after 72 hours. This is consistent with depletion effects from other proteins involved in export, such as Thoc2 or Thoc5 of the THO complex (Wang et al. 2013). However, NXF1 mRNAs stay mostly unperturbed throughout the course of mESC differentiation, indicating that the export factor is most likely involved in localization of mRNAs encoding proteins important for both pluripotency and differentiation (Duren et al. 2020). An interesting experiment to further explore this idea would be to examine transcripts that associate with NXF1 throughout the course of differentiation.

MATERIALS AND METHODS

Cell lines and tissue culture

The E14 mouse embryonic stem cell (mESC) line (Hooper et al. 1987) was thawed and cultured on mitotically-inactivated mouse embryonic fibroblasts (MEFs) (CF1, Applied StemCell, Inc.) in tissue culture wells and plates that were coated with 0.1% gelatin. Prior to performing any experiments, the mESCs were passaged on MEFs at least twice before culturing them on gelatin-coated plates without MEFs. mESC media consisted of DMEM (Fisher Scientific) supplemented with 15% ESC-qualified fetal bovine serum (Thermo Fisher Scientific), 1x non-essential amino acids (Thermo Fisher Scientific), 1x GlutaMAX (Thermo Fisher Scientific), 1x ESC-qualified nucleosides (EMD Millipore), 0.1 mM β-Mercaptoethanol (Sigma-Aldrich), and 10^3 units/ml ESGRO leukemia inhibitor factor (LIF) (EMD Millipore). The same culture conditions were used for maintaining mESCs containing a FRT site for Flp recombinase-mediated DNA recombination and generation of a cell line expressing 1x FLAG-tagged NXF1. The wildtype "Flp-In mESCs" were a kind gift from Kathrin Plath at the University of California, Los Angeles. Rex1GFP mESCs were a kind gift from Kathrin Plath, and the media consisted of the N2B27 base (47.4% DMEM/F12 (v/v) (Thermo Fisher), 47.4% Neurobasal (v/v) (Thermo Fisher), 0.95x N-2 supplement (Thermo Fisher Scientific), 0.95x B-27 supplement, minus Vitamin A (Thermo Fisher Scientific), 0.95x GlutaMAX (Thermo Fisher Scientific), 0.95x non-essential amino acids (Thermo Fisher Scientific), 0.95x Pen/Strep (Thermo Fisher Scientific), 0.1 mM β-Mercaptoethanol (Sigma-Aldrich)) supplemented with 2iL (3 μM CHIR99021 (Tocris), 0.4 μM PD0325901 (Tocris), 10^3 units/ml ESGRO LIF (EMD Millipore)).

Knockdown of NXF1 in mESCs

siRNAs that target NXF1 (s79093, Thermo Fisher Scientific), eGFP (Silencer Select AM4626), and random sequences (AM4611, Thermo Fisher Scientific) were transfected into mESCs twice using Lipofectamine RNAiMAX (Invitrogen). The first transfection was performed

while the cells were in suspension, and the second transfection was performed 24 hours after when the cells had already attached to the surface of the cell culture dish. The cells were harvested and fractionated 24 hours after the second transfection.

Subcellular fractionation, RNA extraction and library preparation

Total RNA from mESCs after siRNA transfections was isolated from the cytoplasmic, soluble nuclear, and chromatin pellet compartments as described previously (Pandya-Jones and Black 2009; Wuarin and Schibler 1994; Yeom and Damianov 2017; Yeom et al. 2021). Briefly, the cells were washed twice with 1X PBS/1 mM EDTA (pH 8.0) and trypsinized to collect the cell pellet by centrifugation. 0.5 - 1 X 10^7 cells from the pellet were transferred into a 2.0 mL low adhesion microcentrifuge tube (USA Scientific) and washed twice with 1X PBS/1 mM EDTA (pH 8.0) to remove debris from prematurely-lysed cells. The cells were then incubated first in ice-cold hypotonic buffer (10 mM Tris–HCl [pH 7.5], 15 mM KCl, 1 mM EDTA [pH 8.0], 0.15 mM Spermine, 0.5 mM Spermidine, 0.5 mM DTT, and 1X Protease inhibitor) for 5 minutes and lysed in ice-cold lysis buffer (10 mM Tris–HCl [pH 7.5], 15 mM KCl, 1 mM EDTA [pH 8.0], 0.15 mM Spermine, 0.5 mM Spermidine, 0.5 mM DTT, and 1X Protease inhibitor, and 0.0375% Igepal CA-630) for 1 minute. The resulting lysate was immediately layered on top of a chilled sucrose cushion (hypotonic buffer with 24% (w/v) sucrose without detergent) and centrifuged for 10 minutes, 4 °C, 6000 x g. 10% of the supernatant, which is the cytoplasmic fraction, was used for immunoblot to check for contamination from nuclear materials, and the rest was mixed with Trizol LS (Invitrogen) to extract cytoplasmic RNA. The nuclear pellet was washed once with 1X PBS/1 mM EDTA (pH 8.0) before being resuspended in chilled glycerol buffer (20mM Tris-HCl [pH 7.9], 75 mM NaCl, 0.5 mM EDTA [pH 8.0], 50% glycerol (v/v), 0.85 mM DTT, and 0.125 mM PMSF). Equal volume of cold nuclear lysis buffer (20 mM HEPES [pH 7.6], 7.5 mM MgCl2, 0.2 mM EDTA [pH 8.0], 300 mM NaCl, 1M urea, 1% Igepal CA-630, 0.3 mM Spermine, 1.0 mM Spermidine, 1× Protease inhibitor) was then added and the mixture incubated on ice for 2

minutes after 2 X 2 s of vortexing. After the 2 minutes incubation period, 1/8 of the nuclear lysate was transferred to a separate 1.5 mL microcentrifuge tube for each sample. After centrifugation for 2 minutes, 4 °C, 6000 x g, the supernatant (soluble nuclear fraction) from both tubes were pooled. 10% of the pooled supernatant was used to check its purity by western blot, and the rest was mixed with Trizol LS to extract nucleoplasmic RNA. The resulting two insoluble nuclear pellets (one large, one small) were washed once with 1X PBS/1 mM EDTA (pH 8.0) before incubation of the large pellet in Trizol at 50 °C until the pellet was completely solubilized to extract chromatin-associated RNA (from 7/8 nuclear lysate). After washing, the small pellet (from 1/8 nuclear lysate) was resuspended in 5% SDS sample buffer (5% SDS, 30% glycerol (v/v) and 150 mM Tris–HCl [pH 8.0]) at 55 °C for 15 minutes to extract proteins from the chromatin fraction and checked for purity. Each fractionation experiment was performed in triplicate.

RNA from each fraction was treated with DNase I (Takara) followed by phenol extraction. Following QC, poly(A)+ libraries were generated using the TruSeq Stranded mRNA Library Prep Kit (Illumina). Note that the sum of the RNAs from the chromatin, nucleoplasmic and cytoplasmic fractions does not represent total cellular RNA when combined due to extensive washing steps during fractionation.

Immunoblotting

To check for success of subcellular fractionation, protein samples from each fraction were loaded on a 4-12% Bis-Tris SDS-PAGE gel (NuPAGE) by equal fractions of their original volume. Transfer to a PVDF membrane was conducted using the Trans-Blot Semi-Dry transfer apparatus (Bio-Rad). After incubation of the membrane with primary and fluorescent (Cy3 or Cy5) secondary antibodies, images were taken on a Typhoon Imager (GE Healthcare). For detection using HRP-conjugated secondary antibodies, LAS-3000 Imaging System (Fuji) and iBright 1500 (Invitrogen) were used. To check for success of NXF1 depletion and abundance of

mESC pluripotency factors, protein samples for western blotting were prepared from whole cell lysates using RIPA buffer and sonication of the insoluble chromatin using Bioruptor Pico Plus sonicator (Diagenode). Quantification of the lysates was performed using Pierce BCA Protein Assay Kit (Thermo Fisher Scientific). Antibodies used in this study are listed in Table 1.

mESC proliferation assay

Throughout the course of the 48 hours NXF1 knockdown, mESCs at hour 0, 24 and 48 were extracted and their concentrations determined using Countess II Automated Cell Counter (Thermo Fisher Scientific).

Gene ontology analysis

Gene ontology analysis was performed using PANTHER (Protein ANalysis Through Evolutionary Relationships) (Ashburner et al. 2000) and Metascape (Zhou et al. 2019). For accurate gene enrichment analysis, all genes expressed in the mESCs with a TPM > 1 were utilized as background gene list.

Preparation of iCLIP-seq libraries after serum stimulation

mESCs expressing 1X FLAG-tagged NXF1 were plated on tissue culture plates with mESC media overnight. Following their attachment, the cells were incubated in mESC media excluding serum for 24 hours. For “0m” samples, the cells were irradiated and harvested after the 24 hours of serum starvation. For “60m” samples, the cells were incubated in mESC media with 15% serum post-starvation for an hour, irradiated and harvested. The harvested cells were then processed for iCLIP-seq library preparation as previously described (Damianov et al. 2016).

Crosslinking RNA immunoprecipitation

Cells were UV crosslinked and harvested. After complete lysis, the cell lysate was used to perform immunoprecipitation using anti-FLAG antibodies immobilized to Dynabeads. After 5X washing of the immunoprecipitated FLAG-NXF1-RNA complexes with WB_{750} from the iCLIP protocol, they were washed twice with WB_{150}. The complexes were eluted using 30 ul of the Dynabeads elution buffer (100 mM Tris pH 7.5, 0.6% SDS, 5 mM EDTA, 50 mM DTT, 50 ng/ul YtRNA) at 85°C for 10 minutes with constant shaking. To digest the associated proteins, the eluted complexes were incubated in deproteinization buffer (270 ul containing 100 mM Tris pH 7.5, 55.56 mM NaCl, 10.56 mM EDTA, 2.22 ug/ul Proteinase K) at 55°C for 30 minutes. The same procedure was done for input by adding the buffer until 300 ul. After pre-incubation of the iCLIP PK-urea buffer (100 mM Tris-HCl pH 7.5, 50 mM NaCl, 10 mM EDTA, 7 M Urea, and 0.5 µg/µl proteinase K) for 2 min at 55 ˚C, 300 ul of the buffer was added to the samples to the final volume of 600 ul and incubated at 55°C for an additional 30 minutes. Following precipitation with acid phenol:chloroform and isopropanol overnight, the RNA-protein fragments were resuspended in 8.2 ul RNase-free water, and 8.0 ul was used to perform DNA digestion with 1 ul of DNase I (Ambion) and 1 ul of its 10X buffer in a PCR tube at 37 °C for 20 minutes. At the end of the reaction, 1 ul of 35 mM EDTA was added and the reaction incubated at 65 °C for 10 minutes to inactivate the DNase I. With the resulting volume of 11 ul, reverse transcription with SuperScript III (Invitrogen) was performed with the final reaction volume of 20 ul with random hexamers and manufacturer-recommended thermocycler settings. 1 – 2 ul of the cDNA was used to amplify transcripts crosslinked to FLAG-NXF1.

Flow cytometry of Rex1GFP mESCs and WT mESCs after NXF1 knockdown

For cell cycle analysis, mESCs underwent 48 hours of NXF1 knockdown as described above. For GFP flow cytometry, the cells underwent 72 hours of knockdown with replacement of media at 48 hours with fresh mESC media. After the knockdown, the cells were washed once with PBS and harvested by incubation for 5 minutes at 37 °C with Accutase (Innovative Cell

Technologies). The cell pellet was washed twice with cold PBS before proceeding to flow cytometry using the LSR II Flow Cytometer (BD Biosciences) and analysis using the FACSDiva software (BD Biosciences).

RNA sequencing and alignment.

The libraries were subjected to 75 nt paired-end sequencing at the UCLA Neuroscience Genomics Core on an Illumina HiSeq 4000 to generate about 25 million mapped reads per sample (each fraction constitutes a sample). RNA-seq data was aligned using STAR (version 020210) (Dobin et al. 2013) on mm10. Sequencing reads and mapping results are summarized in Table 2. Reads per million read values (RPM) were calculated and displayed in UCSC Genome Browser sessions including strand information. TPM values obtained from Kallisto (version 0.43.0) were used in gene expression analyses (Bray et al. 2016).

Ratio analysis

To analyze differential compartmentalization of RNAs, genes were selected that had chromatin, nucleoplasm and cytoplasm expression greater or equal to TPM of 1 as reported by Kallisto and had read counts greater than zero in the cytoplasmic fractions as measured by FeatureCount. DESeq2 was used to measure fold change in read counts between two fractions by calculating the average read count among replicates of one fraction divided by the average read counts of the other fraction. The ratios were log 2 transformed and analyzed using ggplot2 (Figure 2A).

iCLIP-seq data analysis

The libraries were subjected to single-end 100 bp sequencing on a NovaSeq 6000 (Illumina) at the UCLA Broad Stem Cell Research Center. Sequencing reads and mapping results are summarized in Table 3. Data analyses were performed with iCount with few

modifications. In brief, PCR duplicate iCLIP reads were removed using random barcodes. Unique reads were mapped to mm10 using STAR, allowing two mismatches. Mapped reads were assigned to the longest transcripts in the Known Gene table (Hsu et al., 2006) and divided into 5′ UTR, CDS, intron, and 3′ UTR regions.

Table 1

Antibodies used in this study		
Target	**Antibody name (host/source)**	**Detection method**
TUBA1A	alpha-tubulin DM1A (mouse/Millipore, CP06)	Fluorescence
SNRNP70	SNRNP70 (rabbit/DB lab made)	Fluorescence
Histone H3.1	Histone H3 (rabbit/Abcam, ab1791)	Fluorescence
GAPDH	GAPDH 6C5 (mouse/Thermo Fisher Scientific, AM4300)	Fluorescence
NXF1	Anti-NXF1 antibody [EPR8010] (rabbit/Abcam, ab129171)	Chemiluminescence
FLAG tag	Anti Flag M2 (mouse/Sigma-Aldrich, F3165)	Immunoprecipitation
Nanog	eBioMLC51 (rat/Invitrogen, 14-5761-80)	Chemiluminescence
OCT4	OCT4 3H8L6 (rabbit/Invitrogen, 701756)	Fluorescence
Sox2	MAB2018 (mouse/R&D Systems, MAB2018)	Chemiluminescence
Rabbit IgG	Anti-Rabbit IgG Antibody (Cy3®) (goat/GE Healthcare, 95040-030)	
Mouse IgG	Anti-Mouse IgG Antibody (Cy3®) (goat/GE Healthcare, 95040-042)	
Rabbit IgG	Anti-Rabbit IgG Antibody (Cy5®) (goat/GE Healthcare, 95040-050)	
Mouse IgG	Anti-Mouse IgG Antibody (Cy5®) (goat/GE Healthcare, 95040-046)	
Rabbit IgG	ECL Rabbit IgG, HRP-linked whole Ab (donkey/GE Healthcare, NA934)	
Mouse IgG	ECL Mouse IgG, HRP-linked whole Ab (sheep/GE Healthcare, NXA931)	
Primers used in this study		
Target	**Forward primer sequence (5' - 3')**	**Reverse primer sequence (5' - 3')**
Egr1	TATGAGCACCTGACCACAGAG	GCTGGGATAACTCGTCTCCA
Jun	CCTTCTACGACGATGCCCTC	GGTTCAAGGTCATGCTCTGTTT
Krt17	ACCATCCGCCAGTTTACCTC	CTACCCAGGCCACTAGCTGA

Table 2

	siGFP CH	siGFP NP	siGFP CY	siNXF1 CH	siNXF1 NP	siNXF1 CY
Number of input reads	81,327,245	75,047,851	71,520,510	74,195,330	77,824,090	71,712,147
Average input read length	150	150	150	150	150	150
Uniquely mapped reads number	69,424,413	62,995,174	60,231,857	63,589,718	65,942,922	60,234,606
Uniquely mapped reads %	85.36	83.94	84.22	85.71	84.73	83.99
Average mapped length	150	150	150	150	150	150
Number of splices: Total	22,606,758	36,445,083	41,417,459	17,749,105	38,486,515	41,838,478
Number of splices: Annotated (sjdb)	22,292,046	36,206,442	41,184,014	17,393,400	38,213,414	41,585,085
Number of splices: GT/AG	22,303,052	35,995,463	40,879,250	17,503,569	38,022,481	41,310,163
Number of splices: GC/AG	205,254	360,049	443,796	158,583	371,311	432,720
Number of splices: AT/AC	26,707	42,911	50,333	20,691	46,083	51,059

Table 3

	FLAG-NXF1 0m rep 1	FLAG-NXF1 0m rep 2	FLAG-NXF1 0m rep 3	No FLAG 0m rep 1	No FLAG 0m rep 2	No FLAG 0m rep 3	FLAG-NXF1 60m rep 1	FLAG-NXF1 60m rep 2	FLAG-NXF1 60m rep 3	No FLAG 60m rep 1	No FLAG 60m rep 2	No FLAG 60m rep 3
sequnce data			469065936							452285558		
original data	26101531	26135019	20169006	3465142	3406996	2525354	7642424	14803730	15510097	3379789	3146550	1867974
unique read for mapping	1344822	1807910	1074543	133073	168480	118816	399280	969990	798028	174439	202156	93533
star alignment												
Number of input reads	1344822	1807910	1074543	133073	168480	118816	399280	969990	798028	174439	202156	93533
Average input read length	40	41	41	43	45	46	43	43	41	42	41	42
UNIQUE READS:												
Uniquely mapped reads number	634909	908317	502408	42809	42158	30610	151669	391021	321483	51407	54389	23197
Uniquely mapped reads %	47.21%	50.24%	46.76%	32.17%	25.02%	25.76%	37.99%	40.31%	40.28%	29.47%	26.90%	24.80%
Average mapped length	39.13	40.97	40.28	38.41	41.5	41.85	42.17	42.48	39.89	39.98	38.68	38.16
Number of splices: Total	16207	21072	15873	1386	1662	1240	4671	11383	10595	2274	2025	781
Number of splices: Annotated (sjdb)	10092	13913	11073	730	876	604	3016	7716	6729	1332	683	286
Number of splices: GT/AG	14663	19021	14484	1253	1528	1117	4247	10293	9574	2074	1831	650
Number of splices: GC/AG	462	468	354	59	58	40	171	268	275	93	87	83
Number of splices: AT/AC	27	44	36	0	0	1	10	17	23	2	1	0
Number of splices: Non-canonical	1055	1539	999	74	76	82	243	805	723	105	106	48
Mismatch rate per base, %	2.57%	2.16%	2.45%	3.08%	2.88%	2.86%	2.81%	2.62%	2.81%	3.05%	2.95%	3.15%
Deletion rate per base	0.07%	0.09%	0.10%	0.03%	0.06%	0.08%	0.08%	0.09%	0.07%	0.05%	0.07%	0.03%
Deletion average length	1.78	1.73	1.77	1.7	1.64	1.55	1.77	1.68	1.69	1.58	1.82	1.79
Insertion rate per base	0.01%	0.01%	0.01%	0.01%	0.00%	0.01%	0.01%	0.01%	0.01%	0.02%	0.01%	0.01%
Insertion average length	1.2	1.37	1.24	2.21	1.12	1.82	1.23	1.28	1.23	2.07	1.22	1.09
MULTI-MAPPING READS:												
Number of reads mapped to multiple loci	0	0	0	0	0	0	0	0	0	0	0	0
% of reads mapped to multiple loci	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Number of reads mapped to too many loci	171144	216747	127169	11475	10333	7602	37747	91245	83016	12380	14028	5741
% of reads mapped to too many loci	12.73%	11.99%	11.83%	8.62%	6.13%	6.40%	9.45%	9.41%	10.40%	7.10%	6.94%	6.14%
UNMAPPED READS:												
Number of reads unmapped: too many mismatches	221423	276682	173085	21484	22126	17040	74201	165060	129055	29083	26385	11947
% of reads unmapped: too many mismatches	16.46%	15.30%	16.11%	16.14%	13.13%	14.34%	18.58%	17.02%	16.17%	16.67%	13.05%	12.77%
Number of reads unmapped: too short	273897	336190	228351	44642	72411	48543	117374	276918	229685	67985	90734	43436
% of reads unmapped: too short	20.37%	18.60%	21.25%	33.55%	42.98%	40.86%	29.40%	28.55%	28.78%	38.97%	44.88%	46.44%
Number of reads unmapped: other	43449	69974	43530	12663	21452	15021	18289	45746	34789	13584	16620	9212
% of reads unmapped: other	3.23%	3.87%	4.05%	9.52%	12.73%	12.64%	4.58%	4.72%	4.36%	7.79%	8.22%	9.85%
CHIMERIC READS:												
Number of chimeric reads	0	0	0	0	0	0	0	0	0	0	0	0
% of chimeric reads	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

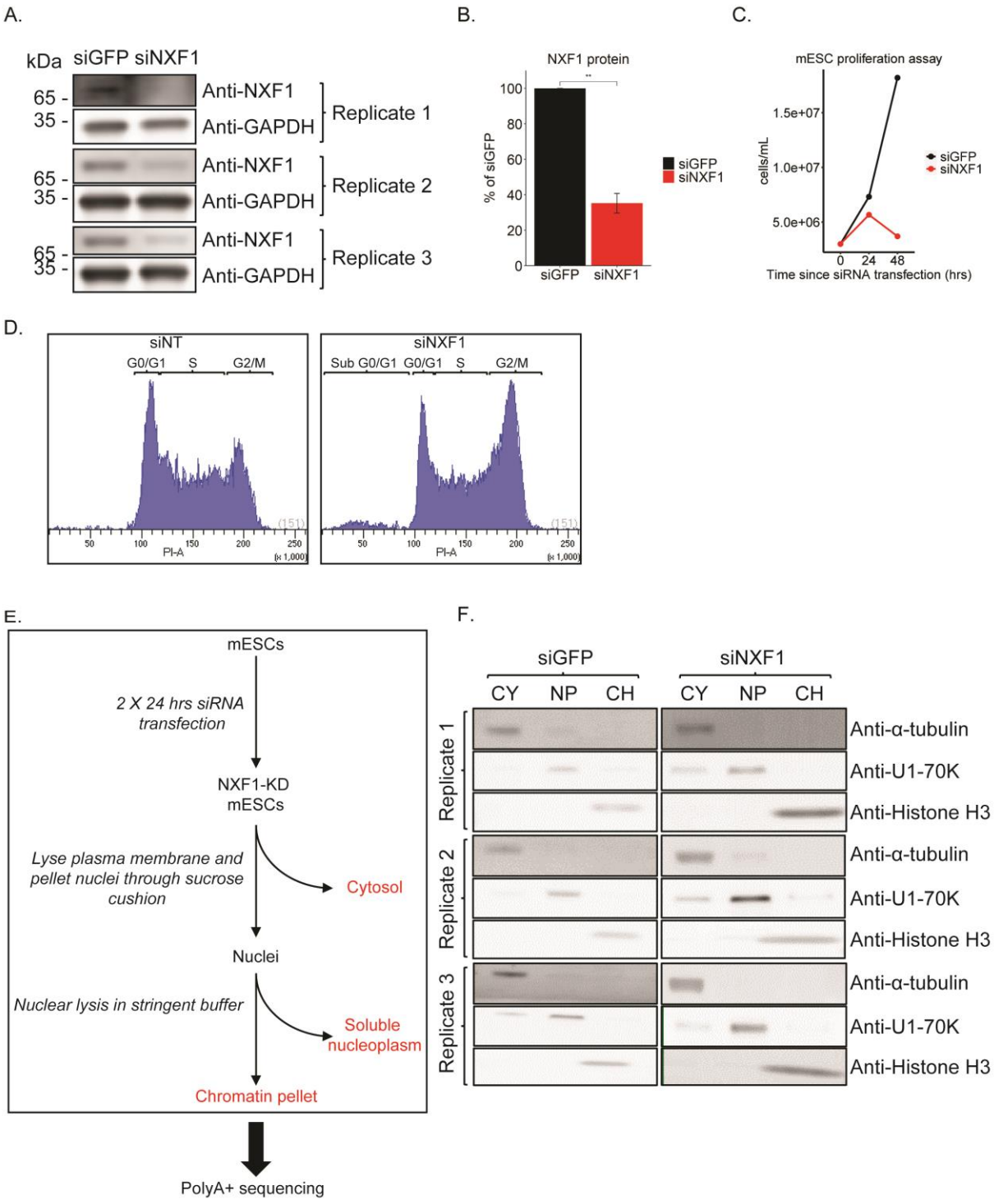

A.
kDa siGFP siNXF1
65 -
35 -
Anti-NXF1
Anti-GAPDH
Replicate 1
Replicate 2
Replicate 3
B.
NXF1 protein
% of siGFP
siGFP
siNXF1
C.
mESC proliferation assay
cells/mL
Time since siRNA transfection (hrs)
D.
siNT
G0/G1 S G2/M
siNXF1
Sub G0/G1 G0/G1 S G2/M
PI-A
E.
mESCs
2 X 24 hrs siRNA transfection
NXF1-KD mESCs
Lyse plasma membrane and pellet nuclei through sucrose cushion
Cytosol
Nuclei
Nuclear lysis in stringent buffer
Soluble nucleoplasm
Chromatin pellet
PolyA+ sequencing
F.
siGFP
siNXF1
CY NP CH
Anti-α-tubulin
Anti-U1-70K
Anti-Histone H3

Figure 1.

Isolation of RNAs from three subcellular compartments after NXF1 knockdown

A. Immunoblot showing depletion of NXF1 in the three biological replicates.

B. Quantification of A as a percentage of control. For siNXF1, mean NXF1 level is plotted. Error bar is SEM; Student's t-test; **: $p \leq 0.01$.

C. Cell proliferation assay of mESCs over the course of NXF1 depletion.

D. Cell cycle analysis of NXF1-depleted mESCs using flow cytometry.

E. Experimental outline for extracting chromatin-associated, nucleoplasmic and cytoplasmic RNAs from NXF1-depleted mESCs for polyA+ RNA sequencing. The stringent buffer was composed of high salt, urea and non-ionic detergent.

F. Immunoblot showing fractionation quality of GFP- (negative control) and NXF1-depleted cells using antibodies against diagnostic proteins for each compartment (α-tubulin for cytoplasm, U1-70K for nucleoplasm, and histone H3 for chromatin). CY = cytoplasm, NP = nucleoplasm, CH = chromatin.

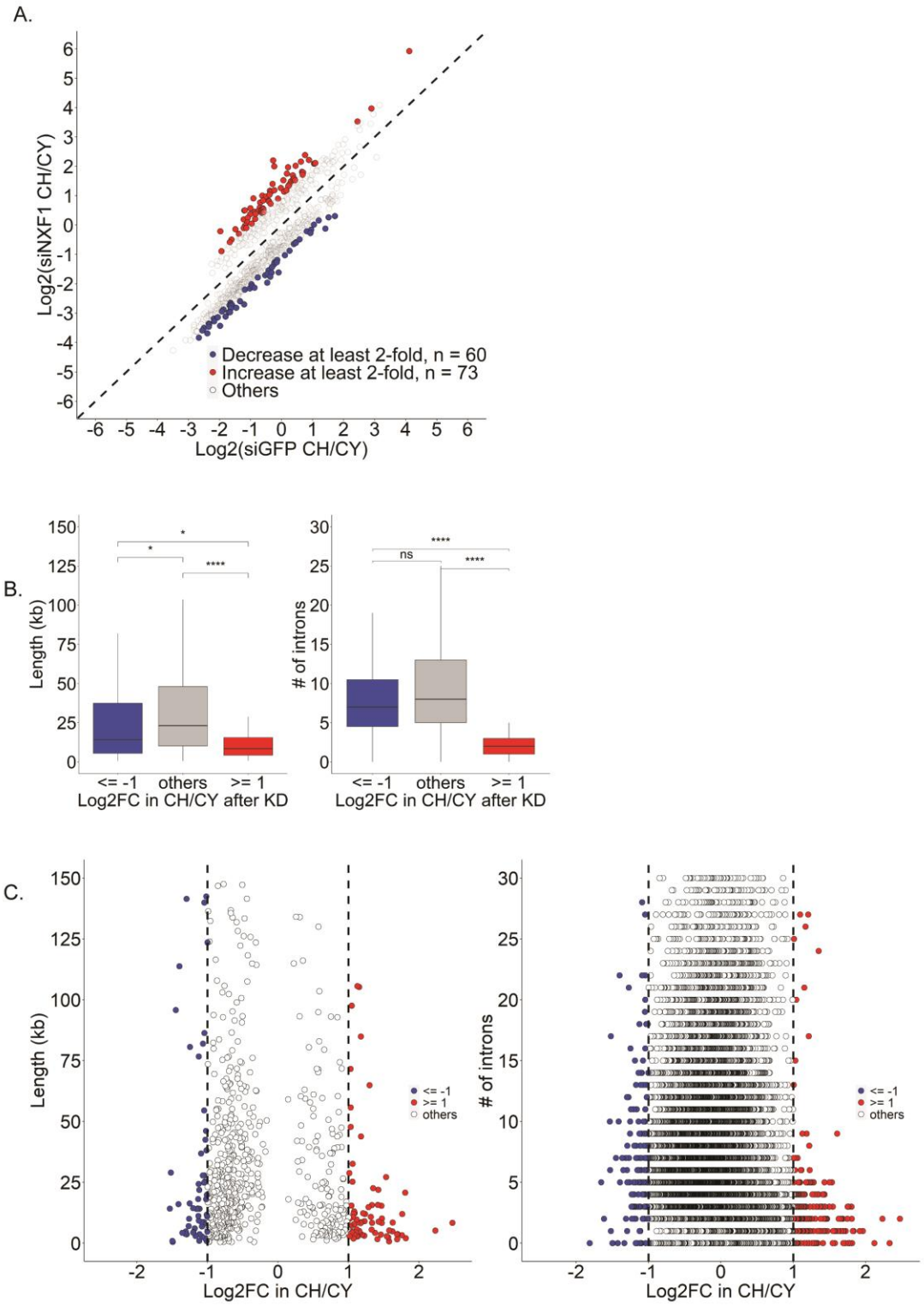

A.
Log2(siNXF1 CH/CY)
Log2(siGFP CH/CY)
Decrease at least 2-fold, n = 60
Increase at least 2-fold, n = 73
Others
B.
Length (kb)
# of introns
<= -1
others
>= 1
Log2FC in CH/CY after KD
ns
C.
Length (kb)
# of introns
Log2FC in CH/CY
<= -1
>= 1
others

Figure 2.

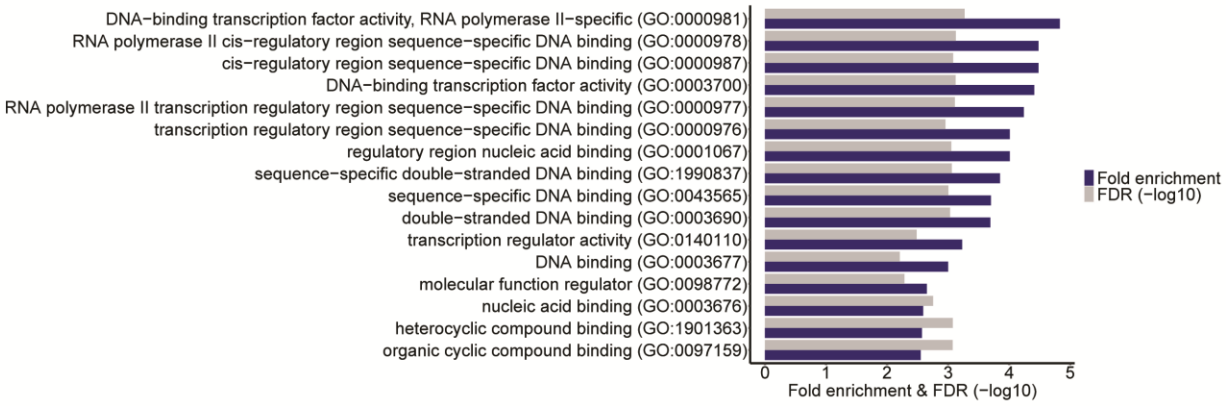
NXF1 most strongly regulates export of RNAs that are short and have low number of introns

A. Scatterplot depicting relationship between CH/CY values of siGFP and siNXF1 (log2-transformed). Each gene has a CH/CY value of before (siGFP) and after (siNXF1) protein depletion. Red-colored points are genes that become at least 2-fold more chromatin-associated after knockdown, while blue-colored points are genes that become at least 2-fold more cytoplasm-enriched after knockdown. Only genes that have change in the ratio with $p < 0.05$ are plotted.

B. Box plots showing relationship between log-2 fold change in CH/CY after NXF1 knockdown and length or number of introns of the respective genes. The upper and lower hinges displayed represent the 25th and 75th percentiles; Wilcoxon rank-sum test; ns: $p > 0.05$, *: $p \leq 0.05$, ****: $p \leq 0.0001$. Only genes that have change in the ratio with $p < 0.05$ are plotted.

C. Log-2 fold change of all genes from the ratio analysis (Figure 2B) is plotted against their number of introns or length.

A.


DNA−binding transcription factor activity, RNA polymerase II−specific (GO:0000981)
RNA polymerase II cis−regulatory region sequence−specific DNA binding (GO:0000978)
cis−regulatory region sequence−specific DNA binding (GO:0000987)
DNA−binding transcription factor activity (GO:0003700)
RNA polymerase II transcription regulatory region sequence−specific DNA binding (GO:0000977)
transcription regulatory region sequence−specific DNA binding (GO:0000976)
regulatory region nucleic acid binding (GO:0001067)
sequence−specific double−stranded DNA binding (GO:1990837)
sequence−specific DNA binding (GO:0043565)
double−stranded DNA binding (GO:0003690)
transcription regulator activity (GO:0140110)
DNA binding (GO:0003677)
molecular function regulator (GO:0098772)
nucleic acid binding (GO:0003676)
heterocyclic compound binding (GO:1901363)
organic cyclic compound binding (GO:0097159)
0
1
2
3
4
5
Fold enrichment & FDR (−log10)
Fold enrichment
FDR (−log10)

B.

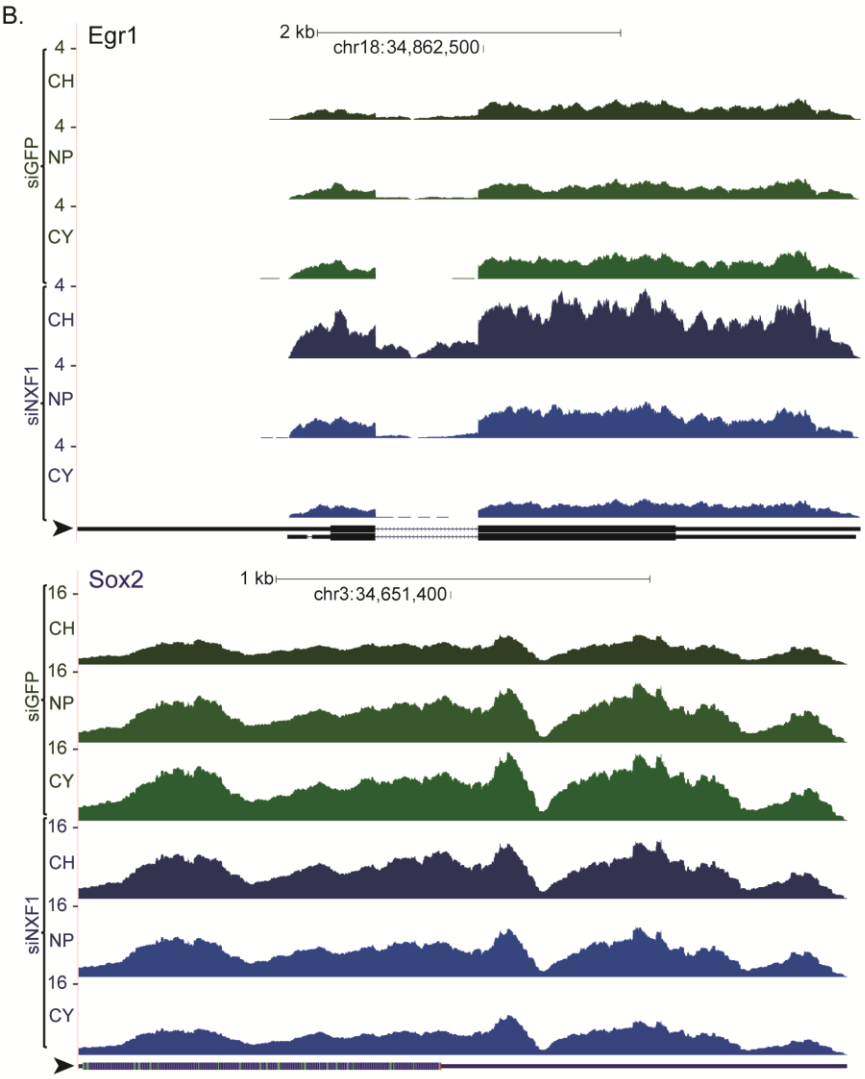

Egr1
2 kb
chr18:34,862,500
siGFP
siNXF1
CH
NP
CY
4
Sox2
1 kb
chr3:34,651,400
16

Figure 3.

NXF1 selectively regulates export of transcripts encoding DNA binding proteins

A. Gene ontology analysis of genes that become at least 2-fold chromatin-enriched over cytoplasm after knockdown using PANTHER for molecular function. Results only display terms with a FDR of less than 0.05.

B. UCSC Genome Browser sessions showing chromatin enrichment of Egr1 (top) and Sox2 (bottom) transcripts after NXF1 depletion. Reads in green represent those from the control libraries (siGFP) while ones in blue represent those from the knockdown libraries (siNXF1). Black arrows next to the gene body diagrams indicate direction of transcription. Read numbers are in RPM.

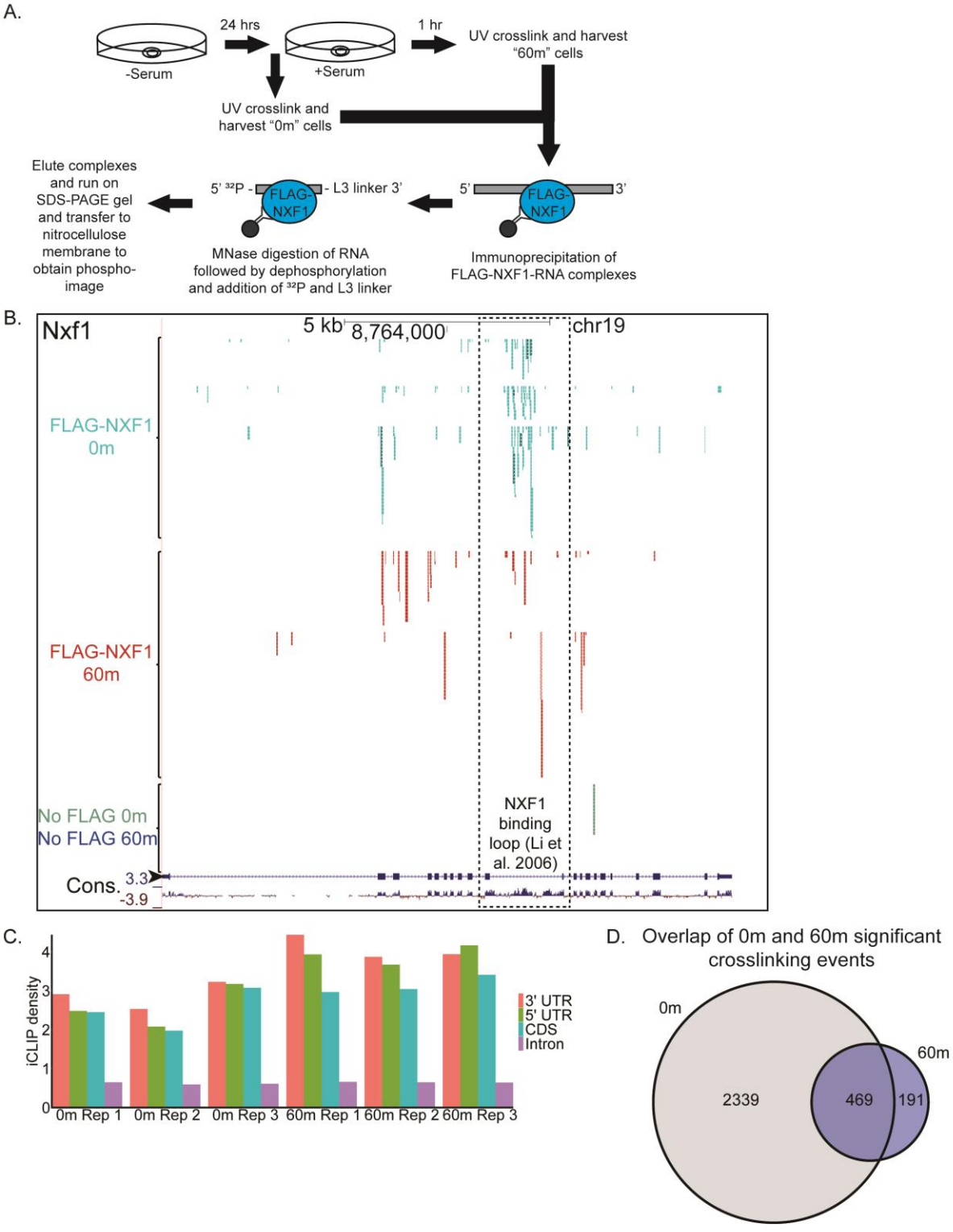

A.
24 hrs
1 hr
-Serum
+Serum
UV crosslink and harvest "60m" cells
UV crosslink and harvest "0m" cells
Elute complexes and run on SDS-PAGE gel and transfer to nitrocellulose membrane to obtain phospho-image
5' 32P -
FLAG-NXF1
- L3 linker 3'
5'
3'
MNase digestion of RNA followed by dephosphorylation and addition of 32P and L3 linker
Immunoprecipitation of FLAG-NXF1-RNA complexes
B.
Nxf1
5 kb
8,764,000
chr19
FLAG-NXF1 0m
FLAG-NXF1 60m
No FLAG 0m
No FLAG 60m
NXF1 binding loop (Li et al. 2006)
Cons.
3.3
-3.9
C.
iCLIP density
0
1
2
3
4
0m Rep 1
0m Rep 2
0m Rep 3
60m Rep 1
60m Rep 2
60m Rep 3
3' UTR
5' UTR
CDS
Intron
D. Overlap of 0m and 60m significant crosslinking events
0m
60m
2339
469
191

Figure 4.

iCLIP-seq to interrogate NXF1 binding sites in the mESC transcriptome

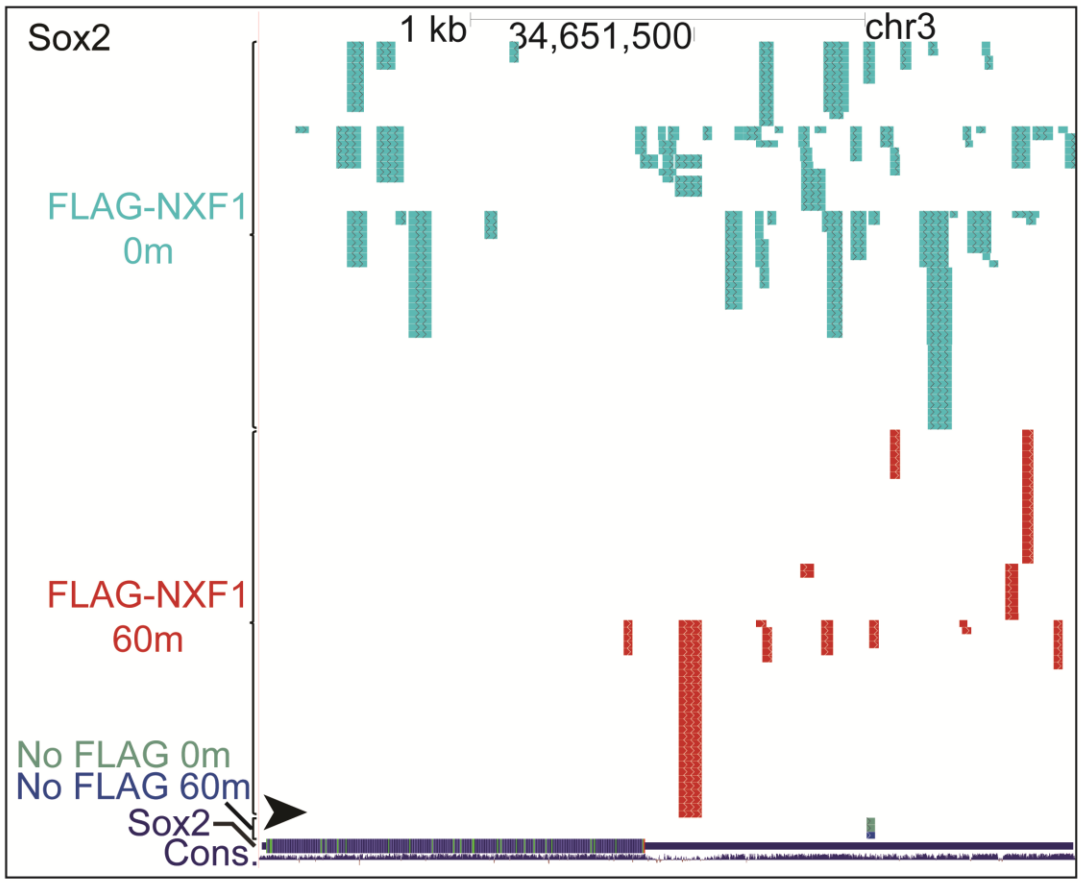
A. Workflow of iCLIP-seq library preparation. Initially, cells were passaged onto gelatin-coated plates without feeder cells in mESC medium. After attachment, the cells were starved from serum for 24 hours. Following serum depletion, “0m” cells were UV-crosslinked and harvested for immunoprecipitation, while “60m” cells were incubated in serum+ medium for one hour until crosslinking, harvesting and immunoprecipitation.

B. UCSC Genome Browser showing FLAG-NXF1 iCLIP peaks at intron 10 of the NXF1 transcript (Li et al., 2006). Black arrow indicates direction of transcription.

C. Crosslinking events from iCLIP-seq for each sample (0m and 60m, both in triplicate) are displayed after normalization using total feature length and library size.

D. Venn diagram showing significant events that overlap between 0m and 60m.

A.


Sox2
1 kb
34,651,500
chr3
FLAG-NXF1
0m
FLAG-NXF1
60m
No FLAG 0m
No FLAG 60m
Sox2
Cons.

B.

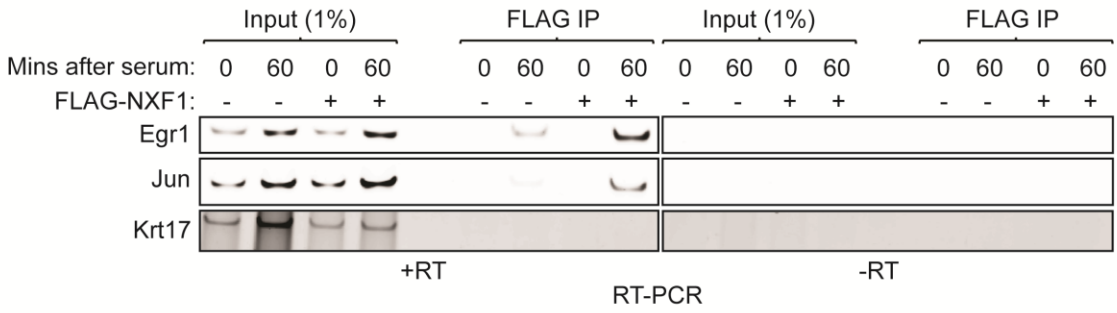
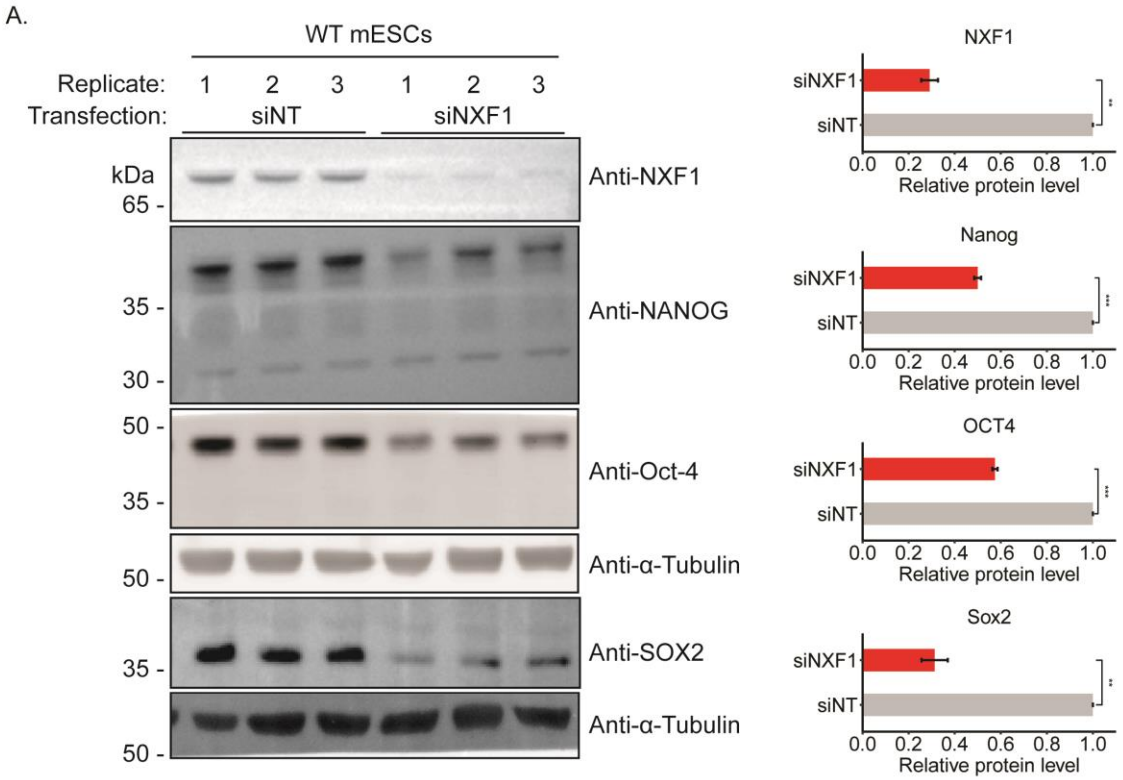

Input (1%)
FLAG IP
Input (1%)
FLAG IP
Mins after serum: 0 60 0 60 0 60 0 60 0 60 0 60 0 60 0 60
FLAG-NXF1: - - + + - - + + - - + + - - + +
Egr1
Jun
Krt17
+RT
-RT
RT-PCR

Figure 5.

Transcription induction of genes by serum withdrawal followed by its reintroduction allows for discovery of novel NXF1 targets

A. UCSC Genome Browser session of Sox2 from FLAG-NXF1 iCLIP-seq. Black arrow indicates direction of transcription.

B. RT-PCR polyacrylamide gel showing amplified Egr1, Jun and Krt17 (negative control) mRNAs using cDNAs from RNAs extracted after UV crosslinking and immunoprecipitation with anti-FLAG antibodies against FLAG-NXF1.


A.
WT mESCs
Replicate: 1 2 3 1 2 3
Transfection: siNT siNXF1
kDa
65 -
35 -
30 -
50 -
35 -
50 -
35 -
50 -
Anti-NXF1
Anti-NANOG
Anti-Oct-4
Anti-α-Tubulin
Anti-SOX2
Anti-α-Tubulin
NXF1
Nanog
OCT4
Sox2
siNXF1
siNT
0.0 0.2 0.4 0.6 0.8 1.0
Relative protein level

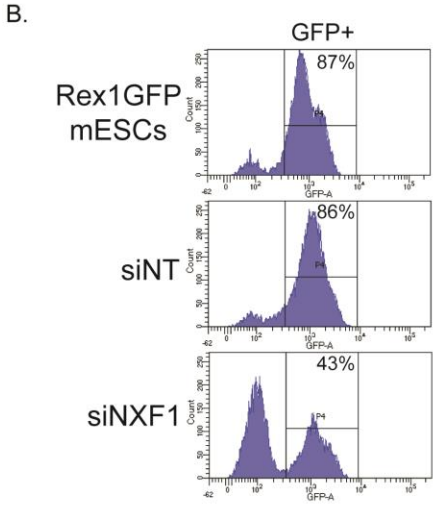
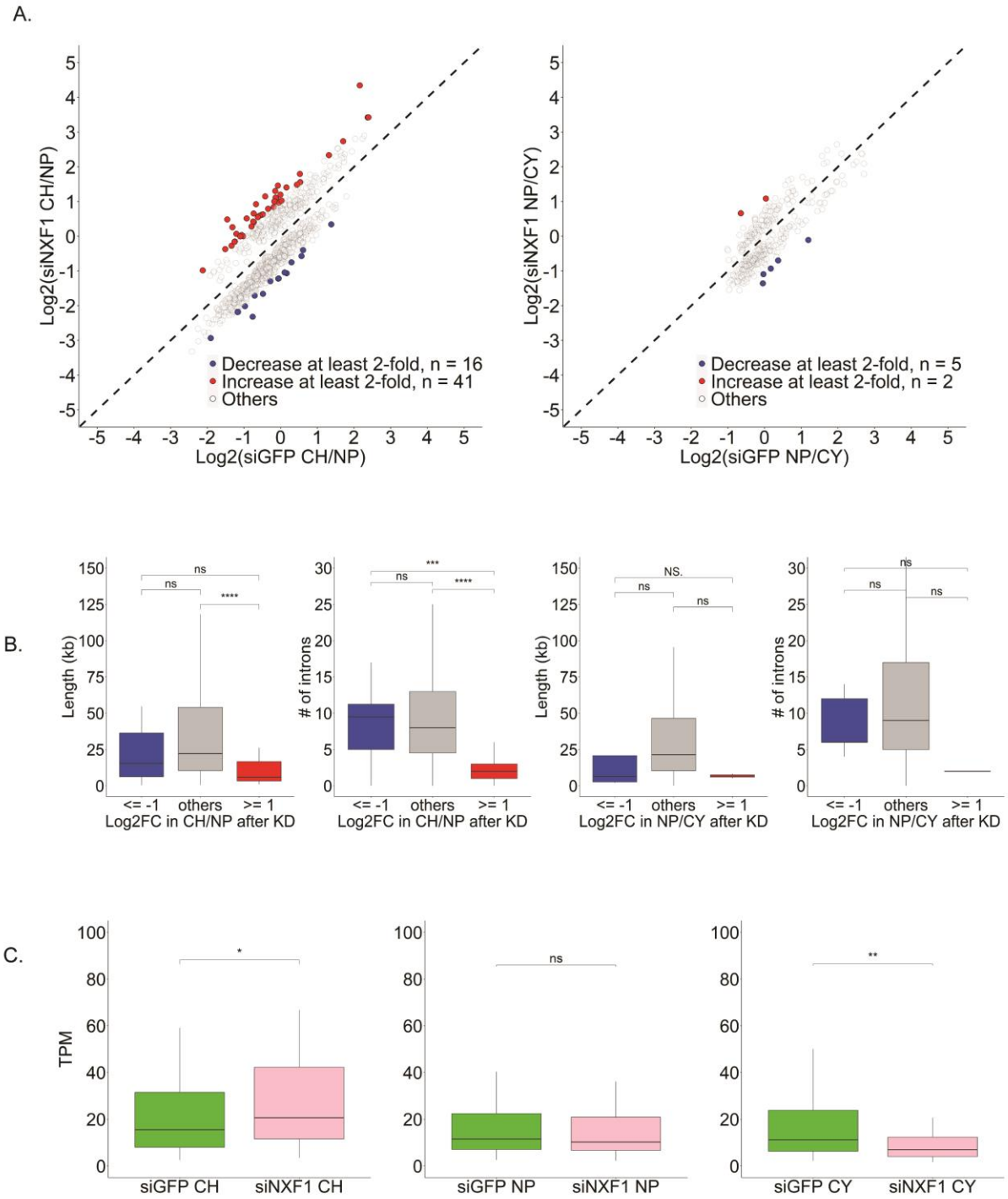

B.
GFP+
Rex1GFP mESCs
87%
siNT
86%
siNXF1
43%
Count
GFP-A

Figure 6.

NXF1 depletion in mESCs causes loss of pluripotency

A. Immunoblot showing loss of pluripotency factors upon 48 hours of NXF1 knockdown. NT = non-targeting. Levels of depletion compared to siNT (1.0 relative protein level) are quantified on the right. Mean protein levels are plotted. Error bars are SEM; Student's t-test; **: $p \leq 0.01$, ***: $p \leq 0.001$.

B. Flow cytometry analysis of GFP signal from Rex1GFP mESCs 72 hours post-siRNA transfections as indicated. "Rex1GFP mESCs" flow data shows GFP signal from Rex1GFP mESCs that had been growing for 72 hours without transfection. GFP+ gates are shown as enclosed areas in the plots in between two vertical lines.


A.
Log2(siNXF1 CH/NP)
Log2(siGFP CH/NP)
Decrease at least 2-fold, n = 16
Increase at least 2-fold, n = 41
Others
Log2(siNXF1 NP/CY)
Log2(siGFP NP/CY)
Decrease at least 2-fold, n = 5
Increase at least 2-fold, n = 2
Others
B.
Length (kb)
# of introns
<= -1 others >= 1
Log2FC in CH/NP after KD
Log2FC in NP/CY after KD
ns
****
***
NS.
C.
TPM
siGFP CH
siNXF1 CH
siGFP NP
siNXF1 NP
siGFP CY
siNXF1 CY

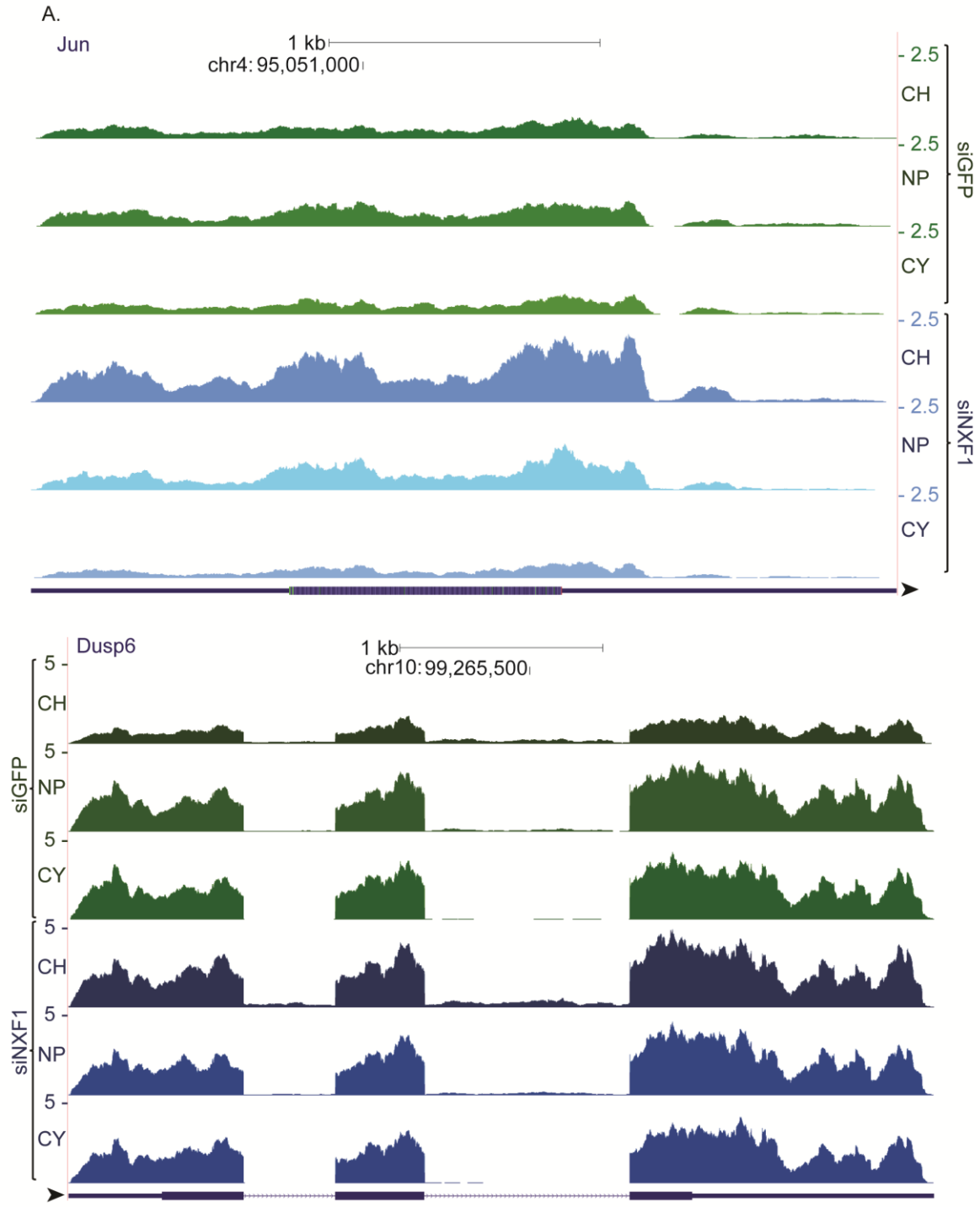
Supplemental Figure S1.

NXF1 most strongly regulates export of RNAs that are short and have low number of introns

A. Scatterplots illustrating relationship between CH/NP (left) or NP/CY (right) values of siGFP and siNXF1 (log2-transformed). Each gene has its own CH/NP and NP/CY value of before (siGFP) and after (siNXF1) protein depletion. Red-colored points are genes that become at least 2-fold more chromatin- or nucleoplasm-enriched after knockdown, while blue-colored points are genes that become at least 2-fold more nucleoplasm- or cytoplasm-enriched after knockdown. Only genes that have change in the ratio with $p < 0.05$ are plotted.

B. Box plots showing relationship between log-2 fold change in CH/NP (left) or NP/CY (right) after NXF1 knockdown and length or number of introns of the respective genes. The upper and lower hinges displayed represent the 25th and 75th percentiles; Wilcoxon rank-sum test; ns: $p > 0.05$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$. Only genes that have change in the ratio with $p < 0.05$ are plotted.

C. Box plots comparing CH, NP and CY TPM of genes before and after NXF1 depletion. Only genes with log-2 fold change in CH/CY of 1 or greater are plotted. Only genes that have change in the ratio with $p < 0.05$ are plotted. The upper and lower hinges displayed represent the 25th and 75th percentiles; Wilcoxon rank-sum test; ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ****: $p \leq 0.0001$.

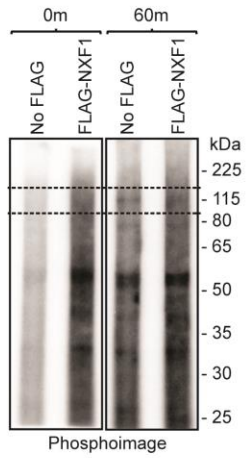

A.
Jun
1 kb
chr4: 95,051,000
2.5
CH
2.5
NP
2.5
CY
siGFP
2.5
CH
2.5
NP
2.5
CY
siNXF1
Dusp6
1 kb
chr10: 99,265,500
5
CH
5
NP
5
CY
siGFP
5
CH
5
NP
5
CY
siNXF1

Supplemental Figure S2.

NXF1 regulates export of transcripts encoding proteins involved in the immediate early response

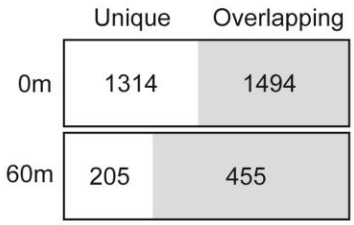
A. UCSC Genome Browser sessions showing Jun and Dusp6 mRNAs being sequestered in the chromatin fraction after NXF1 knockdown. Black arrows indicate direction of transcription.

A.

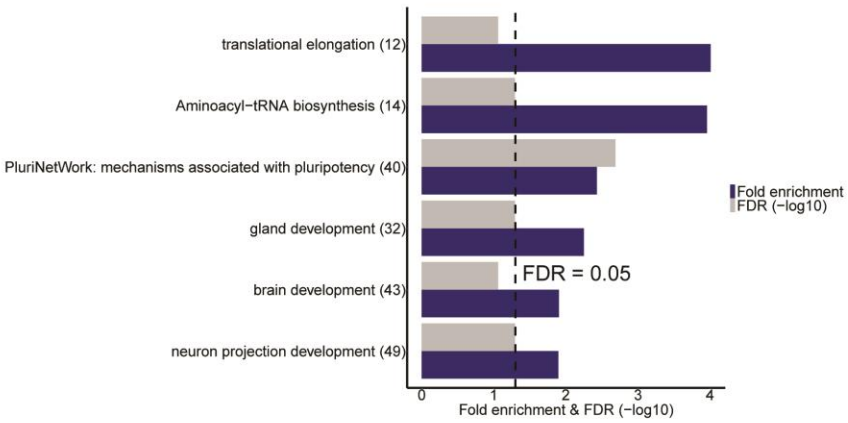


B.

Comparison with NXF1 iCLIP dataset from Müller-McNicoll et al., 2016



C.



Supplemental Figure S3.

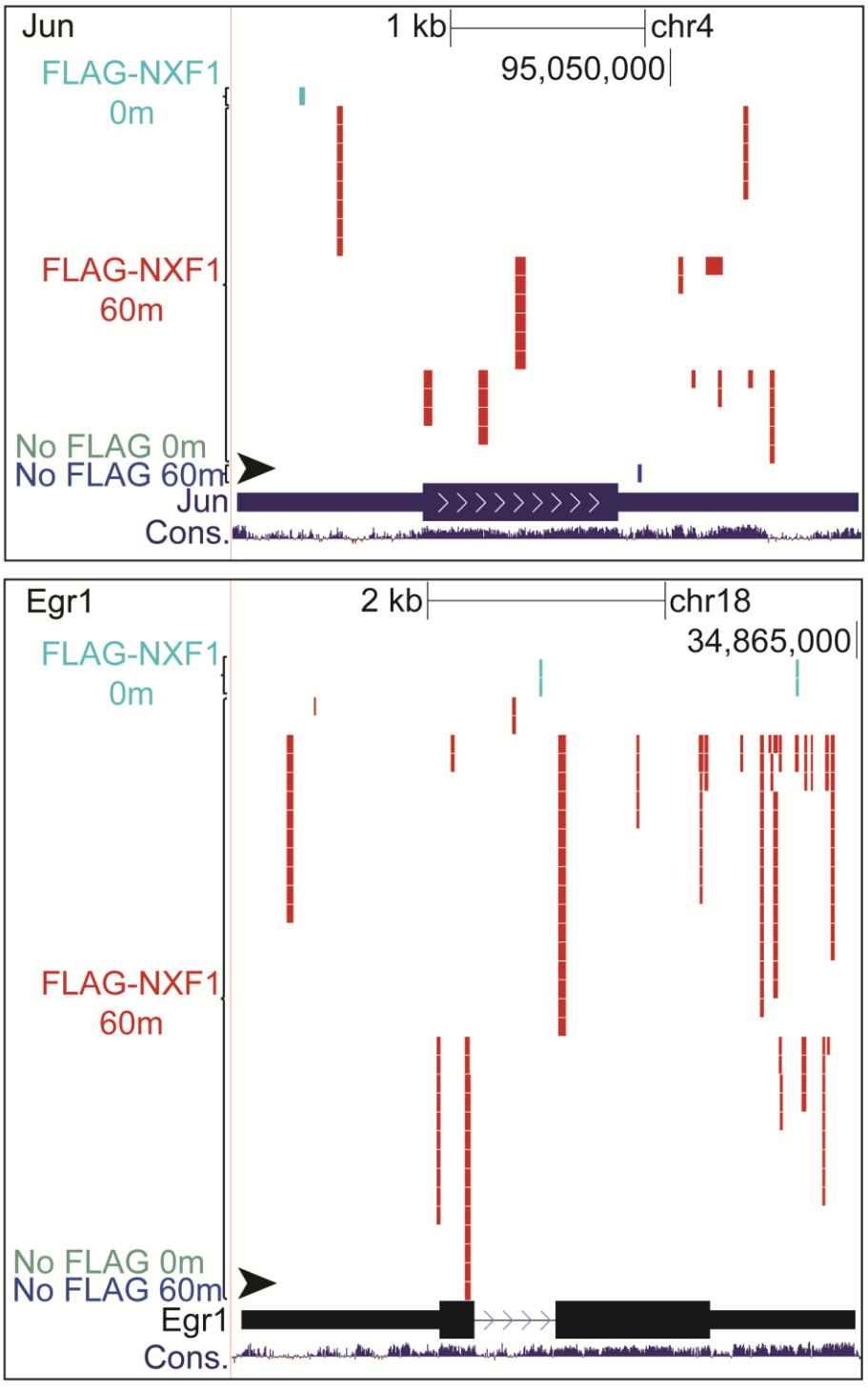
FLAG-NXF1 iCLIP-seq reveals the protein binding to NXF1 mRNA at intron 10

A. Autoradiographs of FLAG-NXF1-RNA complexes using SDS-PAGE. Complexes that are ~20-50 kDa heavier than FLAG-NXF1 (~71 kDa) were extracted from the membrane for library preparation (regions bordered by dotted lines). “No FLAG” is negative control sample from FLAG-IP using lysate of cells not expressing FLAG-NXF1. “0m” and “60m” indicate time elapsed since serum stimulation.

B. Overlap of 0m and 60m significant crosslinking events with those from Müller-McNicoll et al., 2016.

C. Gene ontology analysis of genes containing significant crosslinking events from either 0m or 60m group using Metascape. Events are considered significant when they occur in at least two biological replicates. Dotted vertical line represents FDR = 0.05.

A.

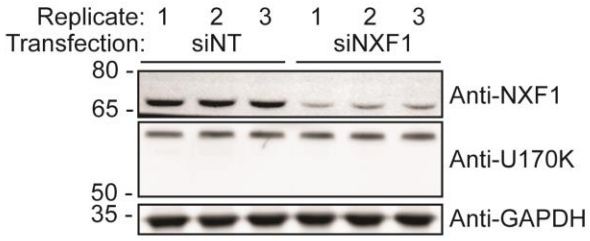

Jun
1 kb
chr4
95,050,000
FLAG-NXF1
0m
FLAG-NXF1
60m
No FLAG 0m
No FLAG 60m
Jun
Cons.
Egr1
2 kb
chr18
34,865,000
FLAG-NXF1
0m
FLAG-NXF1
60m
No FLAG 0m
No FLAG 60m
Egr1
Cons.

Supplemental Figure S4.

Transcription induction of genes by serum withdrawal followed by its reintroduction allows for discovery of novel NXF1 targets

A. UCSC Genome Browser sessions of Jun and Egr1 from FLAG-NXF1 iCLIP-seq. Black arrows indicate direction of transcription.

A.


Replicate: 1 2 3 1 2 3
Transfection: siNT siNXF1
80 -
65 -
50 -
35 -
Anti-NXF1
Anti-U170K
Anti-GAPDH

Supplemental Figure S5.

NXF1 depletion in mESCs causes loss of pluripotency

A. Immunoblot of more housekeeping proteins to show that the reduction in protein levels of pluripotency factors isn’t due to global reduction in protein levels.

REFERENCES

Aksenova, Vasilisa, Alexandra Smith, Hangnoh Lee, Prasanna Bhat, Caroline Esnault, Shane Chen, James Iben, et al. 2020. “Nucleoporin TPR Is an Integral Component of the TREX-2 MRNA Export Pathway.” *Nature Communications* 11 (1): 4577. https://doi.org/10.1038/s41467-020-18266-2.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Bahrami, Shahram, and Finn Drabløs. 2016. “Gene Regulation in the Immediate-Early Response Process.” *Advances in Biological Regulation* 62 (September): 37–49. https://doi.org/10.1016/j.jbior.2016.05.001.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Damianov, Andrey, Yi Ying, Chia-Ho Lin, Ji-Ann Lee, Diana Tran, Ajay A. Vashisht, Emad Bahrami-Samani, et al. 2016. “Rbfox Proteins Regulate Splicing as Part of a Large Multiprotein Complex LASR.” *Cell* 165 (3): 606–19. https://doi.org/10.1016/j.cell.2016.03.040.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Duren, Zhana, Xi Chen, Jingxue Xin, Yong Wang, and Wing Hung Wong. 2020. “Time Course Regulatory Analysis Based on Paired Expression and Chromatin Accessibility Data.” *Genome Research* 30 (4): 622–34. https://doi.org/10.1101/gr.257063.119.

Hautbergue, Guillaume M., Ming-Lung Hung, Alexander P. Golovanov, Lu-Yun Lian, and Stuart A. Wilson. 2008. “Mutually Exclusive Interactions Drive Handover of MRNA from Export Adaptors to TAP.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (13): 5154–59. https://doi.org/10.1073/pnas.0709167105.

Herschman, Harvey R. 1991. “Primary Response Genes Induced by Growth Factors and Tumor Promoters.” *Annual Review of Biochemistry* 60 (1): 281–319. https://doi.org/10.1146/annurev.bi.60.070191.001433.

Hocine, Sami, Robert H. Singer, and David Grünwald. 2010. “RNA Processing and Export.” *Cold Spring Harbor Perspectives in Biology* 2 (12): a000752. https://doi.org/10.1101/cshperspect.a000752.

Hooper, M., K. Hardy, A. Handyside, S. Hunter, and M. Monk. 1987. “HPRT-Deficient (Lesch-Nyhan) Mouse Embryos Derived from Germline Colonization by Cultured Cells.” *Nature* 326 (6110): 292–95. https://doi.org/10.1038/326292a0.

Huang, Y., and J. A. Steitz. 2001. “Splicing Factors SRp20 and 9G8 Promote the Nucleocytoplasmic Export of MRNA.” *Molecular Cell* 7 (4): 899–905. https://doi.org/10.1016/s1097-2765(01)00233-7.

Kalkan, Tüzer, Nelly Olova, Mila Roode, Carla Mulas, Heather J. Lee, Isabelle Nett, Hendrik Marks, et al. 2017. “Tracking the Embryonic Stem Cell Transition from Ground State Pluripotency.” *Development (Cambridge, England)* 144 (7): 1221–34. https://doi.org/10.1242/dev.142711.

Katahira, Jun. 2015. “Nuclear Export of Messenger RNA.” *Genes* 6 (2): 163–84. https://doi.org/10.3390/genes6020163.

Lee, Eliza S., Eric J. Wolf, Sean S. J. Ihn, Harrison W. Smith, Andrew Emili, and Alexander F. Palazzo. 2020. “TPR Is Required for the Efficient Nuclear Export of MRNAs and LncRNAs from Short and Intron-Poor Genes.” *Nucleic Acids Research* 48 (20): 11645–63. https://doi.org/10.1093/nar/gkaa919.

Li, Ying, Yeou-Cherng Bor, Mark P. Fitzgerald, Kevin S. Lee, David Rekosh, and Marie-Louise Hammarskjold. 2016. “An NXF1 MRNA with a Retained Intron Is Expressed in Hippocampal and Neocortical Neurons and Is Translated into a Protein That Functions as an Nxf1 Cofactor.” *Molecular Biology of the Cell* 27 (24): 3903–12. https://doi.org/10.1091/mbc.E16-07-0515.

Li, Ying, Yeou-Cherng Bor, Yukiko Misawa, Yuming Xue, David Rekosh, and Marie-Louise Hammarskjöld. 2006. “An Intron with a Constitutive Transport Element Is Retained in a Tap Messenger RNA.” *Nature* 443 (7108): 234–37. https://doi.org/10.1038/nature05107.

Moore, Melissa J., and Nick J. Proudfoot. 2009. “Pre-MRNA Processing Reaches Back to Transcription and Ahead to Translation.” *Cell* 136 (4): 688–700. https://doi.org/10.1016/j.cell.2009.02.001.

Müller-McNicoll, Michaela, Valentina Botti, Antonio M. de Jesus Domingues, Holger Brandl, Oliver D. Schwich, Michaela C. Steiner, Tomaz Curk, Ina Poser, Kathi Zarnack, and Karla M. Neugebauer. 2016. “SR Proteins Are NXF1 Adaptors That Link Alternative RNA Processing to MRNA Export.” *Genes & Development* 30 (5): 553–66. https://doi.org/10.1101/gad.276477.115.

Okamura, Masumi, Haruko Inose, and Seiji Masuda. 2015. “RNA Export through the NPC in Eukaryotes.” *Genes* 6 (1): 124–49. https://doi.org/10.3390/genes6010124.

Pandya-Jones, Amy, and Douglas L. Black. 2009. “Co-Transcriptional Splicing of Constitutive and Alternative Exons.” *RNA (New York, N.Y.)* 15 (10): 1896–1908. https://doi.org/10.1261/rna.1714509.

Pasquinelli, A. E., R. K. Ernst, E. Lund, C. Grimm, M. L. Zapp, D. Rekosh, M. L. Hammarskjöld, and J. E. Dahlberg. 1997. “The Constitutive Transport Element (CTE) of Mason-Pfizer Monkey Virus (MPMV) Accesses a Cellular MRNA Export Pathway.” *The EMBO Journal* 16 (24): 7500–7510. https://doi.org/10.1093/emboj/16.24.7500.

Reed, Robin, and Hong Cheng. 2005. “TREX, SR Proteins and Export of MRNA.” *Current Opinion in Cell Biology* 17 (3): 269–73. https://doi.org/10.1016/j.ceb.2005.04.011.

Viphakone, Nicolas, Guillaume M. Hautbergue, Matthew Walsh, Chung-Te Chang, Arthur Holland, Eric G. Folco, Robin Reed, and Stuart A. Wilson. 2012. “TREX Exposes the RNA-Binding Domain of Nxf1 to Enable MRNA Export.” *Nature Communications* 3: 1006. https://doi.org/10.1038/ncomms2005.

Viphakone, Nicolas, Ian Sudbery, Llywelyn Griffith, Catherine G. Heath, David Sims, and Stuart A. Wilson. 2019. “Co-Transcriptional Loading of RNA Export Factors Shapes the Human Transcriptome.” *Molecular Cell* 75 (2): 310-323.e8. https://doi.org/10.1016/j.molcel.2019.04.034.

Wang, Li, Yi-Liang Miao, Xiaofeng Zheng, Brad Lackford, Bingying Zhou, Leng Han, Chengguo Yao, et al. 2013. “The THO Complex Regulates Pluripotency Gene MRNA Export and Controls Embryonic Stem Cell Self-Renewal and Somatic Cell Reprogramming.” *Cell Stem Cell* 13 (6): 676–90. https://doi.org/10.1016/j.stem.2013.10.008.

Wickramasinghe, Vihandha O., Robert Andrews, Peter Ellis, Cordelia Langford, John B. Gurdon, Murray Stewart, Ashok R. Venkitaraman, and Ronald A. Laskey. 2014. “Selective Nuclear Export of Specific Classes of MRNA from Mammalian Nuclei Is Promoted by GANP.” *Nucleic Acids Research* 42 (8): 5059–71. https://doi.org/10.1093/nar/gku095.

Wuarin, J., and U. Schibler. 1994. “Physical Isolation of Nascent RNA Chains Transcribed by RNA Polymerase II: Evidence for Cotranscriptional Splicing.” *Molecular and Cellular Biology* 14 (11): 7219–25. https://doi.org/10.1128/mcb.14.11.7219-7225.1994.

Yeom, Kyu-Hyeon, and Andrey Damianov. 2017. “Methods for Extraction of RNA, Proteins, or Protein Complexes from Subcellular Compartments of Eukaryotic Cells.” *Methods in Molecular Biology (Clifton, N.J.)* 1648: 155–67. https://doi.org/10.1007/978-1-4939-7204-3_12.

Yeom, Kyu-Hyeon, Zhicheng Pan, Chia-Ho Lin, Han Young Lim, Wen Xiao, Yi Xing, and Douglas L. Black. 2021. “Tracking Pre-MRNA Maturation across Subcellular Compartments Identifies Developmental Gene Regulation through Intron Retention and Nuclear Anchoring.” *Genome Research* 31 (6): 1106–19. https://doi.org/10.1101/gr.273904.120.

Zhou, Yingyao, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. “Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets.” *Nature Communications* 10 (1): 1523. https://doi.org/10.1038/s41467-019-09234-6.

Zuckerman, Binyamin, Maya Ron, Martin Mikl, Eran Segal, and Igor Ulitsky. 2020. “Gene Architecture and Sequence Composition Underpin Selective Dependency of Nuclear Export of Long RNAs on NXF1 and the TREX Complex.” *Molecular Cell* 79 (2): 251-267.e6. https://doi.org/10.1016/j.molcel.2020.05.013.