Title
Machine Learning Strategies for Alternative Splicing

Author
Pan, Zhicheng

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

MACHINE LEARNING STRATEGIES FOR STUDYING ALTERNATIVE SPLICING

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Zhicheng Pan

2021

ABSTRACT OF THE DISSERTATION

MACHINE LEARNING STRATEGIES FOR STUDYING ALTERNATIVE SPLICING

by

Zhicheng Pan

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2021

Professor Yi Xing, Chair

Alternative splicing (AS) is a fundamental biological process that diversifies the transcriptomes and proteomes. Aberrant splicing is the main cause of rare diseases and cancers. Our understanding of AS is far from complete, resulting in a limited comprehension of phenotypic effects of splicing dysregulation. Recent advances in next-generation sequencing (NGS) technologies have revolutionized the discoveries of AS. There are considerable efforts put into generating a large compendium of RNA-seq datasets. These datasets offer an opportunity to study the regulation of AS in tissues, cell stages, and perturbation of biological conditions at unprecedented resolutions and scales. However, utilizing the large number of datasets to make biological discoveries remains a challenge. In this dissertation, we developed machine-learning-based strategies to integrate various

types of RNA-seq datasets and transform them into biological knowledge, thereby enabling discoveries towards regulatory mechanisms and functional consequences of AS.

In the first part of the dissertation, we report a deep-learning-based computational framework, Deep-learning Augmented RNA-seq analysis of Transcript Splicing (DARTS), that utilizes the Bayesian integration of deep-learning-based predictions with empirical RNA datasets to make inference of differential alternative splicing between biological samples. RNA sequencing (RNA-seq) analysis of alternative splicing is largely limited by depending on high sequencing coverage. DARTS transforms large amounts of publicly available RNA-seq datasets into biological knowledge of how splicing is regulated through deep learning, thus enabling researchers to better characterize alternative splicing inaccessible from RNA-seq datasets with modest coverage.

In the second part of the dissertation, we present a computational tool, Systematic Investigation of Retained Introns (SIRI), to quantify unspliced introns and describe a deep-learning-based computational framework to investigate the sequence preferences of different intron groups across subcellular locations. Steps of mRNA maturation occur in distinct cellular locations, while subcellular distribution of processed and unprocessed transcripts often miss in transcriptomic analyses. We employed SIRI to measure intron levels in subcellular locations across cell development and identified four intron groups that have disparate patterns of RNA enrichment across subcellular locations. Through the deep-learning based framework, we identified a set of triplet motifs and sequence conservation patterns that are predictive of intron behavior among biological conditions.

In the third part of the dissertation, we exhibit a deep-learning-based tissue-specific framework, individualized Deep-learning Analysis of RNA Transcript Splicing (iDARTS), for predicting splicing levels. The rapid accumulation of RNA-seq datasets matched with whole exome or genome sequencing yields enormous variants underlying diseases, traits, and cancer. Interpreting the functional consequences of these variants remains a challenge in disease diagnostics and precision medicine. iDARTS leverages the publicly available RNA-seq datasets to model the cis RNA sequence features and trans RNA binding protein levels determinants of AS, allowing precise predictions of genetic splice-altering variants. We demonstrated that predicted splice-altering variants are functionally relevant and related to cancer development when analysing ~10 million intronic and exonic variants with iDARTS. Applying iDARTS to interpret functional consequences of variants of uncertain significance in clinical studies, we found that predicted splice-altering variants are ten times enriched in pathogenic categories over benign categories. Our results indicate that iDARTS will benefit large-scale screening disease-implicated variants, thus improving disease diagnosis and enabling discoveries for precision medicine.

In the fourth part of the dissertation, we study the underlying mechanisms of N6-methyladenosine (m^6A) RNA modification by investigating the biological consequences of arginine methylation of METTL14 through transcriptome-wide profiling of m^6A. Arginine methylation of METTL14 controls m^6A deposition in mammalian cells. Mouse embryonic stem cells (mESCs) expressing arginine methylation-deficient METTL14 exhibit significantly reduced global m^6A levels. These arginine methylation-dependent m^6A sites identified from transcriptome-wide analysis are associated with enhanced translation of genes essential for the repair of DNA interstrand crosslinks. Collectively, these findings

reveal important aspects of m^6A regulation and new functions of arginine methylation in RNA metabolism.

The dissertation of Zhicheng Pan is approved.

Douglas L. Black

Kathrin Plath

Yingnian Wu

Yi Xing, Committee Chair

University of California, Los Angeles

2021

To Wenbo, Mom, and Dad

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my advisor, Dr. Yi Xing, who has provided tremendous support and guidance for my research studies and career development. He taught me everything about how to conduct research critically and rigorously. It is his patience, encouragement, and wisdom that motivate me to pursue my research interests. He is the best mentor I can ever ask for, making my journey of research and discovery full of fun and fulfillment.

I would also like to thank my committee members, Douglas Black, Kathrin Plath, and Yingnian Wu, for their enormous support and insightful advice. I am very grateful for having them on my committee.

I am also more than thankful to my collaborators, Kay Yeom and Zijun Zhang from UCLA, Zhihao Wang and Yanzhong (Frankie) Yang from the City of Hope, and Lan Lin from the University of Pennsylvania, for their invaluable help and support.

My life as a Ph.D. student would not be so wonderful without my friends and colleagues in the Xing lab: Zijun Zhang, Yang Pan, Yida Zhang, Chengyang Wang, Yuanyuan Wang, Yuan Gao, Yongbo Wang, Yan Gao, Ruijiao Xin, Jingkai Wang, Samir Adhikari, Levon Demirdjian, Eddie Park, Xinyuan Chen, Harry Yang, Yang Guo, Amal Katrib, Yang Xu, Emad Bahrami-Samani, Shaofang Li, Zhixiang Lu, Shayna Stein, Yu-ting Tseng, Yungang Xu, Kadash-Edmondson, Beatrice Zhang, Siwei Luo, and Robert Wang.

Last but not least, I would love to thank my parents for their love throughout my life. They have been unconditionally supporting and trusting me. Thank you to my fiancée

Wenbo. Having you by my side makes me the happiest and luckiest person in the world. I love you all deeply every moment.

VITA

EDUCATION

Visiting Scholar, Center for Computational Genomics and Medicine
University of Pennsylvania 2018-2021

Graduate Student Researcher, Bioinformatics IDP
University of California, Los Angeles 2015-2021

Teaching Assistant, MIMG 180B: Scientific Analysis and Communication,
University of California, Los Angeles 2018

B.S. in Bioinformatics, College of Life Sciences
Huazhong University of Science and Technology 2009-2013

HONORS AND AWARDS

University fellowship, University of California, Los Angeles 2015-2016

Outstanding Graduate Awards, Huazhong University of Science and Technology 2013

Asia Regional Bronze Medal, The iGEM competition 2012

National Scholoarship, China Ministry of Education 2010-2011

SELECTED PUBLICATIONS

† co-first author

1. K Yeom†, **Z Pan**†, C Lin, H Lim, W Xiao, Y Xing, DL Black (2021). Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring. *Genome Research*, gr. 273904.120.

2. Z Wang†, **Z Pan**†, S Adhikari, B Harada, L Shen, W Yuan, T Abeywardena, Q AI-Hadid, JM Stark, C He, L Lin, Y Yang (2021). m6A deposition is regulated by PRMT1-mediated arginine methylation of METTL14 in its disordered C-terminal region. *The EMBO Journal*, e106309.

3. Z Zhang†, **Z Pan**†, Y Ying, Z Xie, S Adhikari, J Phillips, RP Carstens, DL Black, Y Wu, Y Xing (2019). Deep learning-augmented RNA-seq analysis of transcript splicing. *Nature Methods*, 16 (4), 307.

4. E Park, **Z Pan**, Z Zhang, L Lin, Y XING (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *The American Journal of Human Genetics*, 102 (1), 11-26.

5. **Z Pan**†, B Wang†, Y Zhang, Y Wang, S Ullah, R Jian, Z Liu, Y Xue (2015). dbPSP: a curated database for protein phosphorylation sites in prokaryotes. *Database*, 2015.

6. **Z Pan**†, Z Liu†, H Cheng, Y Wang, T Gao, S Ullah, J Ren, Y Xue (2014). Systematic analysis of the in situ crosstalk of tyrosine modifications reveals no additional natural selection on multiply modified residues. *Scientific Reports*, 4, 7331.

1 INTRODUCTION

The central dogma of molecular biology describes an information flow from DNA to messenger RNA to protein[1]. Since the last decades, the diagram has been largely expanded by showing diverse regulations of genes. After transcription initiation, the maturation of pre-messenger RNA (pre-mRNA) involves co-transcriptional and post-transcriptional regulations including 5' capping, splicing, 3' polyadenylation, RNA editing, and RNA modifications such as N6-methyladenosine (m^6A) at the subcellular level. Of these transcriptional regulations, alternative splicing (AS) is a major biological process where introns are removed, and exons are selectively joined together to form mature transcripts[2]. AS enables the production of multiple isoforms from a single gene, thus greatly diversifying the transcriptome and proteomes[3]. AS exhibits a tissue-specific and developmental stage-specific manner and is regulated via recognizing cis-elements and trans-acting factors that bind to the cis-elements[4]. The dysregulation of AS underlies rare diseases and cancers[5,6].

With the advent of next-generation sequencing (NGS), biological technologies including RNA sequencing[7] (RNA-seq) have transformed the discoveries of AS in tissues,

cell developmental stages, and various biological conditions. These enormous datasets offer great opportunities to study the regulation of AS while posing a challenge for the development of computational approaches.

Machine learning is a data-driven computational approach that learns functional relationships from data[8]. The advantage of machine learning over conventional statistical methods is that it does not need strong assumptions or data structures, which are poorly defined especially in biological datasets. Machine learning consists of two main subtypes: unsupervised learning and supervised learning[9,10]. The difference between unsupervised learning and supervised learning is whether labelled data is needed. Unsupervised learning often refers to clustering which aims to find a set of clusters with similar unlabelled data points while supervised learning learns the mapping functions from the input data to output labels. Over the past ten years, striking advances in machine learning have drastically changed how researchers utilize big data to make significant discoveries. Profound progress in image, speech, and languages has shown a promising power of machine learning in transforming large-scale datasets into knowledge base[11,12].

Motivated by the scalability to large-scale datasets and the capability of machine learning to learn complex functions, Chapters 2, 3, and 4 of this dissertation present novel machine-learning-based approaches for studying various types of transcriptome sequencing data. In detail, Chapter 2 describes a novel computational and statistical framework, Deep-learning Augmented RNA-seq analysis of Transcript Splicing (DARTS)[13], for inferring differential splicing events between biological conditions. Consortium efforts such as ENCODE[14] project and Roadmap[15] project have released diverse RNA-seq datasets in different tissues and under perturbations of RNA binding proteins (RBPs). Utilizing these

datasets to make biological discoveries has been limited to highly expressed and observable regulatory events. We trained a deep neural network (DNN) that predicts differential alternative splicing between biological conditions by leveraging the publicly available RNA-seq datasets from ENCODE and Roadmap project. We validated the DARTS DNN's accuracy and generalizability in diverse biological conditions. As an informative prior in the statistical model, DARTS DNN significantly improved the confidence of reported differential AS, especially from low-coverage RNA-seq datasets. We further demonstrated that DARTS could accurately and reliably predict AS changes in lowly expressed genes using the cellular models of the epithelial-mesenchymal transition. In conclusion, DARTS provides a generic framework to infer differential AS between biological samples.

Chapter 3 presents a novel computational tool, Systematic Investigation of Retained Introns (SIRI), and a deep-learning-based computational framework to study the regulation of intron retention at the subcellular level[16]. 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 datasets to track mRNA maturation across subcellular locations in mouse embryonic stem cells, neuronal progenitor cells, and postmitotic neurons. Retained introns are more difficult to characterize than other patterns of AS in whole transcriptome RNA-seq data. Overlapping patterns of alternative processing can be mis-called as intron retention by sequence analysis tools[17,18]. In this regard, we developed SIRI that characterizes four types of intron retention events by their overlapping patterns with other exons and introns. Using SIRI, we

reliably quantified intron retention events in different subcellular locations across different cell types. These intron retention events show disparate regulatory patterns across subcellular locations. Through a machine-learning-based approach, X-means, four regulatory groups were defined, 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 further explored the RNA sequence features of the four regulatory groups by utilizing a deep-learning-based computational approach and found triplet motifs and conservations of introns could be predictive of intron behaviour, indicating that particular RNA/protein interactions likely determine the retention properties of these groups.

Chapter 4 shows a novel deep-learning-based framework for predicting tissue-specific AS from arbitrary sequences, called individualized Deep-learning Analysis of RNA Transcript Splicing (iDARTS). Accumulated RNA-seq datasets with matched exome sequencing or whole genome sequencing provide tremendous resources for studying the associations of mis-splicing related variants with disease and traits. Analyses like splicing quantitative trait locus (sQTL)[19,20] help to characterize variants that are associated with the changes of AS. However, identifying causal variants in association studies remains elusive. Clinical studies have been continuously reporting large number of disease-related variants[21,22]. Interpretation of the functional effects of the variants poses a challenge, owing to our limited understanding of gene regulation and splicing. Motivated by the success of DARTS, we expanded the DARTS framework to the iDARTS framework that directly predicts AS levels, thus enabling the prediction of quantitative effects of variants on AS. The iDARTS framework integrates the cis RNA sequence features and trans RBP levels into a

deep-learning model of AS using large-scale RNA-seq datasets from 8,304 samples in 53 tissues from the GTEx[23] project (V7). iDARTS shows a robust, accurate, and generalizable behaviour in predicting tissue-specific exon skipping levels as validated in held-out and independent RNA-seq datasets. The design of iDARTS that directly learns RNA sequence features and trans RBP levels determinates of AS makes it capable of interpreting the potential effects of variants in AS from a tissue-specific perspective. We expect iDARTS could benefit the interpretation and prioritization of variants implicated in diseases and cancers, therefore helping to discover potential candidates for therapeutic interventions.

Chapter 5 studies the biological regulations of arginine methylation of METTL14 via transcriptome-wide profiling of m^{6}A[24]. The m^{6}A RNA modification serves crucial functions in RNA metabolism and involves in co-transcriptional regulation of splicing[25] while the molecular mechanisms underlying the regulation of m^{6}A are not well understood. We establish arginine methylation of METTL14, a component of the m^{6}A methyltransferase complex, as a novel pathway that controls m^{6}A deposition in mammalian cells. Specifically, protein arginine methyltransferase 1 (PRMT1) interacts with, and methylates the intrinsically disordered C terminus of METTL14, which promotes its interaction with RNA substrates, enhances its RNA methylation activity, and is crucial for its interaction with RNA polymerase II (RNAPII). Our findings indicate methylation deficient METTL14 results in globally reducing m^{6}A levels in mouse embryonic stem cells. Transcriptome-wide m^{6}A analysis identified 1,701 METTL14 arginine methylation-dependent m^{6}A sites located in 1,290 genes involved in various cellular processes, including stem cell maintenance and DNA repair. Investigating the RNA sequence features of these m^{6}A sites revealed their predicted preferences of RNA secondary structures such as helix/stem or multi-branched

loops. In summary, these findings show the important role of m^6A regulation and their potential structural preferences from computational approaches.

References

1. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-3 (1970).
2. Sharp, P.A. Split genes and RNA splicing. *Cell* **77**, 805-15 (1994).
3. Nilsen, T.W. & Graveley, B.R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457-63 (2010).
4. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet* **102**, 11-26 (2018).
5. Wang, G.S. & Cooper, T.A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**, 749-61 (2007).
6. Lord, J. & Baralle, D. Splicing in the Diagnosis of Rare Disease: Advances and Challenges. *Front Genet* **12**, 689892 (2021).
7. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-15 (2010).
8. Angermueller, C., Parnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol Syst Biol* **12**, 878 (2016).
9. Tarca, A.L., Carey, V.J., Chen, X.W., Romero, R. & Draghici, S. Machine learning and its applications to biology. *PLoS Comput Biol* **3**, e116 (2007).
10. Libbrecht, M.W. & Noble, W.S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321-32 (2015).
11. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat Med* **25**, 24-29 (2019).
12. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-44 (2015).
13. Zhang, Z. *et al.* Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat Methods* **16**, 307-310 (2019).
14. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
15. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
16. Yeom, K.H. *et al.* Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring. *Genome Res* **31**, 1106-1119 (2021).
17. Wang, Q. & Rio, D.C. JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proc Natl Acad Sci U S A* **115**, E8181-E8190 (2018).
18. Broseus, L. & Ritchie, W. Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput Struct Biotechnol J* **18**, 501-508 (2020).

19. Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun* **8**, 14519 (2017).

20. Zhang, Y. *et al.* Regional Variation of Splicing QTLs in Human Brain. *Am J Hum Genet* **107**, 196-210 (2020).

21. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* **25**, 911-919 (2019).

22. Wai, H.A. *et al.* Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med* **22**, 1005-1014 (2020).

23. Battle, A., Brown, C.D., Engelhardt, B.E. & Montgomery, S.B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

24. Wang, Z. *et al.* m(6) A deposition is regulated by PRMT1-mediated arginine methylation of METTL14 in its disordered C-terminal region. *EMBO J* **40**, e106309 (2021).

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

2 DEEP-LEARNING AUGMENTED RNA-SEQ ANALYSIS OF TRANSCRIPT SPLICING

2.1 Introduction

RNA sequencing (RNA-seq) enables transcriptome-wide profiling of alternative splicing[1,2]. The rapid accumulation of RNA-seq data in public repositories (for example, ENCODE[3], Roadmap Epigenomics[4]) provides unprecedented resources for characterizing alternative splicing across diverse biological states. However, an inherent limitation of RNA-seq is that it is restricted by sequencing depth[5] and cannot reliably quantify splicing in genes with low expression[6].

Motivated by recent successes in the use of machine learning to predict exon-inclusion/skipping levels in bulk tissues or single cells[7-10], we hypothesized that large-scale

RNA-seq resources can be used to construct a deep-learning model of differential alternative splicing. To test this hypothesis, we developed DARTS (deep-learning augmented RNA-seq analysis of transcript splicing). DARTS consists of two core components: a deep neural network (DNN) model that predicts differential alternative splicing between two conditions on the basis of exon-specific sequence features and sample-specific regulatory features, and a Bayesian hypothesis testing (BHT) statistical model that infers differential alternative splicing by integrating empirical evidence in a specific RNA-seq dataset with prior probability of differential alternative splicing (**Figure 2.1a**). During training, large-scale RNA-seq data are analyzed by the DARTS BHT with an uninformative prior (DARTS BHT(flat), with only RNA-seq data used for the inference) to generate training labels of high-confidence differential or unchanged splicing events between conditions, which are then used to train the DARTS DNN. During application, the trained DARTS DNN is used to predict differential alternative splicing in a user-specific dataset. This prediction is then incorporated as an informative prior with the observed RNA-seq read counts by the DARTS BHT (DARTS BHT(info)) for deep learning-augmented splicing analysis.

2.2 Results

2.2.1 Deep-learning based differential splicing prediction of DARTS framework

Unlike methods that use cis sequence features to predict exon splicing patterns in specific samples[7-10], the DARTS DNN predicts differential alternative splicing by incorporating both cis sequence features and messenger RNA (mRNA) levels of trans RNA-binding proteins (RBPs) in two conditions (**Figure 2.1b** and **Supplementary Figure 2.4**). This design

allows the DARTS DNN to consider how altered expression of RBPs affects splicing. We initially focused on exon skipping, the most frequent alternative splicing pattern in animals[6]. We compiled 2,926 cis sequence features and 1,498 annotated RBPs[11] whose mRNA levels were treated as trans RBP features.

To train the DARTS DNN, we used large-scale RBP-depletion RNA-seq data in two human cell lines (K562 and HepG2) generated by the ENCODE consortium[12] (**Figure 2.1c**). We used RNA-seq data of 196 RBPs depleted by short-hairpin RNA (shRNA) in both cell lines, corresponding to 408 knockdown-versus-control pairwise comparisons (**Figure 2.1c**). The remaining ENCODE data, corresponding to 58 RBPs depleted in only one cell line, were excluded from training and used as leave-out data for independent evaluation of the DARTS DNN (**Figure 2.1c**). To generate training labels, we applied DARTS BHT(flat) to calculate the probability of an exon being differentially spliced or unchanged in each pairwise comparison. DARTS BHT(flat) was benchmarked using simulation datasets, and compared favorably to two state-of-the-art statistical models for differential splicing inference, MISO1 and rMATS2 (**Supplementary Figure 2.5 and 2.6**). From the high-confidence differentially spliced versus unchanged exons called by DARTS BHT(flat), we used 90% of labeled events for training and fivefold cross-validation, and the remaining 10% of events for testing (Methods). The performance of the DARTS DNN increased as training progressed, reaching a maximum area under the receiver operating characteristic curve (AUROC) of 0.97 during cross-validation and 0.86 during testing (**Supplementary Figure 2.7**).

To test the general applicability of the DARTS DNN, we used the leave-out data, corresponding to 58 RBPs that had never been seen during training (**Figure 2.1c**). The

trained DARTS DNN showed a high accuracy (average AUROC=0.87) on the leave-out data. We used the leave-out data to compare the DARTS DNN with three alternative baseline methods: the identical DNN structure trained on individual leave-out datasets (DNN), logistic regression with L2 penalty (logistic), and random forest. We trained these baseline methods using fivefold cross-validation in each leave-out dataset. Additionally, we implemented another alternative baseline method by predicting sample-specific exon-inclusion levels (PSI values; percent spliced in, or ψ)[1,10] and then taking the absolute difference of the predicted PSI values between two conditions as the metric for differential splicing, ($\widehat{\psi}_{KD}$ - $\widehat{\psi}_{CTRL}$). The DARTS DNN trained on the large-scale ENCODE data outperformed baseline methods by a large margin in 57 out of 58 experiments (**Figure 2.1d**). The DARTS DNN model trained on individual leave-out datasets was the worst performer, illustrating the importance of training the DARTS DNN on large-scale data comprising diverse perturbation experiments.

2.2.2 DARTS Bayesian framework improves the inference of differential splicing in RNA-seq data by incorporating informative prior from DARTS DNN.

Next, we evaluated the ability of the DARTS framework to infer differential splicing from a specific RNA-seq dataset, by incorporating the DARTS DNN predictions as the informative prior, and observed RNA-seq read counts as the likelihood (DARTS BHT(info)). The posterior ratio of differential splicing consists of two components: the prior ratio, generated by the DARTS DNN on the basis of cis sequence features and expression levels of trans RBPs; and the likelihood ratio, determined by modeling of the biological variation and estimation uncertainty of the splice isoform ratio based on observed RNA-seq read counts. Simulation studies demonstrated that the informative prior improves the inference when

the observed data are limited, for instance, because of low levels of gene expression or limited RNA-seq depth, but does not overwhelm the evidence in the observed data (**Supplementary Figure 2.8**).

We used DARTS BHT(info) and DARTS BHT(flat) to infer cell type-specific splicing events between HepG2 and K562 cell lines. To obtain high-confidence differential and unchanged splicing events between the two cell types, we aggregated all 24 or 28 RNA-seq replicates of HepG2 or K562 from ENCODE and applied DARTS BHT(flat) to this ultra-deep RNA-seq dataset. Next, we applied DARTS BHT(info) and DARTS BHT(flat) to all possible (24×28) pairwise comparisons between individual replicates of HepG2 and K562, and computed the area under the precision recall curve (AUPR) for the two methods. DARTS BHT(info) outperformed DARTS BHT(flat) in all pairwise comparisons, and the performance gain was negatively correlated with the RNA-seq depth of individual replicates (Spearman's $\rho=-0.69$, $P<2.2\times10-16$), with the largest gain coming from comparisons involving low-coverage RNA-seq samples (**Figure 2.2a**). Thus, incorporating the DNN prediction as prior information improves the detection of cell-type-specific splicing events from low-coverage RNA-seq data.

2.2.3 Extending DARTS to diverse cell types and different types of alternative splicing

Next, we determined whether the DARTS DNN can be extended to additional cell types, and how the choice of training datasets influences its performance. We used RNA-seq data from diverse cell types generated by the Roadmap Epigenomics consortium[4]. We performed 253 pairwise comparisons of Roadmap samples by DARTS BHT(flat) to generate training data

for the DARTS DNN. We excluded all pairwise comparisons involving the thymus tissue from training to use as leave-out data for independent evaluation. We trained three DARTS DNN models, using ENCODE data only, Roadmap data only, or both (**Figure 2.2b**). The DARTS DNN trained on ENCODE data exhibited high predictive power for leave-out ENCODE data but modest predictive power for leave-out Roadmap data. Conversely, the DARTS DNN trained on Roadmap data had high predictive power for leave-out Roadmap data but modest predictive power for leave-out ENCODE data. The DARTS DNN trained on combined ENCODE and Roadmap data had the best performance (**Figure 2.2b**).

We extended the DARTS DNN beyond exon skipping to predict other types of alternative splicing patterns. We compiled cis sequence features and trained three DNN models for predicting differential alternative 5′ splice sites, alternative 3′ splice sites, and retained introns. Trained on ENCODE and Roadmap data, these DNN models exhibited a high prediction accuracy in independent leave-out datasets (**Supplementary Figure 2.9**).

2.2.4 Analysis of epithelial–mesenchymal transition by DARTS

Finally, we applied DARTS to study alternative splicing during the epithelial–mesenchymal transition (EMT), a key process in embryonic development and cancer metastasis[13]. We reanalyzed our published time-course RNA-seq data on an inducible H358 lung cancer cell line model of the EMT[14]. We used DARTS BHT(flat) to compare each day to day 0, then assessed the ability of the DARTS DNN to predict high-confidence differential versus unchanged splicing events during the EMT. The DARTS DNN trained on combined ENCODE and Roadmap data had the best performance, followed closely by the DARTS DNN trained on Roadmap data, whereas the DARTS DNN trained on ENCODE data performed least well

(**Figure 2.3a**). This was expected, given that the Roadmap data cover epithelial and mesenchymal cell types. The best prediction accuracy (AUROC=0.82) was achieved by the DARTS DNN trained on combined ENCODE and Roadmap data for the comparison of day 6 versus day 0. As an example, the DARTS DNN predicted the EMT-associated alternative splicing change in PLEKHA1 (**Supplementary Figure 2.10**).

To further assess the DARTS DNN predictions, we compiled 449 'DARTS DNN rescued' events from the comparison of day 6 versus day 0 (**Methods**). A subset of these DARTS DNN rescued events had significantly reduced exon inclusion during the EMT, and their downstream intronic regions were enriched for the consensus motif of the splicing factors epithelial splicing regulatory proteins 1 and 2 (ESRP1 and ESRP2; ref. [15]) (**Figure 2.3b**). A similar pattern of ESRP-motif enrichment was observed for differential splicing events called by DARTS BHT(flat) using RNA-seq data alone (**Figure 2.3b**). By contrast, events that were called significant by DARTS BHT(flat) but fell below the significance threshold (posterior probability<0.9) after incorporation of the informative prior were not enriched for the ESRP motif (**Supplementary Figure 2.11**). ESRPs are epithelial-specific splicing factors, the downregulation of which is a major driver of alternative splicing during the EMT[14]. This observed pattern of ESRP-motif enrichment is consistent with ESRP binding downstream of alternative exons enhancing exon inclusion[13]. To extend our DARTS analysis of the H358 EMT system, we carried out paired-end RNA-seq of the PC3E and GS689 prostate-cancer cell lines, which have contrasting epithelial versus mesenchymal characteristics[2,16]. The DARTS DNN scores of these two EMT systems were highly correlated (Spearman's ρ=0.87, P<2.2×10−16; **Figure 2.3c**), suggesting that the DARTS DNN can capture a core EMT splicing signature.

2.2.5 The analysis of RASL-seq reveals the ability of DARTS to uncover alternative splicing changes in genes with low expression

To assess whether DARTS can uncover bona fide differential splicing events from genes expressed at low levels, we carried out targeted splicing profiling using the RNA-mediated oligonucleotide annealing, selection, and ligation with next-generation sequencing (RASL-seq) technology[17] and estimated the absolute difference of PSI values (RASL−|ΔPSI|) for 1,058 alternative splicing events between PC3E and GS689 (**Methods**). We restricted our further analysis to events where RASL−|ΔPSI|<0.3. As expected, alternative splicing events called as differential or unchanged by RNA-seq data alone (by DARTS BHT(flat)) had the highest or lowest RASL−|ΔPSI| values, respectively (**Figure 2.3d**). For the remaining events called as inconclusive by DARTS BHT(flat), we compiled DARTS-DNN-predicted differential events and unchanged events, with high and low DARTS DNN scores (false positive rate or FPR<5% and >80%), respectively. DARTS-DNN-predicted differential events had significantly greater RASL−|ΔPSI| values than DARTS-DNN-predicted unchanged events (P=0.035, one-sided Wilcoxon test), with the former group similar to the RNA-seq differential events and the latter group similar to the RNA-seq unchanged events (**Figure 2.3d**). DARTS-DNN-predicted differential events were in genes with significantly lower expression levels (P=0.001, two-sided Wilcoxon test) and had significantly lower RNA-seq coverage (P=2.1×10−7, two-sided Wilcoxon test) compared with differential events called by DARTS BHT(flat) (**Supplementary Figure 2.12a,b**). Collectively, among the events analyzed by RASL-seq, DARTS DNN predicted 52 additional differential splicing events beyond the 77 events called with RNA-seq data alone. Moreover, on RNA-seq inconclusive events with high or low DARTS DNN scores, we used RASL-seq to define the ground truth

with RASL−|ΔPSI|>5% as differential and RASL−|ΔPSI|<1% as unchanged. We benchmarked the performance of DARTS BHT(info), DARTS BHT(flat), DARTS DNN, rMATS[2] and SUPPA2[18] that adopted alignment-based versus alignment-free strategies for quantifying splicing with RNA-seq data. DARTS BHT(info) consistently outperformed baseline methods that use RNA-seq data alone to call differential splicing (**Supplementary Figure 2.12c,d**). These data suggest that by combining deep-learning predictions with empirical evidence in user-specific RNA-seq data, DARTS can uncover alternative splicing changes in genes with low expression and expand the findings beyond a conventional RNA-seq splicing analysis.

2.3 Discussion

We report DARTS, a deep-learning augmented statistical framework for RNA-seq analysis of differential alternative splicing. DARTS utilizes the enormous RNA-seq resources including diverse cell types with perturbations of expressions of RBPs to predict differential alternative splicing in two conditions using cis-elements of alternative splicing events and the expression of trans-RBPs. We demonstrated the benefit of using DARTS to uncover differential splicing events in lowly expressed genes that cannot be captured via conventional RNA-seq analysis. This enables us to investigate the regulation of splicing events in lowly expressed genes potentially implicated in diseases of interest or acting as potential biomarkers or therapeutic targets. The conceptual innovation of iDARTS is it transforms massive amounts of RNA-seq data into a knowledge base of how splicing regulates through deep learning, which can benefit individual studies by expanding the

pools of alternative splicing events under investigation. DARTS is an open resource software available at https://github.com/Xinglab/DARTS.

2.4 Methods

2.4.1 DARTS Bayesian hypothesis testing framework

We developed DARTS BHT, a Bayesian statistical framework to determine the statistical significance of differential splicing events or unchanged splicing events between RNA-seq data of two biological conditions. The DARTS BHT framework was designed to integrate deep-learning-based prediction as prior and empirical evidence in a specific RNA-seq dataset as likelihood. We start by modelling the simplest case, that is, testing the difference in exon-inclusion levels (PSI values) between two conditions without replicates (one sample per condition; for model with replicates, see Appendix):

$$I_{ij} \mid \psi_{ij} \sim \text{Binomial}(n = I_{ij} + S_{ij}, p = f_i(\psi_{ij}))$$

$$\psi_{i1} = \mu_i$$

$$\psi_{i2} = \mu_i + \delta_i$$

$$\mu_i \sim \text{Unif}(0,1)$$

$$\delta_i \sim N(0, \tau^2)$$

Where I_{ij}, S_{ij} and ψ_{ij} are the exon inclusion read count, the exon skipping read count, and the exon inclusion level for exon i in sample group $j \in (1,2)$, respectively; f_i is the length normalization function for exon i that accounts for the effective lengths of the exon inclusion and skipping isoforms (4); μ_i is the baseline inclusion level for exon i, and δ_i is the

expected difference of the exon inclusion levels between the two conditions. The goal of the differential splicing analysis is to test whether the difference in exon inclusion levels between the two conditions δ_i exceeds a user-defined threshold c (e.g. 5%) with a high probability, i.e.

$$P(|\delta_i| > c | I_{ij}, S_{ij}) \approx 1$$

In Bayesian statistics, this test can be approached by assuming a "spike-and-slab" prior for the parameter of interest δ. The spike-and-slab prior is a two-component mixture prior distribution, with the "spike" component depicting the probability of the model parameter δ being constrained around zero, and the "slab" component depicting the unconstrained distribution of the model parameter δ.

In the DARTS BHT statistical framework, we impose a spike prior H_0 with a small variance $\tau = \tau_0$, such that the probability of δ concentrates around 0, to account for random biological or technical fluctuations in PSI values between two biological conditions for unchanged splicing events. We impose a slab prior H_1 with a much larger variance $\tau = \tau_1$ to model the difference in PSI values between two conditions for differential splicing events. We set $\tau_0 = 0.03$, corresponding to 90% density constrained within $\delta \in [-0.05, 0.05]$, and $\tau_1 = 0.3$; we note that the final inference is robust to choice of τ values (for more details, see Appendix and **Supplementary Figure 2.13**). The posterior probability of a splicing event being generated by the two models can be written as:

$$P(H_1 | I_{ij}, S_{ij}) = \frac{1}{Z} P(H_1) \cdot P(I_{ij}, S_{ij} | H_1)$$

$$P(I_{ij}, S_{ij} | H_1) = \int_{\delta} \int_{\mu} P(I_{ij}, S_{ij} | \mu_i, \delta_i) \cdot P(\mu_i, \delta_i | H_1) d\mu_i \, d\, \delta_i$$

$$P(H_0|I_{ij}, S_{ij}) = \frac{1}{Z} P(H_0) \cdot P(I_{ij}, S_{ij}|H_0)$$

$$P(I_{ij}, S_{ij}|H_0) = \int_{\delta} \int_{\mu} P(I_{ij}, S_{ij}|\mu_i, \delta_i) \cdot P(\mu_i, \delta_i|H_0) d\mu_i \, d\, \delta_i$$

Where $P(H_1)$ is the prior probability of exon i being differentially spliced, determined by exon-specific *cis* features and sample-specific *trans* RBP expression levels in the two biological conditions, which is independent of the observed RNA-seq read counts. $P(H_0)$ = 1 − $P(H_1)$ is the prior probability of exon i being unchanged. $P(I_{ij}, S_{ij}|H_1)$ and $P(I_{ij}, S_{ij}|H_0)$ represent the likelihoods under the model of differential splicing or unchanged splicing respectively. Z is a normalizing constant.

Since we are comparing only two models, we can further re-write the above equation as a factorization of the ratios between prior and likelihood:

$$\frac{P(H_1|I_{ij}, S_{ij})}{P(H_0|I_{ij}, S_{ij})} = \frac{P(H_1)}{P(H_0)} \cdot \frac{P(I_{ij}, S_{ij}|H_1)}{P(I_{ij}, S_{ij}|H_0)}$$

Note that when the prior distribution is flat, i.e. $P(H_0)$ = $P(H_1)$ = 0.5, the above equation is equivalent to a likelihood ratio test, which we refer to as DARTS BHT(flat). When $P(H_0)$ and $P(H_1)$ incorporate an informative prior based on exon- and sample-specific predictive features, we refer to this DARTS BHT model as DARTS BHT(info).

Finally, using the equation above, we can derive the marginal posterior probability $P(\delta_i|I_{ij}, S_{ij})$ for the parameter of interest δ_i as a mixture of the posterior conditioned on the two models:

$$P(\delta_i|I_{ij}, S_{ij}) = P(\delta_i|H_1, I_{ij}, S_{ij}) \cdot P(H_1|I_{ij}, S_{ij}) + P(\delta_i|H_0, I_{ij}, S_{ij}) \cdot P(H_0|I_{ij}, S_{ij})$$

Hence, the final inference is performed on the probability $P(|\delta_i| > c|I_{ij}, S_{ij})$. In our analysis, we set c=0.05 (i.e. a 5% change in exon inclusion level) and call events with $P(|\delta_i| > 0.05 |I_{ij}, S_{ij}) > 0.9$ as significant differential splicing events and $P(|\delta_i| > 0.05 |I_{ij}, S_{ij}) < 0.1$ as significant unchanged splicing events. Events with $0.1 \leq P(|\delta i| > 0.05|Iij, Sij) \leq 0.9$ are deemed as inconclusive. In the following text, we omit the subscripts and use $P(|\delta_i| > c |I_{ij}, S_{ij})$ and $P(|\Delta\psi| > c)$ interchangeably.

2.4.2 DARTS DNN model for predicting differential alternative splicing

A core component of the DARTS BHT framework is a DNN model that generates a probability of differential splicing between two biological conditions using exon and sample-specific predictive features. We designed the DARTS DNN to predict differential splicing of a given exon based on exon-specific cis sequence features and sample-specific trans RBP expression levels in two biological conditions. As noted above, a useful feature of the DARTS BHT framework is its capability to determine the statistical significance of both differential splicing events and unchanged splicing events. Specifically, for a splicing event i in the comparison k between RNA-seq datasets from two distinct biological conditions, let Y_{ik}=1 if this event is differentially spliced (that is, H1 is true) and Y_{ik}=0 if H_0 is true as labels for differential and unchanged splicing events, respectively. The task of predicting differential splicing can be formulated as

$$P(Y_{ik} = 1) = F(Y_{ik}; E_i, G_k)$$

where Y_{ik} is the label for event i in the comparison k; E_i is a vector of 2,926, 2,973, 2,971, and 1,748 cis sequence features for event i, including evolutionary conservation, splice site strength, regulatory motif composition, and RNA secondary structure for skipped exons,

alternative 5′ splice sites, alternative 3′ splice sites, and retained introns, respectively; and G_k is a vector of 2,996 (that is, 1,498×2) normalized gene expression levels of 1,498 RBPs in the two conditions. The prediction of $p(Y_{ik}=1)$ based on the features from any specific RNA-seq dataset can then be incorporated as an informative prior for p(H1) in the DARTS BHT framework. We implemented a deep-learning model (DARTS DNN) to learn the unknown function F that maps the predictive features to splicing profiles (differential versus unchanged). For skipped exons, we designed the DARTS DNN with four hidden layers and 7,923,402 parameters. The configuration of the DNN was as follows: an input layer with 5,922 (that is, 2,926+1,498×2) variables; four fully connected hidden layers with 1,200, 500, 300, and 200 variables and the ReLU activation function; and an output layer with two variables and the Softmax activation function. We implemented the DARTS DNN using Keras (https://github.com/ keras-team/keras) with the Theano back-end.

To mitigate potential overfitting of the DARTS DNN, we added a drop-out probability[19] for connections between hidden layers. Specifically, the variables in the four hidden layers were randomly turned off during the training process with probabilities of 0.6, 0.5, 0.3, and 0.1, respectively. We also added batch normalization layers[20] for all hidden layers to help the model converge and generalize. Finally, we used the RMSprop optimizer to adaptively adjust for the magnitudes of the components of the gradient in this deep architecture and chose 1,000 labeled alternative splicing events as one mini-batch to obtain a more stable gradient. In each mini-batch we balanced the composition of positive and negative labels by adding more positive events in the mini-batch such that positive:negative = 1:3 in the mini-batch. Because there were significantly more negative

(unchanged) events than positive (differential) events, such a balanced composition will provide a gradient for learning the positive events in different mini-batches.

To monitor the training loss and validation loss, we computed the loss every ten mini-batches and saved the current model parameters if the validation loss was lower than that of the previous best model. We trained the DARTS DNN on Tesla K40m.

2.4.3 Processing of ENCODE RNA-seq data and training of the DARTS DNN model

We used a comprehensive RNA-seq dataset from the ENCODE consortium to train the DARTS DNN. The ENCODE investigators have performed systematic shRNA knockdown of more than 250 RBPs in two human cell lines, HepG2 and K562. We downloaded all available (as of May 2017) RNA-seq alignments (ENCODE processing pipeline on the human genome version hg19) for shRNA-knockdown and control samples from the ENCODE data portal (https://www.encodeproject.org/).

We processed the RNA-seq alignments (bam files) using rMATS2 (v.4.0.1). Starting with RNA-seq alignment files, rMATS constructs splicing graphs, detects annotated and novel alternative splicing events, and counts the number of RNA-seq reads for each exon and splice junction. Given the modest depth of the ENCODE RNA-seq data (32 million read pairs per replicate on average), the read counts from the two replicates were pooled together for downstream analyses. We processed the raw RNA-seq reads with Kallisto[21] (v.0.43.0) to quantify gene expression levels using Gencode[22] (v.19) protein-coding transcripts as the index. For each of the two biological conditions in a given comparison (that is, shRNA knockdown and control), we extracted the Kallisto-derived gene-level

transcripts per kilobase million (TPM) values of 1,498 known RBPs11. We normalized the TPM value of each RBP by dividing by its maximum observed TPM value of all comparisons, then used that as the RBP expression feature in the DARTS DNN.

To generate training labels for the DARTS DNN, we applied DARTS BHT(flat) to the ENCODE RNA-seq data. Events with posterior probability $p(|\Delta\psi|>0.05) > 0.9$ were called positive (Y=1). Events with posterior probability $p(|\Delta\psi|>0.05) < 0.1$ were called negative (Y=0). We defined these significant differential splicing events and significant unchanged splicing events as labeled events and used them to train the DARTS DNN.

The vast majority of the RBPs (n=196) in the ENCODE data were knocked down by at least one shRNA in both HepG2 and K562 cell lines, corresponding to a total of 408 comparisons between knockdown and control. We set aside 10% of the labeled positive events and the same number of labeled negative events in each comparison as the testing data for estimating the generalization error of the trained DNN model. For the remaining 90% of the labeled events, we further split them into fivefold cross-validation subsets for the purposes of training, monitoring overfitting, and early stopping. We also collected ENCODE RBP-knockdown experiments performed in only one cell line (either HepG2 or K562; n=58) as leave-out datasets. All labeled events in these leave-out datasets were used only to evaluate the trained DARTS DNN and never during training.

We randomly drew 4 RBPs without replacement for a training batch, and iterated through all 196 RBPs as an epoch. The performance of the DARTS DNN was measured on the basis of the AUROC. The model with the best performance during training and cross-

validation was selected, and subsequently benchmarked using the testing data and leave-out data.

2.4.4 Rank-transformation of the DARTS informative prior

In a typical RNA-seq study, the number of unchanged splicing events can be orders of magnitude larger than the number of differential splicing events, and machine-learning algorithms may be biased to the majority class. To mitigate this potential bias, we used an unsupervised rank-transformation to rescale DARTS DNN scores to derive the informative prior for the DARTS BHT framework. Specifically, we first fit a two component Gaussian mixture model for all the DARTS DNN scores to derive the mean and variance of the two mixed Gaussian components, as well as the posterior probability λ of each DARTS DNN score belonging to a specific component. With the new mean and variance of the two Gaussian components set at $\mu 0$ and $\mu 1$, $\sigma 0$ and $\sigma 1$, respectively, each DARTS DNN score was rank-transformed to the new Gaussian components and then averaged by the weight parameter λ. Finally, to maintain a valid prior probability, we rescaled the transformed DARTS DNN scores to $[\alpha, 1-\alpha]$, where $\alpha \in [0, 0.5)$ sets the desired prior strength for the DARTS BHT framework and a smaller α value corresponds to a stronger strength of the informative prior. With this rescaling scheme, the entire ranks of the DARTS DNN scores are preserved while the potential bias for negative over positive events is reduced. In practice, we set $\mu 0=0.05$, $\mu 1=0.95$, $\sigma 0=\sigma 1=0.1$, and $\alpha=0.05$.

2.4.5 Generalization of the DARTS framework to diverse tissues and cell types

We generalized the DARTS framework to incorporate diverse tissues and cell types by using RNA-seq resources from the Roadmap Epigenomics project[4]. The Roadmap data were processed via the same protocol used for the ENCODE data. We took all Roadmap data with 101bp×2 or 100bp×2 paired-end RNA-seq, and truncated reads from the 101bp×2 datasets to 100bp for rMATS. In total, this represented 23 distinct tissues or cell types. All possible pairwise comparisons (n=253) between these 23 RNA-seq samples were made. Comparisons involving thymus were held out as Roadmap leave-out data, and all remaining comparisons were used as training datasets.

We trained three DARTS DNN models using different training datasets: (1) ENCODE data only, (2) Roadmap data only, and (3) the combination of ENCODE and Roadmap data. We subsequently benchmarked the performances of the three models by using ENCODE or Roadmap leave-out datasets.

2.4.6 DARTS splicing analyses of EMT-associated RNA-seq datasets

We applied the trained DARTS model to study EMT-associated alternative splicing events in two distinct human cell culture systems: H358 lung-cancer cell line induced to undergo EMT through a seven-day time course[14], and PC3E/GS689 prostate-cancer cell lines that had contrasting epithelial versus mesenchymal characteristics[2,16]. For the H358 time-course RNA-seq data (GSE75492), we used DARTS BHT(flat) to compare RNA-seq data from day 1 to day 7 against that for day 0. Splicing events that displayed a high DARTS DNN score of differential splicing (FPR<5%) and a non-trivial splicing change (more than 10% difference in exon inclusion level) but did not pass the significance threshold by DARTS

BHT(flat) using observed RNA-seq read counts alone were defined as DARTS DNN rescued events. We carried out motif analysis by calculating the average percentage of nucleotides covered by any of the top 12 ESRP SELEX-seq hexamer motifs[15] in a 45-bp sliding window. Background sequences were significant unchanged events by DARTS BHT(flat). For the PC3E and GS689 cell lines, we conducted RASL-seq[17] and RNA-seq experiments on the same batch of RNA samples, each with three replicates and on average 125 million read pairs per RNA-seq replicate (raw data deposited as GSE112037). RASL-seq reads were aligned to the pool of target splice junctions in the RASL-seq library using Blat[23]. RASL-PSI values were calculated as I/(I+S), where I is the number of exon-inclusion splice junction reads and S is the number of exon-skipping splice junction reads. Alternative splicing events with total RASL-seq read counts greater than five in every replicate were used for downstream analyses. Gene expression levels of RBPs in the two datasets were quantified with Kallisto v.0.43.0.

2.4.7 RASL-seq library preparation and sequencing

RASL-seq was performed as described[24], with some modifications. Total RNA from PC3E and GS689 cell lines were extracted with Trizol (Thermo Fisher Scientific). RASL-seq oligonucleotides (a gift from X.-D. Fu) were annealed to 1μg of total RNA and then subjected to selection by oligo-dT beads. Paired probes templated by poly(A)+ RNA were ligated and then eluted. We used 5μl of the eluted ligated oligos for eight cycles of PCR amplification using primers F1:

5′-CCGAGATCTACACTCTTTCCCTACACGACGGCGACCACCGAGAT-3′ and

R1: 5′-GTGACTGGAGTTCAGACGTGTGCGCTGATGCTACGACCACAGG-3′. One-third of the resulting PCR products were used in the second round of PCR amplification (nine cycles) with primers F2: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG-3′ and R2: 5′-CAAGCAGAAGACGGCATACGAGAT[index] GTGACTGGAGTTCAGACGTGTGC-3′; indexes used in this study were Illumina indexes D701–D706. The indexed PCR products were pooled and sequenced on a MiSeq with custom sequencing primer 5′-ACACTCTTTCCCTACACGACGGCGACCACCGAGAT-3′ and custom index sequencing primer 5′-TAGCATCAGCGCACACGTCTGAACTCCAGTCAC-3′.

2.4.8 Code availability

The DARTS program, trained model parameters, and predictive features are provided at GitHub (https://github.com/Xinglab/DARTS).

2.4.9 Data availability

The RNA-seq data that support the findings of the deep learning models are available from the ENCODE project (https://www.encodeproject.org/) and the Roadmap Epigenomics project (http://www.roadmapepigenomics.org/). The H358 time-course RNA-seq data were downloaded from GEO accession GSE75492. The PC3E-GS689 RNA-seq data and RASL-seq data can be accessed from GEO under accession GSE112037.

Acknowledgement

We thank X.-D. Fu (UCSD) for the RASL oligos and advice on RASL-seq. This study is supported by National Institutes of Health grants (R01GM088342, R01GM117624,

U01HG007912, and U01CA233074 to Y.X.). Z.Z. is partially supported by a UCLA Dissertation Year Fellowship.

2.5 Figures

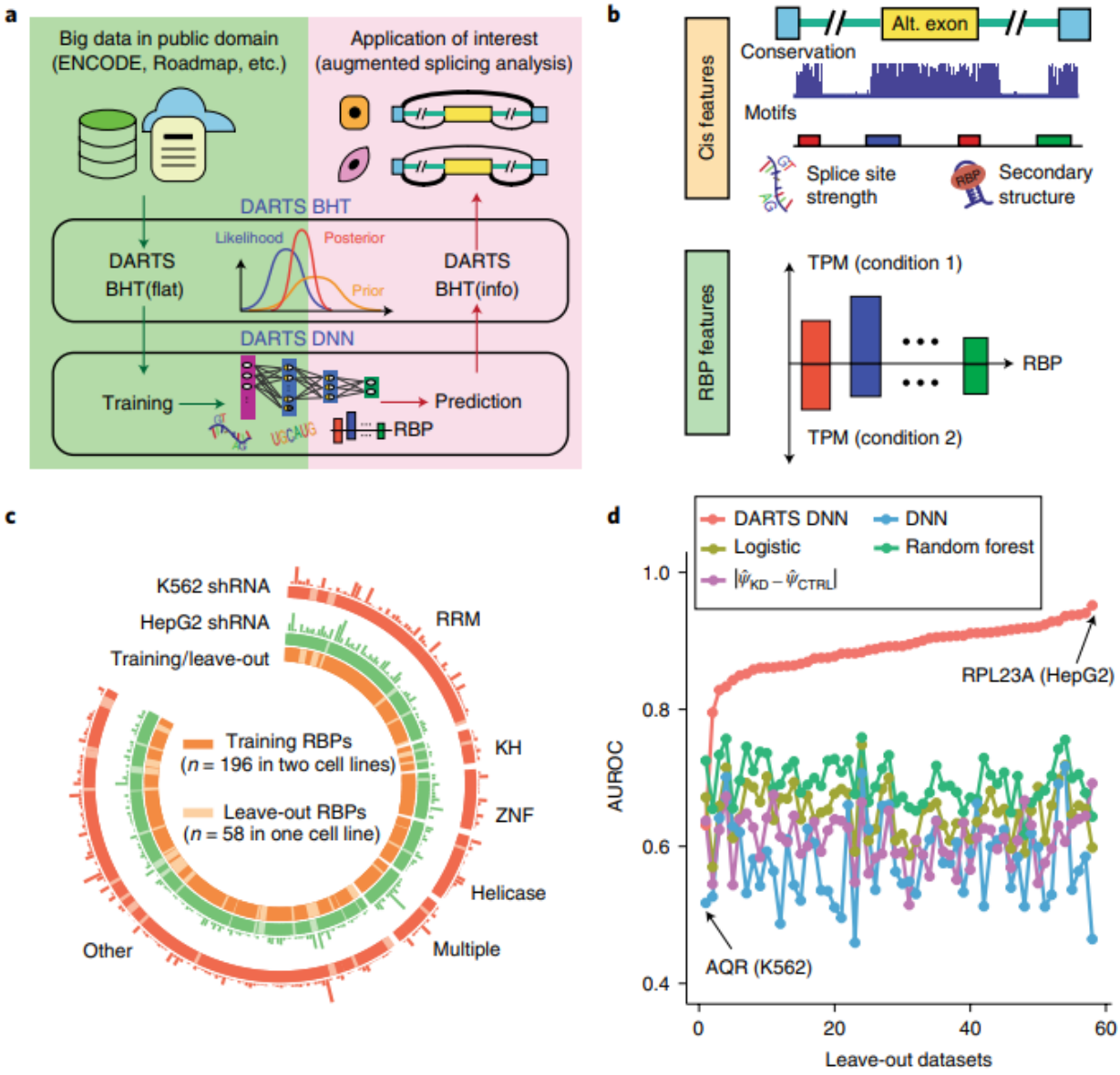


Figure 2.1 The DARTS computational framework.

(a) Overall workflow of DARTS. **(b)** Schematic of the DARTS DNN features, including cis sequence features and trans RBP features. **(c)** Overview of training and leave-out RBPs, and the number of significant differential splicing events called by DARTS BHT(flat) on the ENCODE data (illustrated by bar charts above the outer and middle circles). We used 196

RBPs knocked down in both the K562 and HepG2 cell lines for training (orange), while the remaining 58 RBPs knocked down in only one cell line were leave-out data (light orange) (illustrated in the inner circle). RRM, RNA recognition motif; KH, K homology; ZNF, zinc finger. **(d)** Comparison of the DARTS DNN with baseline methods in leave-out datasets. KD, knockdown; CTRL, control; RPL23A, ribosomal protein L23a; AQR, aquarius intron-binding spliceosomal factor.

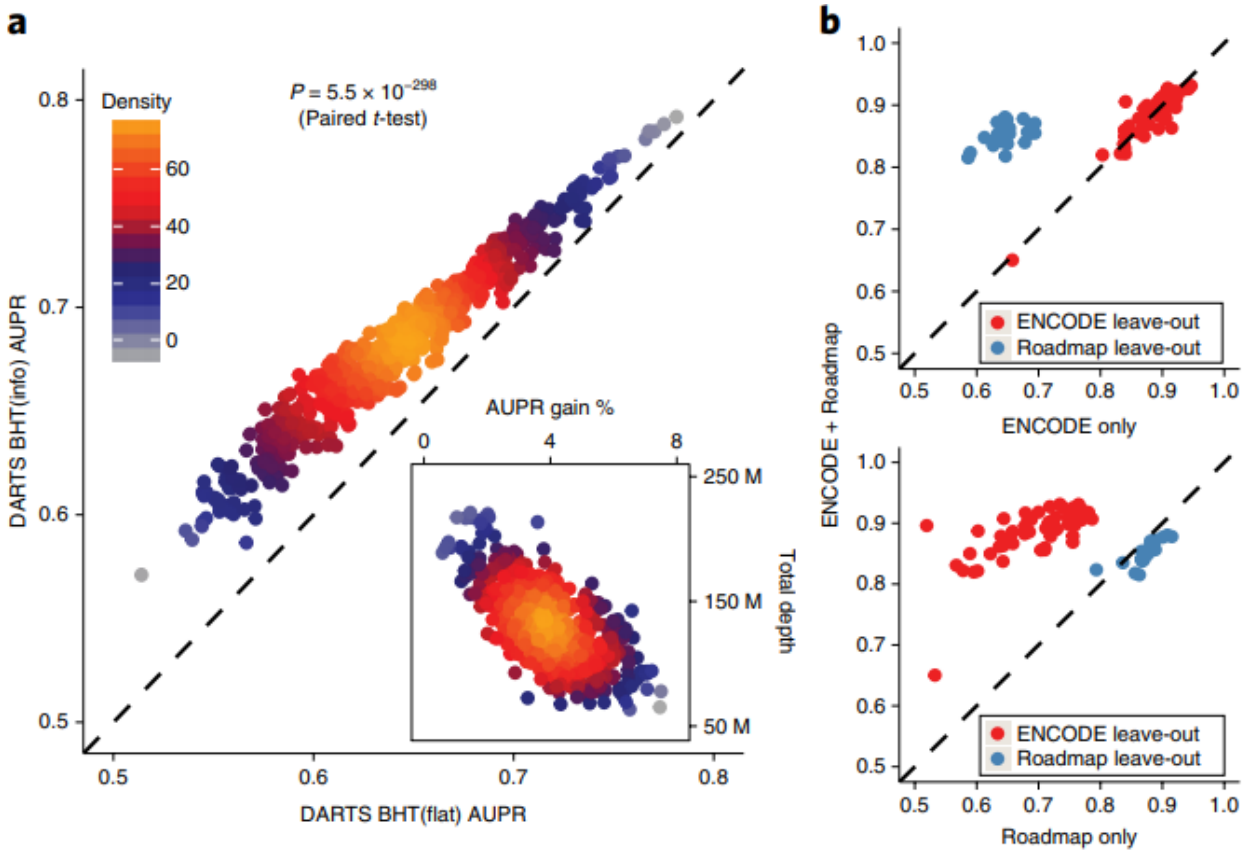


Figure 2.2 Performance evaluation of the DARTS BHT framework, and the influence of training datasets on the performance of the DARTS DNN.

(a) The performance of DARTS BHT(info) versus DARTS BHT(flat) in the cell-type-specific differential splicing analysis of HepG2 and K562 (two sided paired t-test; n= 672 pairwise comparisons). The performance gain by DARTS BHT(info) is plotted against the RNA-seq depth in pairwise comparisons of individual replicates (inset). **(b)** AUROC values of the DARTS DNN trained on both ENCODE and Roadmap data, ENCODE data only, or Roadmap data only when applied to ENCODE or Roadmap leave-out data.

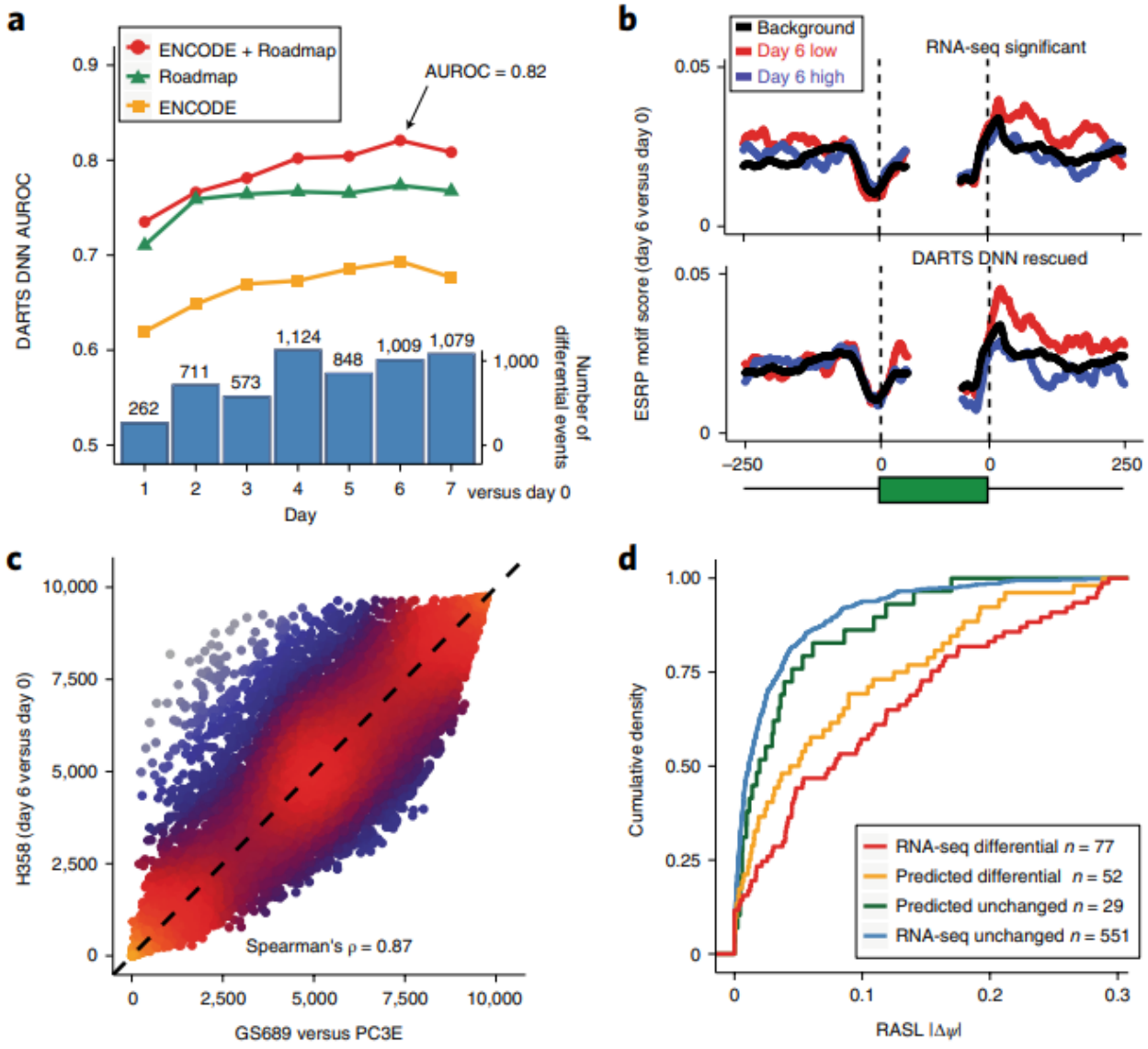
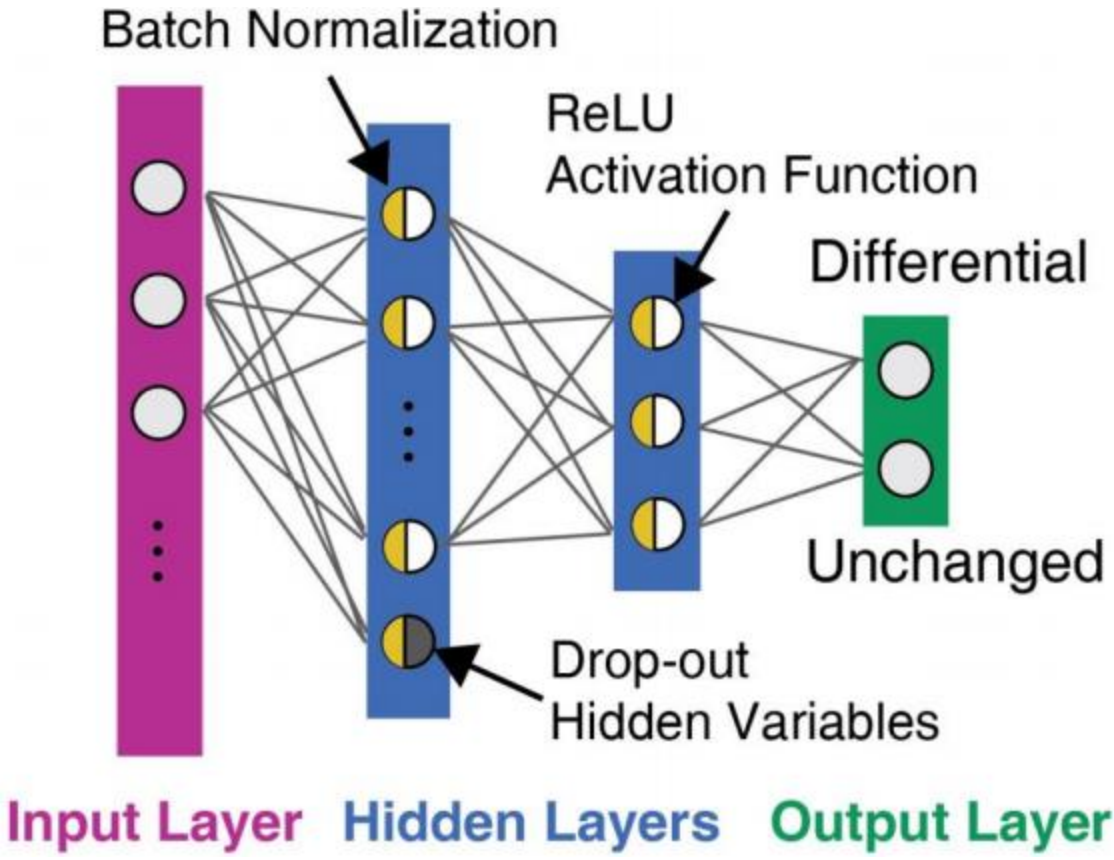


Figure 2.3 DARTS analysis of alternative splicing during the EMT.

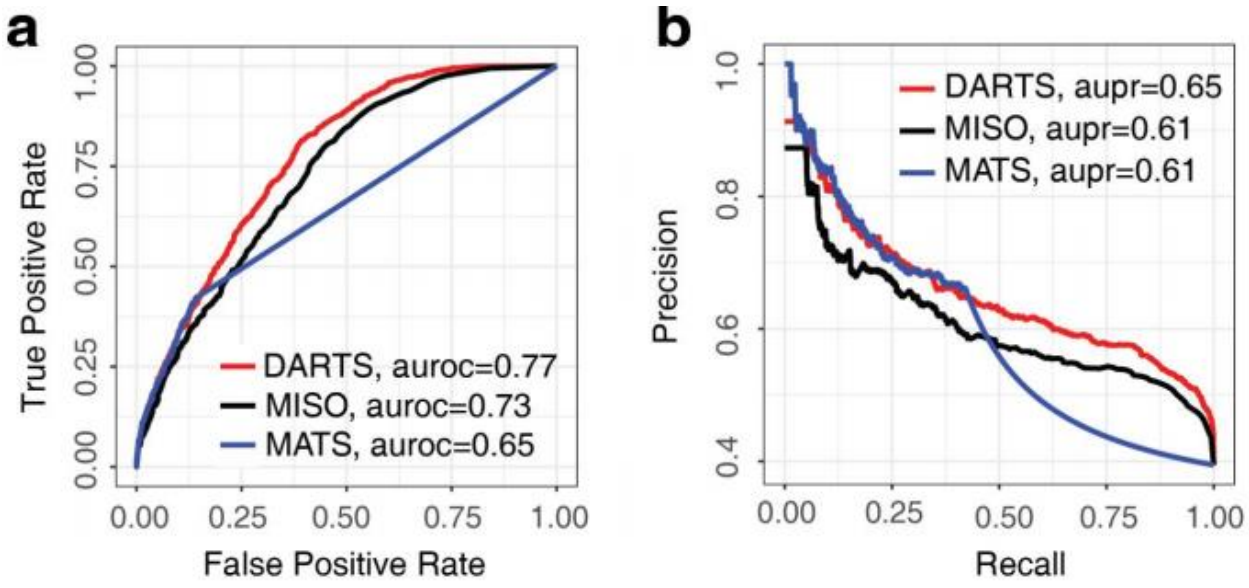
(a) The performance of the DARTS DNN on the time-course RNA-seq data of an inducible H358 lung cancer cell line model of the EMT. The numbers of differential splicing events called by DARTS BHT(flat) are shown as bar plots at the bottom. **(b)** Meta-exon-motif analysis of the ESRP motif for RNA-seq differential events called by DARTS BHT(flat) and DARTS DNN rescued events in the comparison of day 6 versus day 0. **(c)** DARTS DNN predictions for the H358 EMT time course (day 6 versus day 0) and in GS689 versus PC3E.

Plotted are the ranks of predicted DARTS DNN scores. **(d)** RASL-seq validation of RNA-seq called events and DARTS DNN predicted events. Plotted are the cumulative density functions of the RASL – |ΔPSI| values of RNA-seq inconclusive events with high DARTS DNN scores (FPR 80%; n= 29 events) (green line).



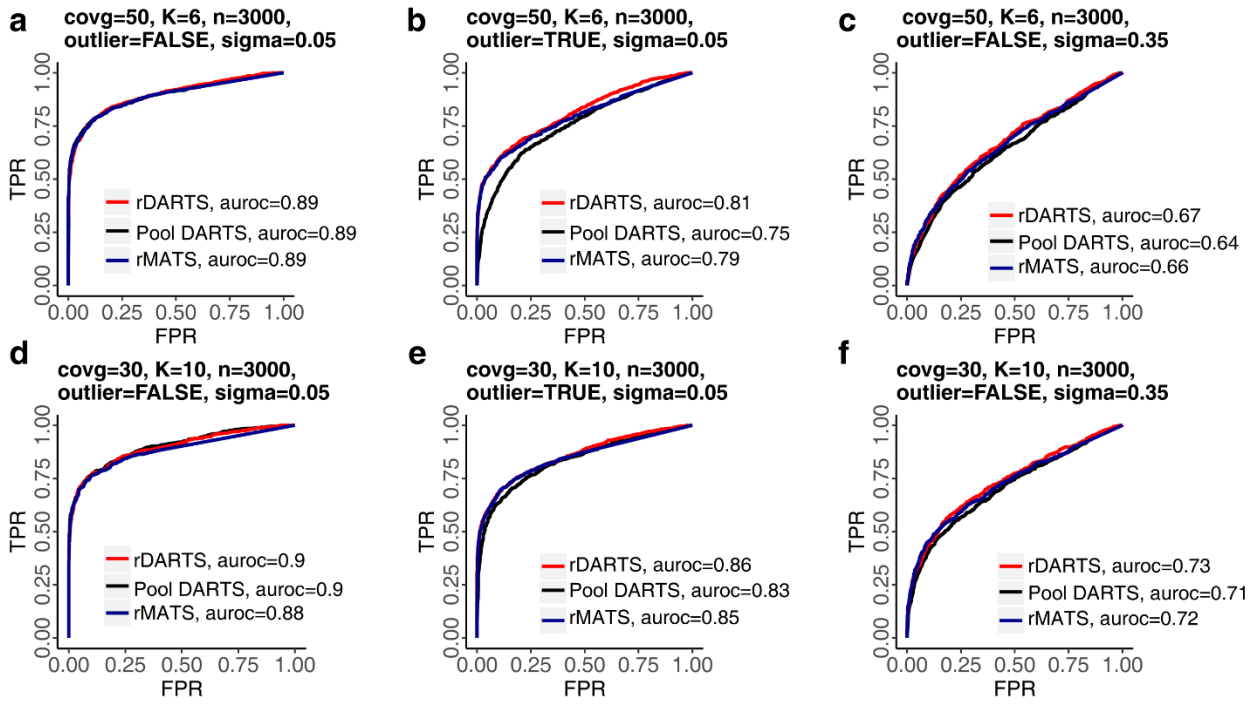
Supplementary Figure 2.4 Schematic overview of the DARTS DNN model.

The DARTS DNN model consists of four hidden layers and 7,923,402 parameters. Batch normalization and drop-out of hidden variables are implemented during training to mitigate overfitting.



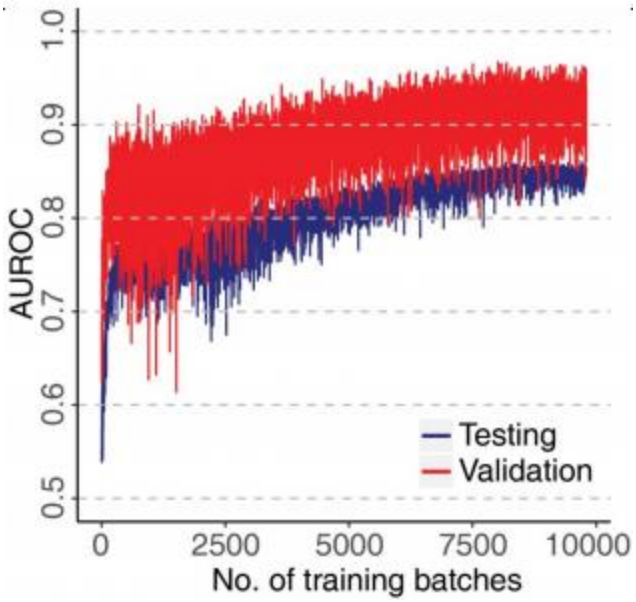
Supplementary Figure 2.5 Performance comparison of DARTS BHT(flat), MISO, and MATS using simulated RNA-seq data generated by Flux simulator.

We derived the transcriptome profiles from a real RNA-seq dataset with widespread splicing changes (E-MTAB-1147; knockdown of splicing factor HNRNPC in the HeLa cell line), and plugged into Flux simulator as ground-truth to simulate RNA-seq reads. Then **(a)** AUROC and **(b)** AUPR were computed for each statistical method by labelling the exon skipping events with ground-truth $|\Delta\psi|>0.05$ as positive and $|\Delta\psi|\leq 0.05$ as negative (for details, see Supplementary Notes). DARTS BHT(flat) performs favorably to MISO and MATS.



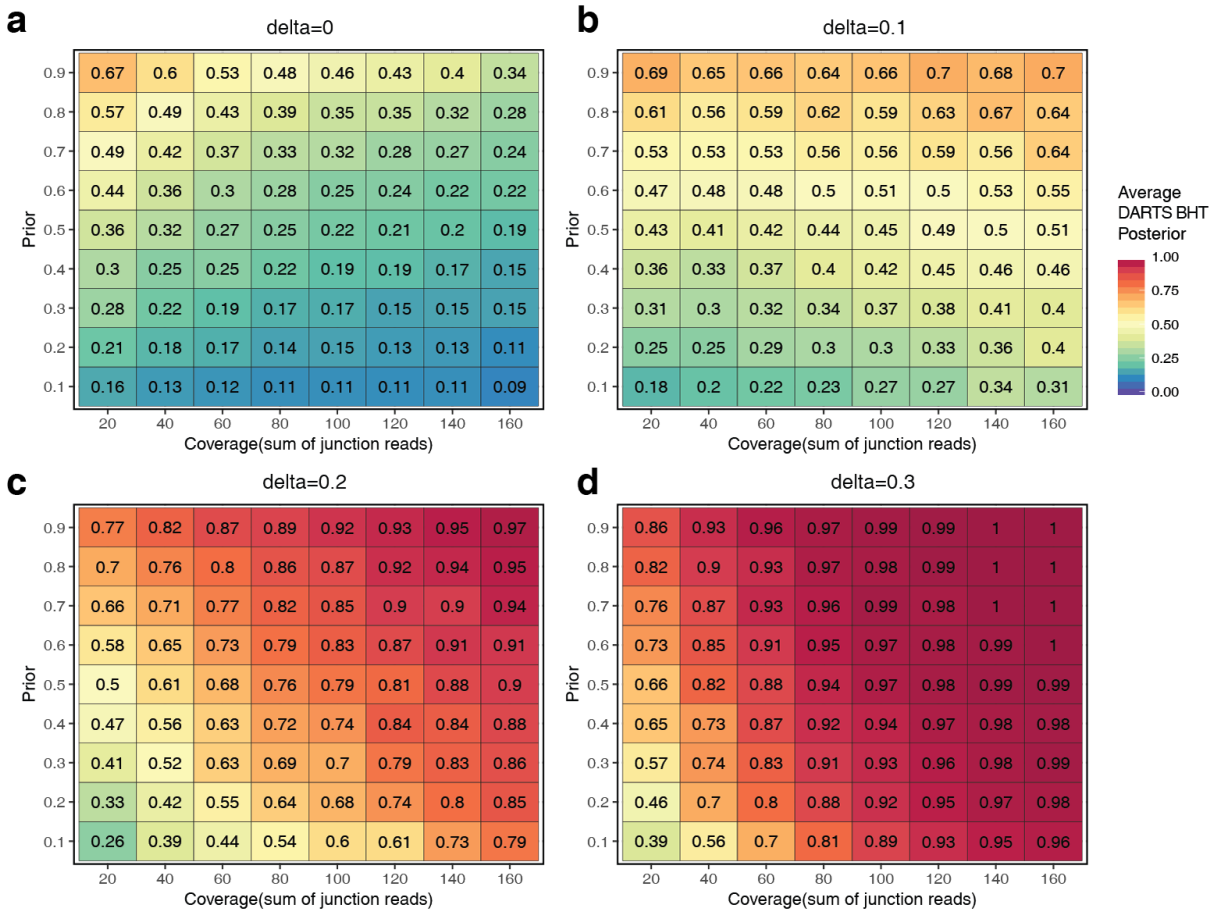
Supplementary Figure 2.6 Performance comparison of DARTS BHT(flat) with replicates versus DARTS BHT(flat) on pooled data and rMATS with replicates.

We fixed the total RNA-seq read counts (coverage per replicate x number of replicates) while varying the number of replicates (K), within group variance (sigma), and whether there is one outlier sample. The replicate DARTS model (rDARTS) outperforms DARTS on pooled data when there exists outlier samples **(b,e)** or when the within-group variance is large **(c,f)**.



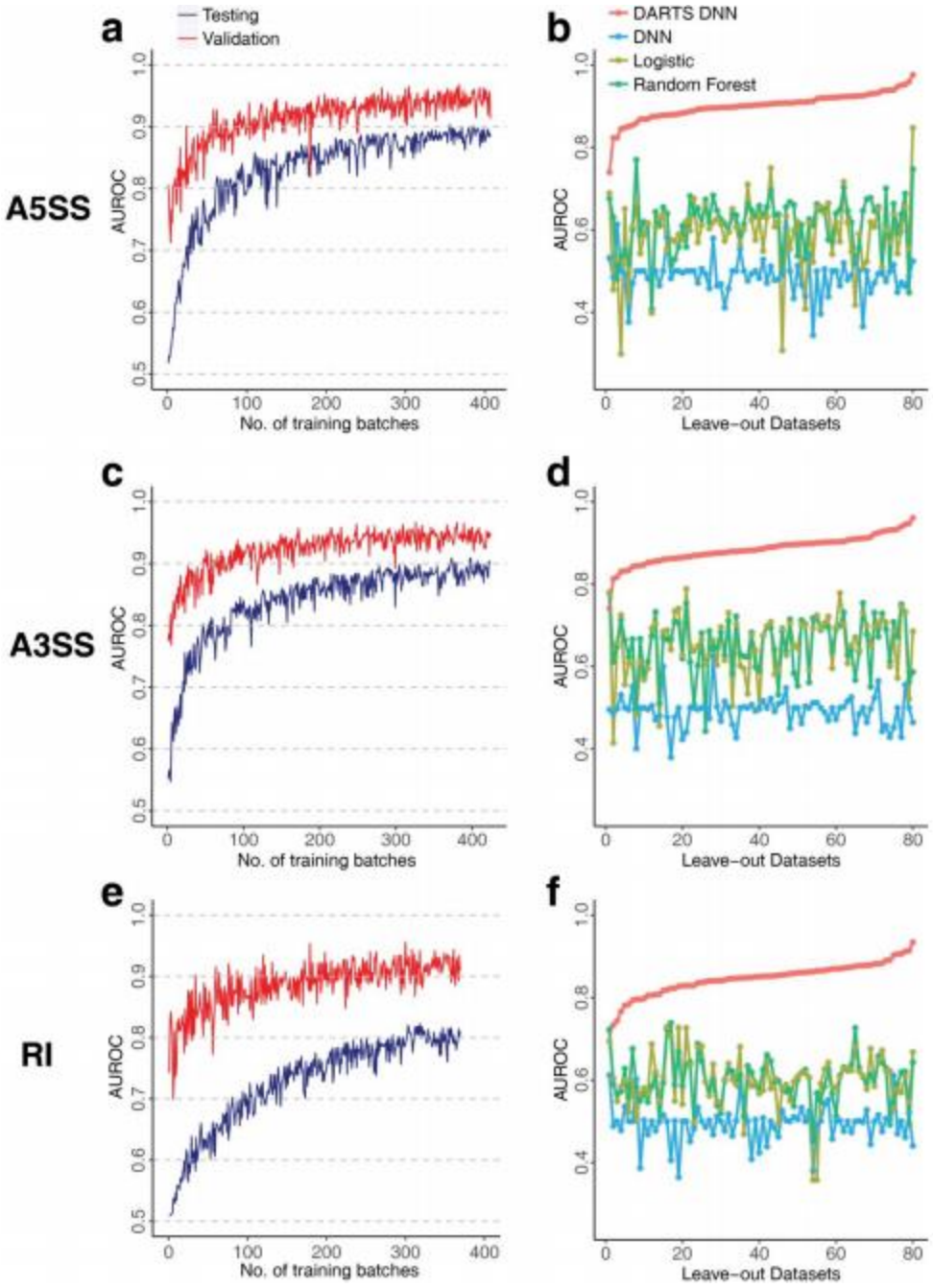
Supplementary Figure 2.7 The performance of the DARTS DNN during cross-validation and testing as training progressed.

The maximum AUROC was 0.97 during cross-validation and 0.86 during testing.



Supplementary Figure 2.8 Relationship of DARTS posterior, prior, and the amount of observed RNA-seq read counts.

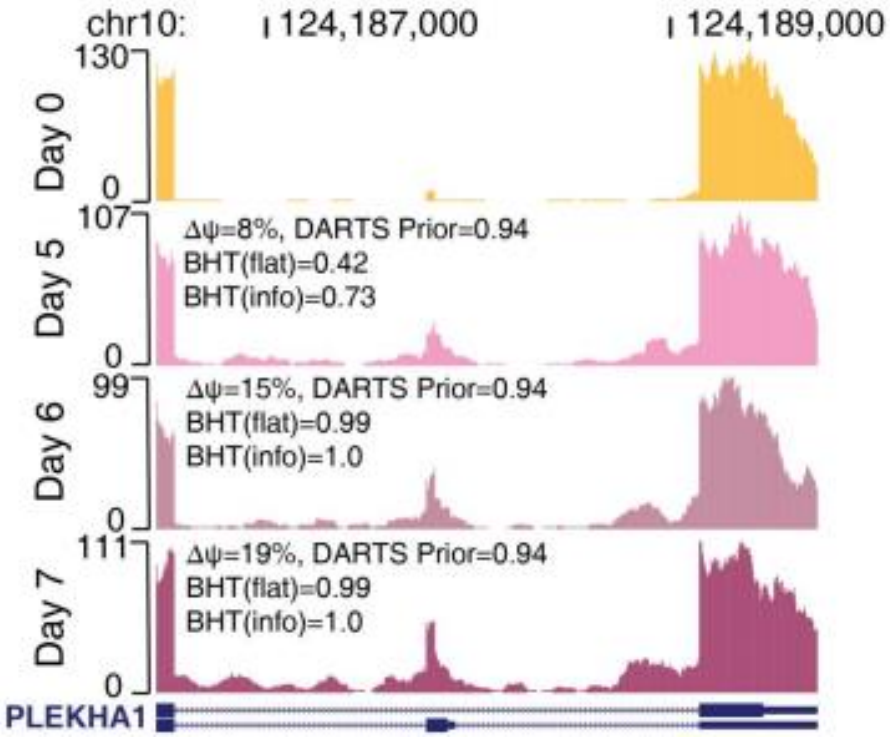
For a fixed absolute PSI difference between the two conditions, i.e. the effect size (denoted as delta), posterior probability P(|δ|>0.05|I,S) was computed from simulated data by varying the prior probability and the amount of read counts. The prior's effect on DARTS posterior diminished when the observed read counts were large (>100) and/or with large effect size (delta=0.3). For events with moderate or low read counts, a strong informative prior improves the inference accuracy.



Supplementary Figure 2.9 Application of the DARTS DNN to different classes of alternative splicing patterns.

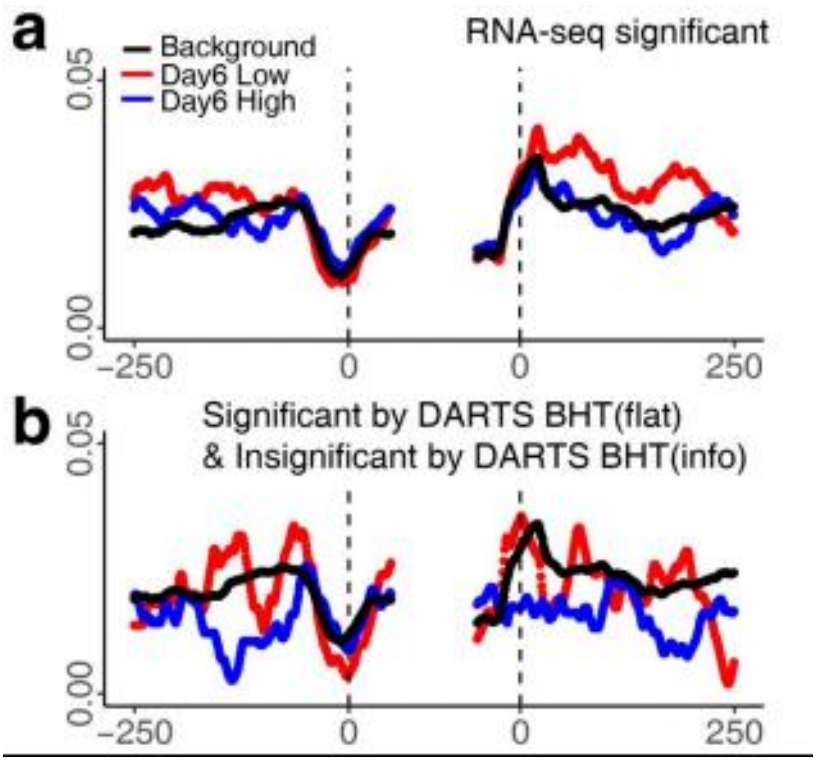
(a, c, e) The performance of the DARTS DNN on validation and testing data as training progresses for alternative 5' splice sites (A5SS), alternative 3' splice sites (A3SS), and

retained introns (RI) as measured by AUROC. **(b, d, f)** Comparison of the DARTS DNN with baseline methods in independent leave-out datasets. DARTS DNN outperforms baseline methods trained on individual leave-out datasets by a large margin. Note that in these analyses the DARTS DNN is trained using combined ENCODE + Roadmap RNA-seq datasets, with certain pairwise comparisons held-out for benchmarking as independent leave-out datasets.



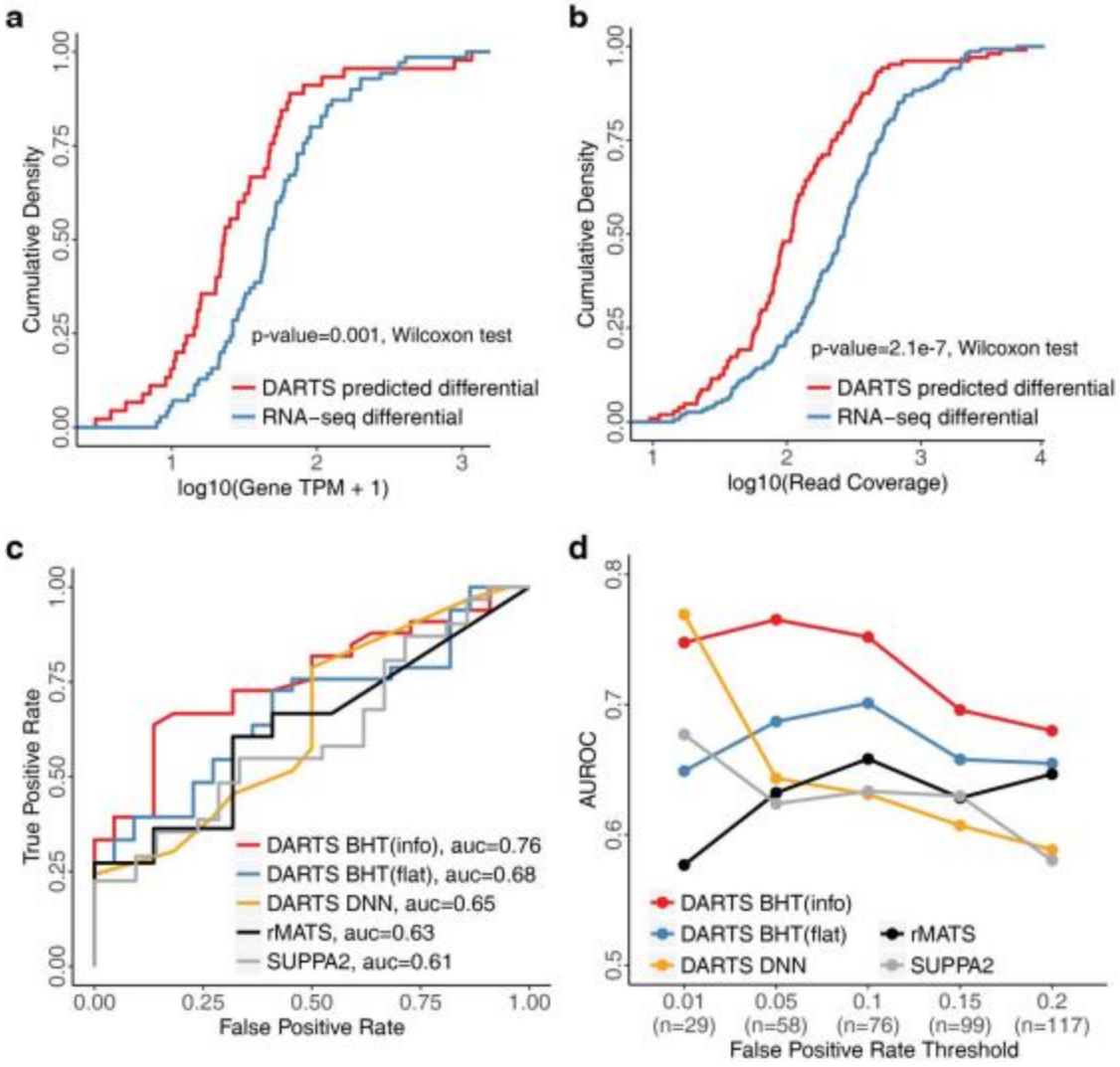
Supplementary Figure 2.10 An example of the DARTS DNN prediction for the PLEKHA1 gene in the H358 EMT time-course RNA-seq data.

The genome browser view represents aggregated RNA-seq signals from three biological replicates. The DARTS DNN score for this exon is 0.94 in day 5 versus day 0, increasing the posterior probability of differential splicing to 0.73 over 0.42 when using RNA-seq data alone. The differential splicing pattern of this exon was apparent throughout the time course and was previously validated by RT-PCR.



Supplementary Figure 2.11 Meta-exon motif analysis of the ESRP motif.

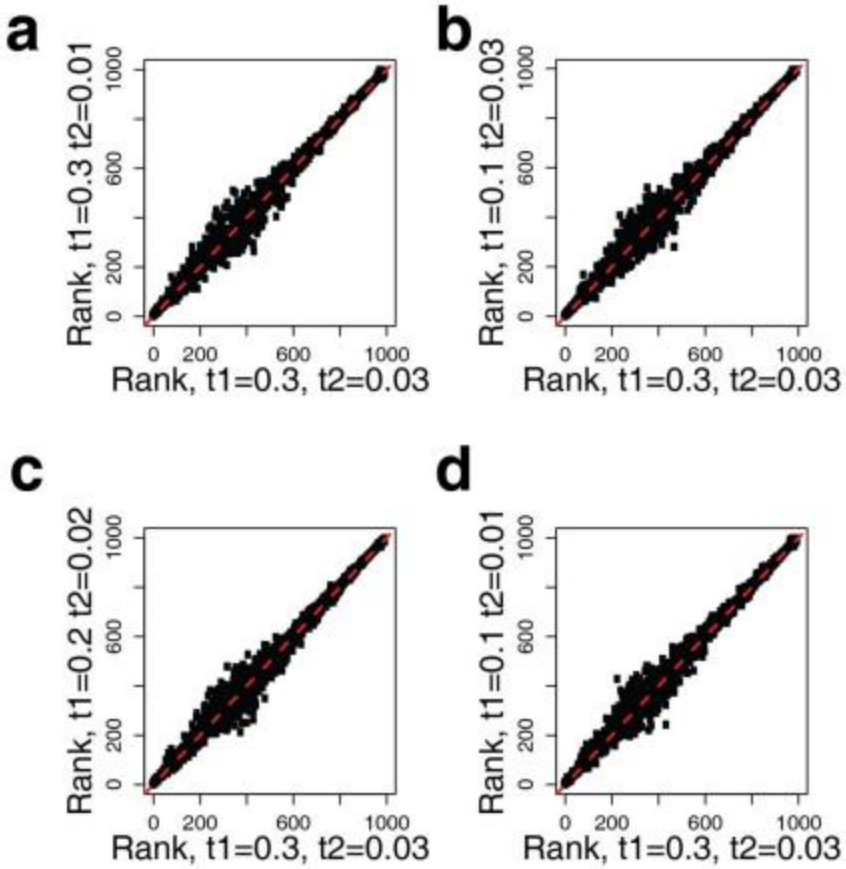
(a,b) For the comparison of day 6 versus day 0 on the H358 time-course RNA-seq data, we calculated ESRP motif scores for **(a)** all DARTS BHT(flat) significant events and **(b)** DARTS BHT(flat) significant events that become insignificant in DARTS BHT(info). The latter set of events does not have enrichment of the ESRP motif.



Supplementary Figure 2.12 Characteristics of the DARTS DNN predicted events.

(a,b) The cumulative density function of **(a)** gene expression levels (TPM values) and **(b)** RNA-seq read coverage for DARTS-DNN-predicted differential events and RNA-seq differential events. The DARTS-DNN predicted differential events are from genes with significantly lower expression levels and have significantly lower RNA-seq read coverage compared with that of RNA-seq differential events (two-sided Wilcoxon test). c, DARTS BHT(info) outperforms baseline methods that use RNA-seq data alone to call differential splicing (DARTS BHT(flat), rMATS, and SUPPA2), as benchmarked using ground truth

defined by RASL-seq. d, DARTS BHT(info) outperforms baseline methods at different FPR thresholds for DARTS-DNN-predicted differential events (n represents the number of alternative splicing events), with the maximum gain observed for the most confidently predicted events with FPR = 1%.



Supplementary Figure 2.13 Ranking by DARTS BHT on simulated data when using different t1 and t2 values.

The results of DARTS BHT are robust to different choices of parameters, especially for the inference of differential alternative splicing events (upper right corner in each panel).

2.6 Appendix

2.6.1 DARTS BHT statistical modelling

Benchmarking DARTS BHT on simulated data

Generation of benchmark dataset

In order to better represent the variability inherent in real experimental datasets, while knowing the ground-truth, we employed the flux-simulator[25] software (v1.2.1) to simulate RNA-seq reads. Flux-simulator is a specialized simulation software program that models RNA-seq experiments using a set of modules for different experimental procedures, including RNA fragmentation, library preparation and high-throughput sequencing. The major advantage of simulating data using flux-simulator as opposed to directly drawing reads from a statistical distribution is that the former approach takes the variances/noises at different stages into consideration, whereas the latter assumes all reads are generated by a simple stochastic process and are counted correctly; hence, our approach better captures the real-world variances compared to a naïve simulator.

Flux-simulator simulates RNA-seq reads based on a given molecular profile that contains "number of molecules" for each transcript. We derived the molecular profile from a previously published dataset, E-MTAB-1147 from Array Express, which is an RNA-seq experiment of HeLa cell line upon hnRNPC knockdown[26]. We chose this dataset because our previous analysis had demonstrated that it contained abundant splicing changes[27], and that its sequencing depth was sufficiently deep to ensure robust estimation of the transcript expression. We used Kallisto4 (v 0.43.0) to estimate the transcript TPM from the raw reads using Gencode5 V19 as reference GTF. The transcript TPM was subsequently

converted to number of molecules by fixing the total number of molecules at 5,000,000 (default setting in flux-simulator) and rounding fractional molecules to the nearest integer.

Taking the customized molecular profile, we ran Flux-simulator using: the fragment distribution derived from the above experiment, sequencing read length equal to 72bp with 100 million paired-end reads, and leaving other parameters at their default settings. Next, we ran STAR (v 2.5.2a) to map the reads to the hg19 version of genome with Gencode v19 as gene annotation file. The resulting outputs were two alignment bam files corresponding to the profiles derived from Control and hnRNPC knockdown.

Evaluation of DARTS BHT, MISO and MATS

We processed the alignment files with rMATS[2] (v4.0.1) to count the junction-spanning reads with Gencode v19 as reference annotation. The inclusion junction counts and skipping junction counts for all detected events were then fed into the DARTS BHT model with a flat prior as input. We ran DARTS BHT with τ_1 = 0.3, τ_0= 0.03 and testing for C=0.05. The output of DARTS BHT was subsequently benchmarked using the true delta-PSI values between two conditions.

Note that we only considered simple skipping events in the simulation study, because the complex events are often combinations of multiple alternative splicing events and the true PSI values are often ambiguous to define and hence hard to compute. We define the simple events as events with a unique one-to-one mapping for the 5'- and 3'- splice sites of the middle skipping exon, upstream exon 5'- to middle exon 3'- splice site, and middle exon 5'- to downstream 3'- splice site. After filtering for simple events, we had 7,678 simple exon-skipping events out of 16,676 exons detected by rMATS.

As a comparison, we also ran MISO[1] (v0.5.3) and MATS (v4.0.1) on the simulated datasets. To run MISO exoncentric analysis, we downloaded the Human genome (hg19) annotation file v1.0 from the MISO website (https://miso.readthedocs.io/en/fastmiso/annotation.html) and built the index for MISO using the Skipping Events (SE) in the annotated folder by "index_gff". Next, we ran MISO to quantify the splicing level under each condition then used "compare_miso" to compute the Bayes Factor for the skipping events in MISO annotation files. We ran the MATS statistical model with setting C=0.05 on the read counts generated by rMATS.

Since MISO analyzes its own internal skipping events annotation which is different from the simple events definition in DARTS and MATS, we took the intersection of the events from these software programs. There were 3,407 common events between the two software programs, with 1,344 events' absolute delta-PSI larger than 5% which we labeled as positive. We measured the accuracy of DARTS BHT, MISO and MATS by AUROC and AUPR. As shown in **Supplementary Figure 2.5**, DARTS compares favorably to MISO and MATS, demonstrating its superior inference power to the state-of-the-art splicing inference tools when using only empirical RNA-seq data.

DARTS BHT statistical model for unpaired or paired replicates

Illustration of replicate DARTS BHT statistical model

Thanks to the rapid development of sequencing technology, it has become practical and common for transcriptomic studies to carry out RNA-seq experiments with multiple replicates to quantify biological variances and improve reproducibility. Previously we had demonstrated that pooling the reads from different replicates is not recommended[2].

Motivated by the replicate analysis, we sought to develop the replicate DARTS model that considers replicates in its likelihood function while still being capable of taking the informative prior into account.

Following the notations in the DARTS main text, we extend the DARTS BHT model to include read counts from different replicates into the following hierarchical model (we are abusing the subscripts k here to index replicate; whereas k was used to index different experimental conditions in the main text):

$$I_{ijk}|\psi_{ijk} \sim Binomial(n = I_{ijk} + S_{ijk}, p = f_i(\psi_{ijk}))$$

$$\psi_{ijk} = \mu_i + 1(j = 2) \cdot \delta_i + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma^2)$$

$$\mu_{ik} = \mu_i + \varepsilon_k, \qquad \mu_{ik} \sim N(\mu_i, \sigma^2)$$

$$\mu_i \sim Unif(0,1)$$

$$\delta_i \sim N(0, \tau^2)$$

I_{ijk}, S_{ijk} and ψ_{ijk} are the inclusion read counts, the skipping read counts and the exon inclusion level for exon i, sample group j=1,2, in replicate k; f_i is the length normalization function for exon i; μ_i is the baseline inclusion level for exon i, and δ_i is the difference of the exon inclusion levels between the two conditions. Without loss of generality, we let $\psi_{i1k} = \mu_i + \epsilon_{ik}$, $\psi_{i2k} = \mu_i + \delta_i + \epsilon_{ik}$; that is, we assume that the effect size δi is the same across different replicates; and that ψijk values in each replicate k have a random replicate-specific deviation from the group mean μ_i by ϵ_{ik}. The term ϵ_{ik} captures the within group variance of PSI values in different replicates and has an expectation of 0.

It is worthwhile to point out that the above replicate DARTS framework is applicable for both paired replicates and unpaired replicates. The subscript k indexes for the samples from different/same origins. For paired replicates, the two paired observations under the two corresponding conditions are indexed with the same k, and should therefore share the same starting point/baseline level of $\mu_{ik} = \mu_i + \epsilon_{ik}$, while only differing by the amount δ_i caused by the treatment. For unpaired replicates, each sample is indexed with a different k, hence the baseline level μ_{ik} was drawn independently from $N(\mu_i, \sigma^2)$ and there is no covariance between samples in the two groups.

Simulated read counts and evaluation

Next, we simulated read counts by drawing reads from binomial distributions. We did not use flux-simulator for this analysis because it is non-trivial to define the within group variances at the "number of molecules" level; instead, we imposed a normal distribution to the simulated group mean PSI value, then drew read counts from this hierarchical generating process.

We performed extensive simulation studies using different combinations of parameters. Specifically, we set the model parameters equal to the following values: $\delta \in$ the within group variance, smaller values of σindicated more consistent patterns across replicates; K the number of replicates, more replicates would help better capture the within group variance; $n \in$,the coverage of each replicate, deeper coverage would help estimation of sample-wise PSI; presence of outlier, outlier PSI value was draw randomly from [0,1] to represent one unrelated sample out of the all replicates. We benchmarked the performances of pooled DARTS, replicate DARTS (rDARTS), and rMATS, using the AUROC

and AUPR. To obtain a reliable performance estimate, we randomly sampled n=3,000 events under each simulation configuration, with the expected differentially spliced events (positive cases) at 50%.

As shown in **Supplementary Figure 2.6**, replicate DARTS showed a consistent gain in power under two specific situations, regardless of the number of replicates: i) when the within-group variance σ is large, and ii) when there is an outlier sample. This is consistent with our previous observation in the rMATS paper. Notably, in all simulations, we fixed the total coverage at 300, i.e. when K=6, each sample has 50 read counts per event; when K=10, each sample has 30 read counts per event. Such configurations emulate a fixed sample-size budget, where researchers hope to get the best scientific outcomes using the optimal experimental design. It is not surprising that increasing the number of 6 replicates by 4 would significantly reduce the loss of power caused by introducing 1 outlier sample. The same effect was true for larger within-group variances, demonstrating the better group variance estimation captured by more replicates with less coverage per replicate. In all comparisons, the replicates DARTS model outperforms the pooled DARTS model under certain conditions, while inflicting no loss of power under regular conditions. Hence, we recommend using the replicate DARTS model whenever possible, and advise against pooling reads from replicates.

Technical notes on statistical model optimization

Laplacian approximation

The optimization of the DARTS model involves two major steps: i) calculating the Bayes Factor of two competing models/hypotheses, ii) sampling the posterior distribution given

the non-conjugate priors. In this part we will first deal with the calculation of the Bayes Factor, where we utilized Laplace's method to approximate the intractable integrals.

Following the notation in the Method section, the essence of DARTS BHT with flat prior is the ratio of the integrated likelihood function, also known as the Bayes Factor. In the DARTS model, the integrated likelihood function takes the form of

$$P(I_{ij}, S_{ij} \mid H_n) = \iint_{\Theta_n} P(I_{ij}, S_{ij} \mid \mu_i, \delta_i) \cdot P(\mu_i, \delta_i \mid H_n) d\mu_i d\delta_i$$

$$\propto \int_{-\infty}^{+\infty} \int_{0}^{1} f_i(\psi_{i1})^{I_{i1}} \cdot \left(1 - f_i(\psi_{i1})\right)^{S_{i1}} \cdot f_i(\psi_{i2})^{I_{i2}} \cdot \left(1 - f_i(\psi_{i2})\right)^{S_{i2}} \cdot \mathbf{1}(|2 \cdot (\mu_i + \delta_i - 0.5)| < 1) \cdot e^{-\delta^2/\tau_n^2} d\mu_i d\delta_i$$

$$= \iint_{\Theta_n} g(\mu_i, \delta_i; I_{ij}, S_{ij}) d\mu_i d\delta_i$$

$$= \iint_{\Theta_n} \exp\left(g_1(\mu_i, \delta_i; I_{ij}, S_{ij})\right) d\mu_i d\delta_i$$

The above integral cannot be solved in closed form. Instead, we employ Laplace's method to approximate the integral. Let $g_1 = \log g$ be the log posterior density function, the Laplacian approximation can be viewed as the Gaussian approximation to any (posterior) distribution that is smooth and well-peaked around its maximal point. To implement Laplacian approximation for DARTS BHT, we compute both the maximal point of the posterior probability as well as the local curvature/Hessian matrix around the maximal point using the "optim" function in R by feeding its objective function and the gradient function. Then the approximation for the integral, denoted by Z_n, is

$$Z_n = \log\left(P(I_{ij}, S_{ij} \mid H_n)\right) \approx g_1(\hat{\mu}_i, \delta_i; I_{ij}, S_{ij}) - 0.5 \times \log\left(|\mathrm{H}(\hat{\mu}_i, \hat{\delta}_i)|\right) + \frac{d}{2} \cdot \log(2\pi)$$

$\hat{\mu}_i$, δ_i are the parameter values that maximize posterior probability; $g_1(\hat{\mu}_i, \delta_i; I_{ij}, S_{ij})$ is the log posterior probability function evaluated at maximal point; $\mathrm{H}(\hat{\mu}_i, \hat{\delta}_i)$ is the Hessian matrix of g_1 evaluated at the maximal point; and d is the total number of parameters in$g_1(\cdot)$. Then, the Bayes Factor (BF) is

$$\mathrm{BF} = \frac{P(I_{ij}, S_{ij} \mid H_1)}{P(I_{ij}, S_{ij} \mid H_0)} = \exp(Z_1 - Z_0)$$

MCMC sampling

Next we seek to sample from the posterior distribution of the parameters given the data/observations under a specific hypothesis. Since we do not have the conjugate prior for the likelihood, we employ an MCMC random walk to draw samples from the posterior distribution. Specifically, we designed the transition probability q as a normal distribution with mean equal to the current state and a small variance corresponding to a small step size. For each proposed state, we accept the proposal by a Metropolis-Hasting acceptance probability:

$$\alpha(\theta^t, \theta^{t+1}) = \min\left(1, \frac{q(\theta^t \mid \theta^{t+1}) \cdot g(\theta^{t+1}; I_{ij}, S_{ij})}{q(\theta^{t+1} \mid \theta^t) \cdot g(\theta^t; I_{ij}, S_{ij})}\right)$$

$g(\theta^{t+1}; I_{ij}, S_{ij})$ is the posterior probability function defined in subsection 1.3.1, and $q(x|y)$ is the transition probability from state y to state x. Note that to maintain the domain of $\psi_{ij} \in [0,1]$, out of domain parameter values were truncated by setting the likelihood function to zero.

In order to shorten the burn-in period, we initialize the Markov Chain at $\hat{\theta}$, i.e. the optimal point obtained from the previous step while computing the Bayes Factor. Moreover, such an initialization ensures that the starting state is close to where the target probability density is concentrated, especially when there are multiple replicates and the target probability density is in high-dimensional space. The initialization scheme can greatly shorten the burn-in period.

Under the above configurations, we noticed that in practice, drawing 1500 samples with a burn-in period of 100 and 10 thinning achieved good balance between estimation accuracy and running time.

Justification on different values of τ parameter

In DARTS BHT, the choice of the parameter τ specifies the two competing hypotheses: differential splicing and unchanged splicing between two biological conditions. Here we show that since the final inference is performed on the probability of $P(|\Delta\psi| > c)$ marginalizing over the hypotheses, DARTS BHT is robust to different choices of τ_k. We started with an example by comparing the inference results on a set of simulated splicing events (n=1000) when setting $\tau_1 = 0.3$, $\tau_2 = 0.03$ (default setting in our paper) with $\tau_1 = 0.4$, $\tau_2 = 0.02$ (alternative setting here). We observed the ranks of the final inference $P(|\Delta\psi| > c)$ under these two settings were highly consistent (Spearman's rho=0.99), demonstrating the robustness of DARTS BHT to difference choices of τ. Additionally, comparing the actually posterior probability of these two settings, we observed the values were highly similar for $P(|\Delta\psi| > c) \approx 1$, where is the major region of interest for inference of differential splicing. The alternative setting has a negative bias (more conservative) around $P(|\Delta\psi| > c)$

≈ 0 due to stronger regularization effect from a smaller $\tau_2 = 0.02$. This will allow users to reflect their beliefs on data quality through the choices of τ as regularization strength. For example, when data is noisy, users would preferably specify the alternative setting over our default setting. To further understand the impact of the parameter τ, we examined another four alternative settings of τ using various combinations of different τ values. Indeed, the inference results are robust in different scenarios, especially for the ranking/inference of differential alternative splicing events (upper right corner of each panel in **Supplementary Figure 2.13**). The model of DARTS BHT is designed to be robust to different specifications as well as flexible enough to account for different dataset-specific requirements.

Running time analysis

The computation of the DARTS BHT model is demanding because of the random sampling of the non-conjugate posterior. Compared to conventional inference methods that only estimate point estimates for the parameters of interest, the DARTS BHT model needs to derive the whole posterior probability distribution using an MCMC sampling. Hence, we re-wrote the MCMC sampler in Rcpp[28]. The source code was compiled during the installation of the DARTS R package and the resulting speed gain was around 10-fold. We also tuned the MCMC sampling (see subsection 1.3.2) to shorten the burn-in period.

In general, the optimized optimization procedure runs in a reasonable amount of time. For the DARTS BHT without replicate mode, an individual event takes 0.23s wall-clock time on average to finish the optimization on an Intel i7-4790 3.60GHz CPU. For the DARTS BHT with replicate, the running time scales linearly with the number of replicates

for an individual event. In our benchmarking, an event with 6 replicates takes around 1.38s and an event with 10 replicates takes 2.07 on average.

2.6.2 DARTS DNN Machine learning

Sequence feature extraction and normalization

The DARTS DNN cis sequence features are built upon a previous report[7] that curated 1,393 RNA features. Furthermore, we expanded the feature set by including 1,533 additional features on RBP binding motifs and conservation scores. We compiled cis sequence features for four different types of alternative splicing events, i.e. exon skipping, alternative 5' splice sites, alternative 3' splice sites, and retained introns. Below we briefly describe all the feature annotations of exon skipping events as an example; the full lists of all cis sequence features for the four types of alternative splicing events are publicly available in the GitHub repository.

For each exon skipping events, let C1, A, and C2 be the upstream exon, skipping exon and downstream exon respectively. I1 denotes the intron region between C1 and A, and I2 denotes the intron region between A and C2. The DARTS DNN cis features are grouped by the following generic categories:

1) Exon length and ratio of length of exons and introns.

2) Nucleosome occupancy scores are computed using NuPoP[29] for the skipping exon and flanking introns. The features are defined as predicting the nucleosome positioning in the first 100 nucleotides of each intron and in the first and last 50 nucleotides of skipping exons.

3) The definition of translatability is whether a sequence can be translated without stop codons under three different reading frames. We are evaluating translatability of C1, C1-C2, C1-A, C1-A-C2.

4) We include 111 curated RBP-binding motifs and count motifs in each of the 7 intronic and exonic regions. In addition to the counting procedure, we also download the RBP binding PSSM matrix from RBPmap[30] and calculate the PSSM scores of each RBP-binding profile.

5) We run two different tools, one from Itoh et al.[31], and maxent[32], to estimate the splicing strength between the three exon-exon junctions: C1-C2, C1-A and A-C2.

6) Conservation scores are computed as average conservation score of the first and last 100 nucleotides of intron I1 and intron I2. The conservation scores are downloaded from UCSC phastCons46way.

7) The secondary structure score is predicted by the maximum availability of intron regions using RNAfold[33].

8) Short motifs are integrated from Xiong et al[7].

9) Alu repeats annotation is downloaded from UCSC genome browser. Features are defined as counts of Alu repeats on the plus and minus strand of two intronic regions.

10) ESE (exon splicing enhancer), and ESS (exon splicing silencer) are from Burge's and Chasin's work[34,35]. ISS (intron splicing silencer) and ISE (intron splicing enhancer) are from Wainberg's work[36].

In total, the number of RNA features was 2,926.

Although certain classifiers (e.g. tree-based models) are robust to the feature scaling, it is important to scale the features for neural networks. We followed the feature scaling method described previously[7] and divided each feature by its maximum absolute value across all training sets. This rescaled the features to [-1,1] while preserving the zero values, which has specific biological indications.

ENCODE data processing

Extraction of junction counts and detection of novel events

Following the descriptions in the Method section, we had downloaded all the alignment files from the ENCODE data portal[3] and processed the bam files with rMATS. Aside from the annotated events in the reference GTF Gencode v19, rMATS detected novel splicing events where edges not annotated in the GTF splicing graph connect two annotated exons. These novel events consist of a large proportion of our training dataset and are crucial for learning the regulatory code between RBP perturbations and alternative splicing. Note that our definition of novel events are novel edges or junction reads that are not present in GTF; we do not detect novel splice sites or novel exons.

RBP expression estimation

The robust performance of DARTS DNN is dependent on the robust estimation of RBP expression levels, given that all sequence features are static. A previous report has demonstrated that 10 million reads per sample was a good depth for differential gene expression analysis[37], hence we reasoned that the gene expression estimates are fairly robust to reduction in sequencing depth, unlike the exon inclusion level estimates that depends on junction spanning read counts. In practice we re-analyzed gene expression

using Kallisto (v.0.43.0) from raw fastq reads downloaded from the ENCODE data portal. We extracted the TPM of all RBPs from the annotated list. The estimated TPM was subsequently divided by the maximum value across all datasets to rescale it range to [0,1].

Implementation of other machine learning strategies and comparison to DARTS DNN

Logistic regression and Random Forest

To benchmark the performance of our trained DARTS DNN model to other machine learning strategies, we implemented two baseline methods, Logistic regression with L2 penalty and Random Forest. Because these baseline methods were unable to scale up to big data (see 2.3.2 below), they were trained and benchmarked on individual ENCODE leave-out datasets by cross-validation. The identical events with their corresponding labels and features were fed into the baseline classifiers through 5-fold cross-validation and we recorded the performance measured by AUROC in each of the validation sets. We implemented the two methods using scikit-learn in python. For the logistic regression, we need to tune one parameter, i.e. the penalty strength, or the inverse of the penalty strength C. This parameter controls the complexity of the classifier and hence the severity of overfitting. We chose C=0.1 for our implementation of logistic regression because in practice such a penalty achieves good reasonable generalization over different datasets. Although logistic regression is easy to interpret and a good baseline method for most classification tasks, it cannot effectively detect high-order interaction terms, diminishing its predictive power for such complex tasks. Another more powerful and robust machine learning strategy we employed as a baseline method was Random Forest. Random Forest is an ensemble learning method where each base classifier is a decision tree that over-fits a

set of bootstrapped training samples with a subset of features. The Random Forest classifier has several desirable properties, including being robust to feature scaling and irrelevant features, and being capable of dividing the feature space more flexibly than more conventional partitioning based classification methods. We tuned the hyper-parameter of Random Forest, i.e. the number of trees in the forest. Typically, the more trees in a random forest, the better predictive power it renders to the ensemble classifier. We noticed that for our datasets, 500 trees achieved the best testing accuracy while increasing the number of trees further did not grant much more gain. As shown in **Figure 2.1d**, Random Forest almost always outperformed Logistic regression given the same training datasets. We can also observe a positive correlation between the performance of Random Forest and Logistic regression, indicating the internal structure of the training data plays an important role in the learning efficiency, despite the fact that the two learning algorithms are based on dramatically different underlying structures. Nevertheless, DARTS DNN showed superior performance compared to the baseline methods, even though these knock-down datasets have never been trained in DARTS DNN. Furthermore, the performance of DARTS DNN does not show strong correlations with the base learners, indicating its generalization over the single datasets to a more generic regulatory code.

Technical notes on DNN training

Below we briefly describe some technical details in training the DARTS DNN model using the ENCODE data. DARTS DNN was implemented in Keras with Theano backend. The DNN model was a 4-hidden layer fully connected neural network with drop-out (with probability 0.6, 0.5, 0.3 and 0.1, respectively) and batch normalization layers, and each

neuron had ReLU (rectifier linear unit) activation function that maps the input vector x to a non-linear output:

$$\text{ReLU}\ (x) = \max(0, w^T x + b)$$

The weight parameter w and bias b are learned through training on labelled samples and minimizing the loss function, which is the binary cross-entropy between the observed labels Y and predictions $\hat{Y}$:

$$\text{L}(Y; Y) = -\frac{1}{\text{N}} \sum_{i=1}^{N} [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)]$$

We optimized the model parameters using the RMSprop optimizer. RMSprop is a variant of the stochastic gradient descent algorithm, which accounts for the recent momentums of the gradient and adaptively adjusts the learning rate. In our experiments, RMSprop works better than other optimizers for most network architectures. Because the training dataset was huge and took too much memory ($>$ 100G) to be loaded at once, we divided the training samples into different data batches by the knock-down experiments. In each data batch, we randomly picked two different RBP knockdown experiments; due to the way the training datasets were constructed, every RBP selected must have been knocked down in both HepG2 and K562 cell lines. Hence in each data batch, we had at least 4 different datasets, sometimes more if this RBP was knocked-down by more than one shRNA in a certain cell line. The pairing of the same RBP in two different cell lines ensured that there was sufficient variance in the RBP expression features, hence facilitating the classifier to learn from the trans-acting factors.

Next we mixed the training skipping-exon events from the data batch, and held-out 20% of these events as validation set, and the remaining 80% as training set. The training set was then split into positive and negative stacks of cases, and we aimed to construct mini-batches of size 400 to feed into training the model sequentially. Because the training set was very imbalanced and the number of negative cases outweighed the number of positive cases, we balanced the composition of each mini-batch by first extracting 100 (25%) positive cases from the positive stack, then compensating 300 (75%) negative cases from the negative stack. Such biased composition of mini-batches will generate the back-propagation of errors from positive cases and reduce strong negative bias caused by the imbalanced data.

To monitor potential overfitting, we computed the validation loss and the prediction AUROC of the current model every 10 mini-batches of training. Due to the imbalanced composition of the datasets, we noticed that using AUROC as the monitoring criteria performed better than the loss function because the loss function could be stuck in a local optima where all cases were classified as negative. We only saved the parameter values of the best performing models on the validation data; by the end of the training for each data batch, we re-loaded the saved model parameter values. The goal of such a configuration was to avoid overfitting to any particular individual data batch while exploring for the global optimal point(s) in the model energy landscape.

2.7 References

1. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-15 (2010).
2. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**, E5593-601 (2014).
3. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
4. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
5. Cieslik, M. & Chinnaiyan, A.M. Cancer transcriptome profiling at the juncture of clinical translation. *Nat Rev Genet* **19**, 93-109 (2018).
6. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet* **102**, 11-26 (2018).
7. Xiong, H.Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
8. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53-9 (2010).
9. Leung, M.K., Xiong, H.Y., Lee, L.J. & Frey, B.J. Deep learning of the tissue-regulated splicing code. *Bioinformatics* **30**, i121-9 (2014).
10. Huang, Y. & Sanguinetti, G. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol* **18**, 123 (2017).
11. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829-45 (2014).
12. Van Nostrand, E.L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711-719 (2020).
13. Warzecha, C.C. *et al.* An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J* **29**, 3286-300 (2010).
14. Yang, Y. *et al.* Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition. *Mol Cell Biol* **36**, 1704-19 (2016).
15. Dittmar, K.A. *et al.* Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Mol Cell Biol* **32**, 1468-82 (2012).
16. Lu, Z.X. *et al.* Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization. *Mol Cancer Res* **13**, 305-18 (2015).
17. Li, H., Qiu, J. & Fu, X.D. RASL-seq for massively parallel and quantitative analysis of gene expression. *Curr Protoc Mol Biol* **Chapter 4**, Unit 4 13 1-9 (2012).

18. Trincado, J.L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* **19**, 40 (2018).

19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

20. Iofe, S.S., C. *In Proc. 32nd International Conference on Machine Learning (eds Bach, F. & Blei, D.)* **448–456**(PMLR/Microtome Publishing, Brookline, MA, USA, 2015).

21. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-7 (2016).

22. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**, S4 1-9 (2006).

23. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).

24. Ying, Y. *et al.* Splicing Activation by Rbfox Requires Self-Aggregation through Its Tyrosine-Rich Domain. *Cell* **170**, 312-323 e10 (2017).

25. Griebel, T. *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res* **40**, 10073-83 (2012).

26. Zarnack, K. *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453-66 (2013).

27. Zhang, Z. & Xing, Y. CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res* **45**, 9260-9271 (2017).

28. Eddelbuettel, D. Seamless R and C++ integration with Rcpp. (2013).

29. Xi, L. *et al.* Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics* **11**, 346 (2010).

30. Paz, I., Kosti, I., Ares, M., Jr., Cline, M. & Mandel-Gutfreund, Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res* **42**, W361-7 (2014).

31. Itoh, H., Washio, T. & Tomita, M. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *RNA* **10**, 1005-18 (2004).

32. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-94 (2004).

33. Bindewald, E. & Shapiro, B.A. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA* **12**, 342-52 (2006).

34. Fairbrother, W.G. *et al.* RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* **32**, W187-90 (2004).

35. Zhang, X.H. & Chasin, L.A. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18**, 1241-50 (2004).

36. Wainberg, M., Alipanahi, B. & Frey, B. Does conservation account for splicing patterns? *BMC Genomics* **17**, 787 (2016).

37. Liu, Y., Zhou, J. & White, K.P. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30**, 301-4 (2014).

3 TRACKING PRE-MRNA MATURATION ACROSS SUBCELLULAR COMPARTMENTS IDENTIFIES DEVELOPMENTAL GENE REGULATION THROUGH INTRON RETENTION AND NUCLEAR ANCHORING

3.1 Introduction

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, while nascent, intron-containing transcripts are limited to small nuclear puncta at the sites of transcription[1,2]. For other genes, unspliced introns may remain after transcript completion but are ultimately excised to allow export[3-5]. These nuclear transcripts are not necessarily found at their gene loci but some polyadenylated transcripts, including many non-coding RNAs, are tightly associated with chromatin[6]. Although proteins affecting processes such as DNA template release, RNA export, and nuclear RNA decay have been identified[5,7], 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[8,9]. 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[8,9]. 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.

The above analyses used a fractionation procedure to enrich for nucleoplasmic or chromatin-associated RNA[10-15]. 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[14]. The cytoplasmic fraction is enriched for mature

mRNA, while the nucleoplasmic fraction contains recently matured transcripts released from the chromatin that have not yet reached the cytoplasm[8,9], as well as some mature mRNAs associated with ER and mitochondria[15]. The chromatin pellet is enriched for nascent RNA bound by elongating RNA Pol II, but also contains substantial polyadenylated RNA, including the Xist non-coding RNA tightly bound to chromatin[16], and the Malat1 non-coding RNA, which is enriched in nuclear speckles that are adjacent to chromatin but only partially in contact with it[17,18].

The consequences of intron retention are diverse and complex to dissect. Splice sites and binding of spliceosomal components can prevent nuclear RNA export[5,19,20]. 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[21,22]. Other introns slow to be excised relative to transcription are ultimately removed and their transcripts exported as fully spliced mRNAs[8,9,23-25]. 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 were named “detained introns” to distinguish them from classical “retained introns” found in cytoplasmic mRNA[26,27]. 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[28]. 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 mis-called as intron retention by sequence analysis tools[29,30].

Many RNA-seq studies have identified conditions leading to higher levels of unspliced introns across the transcriptome[21,31-37]. 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.

3.2 RESULTS

3.2.1 Both coding and non-coding RNAs exhibit defined partitioning between cellular compartments.

To broadly categorize RNAs enriched in different cellular locations, and gain insight into how this compartmentalization might be regulated across cell types, we generated deep RNA-seq data from mouse embryonic stem cells (mESC), a neuronal progenitor cell line derived from embryonic mouse brain (mNPC), and explanted mouse cortical neurons cultured in vitro for 5 days (E15DIV5; mCtx) (**Figure 3.1A**). RNA was isolated from three fractions of each cell: cytoplasm, soluble nucleoplasm, and chromatin pellet as previously described[8,12,14,15]. 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 Figure 3.7A**).

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 is 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. Gene expression markers for each of the three cell types confirmed the expected patterns of ESC, NPC or immature neurons (E15DIV5; **Supplementary Figure 3.7B**). Clustering of gene expression values across all the datasets showed the expected segregation by cell type, fraction, and replicate, for both the poly(A)+ and total RNA libraries (**Supplementary Figure 3.7C**). The resulting 54 datasets constitute an extensive resource for examining multiple aspects of RNA maturation and its modulation during development [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[38]. These 27 datasets are also available from GEO [GSE159971].

Examining read distributions in the different RNA pools and fractions, we found that the housekeeping gene Gapdh (**Figure 3.1B**) 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 (**Figure 3.1B** 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 non-coding RNA Xist, which condenses on the inactive X Chromosome in female cells (**Figure 3.1C**). The mNPCs were isolated from female mice, and Xist is seen to partition almost completely to chromatin in these cells. The poly(A)+ and the total RNA samples yielded very similar patterns of Xist reads indicating that this RNA is largely spliced and polyadenylated[39]. Other non-coding RNAs yielded more complex patterns of subcellular partitioning that changed with cell type. The paraspeckle lncRNA Neat1 is more highly expressed in mESC than mNPC or neurons (**Supplementary Figure 3.8A**). The short polyadenylated form (Neat1_1) predominates in ESC 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 mNPC and neurons, consistent with observations that Neat1 cleavage and polyadenylation may be modulated[40]. Overall, we find that gene transcripts can exhibit 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, RPM (Reads Per Million) values or other read number normalizations of individual genes cannot be directly compared between different subcellular fractions. Using qRT-PCR in mESC 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. 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. 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[41] 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 whose TPM (Transcripts Per Million) value in chromatin was over the median and which had read counts greater than 0 in the cytoplasm (13,036 genes), this chromatin partition index was distributed over a 100 fold range centered on 1 (Log2=0). Thus, a typical gene showed equal normalized read counts in chromatin and cytoplasm (**Figure 3.1D**). 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 (**Figure 3.1D**), Gapdh RNAs partition much more strongly to the cytoplasm than is typical. In the middle of the distribution, Rbfox2 RNAs exhibit slightly fewer reads on chromatin than in the cytoplasm, whereas Cdk8 exhibits 2 to 3 fold more chromatin reads (**Figure 3.1D**). 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. Comparing qRT-PCR quantification for select genes to their chromatin partition indices, we found that RNAs from genes exhibiting a partition index above 3.6 were actually

more abundant in chromatin than the cytoplasm. This included about 3 % of protein coding genes. At the right edge of the curve, the 400 most chromatin enriched transcripts included the expected non-coding RNAs, such as pri-miRNAs, snoRNAs, and lncRNAs, but also many protein-coding genes, including Clcn2, Ankrd16, and Gpc2 (**Supplementary Figure 3.8C, 3.8D, 3.12B**), 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 (**Supplementary Figure 3.8C, 3.8D, 3.12B**).

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 non-coding RNA is well expressed in mESC and neurons but not in mNPC (**Supplementary Figure 3.8B**). 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 mESC. This intron is absent from the RNA in mESC where it is apparently efficiently spliced. By 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 (**Supplementary Figure 3.8B**). This is consistent with observations that perturbations causing a host transcript to be released from chromatin reduce DROSHA cleavage and miRNA expression[13,42]. The mESC small RNA data was previously used to examine expression of primary MiR-124a-1 in mESC whose processing is blocked by PTBP1 in the chromatin fraction[38]. 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 exhibits more splicing in neurons than in mESC. Thus, an additional product from the gene, possibly MiR-1906-1, may also be

differentially regulated between mESC 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.

3.2.2 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 prior to transcript completion, while some introns with slow kinetics will be removed later. Various studies estimate that 45 to 84 % of introns are cotranscriptionally excised in mammals[3,8,11,43-45]. Several approaches compare read numbers for spliced (exon-exon) and unspliced (exon-intron or intron-exon) junctions in nascent RNA to those in total RNA to measure cotranscriptional excision[10,44,45]. 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[43,46,47]. While sawtooth read densities can be observed on certain introns in the total chromatin RNA pools (**Supplementary Figure 3.9**), these patterns were infrequent and lost on introns shorter than 50kb, many of which are expected to be cotranscriptionally excised[43].

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 (**Figure 3.2A**) reads are observed across all the introns in the total RNA from 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 Sorbs1 exhibits substantial read numbers in poly(A)+ RNA on chromatin that are reduced in RNA from the nucleoplasm and absent from the cytoplasm (**Figure 3.2A**). This intron is presumably excised after cleavage/polyadenylation. While most introns are absent from the polyadenylated RNA and likely spliced cotranscriptionally, there are many transcripts with one or more introns that are highly retained in the polyadenylated chromatin associated RNA (**Figure 3.2A, 3.2B**). 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 values (FI; **Supplementary Figure 3.10A**) by counting reads across exonintron, intron-exon, and exon-exon junctions. Assessing intron retention (IR) by FI value can be confounded by alternative splicing, polyadenylation or transcription initiation events occurring within the intron being measured (**Supplementary Figure 3.10B**) [29,30]. To avoid errors in IR measurements arising from other processes, we defined a set of introns exhibiting a unique Ensembl v91 annotation without alternative processing events

(**Supplementary Figure 3.10B**). This set of 149,333 "unique" introns (U introns) across 28,733 genes was used for subsequent analysis. 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 [48]. 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 where 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 due 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 mESC, these criteria returned 49,629 U introns within 7,672 genes for analysis.

Of the 49,629 U introns being measured, 34,939 introns (within 6,952 genes) exhibited low FI values in the poly(A)+ RNA (FI < 0.1) and are presumably spliced before transcript completion. Conversely, 14,753 introns within 5,550 genes exhibited a FI value greater than or equal to 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 (**Figure 3.2C; Supplementary Figure 3.10C-10F**). On the other hand, the majority of genes (5,550 out of 7,672) have at least one posttranscriptionally spliced intron. Restricting the analysis 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 (**Supplementary Figure 3.10F**). 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 3.2B**. Although in the minority, posttranscriptionally spliced introns are found across a wide range of genes, and often exhibit 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. Using the same parameters to define co- versus posttranscriptional splicing, we found a reversal in the percentages. Of these introns flanking alternative exons, approximately 67 % exhibit high read numbers (FI > 0.1) in the poly(A)+ RNA and thus appear to be excised posttranscriptionally (**Figure 3.2C; Supplementary Figure 3.10E**). This was seen for introns both upstream and downstream of the cassette exon. These data indicate that the majority of regulated splicing events occur with slower kinetics than the excision of typical constitutive introns.

3.2.3 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 (**Figure 3.3A**) [49]. Consistently, in all three cell types the clustering algorithm defined four groups of introns. The largest cluster Group A, containing 49,981 introns in mESC, was almost entirely spliced in all three fractions. Introns in Group B (7,529), exhibited measurable retention in the poly(A)+ RNA from chromatin, but showed nearly complete splicing in the nucleoplasm and cytoplasm (**Figure 3.3A**). Group C introns (1,351), including introns in Zfp598 and Neil3 (**Figure 3.3B**), showed higher FI values in the chromatin and nucleoplasm than Group B, but were almost completely excised from the cytoplasmic RNA. The smallest cluster of only 247 introns in mESC, 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 (**Figure 3.3A**).

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. Group D

introns were also found to be depleted of in-frame premature termination codons (PTC) compared to Groups A, B and C (**Figure 3.3C**), presumably due 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 where all introns could be assigned a group (**Supplementary Figure 3.10G**), RNAs containing at least one Group C intron have a higher average chromatin partition index than transcripts with no Group C intron (**Supplementary Figure 3.10H**). Previous work defined nuclear transcripts in mESC containing what are called detained introns (DI), whose splicing is modulated in cancer and growth control pathways [26,27]. Of 3,150 detained introns, 1,021 were on our U intron list. Of these, 1,000 introns passed the filters for FI measurement and are seen to fall predominantly into Groups B and C, in agreement with the earlier studies (**Figure 3.3D**). However, the 1,021 detained introns were only a subset of the nearly 9,000 retained introns identified in Groups B and C. Similar to the detained introns affecting growth control, as well as inflammatory and neuronal gene introns also identified previously [8,9,24,28,50], these new retained introns could affect cellular function by altering the movement of material through the gene expression pathway.

3.2.4 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 1,387 sequence features from the first and last 300 nucleotides of each intron and from the two flanking exons. For introns less than 300 nucleotides, 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. This feature information was used to train a three-layer deep neural network tasked with predicting whether an intron belonged in Group A, B, C, or D (**Figure 3.4A**).

The performance of the model was assessed using Receiver Operating Characteristic (ROC) curves plotting the false and true positive rates (**Figure 3.4B**). The model was highly predictive in distinguishing Group D introns from A, yielding an Area Under the Curve 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), and Group B and C introns 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 intron retention or its absence and used a t-distributed stochastic neighbor embedding algorithm (t-SNE) to project them onto two dimensions (**Figure 3.4C**). As previously observed, high splice site strength scores were predictive of Groups A and C over D, and also Group A over C [31,51]. 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. 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 (**Figure 3.3C**) and adding a coding segment to the mRNA. Overall, the data indicate that intron retention 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 intron retention such as constitutive transport elements or decoy exons [34,52]. 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. 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.

3.2.5 Intron retention and chromatin association are regulated with neuronal development.

Since the X-means analysis yielded four intron clusters in each cell type, these cluster definitions allow bioinformatic analysis of IR regulation between cell types. While many introns maintain their classification between cell types (**Figure 3.5A**, left), some introns switched their group (**Figure 3.5A**, right). One example is Med22 (**Figure 3.5B**), which

contains a highly retained intron 3 (I3) in all three fractions of mESC (Group D). This intron became more spliced in mNPC 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 intron 3 creates MED22 proteins with different C-terminal peptides that likely alters mediator function in the two cell types. The group switching introns are presumably part of the extensive alternative splicing programs modulated during neuronal development. Examining their Gene Ontology (GO) functions, we found that the 231 genes containing introns highly spliced in mESC 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 mESC 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 Figure 3.11**).

The changes in splicing between mESC, mNPC 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 [53,54]. In ESC and other cells, PTBP1 maintains alternative splicing patterns characteristic of non-neuronal cells, and PTBP1 downregulation is a key step in neuronal differentiation. While the cultured NPC's 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 ESC [55], and PTBP1 regulated retained introns, including the Med22 intron, have been described in a neuronal cell line [56]. We next examined whether additional PTBP1 targets could be identified in the chromatin compartment of mESC.

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 intron 3, which shifted from Group D to Group C with Ptbp1 depletion (**Figure 3.5B**, right). 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 (**Figure 3.5C**). As shown previously with cassette exons, these PTBP1 dependent introns in ESC also change with neuronal differentiation as PTBP1 levels drop (**Figure 3.5C**). These include introns identified previously [56] 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 previously recognized. The transcripts containing these introns may remain in the nucleus, similar to detained introns, 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 mESC [57]. A majority of Group A, B and C introns are predicted to induce NMD, if their parent transcripts were exported to the cytoplasm (**Figure 3.3C**). However, we find that of 871 genes containing PTBP1 dependent retained introns in the chromatin fraction,

only 87 exhibited greater than 10 % transcript upregulation after Upf1 depletion. Thus, the majority of the PTBP1 dependent retained intron transcripts likely stay in the nucleus and will be eliminated by nuclear RNA decay pathways.

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. For the genes in the L, M, and R regions in **Figure 3.1D**, 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.

3.2.6 Posttranscriptional repression of Gabbr1 expression.

For the most part, transcripts enriched in the chromatin fraction of mESC's (**Figure 3.1D**) were only mildly or unaffected by Upf1 depletion. 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 mNPC, but nearly absent in mESC (**Figure 3.6A**). By immunoblot, GABBR1 protein is only observed in neurons (**Figure 3.6C**). In the chromatin fraction of mESC the Gabbr1 precursor RNA is present at high levels that nearly match those seen in mNPC and neurons (**Figure 3.6A**). This Gabbr1 RNA is polyadenylated and most introns are excised, but introns 4 and 5 that exhibit a complex pattern of alternative processing in neurons, are

largely unprocessed in mESC (**Figure 3.6A**). 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 (**Figure 3.6A**). 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 mESC to 1.09 in neurons (**Supplementary Figure 3.12B**).

PTBP1 was previously found to regulate Gabbr1 exon 15 in a neuronal cell line [58]. To assess introns 4 and 5, we examined iCLIP maps of PTBP1 binding in mESC [55], 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 (**Figure 3.6A, 3.6B**). 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 (**Figure 3.6A, 3.6B**). 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 (**Supplementary Figure 3.12C**). GABBR1 protein was also not observed in mESC after Ptbp1 knockdown (**Supplementary Figure 3.12D**). 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 mESC. Gabbr1 is highly transcribed in mESC, 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.

3.3 DISCUSSION

3.3.1 A resource for the analysis of RNA-level gene regulation.

We developed extensive datasets to examine RNA maturation events across cellular location and developmental state. Applying these data to analyze intron retention, we compare total and polyadenylated RNA across subcellular fractions and cell types to define classes of introns exhibiting 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 [59,60]. The isolation of chromatin associated RNA has frequently been used to enrich for nascent pre-mRNAs and other short lived species [12,61,62]. We find that many introns are only observed in the total RNA of this fraction, while others are also present in the polyadenylated RNA. 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 - exhibiting significant intron retention 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 are among the 30 % that appear to be cotranscriptionally excised [10,63,64].

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 [65-67]. These data could also allow more sensitive detection of recursive or back-splicing, and inform studies of regulated RNA export. We have also examined regulated miRNA processing using parallel data from short RNA libraries (GSE159971) [38].

3.3.2 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 prior to 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 exhibiting different levels of retention on chromatin and to some extent the nucleoplasm. A deep neural network 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 (**Figure 3.4B**). These data indicate that Groups D and C are functionally distinct and the features which 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[28,31,34,51,68,69]. 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[50,56,70,71], but there may be many such factors each regulating a subset of introns in a group. The extension of our approach to larger datasets 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. Detained introns were defined as partially spliced introns in transcripts affecting

growth control, whose excision can be modulated by cellular stimuli[25-27]. These detained introns are a subset of the Group B and particularly Group C introns we defined in mESC. Another group of retained introns were shown to be regulated by PTB proteins in a neuronal cell line[56]. Our analytical strategy identified many new PTBP1 dependent introns that remain as chromatin associated transcripts in mESC. In the total cellular polyadenylated RNA of mature primary neuronal cultures[28], 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, about 40 % were in Group D, consistent with the stable introns remaining in cytoplasmic mRNA after transcriptional shutoff. Similar to detained introns, Mauger et al. found that 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.

3.3.3 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 PSD-95 proteins are expressed in ESC and other non-neuronal cells, but through the action of PTBP1 are spliced as isoforms that are subject to NMD[55,58,72-74]. A similar mechanism affects Gabbr1 through regulation of exon 15 by PTBP1[58], but the change in RNA with loss of NMD is small[57]. Most protein-coding transcripts exhibiting

chromatin enrichment were not seen to be upregulated by Upf1 depletion, while 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 exhibit 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 mESC but its splicing is incomplete and its transcript remains in the chromatin compartment.

Gabbr1 is expressed as multiple isoforms[75]. 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 [76]. There is also a short transcript derived from an alternative polyadenylation site in intron 4. A micro exon 5 between these two introns adds a linker into the 1a isoform [76]. This complex intron 4-5 region is largely unprocessed in mESC cells and becomes processed in neurons with the production of cytoplasmic mRNA including exon 5. The depletion of Ptbp1 from mESC leads to multiple changes in Gabbr1 splicing including activation of micro exon 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[77]. Extensive PTBP1 binding to the long non-coding

RNA Xist is required for Xist condensation onto the X Chromosome during X inactivation[16]. PTBP1 also drives the condensation of the long non-coding RNA PNCTR in the perinucleolar compartment, and a similar mechanism may be involved in its interaction with LINE RNAs[78,79]. 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 due to 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.

3.4 Methods

3.4.1 Subcellular fractionation, RNA isolation, and library construction.

Total RNA was isolated from mESCs, mNPCs, and cortical neurons (mCtx) that were fractionated into cytoplasmic, soluble nuclear, and chromatin pellet compartments as described previously[12,14,15,38]. After checking RNA quantity and integrity, RNAs longer than 200 nt (long RNA) and shorter than 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 Supplemental Materials.

3.4.2 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 0 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 log2 of this ratio (**Figure 3.1D**). 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.

3.4.3 Measurement of intron retention

We developed SIRI (Systematic Investigation of Retained Introns), 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[80]. The number of reads mapping to each exon-exon (EE), exon-intron (EI), and intron-exon (IE) junctions were counted to determine the FI (Fraction of intron Inclusion) value of each intron. We selected only introns with a unique intron annotation (U introns) that are not involved in other alternative processing events (**Supplementary Figure 3.10B**). Introns subjected to FI measurement were also required have an intron length greater than and equal to 60 and have a sum of EE + EI + IE reads be

greater than and equal to 20. From this set, IR events with EE reads no fewer than 2 in at least one cell compartment in one cell type were then kept for downstream analysis.

3.4.4 X-means clustering of IR events

X-means clustering was performed using the Pyclustering tool[81] applied to the FI values determined in all three compartments of each cell type (**Figure 3.3A**), with the maximum number of clusters set at 6. The distance matrix for X-means clustering is based on the Dynamic Time Warping (DTW) algorithm[82] for the purpose of investigating directional changes of FI values from chromatin to nucleoplasm to cytoplasm. The Circos plot[83] 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)[84] (**Figure 3.5A**).

3.4.5 Predicting intron retention patterns by deep learning

To apply deep learning to IR group prediction, we constructed a compendium of 1,387 intron features of five types: sequence motifs, transcript features, RNA secondary structure, nucleosome positioning, and conservation. Sequence motif features included splice site consensus sequences, position-specific matrices of RNA-binding proteins, 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 to 70 nt, 70 to 140 nt, 140 to 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)[85] 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)[86] 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 dataset 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 Deep Neural Network (DNN)[87] with these 1,387 features to predict whether introns belong to group A, B, C, and D for each cell type (**Figure 3.3A**). The training was done with fivefold cross-validation with area under the ROC curves on data held-out during training reported for performance evaluation[88]. 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.

3.4.6 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 Figs. 5, 6. Links to the data displayed on the UCSC Genome Browser are here: (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 in https://github.com/Xinglab/intronretention-paper, as well as in Supplemental Code files. The data resources used to reproduce the analysis are available in https://doi.org/10.5281/zenodo.4540589.

ACKNOWLEDGMENTS

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

3.5 Figures

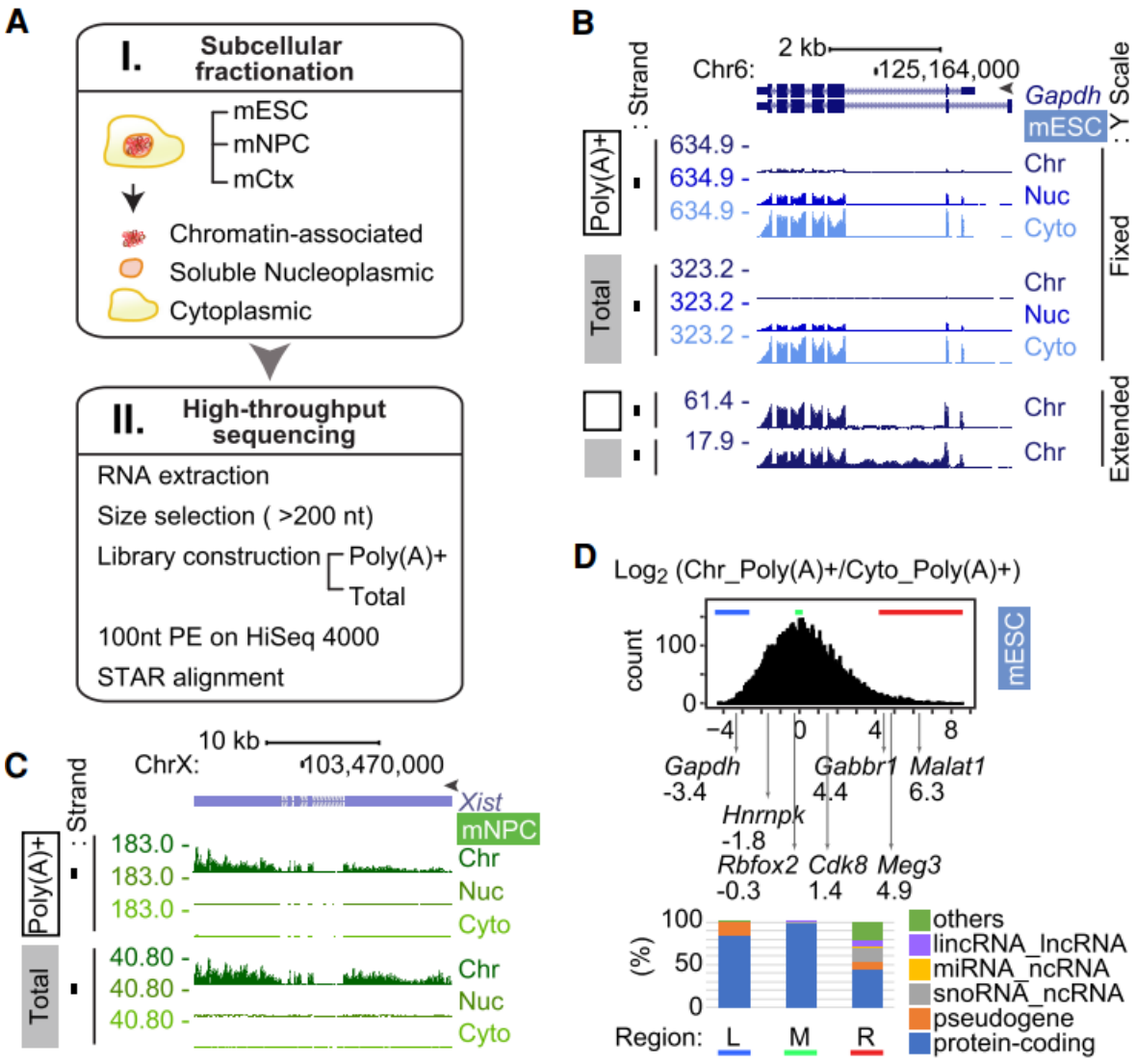


Figure 3.1 RNA partitioning between subcellular compartments.

(A) Workflow used in this study. **(B)** Genome browser tracks of the Gapdh locus in mESC. GENCODE annotated isoforms (M11) are diagrammed at the top. Poly(A)+ RNA (open box), total RNA (grey 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 log2 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.

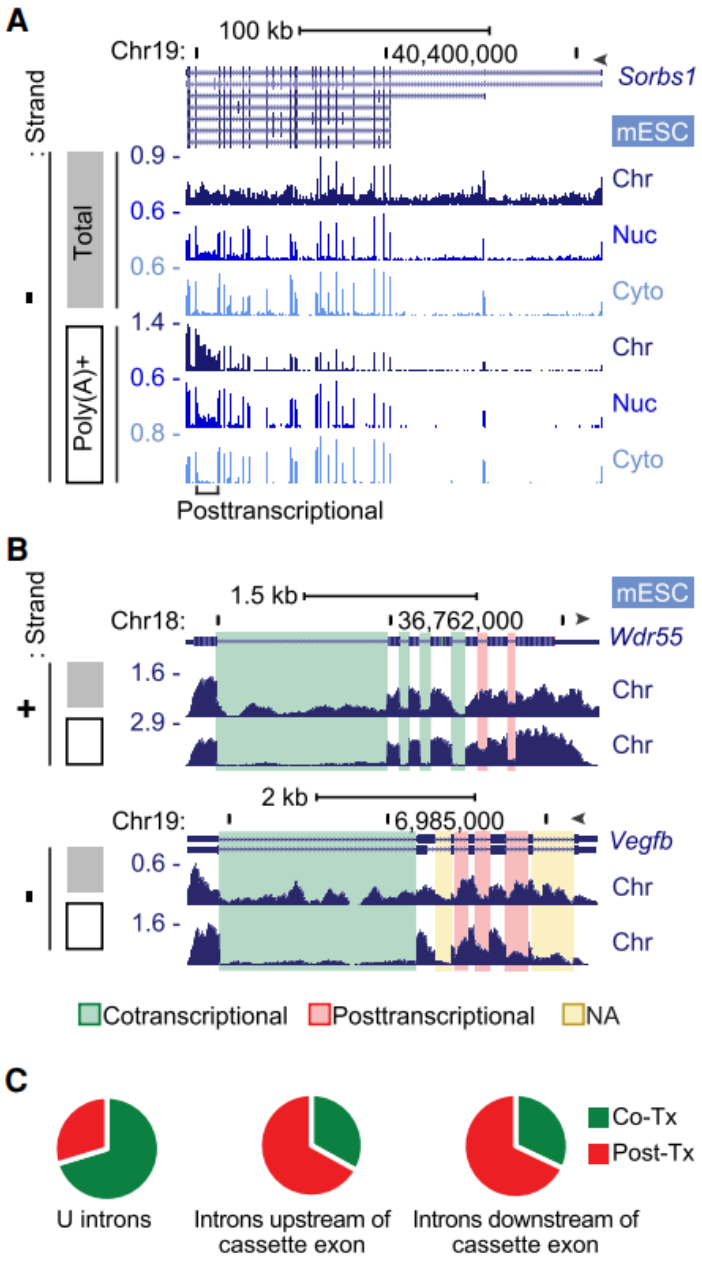


Figure 3.2 Cotranscriptional and posttranscriptional intron excision.

(A) Genome browser tracks of the Sorbs1 locus in mESC. Total chromatin RNA (grey 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 mESC. Total (grey box) and poly(A)+ (open box) are shown, with

cotranscriptionally and posttranscriptionally spliced introns are highlighted in green and red respectively. Yellow highlighted introns were not analyzable due to multiple processing patterns. **(C)** Proportions of co- and posttranscriptional splicing for 49,692 U introns in mESC, using criteria described in **Supplementary Figure 10C-10E**. Introns upstream (2,779) and downstream (2,744) from simple cassette exons were similarly analyzed.

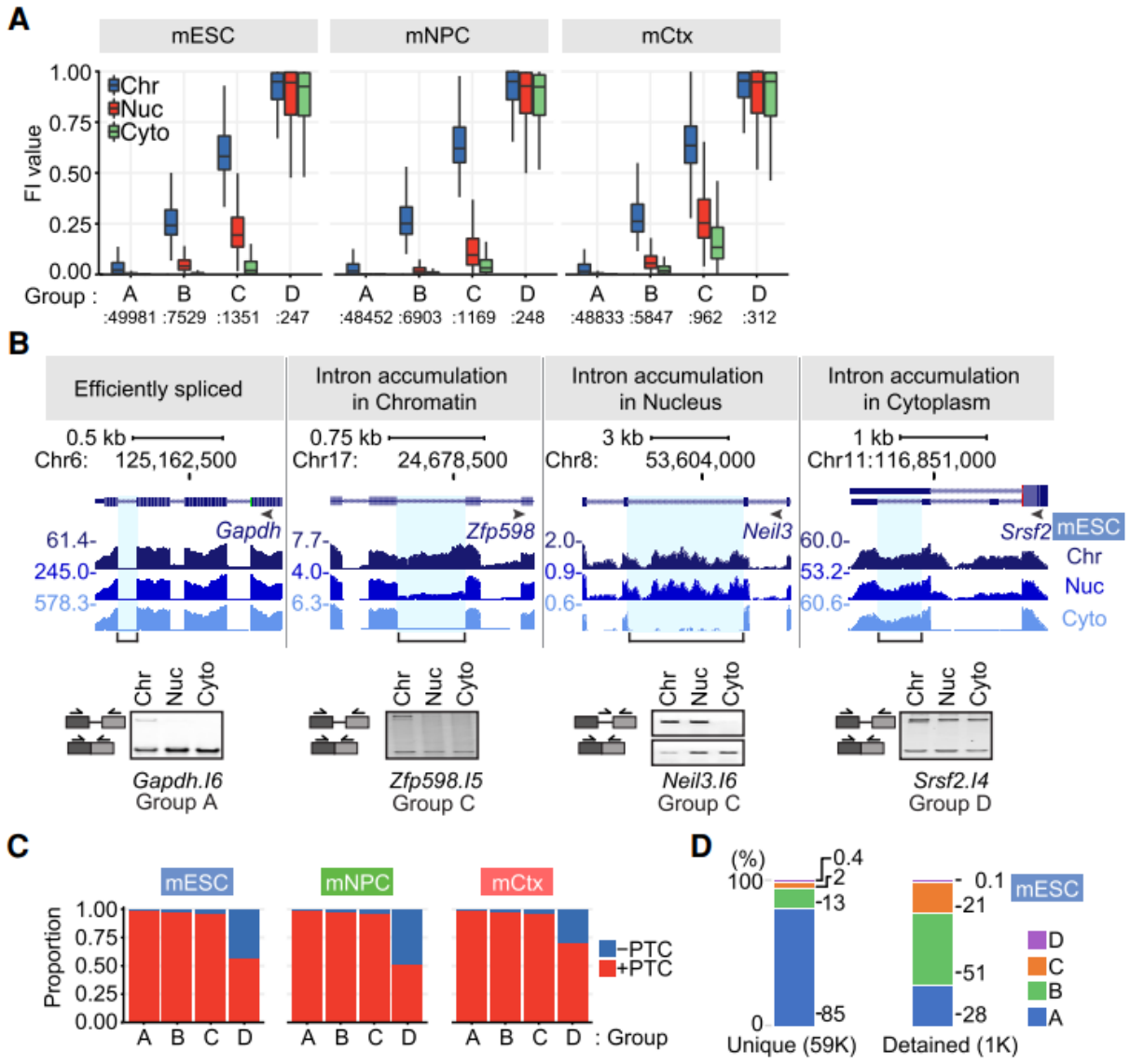


Figure 3.3 Intron Groups defined by their retention level and fractionation behavior.

(A) X-means clustering was applied to intron FI values and fraction enrichment in mESC, mNPC, and mCtx. 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 mESC. Validated introns are indicated by a blue highlight and a bracket below. Gel images are one of 3 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. Percent of introns in each group for U introns from mESC and for detained introns within the U intron set (Boutz et al. 2015).

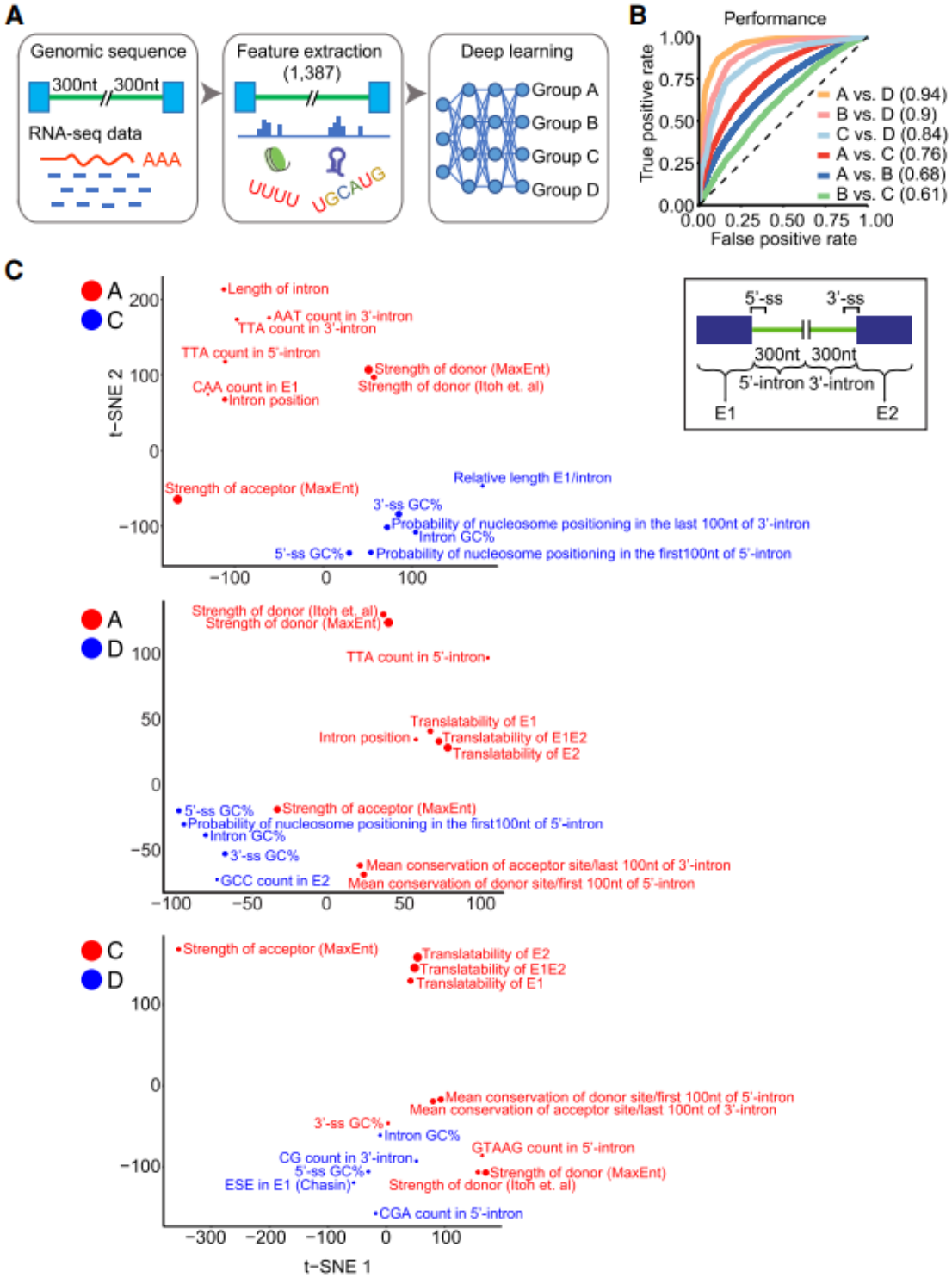


Figure 3.4 Deep Learning Analysis of Intron Groups.

(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.

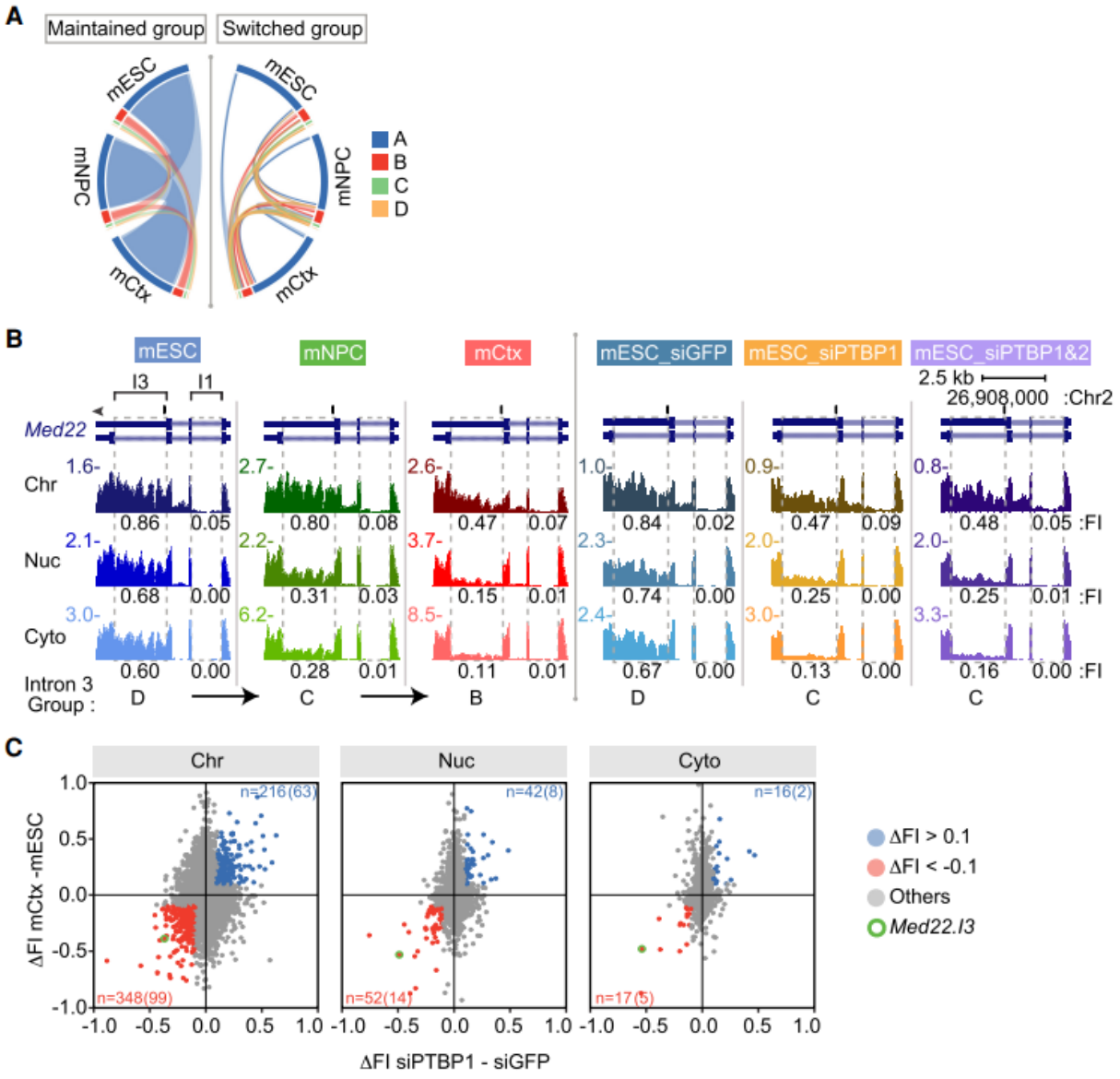


Figure 3.5 Regulation of intron retention and chromatin association during neuronal development.

(A) Circos plot (Krzywinski et al. 2009; Gu et al. 2014) of intron group changes between cell types (mESC, mNPC, and mCtx). 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 mESC

(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 mESC and neurons (mCtx) plotted for each fraction against FI change after Ptbp1 knockdown in mESC. Introns with delta FI less than -0.1 in both conditions are in red, and with delta FI greater than 0.1 in blue. Number of introns showing these changes with the number carrying PTBP1 iCLIP tags in parentheses, are above and below (Linares et al. 2015). Intron 3 of Med22 is circled in green.

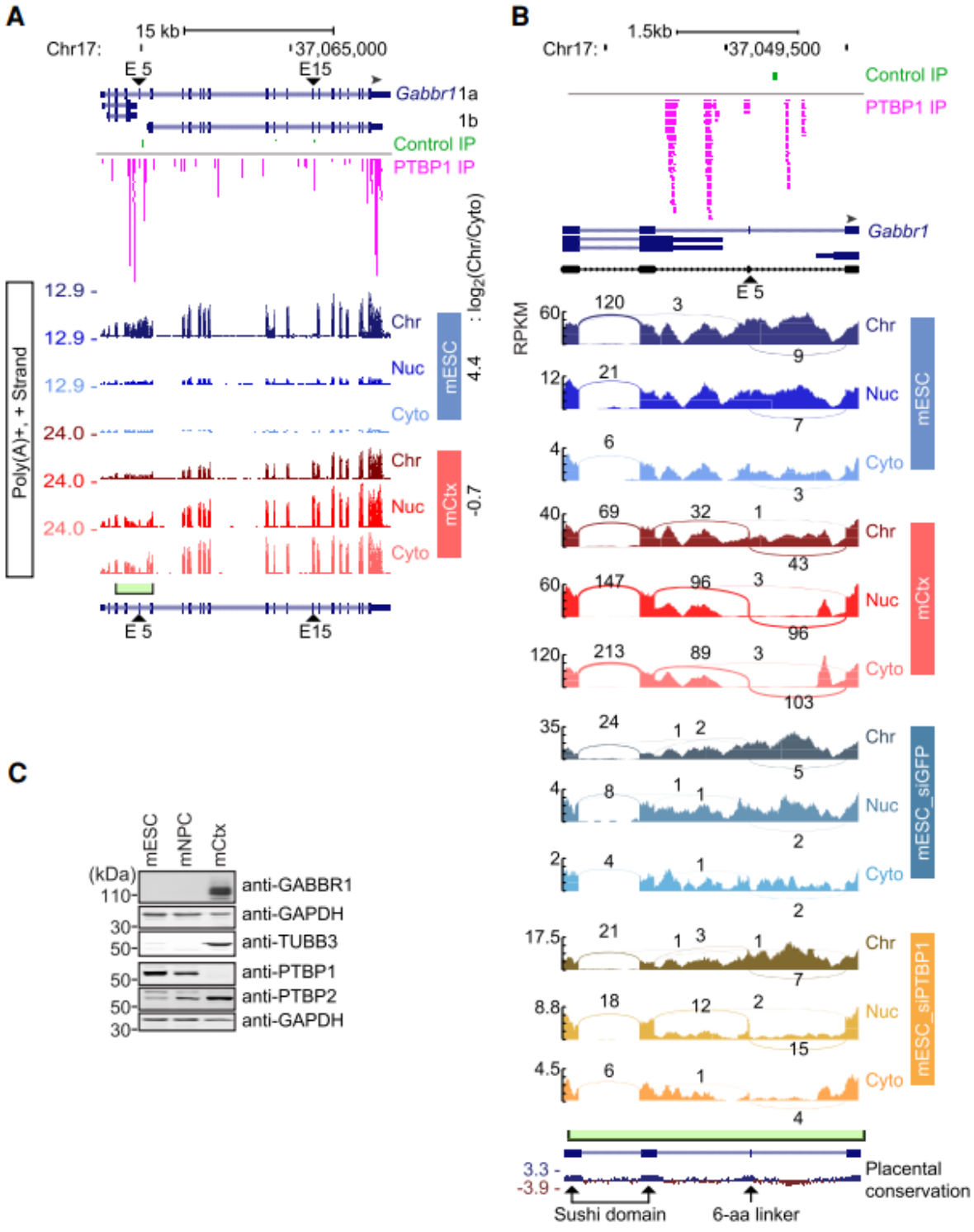
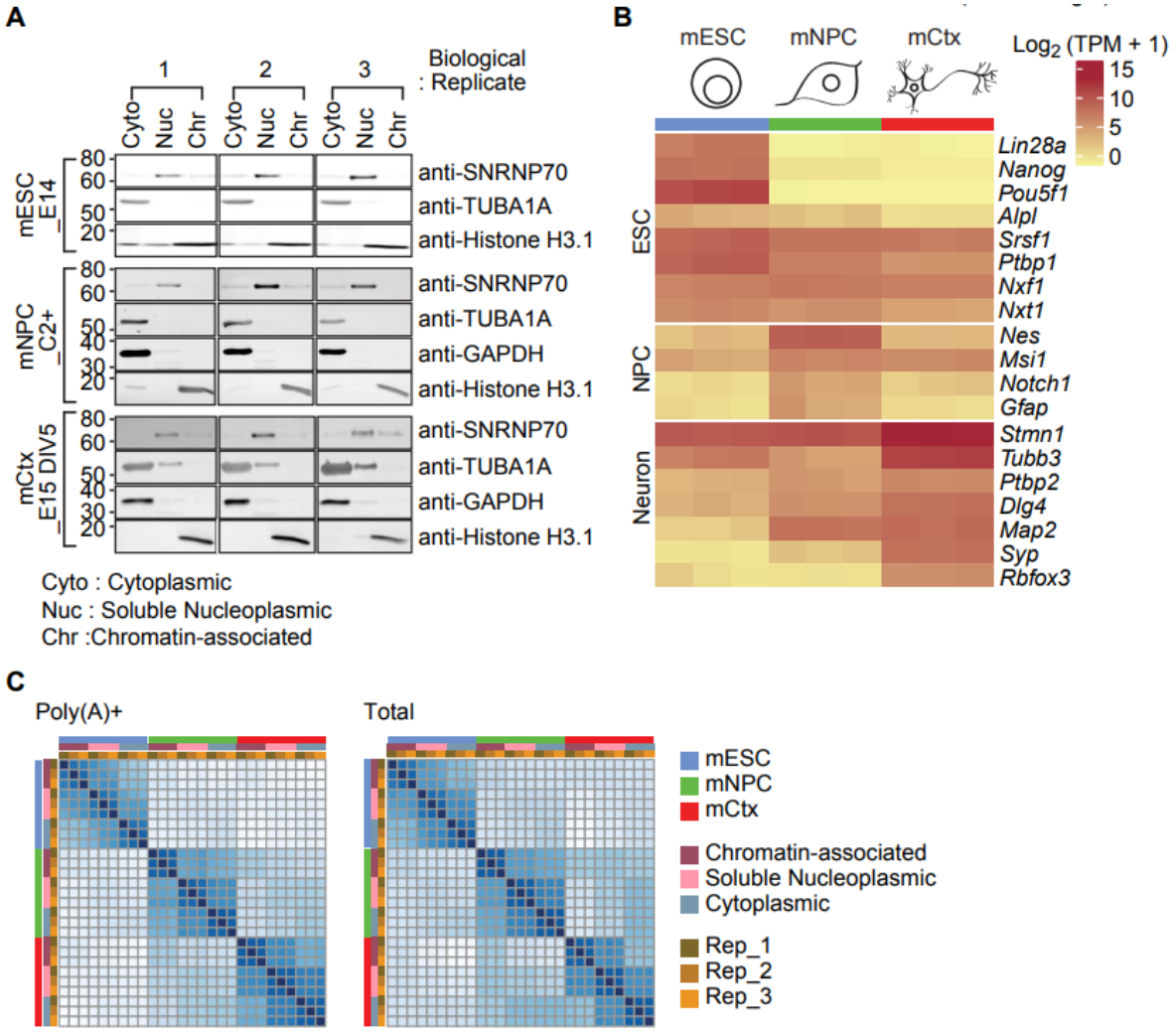


Figure 3.6 Chromatin enrichment and PTBP1 regulation of Gabbr1 transcripts.

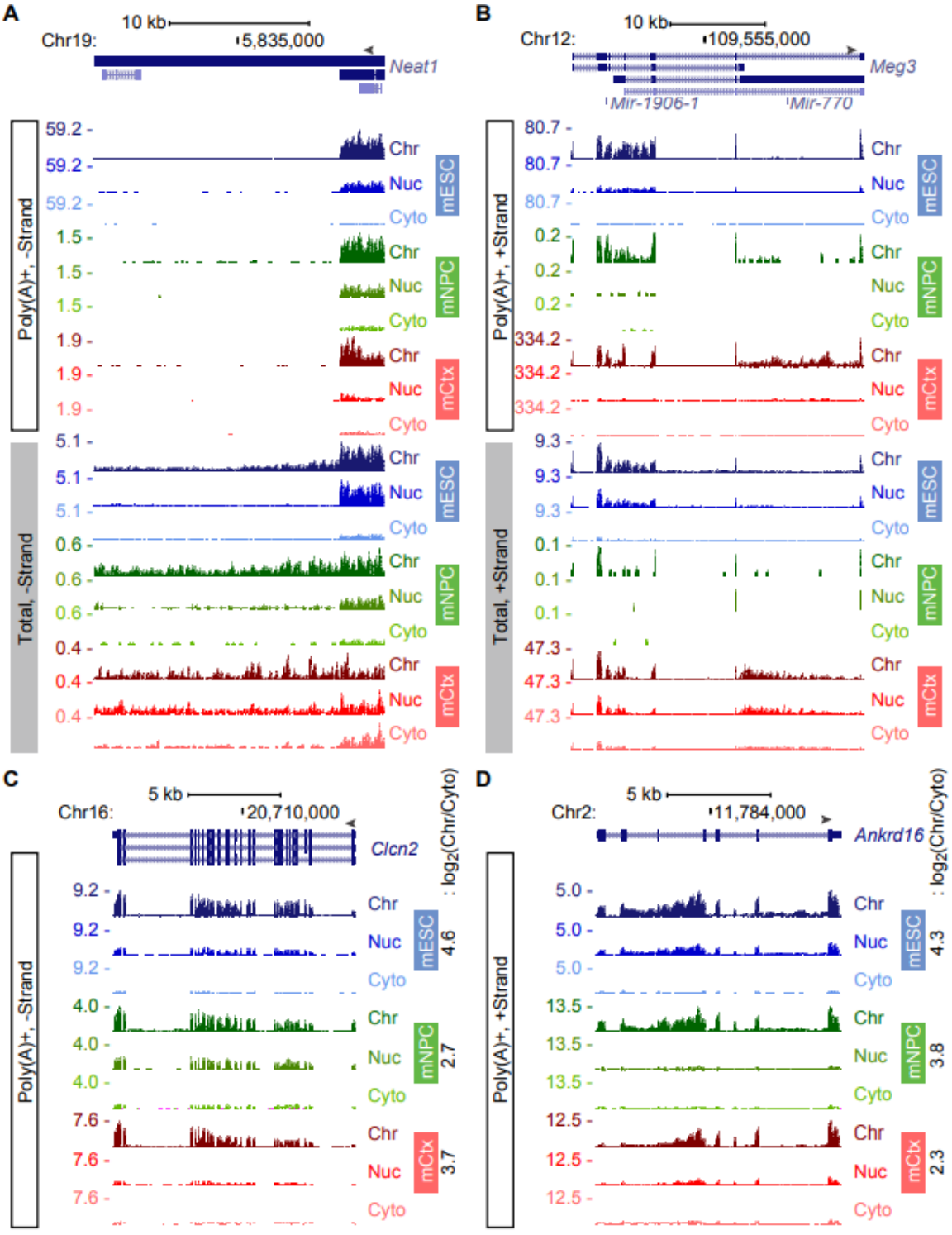
(A) Genome browser tracks of Gabbr1 in mESC and mCtx. PTBP1 iCLIP tags in mESC are plotted above in pink. Y axis indicates the maximum RPM in each cell type. The green box and bracket mark the intron 4-5 region expanded in panel B. PTBP1 responsive exons 5 and 15 are marked with arrowheads. **(B)** Sashimi plots of the Gabbr1 intron 4-5 region in mESC, mCtx, and after Ptbp knockdown in mESC. 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 mESC and cortical neurons.



Supplementary Figure 3.7 Validation of subcellular fractionation, cell type gene expression, and library consistency.

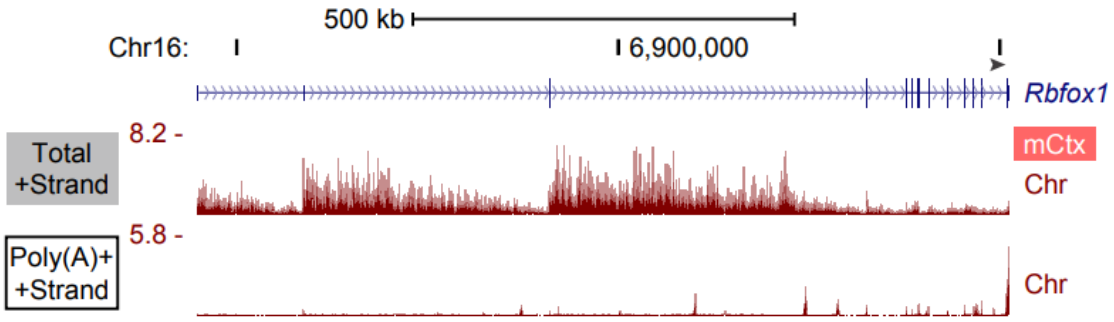
(A) Confirmation of subcellular fractionation. Immunoblot analysis of diagnostic proteins in subcellular fractions. SNRNP70 for soluble nucleoplasm (Nuc), TUBA1A and GAPDH for cytoplasm (Cyto), and Histone H3.1 for chromatin pellet (Chr). Gel images include 3 biological replicates of mouse embryonic 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.


A
10 kb
Chr19:
5,835,000
Neat1
Poly(A)+, -Strand
Total, -Strand
Chr
Nuc
Cyto
mESC
mNPC
mCtx
B
10 kb
Chr12:
109,555,000
Meg3
Mir-1906-1
Mir-770
Poly(A)+, +Strand
Total, +Strand
C
5 kb
Chr16:
20,710,000
Clcn2
: log2(Chr/Cyto)
4.6
2.7
3.7
D
5 kb
Chr2:
11,784,000
Ankrd16
4.3
3.8
2.3

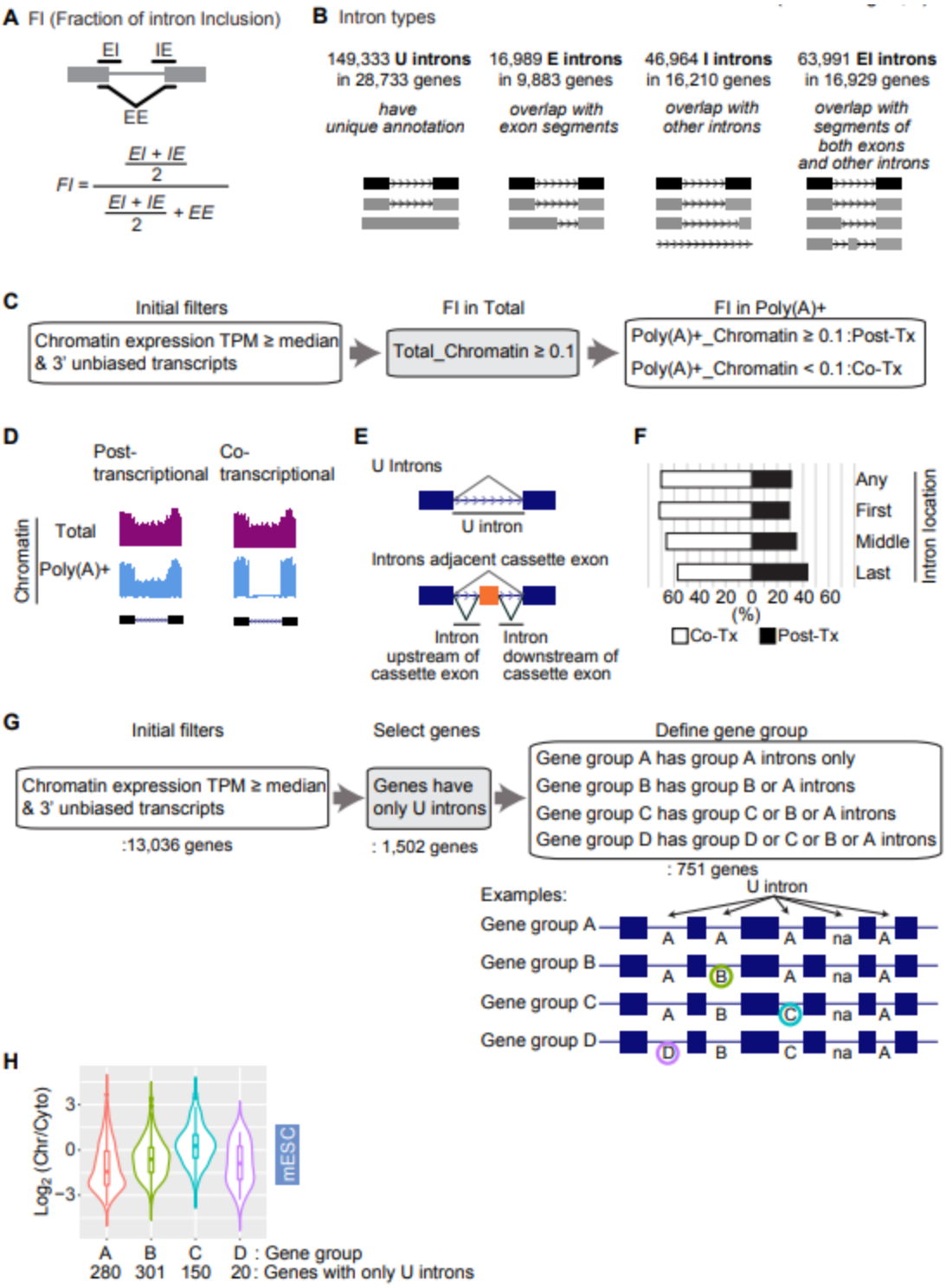
Supplementary Figure 3.8 Example genome browser tracks of non-coding and coding RNAs.

(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.



Supplementary Figure 3.9 Very long introns exhibit declining reads 5' to 3' to create a sawtooth pattern.

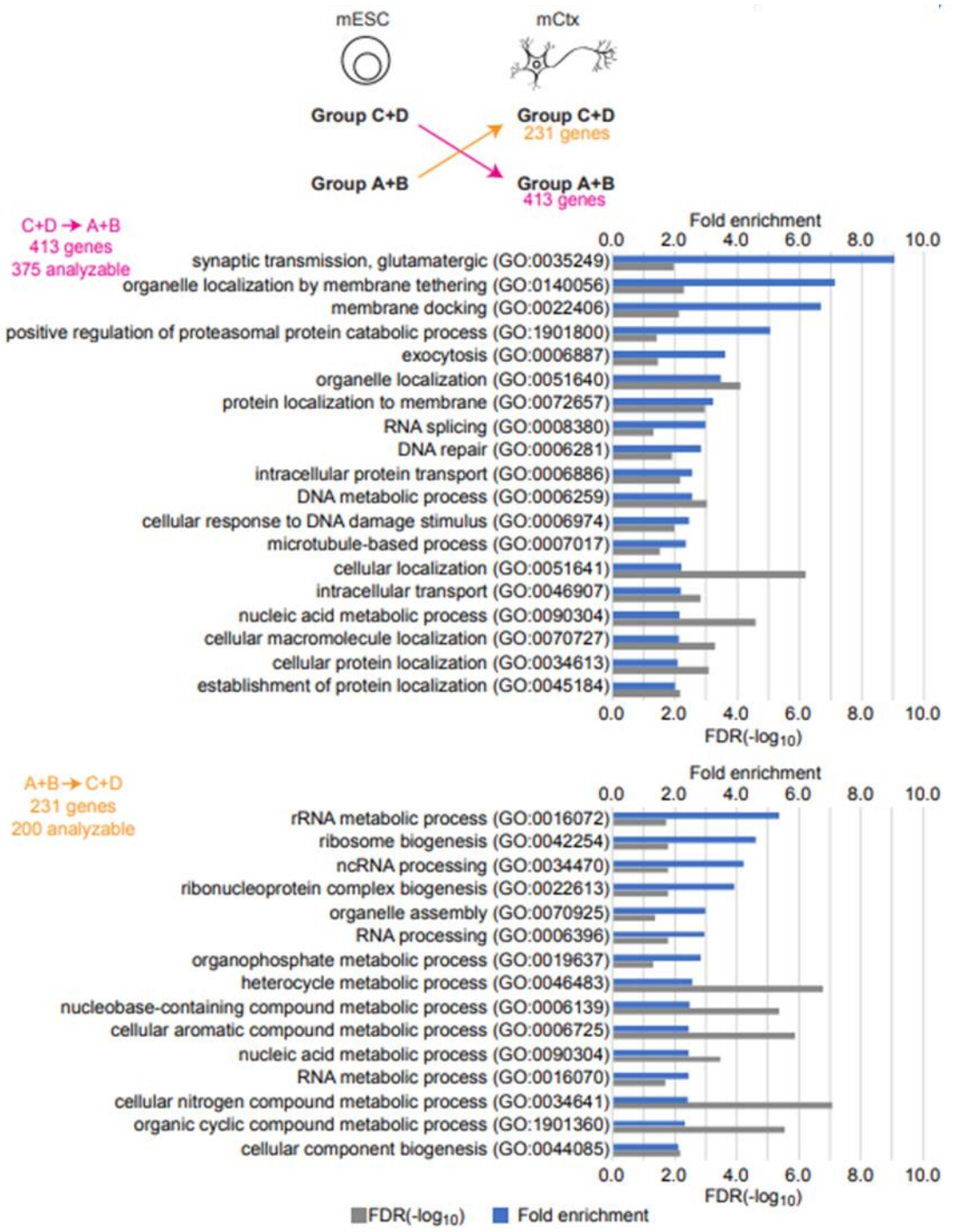
Genome browser tracks of the Rbfox1 locus for poly(A)+ and total libraries. Y axis shows RPM in each library.



Supplementary Figure 3.10 Computational definition of introns and splicing.

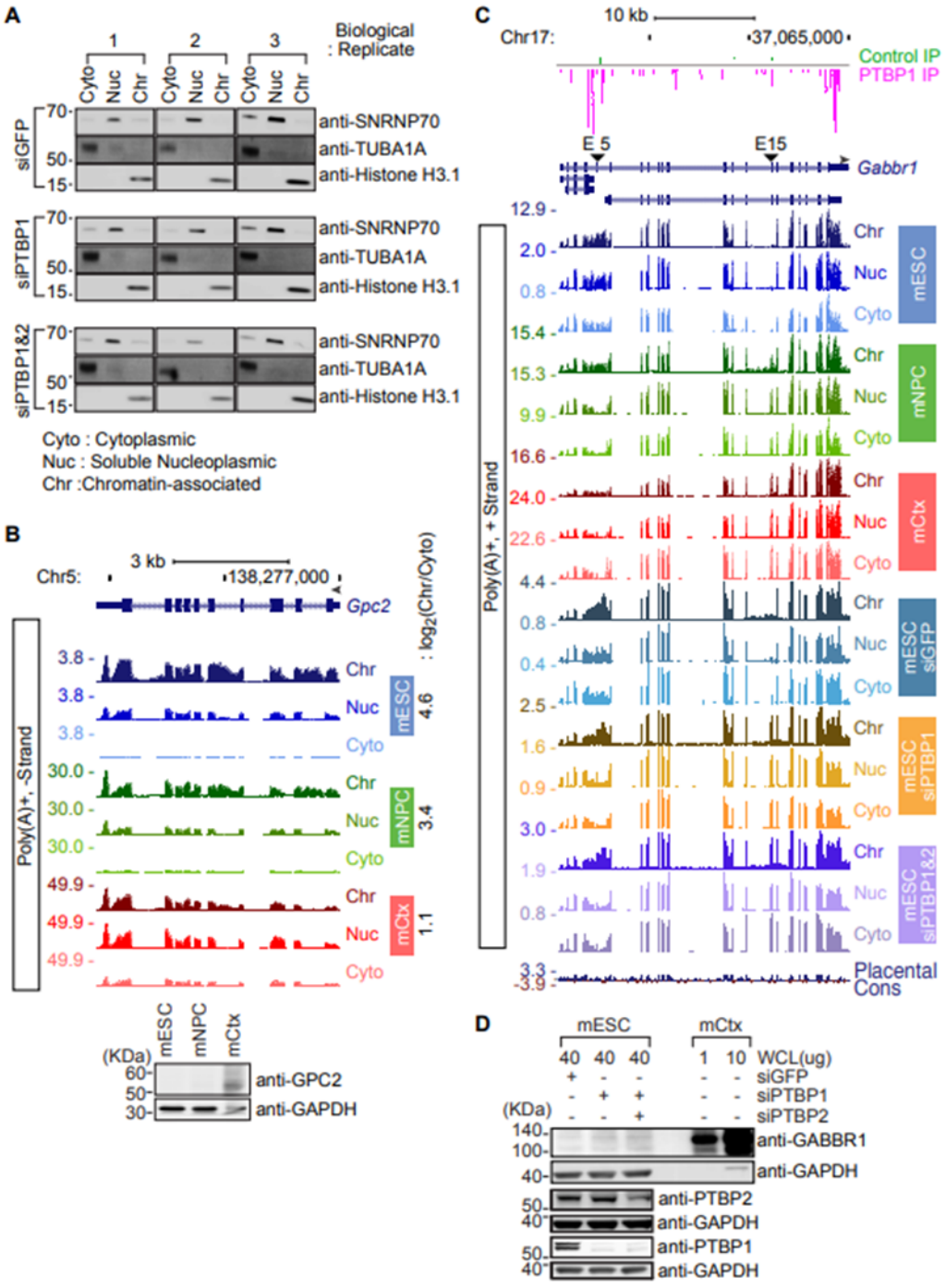
(A) Determination of FI value using read numbers for the exon-intron junction (EI), intronexon junction (IE), and exon-exon junction (EE). **(B)** Introns were categorized as one of four types based on their Ensembl v91 annotation. Introns that are not partly overlapped with either exons or other introns 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 3.2C**. **(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 (Log2(Chr/Cyto)) of transcripts from different gene groups defined above. The number of genes in each gene group is indicated at the bottom.



Supplementary Figure 3.11 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 (-log10) shown in blue and grey bars, respectively.



Supplementary Figure 3.12 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.

3.6 References

1. Vargas, D.Y. *et al.* Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell* **147**, 1054-65 (2011).
2. Coulon, A. *et al.* Kinetic competition during the transcription cycle results in stochastic RNA processing. *Elife* **3**(2014).
3. Girard, C. *et al.* Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat Commun* **3**, 994 (2012).
4. Popp, M.W. & Maquat, L.E. Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu Rev Genet* **47**, 139-65 (2013).
5. Stewart, M. Polyadenylation and nuclear export of mRNAs. *J Biol Chem* **294**, 2977-2987 (2019).
6. Quinn, J.J. & Chang, H.Y. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17**, 47-62 (2016).
7. Schmid, M. & Jensen, T.H. Controlling nuclear RNA levels. *Nat Rev Genet* **19**, 518-529 (2018).
8. Bhatt, D.M. *et al.* Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279-90 (2012).
9. Pandya-Jones, A. *et al.* Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. *Rna* **19**, 811-27 (2013).
10. Herzel, L. & Neugebauer, K.M. Quantification of co-transcriptional splicing from RNA-Seq data. *Methods* **85**, 36-43 (2015).
11. Khodor, Y.L., Menet, J.S., Tolan, M. & Rosbash, M. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *Rna* **18**, 2174-86 (2012).
12. Pandya-Jones, A. & Black, D.L. Co-transcriptional splicing of constitutive and alternative exons. *Rna* **15**, 1896-908 (2009).
13. Pawlicki, J.M. & Steitz, J.A. Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production. *J Cell Biol* **182**, 61-76 (2008).
14. Wuarin, J. & Schibler, U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* **14**, 7219-25 (1994).
15. Yeom, K.H. & Damianov, A. Methods for Extraction of RNA, Proteins, or Protein Complexes from Subcellular Compartments of Eukaryotic Cells. *Methods Mol Biol* **1648**, 155-167 (2017).
16. Pandya-Jones, A. *et al.* A protein assembly mediates Xist localization and gene silencing. *Nature* **587**, 145-151 (2020).
17. Fei, J. *et al.* Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J Cell Sci* **130**, 4180-4192 (2017).

18. Hutchinson, J.N. *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8**, 39 (2007).

19. Garland, W. & Jensen, T.H. Nuclear sorting of RNA. *Wiley Interdiscip Rev RNA* **11**, e1572 (2020).

20. Hautbergue, G.M. RNA Nuclear Export: From Neurological Disorders to Cancer. *Adv Exp Med Biol* **1007**, 89-109 (2017).

21. Jacob, A.G. & Smith, C.W.J. Intron retention as a component of regulated gene expression programs. *Hum Genet* **136**, 1043-1057 (2017).

22. Wegener, M. & Müller-McNicoll, M. Nuclear retention of mRNAs - quality control, gene regulation and human disease. *Semin Cell Dev Biol* **79**, 131-142 (2018).

23. Frankiw, L., Baltimore, D. & Li, G. Alternative mRNA splicing in cancer immunotherapy. *Nat Rev Immunol* **19**, 675-687 (2019).

24. Hao, S. & Baltimore, D. RNA splicing regulates the temporal order of TNF-induced gene expression. *Proc Natl Acad Sci U S A* **110**, 11934-9 (2013).

25. Ninomiya, K., Kataoka, N. & Hagiwara, M. Stress-responsive maturation of Clk1/4 pre-mRNAs promotes phosphorylation of SR splicing factor. *J Cell Biol* **195**, 27-40 (2011).

26. Boutz, P.L., Bhutkar, A. & Sharp, P.A. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev* **29**, 63-80 (2015).

27. Braun, C.J. *et al.* Coordinated Splicing of Regulatory Detained Introns within Oncogenic Transcripts Creates an Exploitable Vulnerability in Malignant Glioma. *Cancer Cell* **32**, 411-426.e11 (2017).

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

29. Wang, Q. & Rio, D.C. JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proc Natl Acad Sci U S A* **115**, E8181-E8190 (2018).

30. Broseus, L. & Ritchie, W. Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput Struct Biotechnol J* **18**, 501-508 (2020).

31. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* **24**, 1774-86 (2014).

32. Edwards, C.R. *et al.* A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood* **127**, e24-e34 (2016).

33. Naro, C. *et al.* An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation. *Dev Cell* **41**, 82-93.e4 (2017).

34. Parra, M. *et al.* 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 (2018).

35. Pimentel, H. *et al.* A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res* **44**, 838-51 (2016).

36. Schmitz, U. *et al.* Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biol* **18**, 216 (2017).

37. Wong, J.J. *et al.* Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**, 583-95 (2013).

38. Yeom, K.H. *et al.* Polypyrimidine tract-binding protein blocks miRNA-124 biogenesis to enforce its neuronal-specific expression in the mouse. *Proc Natl Acad Sci U S A* **115**, E11061-e11070 (2018).

39. Brockdorff, N. *et al.* 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**, 515-26 (1992).

40. Naganuma, T. *et al.* Alternative 3'-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *Embo j* **31**, 4020-34 (2012).

41. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).

42. Liu, H. *et al.* HP1BP3, a Chromatin Retention Factor for Co-transcriptional MicroRNA Processing. *Mol Cell* **63**, 420-32 (2016).

43. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18**, 1435-40 (2011).

44. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**, 1616-25 (2012).

45. Windhager, L. *et al.* Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res* **22**, 2031-42 (2012).

46. Duff, M.O. *et al.* Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature* **521**, 376-9 (2015).

47. Sibley, C.R. *et al.* Recursive splicing in long vertebrate genes. *Nature* **521**, 371-375 (2015).

48. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-7 (2016).

49. Pelleg D, M.A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *In Proceedings of the 17th International Conf. on Machine Learning*, 727–734 (2000).

50. Frankiw, L. *et al.* BUD13 Promotes a Type I Interferon Response by Countering Intron Retention in Irf7. *Mol Cell* **73**, 803-814.e6 (2019).

51. Sakabe, N.J. & de Souza, S.J. Sequence features responsible for intron retention in human. *BMC Genomics* **8**, 59 (2007).

52. Li, Y. *et al.* An intron with a constitutive transport element is retained in a Tap messenger RNA. *Nature* **443**, 234-7 (2006).

53. Keppetipola, N., Sharma, S., Li, Q. & Black, D.L. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. *Crit Rev Biochem Mol Biol* **47**, 360-78 (2012).

54. Vuong, C.K., Black, D.L. & Zheng, S. The neurogenetics of alternative splicing. *Nat Rev Neurosci* **17**, 265-81 (2016).

55. Linares, A.J. *et al.* The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. *Elife* **4**, e09268 (2015).

56. Yap, K., Lim, Z.Q., Khandelia, P., Friedman, B. & Makeyev, E.V. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes Dev* **26**, 1209-23 (2012).

57. Hurt, J.A., Robertson, A.D. & Burge, C.B. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res* **23**, 1636-50 (2013).

58. Makeyev, E.V., Zhang, J., Carrasco, M.A. & Maniatis, T. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27**, 435-48 (2007).

59. Edfors, F. *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* **12**, 883 (2016).

60. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535-50 (2016).

61. Davidson, L., Kerr, A. & West, S. Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. *Embo j* **31**, 2566-78 (2012).

62. Herzel, L., Ottoz, D.S.M., Alpert, T. & Neugebauer, K.M. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat Rev Mol Cell Biol* **18**, 637-650 (2017).

63. Naftelberg, S., Schor, I.E., Ast, G. & Kornblihtt, A.R. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu Rev Biochem* **84**, 165-98 (2015).

64. Saldi, T., Cortazar, M.A., Sheridan, R.M. & Bentley, D.L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J Mol Biol* **428**, 2623-2635 (2016).

65. Seila, A.C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849-51 (2008).

66. Flynn, R.A., Almada, A.E., Zamudio, J.R. & Sharp, P.A. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc Natl Acad Sci U S A* **108**, 10460-5 (2011).

67. Vilborg, A. & Steitz, J.A. Readthrough transcription: How are DoGs made and what do they do? *RNA Biol* **14**, 632-636 (2017).

68. Jaillon, O. *et al.* Translational control of intron splicing in eukaryotes. *Nature* **451**, 359-62 (2008).

69. Dvinge, H. & Bradley, R.K. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med* **7**, 45 (2015).

70. Horan, L., Yasuhara, J.C., Kohlstaedt, L.A. & Rio, D.C. Biochemical identification of new proteins involved in splicing repression at the Drosophila P-element exonic splicing silencer. *Genes Dev* **29**, 2298-311 (2015).

71. Pendleton, K.E. *et al.* The U6 snRNA m(6)A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. *Cell* **169**, 824-835.e14 (2017).

72. Boutz, P.L. *et al.* A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev* **21**, 1636-52 (2007).

73. Spellman, R., Llorian, M. & Smith, C.W. Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol Cell* **27**, 420-34 (2007).

74. Zheng, S. *et al.* PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. *Nat Neurosci* **15**, 381-8, s1 (2012).

75. Kaupmann, K. *et al.* Expression cloning of GABA(B) receptors uncovers similarity to metabotropic glutamate receptors. *Nature* **386**, 239-46 (1997).

76. Vigot, R. *et al.* Differential compartmentalization and distinct functions of GABAB receptor variants. *Neuron* **50**, 589-601 (2006).

77. Lin, Y., Protter, D.S., Rosen, M.K. & Parker, R. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol Cell* **60**, 208-19 (2015).

78. Attig, J. *et al.* Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell* **174**, 1067-1081.e17 (2018).

79. Yap, K. *et al.* A Short Tandem Repeat-Enriched RNA Assembles a Nuclear Compartment to Control Alternative Splicing and Promote Cell Survival. *Mol Cell* **72**, 525-540.e13 (2018).

80. Hunt, S.E. *et al.* Ensembl variation resources. *Database (Oxford)* **2018**(2018).

81. AV, N. PyClustering: Data Mining Library. *J Open Source Software* **4:1230**(2019).

82. Berndt DJ, C.J. 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**, 359–370 (1994).

83. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-45 (2009).

84. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2 (2014).

85. Kerpedjiev, P., Hammer, S. & Hofacker, I.L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 3377-9 (2015).

86. Xi, L. *et al.* Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics* **11**, 346 (2010).

87. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-44 (2015).

88. Pounraja, V.K., Jayakar, G., Jensen, M., Kelkar, N. & Girirajan, S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome Res* **29**, 1134-1143 (2019).

4 INDIVIDUALIZED DEEP-LEARNING ANALYSIS OF RNA TRANSCRIPT SPLICING

4.1 Introduction

Pre-mRNA splicing is a fundamental biological process where introns are excised, and exons are joined to form mature mRNA transcripts[1]. Alternative splicing (AS) enables the production of multiple mRNA isoforms from every single gene by selective usages of splice sites[1,2]. Around 95% of genes undergo AS, hence diversifying the transcriptome and the proteome[3]. AS consists of five main patterns, which are exon skipping, alternative 5' and 3' splice sites, mutually exclusive exons, and intron retention[4-6]. Of the five types, exon skipping is the predominant type of AS[5]. AS exhibits tissue-specific and developmental-stage-specific manner[7], thus increasing the complexities of gene regulations. AS is regulated via recognizing cis-elements and trans-acting factors that bind to the cis-elements[8]. The core cis-elements are the 5' splice site, 3' splice site and the branch point

site. The trans-acting factors are splicing factors that interact with the cis-elements in exons and flanking introns acting as splicing enhancers or splicing silencers[9]. These splicing factors, particularly referred to as RNA binding proteins (RBPs), often contribute to tissue-specific regulations. The examples of tissue-specific splicing factors include RBFOX, NOVA, PTB, and MBNL for regulating AS in neuronal and muscle cells[7].

The defects of AS are frequently observed in human diseases and cancers. For example, a C to T mutation at position 6 in exon 7 of SMN2 causes the skipping of exon 7, thereby producing a truncated and less stable protein and responsible for spinal muscular atrophy[10,11]. It is estimated that 62% of disease-related pathogenic single-nucleotide variants (SNVs) affect splicing[12,13]. Up to 30% of disease-related mutations documented in the Human Gene Mutation Database (HGMD) have been estimated to disrupt splicing[14-17]. Additionally, half of cancer synonymous drivers are predicted to be associated with splicing defects[18].

With the advent of next-generation sequencing (NGS), technologies including RNA sequencing (RNA-seq) have revolutionized the quantification and identification of AS scaled up to a genome-wide level[19]. Accumulated RNA-seq datasets with matched exome sequencing or whole genome sequencing (WGS) provide tremendous resources to study the associations of mis-splicing related variants with diseases and traits[20,21] and even investigate the potential disease diagnosis[22]. Through analysis of splicing quantitative trait locus (sQTL), many genetic variants have been discovered to be associated with AS and treated as potential candidates or biomarkers for diseases or traits[6]. However, identifying causal variants from the analysis of sQTL remains a challenge as multiple variants within the same haplotype block of causal variants are likely to be called significantly[6].

Furthermore, the analysis of sQTL is limited to commonly observable variants in relevant tissues, given the natural property of tissue-specific regulation of AS, leaving large amounts of variants unable to explore their impacts on AS. Moreover, the rapid progress in applying NGS to study disease diagnosis has resulted in an increasing number of variants identified in patients waiting for interpretation. Our incomplete understanding of AS leads to a large proportion of variants of uncertain significance (VUS), thus adversely affecting the yields of disease diagnosis. Considerable efforts such as high-throughput screening assays[23] have been put into reducing the number of VUS by interpreting their functionalities and pathogenicity. Yet experimental validations of splicing variants are impractical and time-consuming, especially for investigating de novo mutations or a combination of mutations in genetic diseases, motivating the development of computational approaches to evaluate the splicing effects of genomic variants.

Progress in predicting splicing has focused on modeling splice sites or splicing levels[24-35]. An inherent limitation of the developed computational tools that score the strength of splice sites is the lack of tissue-specific predictions of splice site usage. In this regard, multiple studies have been conducted to predict tissue-specific splicing patterns of AS by utilizing quantitative splicing microarrays or RNA-seq[27,29,31,35]. In a pioneer study, Barash et al. utilized quantitative splicing microarrays to predict tissue-specific splicing patterns in mouse with cis sequence features[27]. After this work, Xiong et al. developed a Bayesian deep-learning based method (SPANR) for predicting tissue-specific splicing in 16 human tissues[31]. A recent study conducted by Cheng et al. concentrates on predicting the tissue-specific effects of variants on splicing with a restricted number of quantified exons from RNA-seq[35]. However, to date, no tools are publicly available for predicting an

individualized tissue-specific splicing level of any transcript from sequence information and other measurements.

Here we report iDARTS (individualized Deep-learning Analysis of RNA Transcript Splicing), an individualized tissue-specific framework for predicting splicing levels. Inspired by the recent success[34] of using cis sequence features and mRNA expressions of trans RBPs to model differential AS in two conditions, the iDARTS framework integrates 997 cis sequence features including convolutional neural network (CNN) based splice sites predictor and 1,498 annotated trans RBPs to construct a deep-learning model of AS using large-scale RNA-seq datasets from 8,304 samples in 53 tissues from the GTEx[36] project (V7) (**Figure 4.1a**). This framework enables to model the cis-elements and trans-acting factors determinants of splicing patterns in the context of tissue-specificity, thus being capable of inferring causal effects of any common or rare variant on splicing.

We observed highly accurate predictions (R^2 = 0.68) of RNA-seq quantified PSI values for exon skipping events from held out chromosomes during training (**Figure 4.1b**). We further evaluated iDARTS on tissue differences in splicing levels, tissues from independent data resources, and splicing changes quantified by reverse transcription polymerase chain reaction (RT-PCR). All of them showed the accurate, robust, and generalizable behaviors of iDARTS. With these behaviors, we applied iDARTS to prioritize sQTLs identified in the GTEx project (v7). The prioritized variants from iDARTS highly correlated with true sQTL signals, and the predicted effects accurately captured the directionalities of splicing changes. It demonstrated the potential utility of prioritization of casual variants by splicing prediction.

The capability of iDARTS allows us to make splicing predictions with arbitrary sequences in the context of tissue-specificity. To explore the signatures of genetic variations implicated in splicing, we profiled the splicing effects of 9,991,388 variants up to 300nt into introns from gnomAD[37] across 447,248 exon triplets. The unprecedented scale of evaluated variants drastically expands our knowledge of the impact of variants on splicing. Our results highlighted that predicted splicing disrupt variants were strongly depleted in human populations and enriched in cancer genes. Encouraged by the findings that the predicted splicing disrupt variants could be potentially functional relevant, we envision that iDARTS, complementary to experimental approaches and genetic studies, could be of great clinical interest for disease causing variants, as evaluated in interpreting the functional consequences of VUS in clinical studies.

4.2 Results

4.2.1 Deep-learning based individual tissue-specific alternative splicing prediction

We built a framework of iDARTS to predict tissue-specific exon skipping from cis-features and trans-features as described in **Figure 4.1a**. We firstly processed the RNA-seq datasets and VCF file from the GTEx[36] project (v7) in which we obtained the quantifications of AS events and gene expressions of RNA binding proteins (RBPs). In total, 23,764 exon skipping events of 635 individuals across 53 tissues were fed in a deep-learning based model to learn the relations of input features and output PSI values. We split the exon skipping events by chromosomes (excluding chrY) into 5 folds and performed 5-fold cross-validation. Three metrics consisting of R-square, Lin's concordance correlation, and Pearson correlation were employed to evaluate the agreement between observed

predicted PSI and RNA-seq quantified PSI for exon skipping events from held out chromosomes. We observed a good agreement (R^2 = 0.68) between predicted PSI and RNA-seq quantified PSI (**Figure 4.1c**). We also noticed that the model performed the best in Esophagus Mucosa and Adrenal Gland and performed relatively less well in brain cerebellum and brain cerebellar hemisphere (**Figure 4.1d**).

We next sought to investigate the tissue-specific prediction of iDARTS for AS events. As whole blood often is frequently treated as a proxy tissue to investigate the effects of exon skipping events[12,22], we took exon skipping events from other tissues and predicted the PSI values of the exon skipping events by using the expression of RBP profiles from whole blood. We found that most tissues showed an increase in R-square between exon skipping events predicted with tissue-specific expression of RBP profiles and the expression of RBP profiles from whole blood (**Figure 4.2a**). Intriguingly, the predictions of exon skipping events from brain tissues benefited a lot with brain tissue-specific expression of RBP profiles compared to expression of RBP profiles from whole blood with at least 2% increment of R-square. The observation showed that iDARTS has successfully learned the tissue-specific manner of exon skipping.

Furthermore, we evaluated the performance of iDARTS on external datasets. In this regard, we downloaded the RNA-seq datasets from Roadmap project[38] and processed the RNA-seq similarly. iDARTS was used to predict the PSI values of exon skipping in a tissue-specific manner. We achieved good performances (**Figure 4.2b**) for most tissue types except for embryonic stem cells. The result showed that iDARTS could be applied to other datasets. A statement noted here is that iDARTS can predict the level of exon skipping when

expression of RBP profiles is unavailable. Instead, iDARTS will output the predictions of AS events with the expression of RPB profiles curated from GTEx.

In addition to the validation of iDARTS on RNA-seq data, we evaluated the performance of iDARTS on predicting PSI values of the exon 7 from SMN1 and SMN2. We extracted the effects of substitutions within exon 7 measured by RT-PCR from previous works[11,30,39-41]. In total, we collected 118 data points containing 71 different single substitutions and 47 different combinations of multiple substitutions. For each data point, we reported the predicted effects as the largest value of ΔPSI across the 53 tissues. We then compared the predictions to the experimentally estimated ΔPSI of the 118 data points. A good spearman correlation (R = 0.80) was observed (**Supplementary 4.2**) between the experimentally quantified ΔPSI and predicted ΔPSI. We also benchmarked the performance in comparison with the tool SPANR[31]. As only maximum ΔPSI was reported for every single substitution in SPANR, we only compared the predictions of data points with single substitutions from iDARTS and SPANR. We found that iDARTS achieved better concordance with experimentally quantified ΔPSI than SPANR. It is noted that both iDARTS and SPANR tend to underestimate the effects of the substitutions even with a good correlation with experimentally quantified ΔPSI. These results demonstrate the ability of iDARTS when applying to investigate the potential effects of single or multiple substitutions.

In summary, we evaluated the performance of iDARTS on held out chromosomes in GTEx project, tissue-specific exon skipping prediction, and external data resources. The results from these analyses showed a good performance of iDARTS on predicting tissue-specific exon skipping events. iDARTS can be used in a variety of scenarios with or without

the expression of customized RBP profiles. Besides, iDARTS can also be utilized to evaluate the effects of variants on exon skipping events. Thus, we applied iDARTS to study the effects of variants including sQTL variants, variants in human populations, and variants implicated with cancer and diseases. We exhibit the benefit of applying iDARTS to potentially identify causal variants for traits and reduce clinical uncertainty significance of variants implicated with diseases below.

4.2.2 Prioritizing the effect of genomic variants on tissue-specific alternative splicing

Identification of causal splicing quantitative traits loci remains challenging due to linkage disequilibrium[6,42]. We took the exon skipping event harbored in FERIL4 in stomach as an example. Three variants having significant p-values are very close to each other (**Figure 4.3a**). It brought us great trouble to determine which one may be the causal one as they all are significant. We provided a way to disentangle this problem by prioritizing them with the predictions made by iDARTS. From the predictions of the ΔPSI of iDARTS, we found the first variant with the most significant p-value had a -0.25 decrease by comparing the reference alleles and alternative alleles. The evidence together suggested the first variant may be the potential causal variant.

To investigate the benefit of using iDARTS prediction in helping sQTL analysis, we applied iDARTS to tissue-specific sQTLs with significant p-values within 300nt of skipping exons of exon skipping events and their nearby variants in the same exon skipping events. To select a list of reliable sQTLs, we required at least 10 samples available for at least two genotypes of each sQTL and, that the differences of median PSI values between two

homozygous alleles are no less than 0.05 and homozygous allele and heterozygous allele are no less than 0.025. We further removed exon skipping events that were used for training the DNN model. We observed that iDARTS predicted the variants with significant p-value having larger effects on exon skipping than variants with insignificant p-values (**Figure 4.3b**). This concluded that iDARTS preferentially prioritized variants with true sQTL signal.

Next, we explored the effects of iDARTS on prioritizing SNVs. The problem behind prioritizing SNVs is that multiple SNVs on the same exon skipping events were significant based on their p-value due to their distances close to the cause SNVs. Under the assumption that variants with significant p-value tend to be the causal variants, we selected the exon skipping events with sQTLs within 300nt of the exon-intron boundary and exonic regions of the skipping exons. To avoid ambiguity, we removed exon skipping events with at least two variants with the same most significant p-value. We ranked the variants with the absolute ΔPSI predicted by iDARTS with tissue-specific RBP expression for each selected exon skipping event. The proportion was calculated as the sQTLs being ranked the first over all sQTLs by varying the thresholds of the absolute maximum predicted ΔPSI for all variants in each exon skipping events. We also randomly ranked variants within each exon skipping event with sQTL served as control and calculated the proportion. Given each threshold, the analysis was performed with at least five observed sQTLs. In comparison with random ranked variants, iDARTS preferentially ranked sQTLs the first by varying the thresholds of the absolute maximum predicted ΔPSI (**Figure 4.3c**). The trend was going up in general with the increasing absolute maximum predicted ΔPSI (**Figure 4.3c**). We expected the strongly predicted effect variants to be highly enriched in bona fide causal

sQTLs. Among our identified sQTLs from the GTEx, iDARTS correctly predicted at least 92% of the variants with absolute predicted ΔPSI >= 0.2 for all 53 tissues (**Figure 4.3d**). The results indicated that a higher proportion of sQTLs with larger predicted ΔPSI are tending to be causal compared to background with the underline assumption that the effect of sQTLs direction of non-causal variants is independent of the predicted directions. Thus, iDARTS serves a unique role to help to interpret and prioritize genomic variants related to exon skipping.

4.2.3 Predicted splicing disruption variants are strongly depleted in human population and enriched in cancer genomes.

Connecting predicting the effects of variants on exon skipping and interpreting functionalities poses a challenge. Perturbations of different exon skipping events may result in distinct functional consequences due to the intricate biological processes[43-45]. As iDARTS can predict the effects of all possible variants on exon skipping in a genome-wide scale, it enables us to explore the signatures of splicing disruption variants in the context of natural selections and cancers.

To facilitate a genome-wide study of the effects of SNVs on exon skipping events, we curated 444,248 exon triplets from GENCODE v26lift37. For scoring the effects of each SNV, we first mapped the SNV to all possible exon triplets of which the SNV could be found within 300nt of exon-intron boundaries and exonic regions of skipping exons. Then we applied iDARTS to predicting the splicing levels with and without the SNV for each exon triplet. The difference in predicting splicing level, ΔPSI was computed for each tissue of each exon triplet. We took the largest value of ΔPSI across all tissues from all mapped exon

triplets as the potential effects of SNVs on exon skipping. The large splicing disruption variants (LSDVs) were defined to have $|\Delta PSI| \geq 0.1$.

Next, we predicted the effects of 9,991,388 variants across 447,248 exon triplets from gnomAD[37] to investigate natural selections of splicing disruption variants. gnomAD variants predicted by iDARTS are broadly spreading across the intronic and exonic regions of skipping exons and having larger effects on exon skipping near splice sites (**Figure 4.4a**). As expected, we found a significant reduction of the proportions of LSDVs from extremely rare (allele frequency $< 0.001\%$) to common (allele frequency $\geq 5\%$) variants (**Figure 4.4b**). The depletion of LSDVs in common variants relative to rare or extremely rare variants suggests that predicted LSDVs may be functionally important and undergoing strong purifying selections. Furthermore, we found that the proportion of LSDVs in functionally important genes (intolerant with possibility of loss-of-function ≥ 0.9) is significantly lower than that in tolerant genes which indicates that predicted LSDVs are potentially functional relevant (**Figure 4.4c**).

LSDVs predicted from human population genomes have demonstrated the advantages of using iDARTS in understanding the potentially functional roles of LSDVs from the perspective of population genetics. We then sought to study the functional importance of LSDVs of somatic mutations in cancers. We downloaded both coding and noncoding somatic mutations from COSMIC v91[46] and predicted the effects of variants. There are two types of somatic mutations in COSMIC database, one is recurrent (found in at least 3 samples from all tumor samples collected in COSMIC) type and the other is non-recurrent (only found in one sample) type. In contrast to non-recurrent cancer mutations, recurrent cancer mutations are more likely to be driver cancer mutations. To evaluate the

extent of driver cancer mutations on splicing, we obtained the number of LSDVs in recurrent and non-recurrent cancer mutations. We found LSDVs are significantly enriched in recurrent cancer mutations with odds ratio of 1.50 (**Supplementary Figure 4.3a**). These results indicate the advantage of interpreting the functions of mutations by using iDARTS in cancer genomes. For example, a somatic variant chr17:7579312 C>A in gene TP53 is found in a variety of cancer types including adenocarcinoma, hepatocellular carcinoma, acute lymphoblastic B cell leukemia, and endometrioid carcinoma[46]. It has been predicted to be highly deleterious and pathogenic by both CADD and FATHMM-MKL[47,48]. Previous studies indicated that this variant disrupts splicing of the TP53 mRNA[49]. We found it could also change the splicing by at least 31% predicted by iDARTS. The additional evidence from the view of splicing helps to understand the functions of the cancer mutation.

4.2.4 Utilizing iDARTS to reduce variants of uncertain significance in clinical interpretation

Splicing variants have also been recognized as significant contributors to diseases while are frequently underappreciated by disease diagnostic processes[50]. Our limited understanding of alternative splicing poses a challenge of evaluating clinical significances of splicing relevant variants, leading a large proportion of variants being annotated as variants of uncertain significance (VUS)[12,51]. The capability of iDARTS in identifying putatively functional important variants evidenced from the aspects of evolutionary constraint, gene function constraint, and cancer mutations makes it possible to explore the disease implicated variants acting upon splicing. In this regard, we used iDARTS to investigate the

utility of iDARTS for predicting the pathogenicity of SNVs and its potential application of reducing the number of VUS in clinical studies.

We selected a set of credible SNVs having stars ranging from 2 to 4 with the clinical significances annotated as “Pathogenic” and “Benign” in ClinVar. More specifically, the SNVs that are rated 2 stars or more are in practice guideline, or reviewed by expert panel, or annotated by at least two submitters with the same interpretation[51]. To avoid ascertainment bias, we classified the SNVs into four different gene features which are dinucleotide, splicing window, exonic region, and intronic region based on the distances of the SNVs from the skipping exons. Each of the SNVs is only assigned to one of the gene features. We then compared the effects of the pathogenic SNVs against the effects of the benign SNVs per gene feature respectively. We found that pathogenic SNVs and benign SNVs can be significantly separated based on the effects of the SNVs predicted by iDARTS for all gene features (**Figure 4.5a**). The result is indicative of the usefulness of iDARTS in the understanding of the pathogenicity of variants in disease. However, pathogenic and benign SNVs that are in intronic regions are barely though significantly separated by the effects predicted by iDARTS. This observation implied difficulty when predicting the effects of deep intronic variants. Furthermore, we benchmarked the performance of iDARTS against other popular splicing predicting tools including SPANR[31], SpliceAI[33], and MMSplice[32] on predicting the pathogenicity of SNVs. We found that iDARTS performs much better than any of these tools in splicing window and intronic regions but performs worst in dinucleotide region and slightly worse in exonic region (**Supplementary Figure 4.4**). Overall, iDARTS performs best of all compared tools but only slightly better than SPANR in terms of predicting the pathogenicity of SNVs.

We then explored whether iDARTS could be used to interpret the potential pathogenicity of variants being annotated as VUS. To evaluate the performance of iDARTS on predicting the potential pathogenicity of VUS, we took the VUS that could be mapped to within 300nt of exon-intron boundary and exon regions of all internal exons from the exon triplets from ClinVar 20170516 and found their reclassifications in ClinVar 20210511. Around 5.3% of these VUS in ClinVar 20170516 can be reclassified in the clinically significant categories of Pathogenic, Benign, Likely Pathogenic, and Likely Benign in ClinVar 20210511 (**Figure 4.5b**). The reclassifications of these VUS provide a great opportunity of validating the predictions of iDARTS on VUS in ClinVar 20170516. Thus, we predicted the effects of these VUS similarly as we did for gnomAD variants. We defined the variants with $|\Delta PSI| \geq 0.05$ as splicing disrupt variants. We observed a great enrichment of predicted splicing disrupt variants in Pathogenic (2-4 star), Pathogenic (0-1 star), or Likely Pathogenic with the proportions 20%, 31%, and 12% compared to Benign (2-4 star), Benign (0-1 star) or Likely Benign with the proportions 2%, 2%, and 3% (**Figure 4.5b**). The result suggests that predicted splicing disrupt variants may be potential pathogenic. And the reduction of the number of VUS via iDARTS makes it possible for improving disease diagnostics and allowing evidence-based treatments to be conducted.

4.3 Discussion

In this work, we developed a deep-learning-based framework for modeling the cis sequence features and trans-acting factors determinants of AS in the context of tissue-specificity. The cis sequence features include 997 sequence features from previous studies[31,34], the hexamer-level scores of splicing enhancers and silencers learned from 2

million random sequences of a massive parallel report assay[30], and the CNN based predictors of the strength of splice sites. The trans-acting factors consist of 1,498 mRNA expressions of trans-RBPs. Both the cis- and trans-features characterize our best knowledge of AS. Our framework learned the mapping function from 2,495 features to splicing patterns by leveraging 8,304 RNA-seq datasets from 635 individuals in 53 tissues from the GTEx project (V7) that provides enormous amounts of quantified AS events as well as unprecedented variations of splicing in terms of genetic backgrounds from different individuals and tissues. Our framework can predict individual's tissue-specific, exon skipping-specific splicing level and can also be used in a variety of scenarios with or without the expression of individualized RBP profiles. We anticipate our framework could be of great benefits to studying splice-altering SNVs or combination of SNVs in a genome-wide scale, inferring causal effects of variants in genetic studies, and interpreting potentially functional consequences of variants in disease.

Inferring causal variants from associations in population-based genetic studies remains a challenge. We demonstrated the performance of iDARTS in prioritizing sQTLs. The design of iDARTS framework enables us to predict the effects of variants directly from sequence information, thus capable of identifying causal variants. Additionally, our model does not use disease or traits-related variants collected in database or allele frequencies, thereby avoiding ascertainment bias. Our model could potentially be coupled with statistical methods in the case of rare diseases when available samples are not enough for drawing statistical significances of variants of interests.

Through genome-wide profiling of the effects of variants on splicing, we found that the predicted splicing disrupt variants undergo evolutionary constraints. Furthermore, our

results showed that iDARTS could be helpful to interpret the functionalities of VUS in diseases. These findings shed light on using iDARTS to study genetic determinants of diseases. Using in-silico tools to discover candidate disease implicated variants has been popular in clinical studies[50]. For example, ATP7B variant c.1934T > G has been reported to be a pathogenic variant with a limited understanding of its functional consequences[52]. Merico et al. combined the evidence from in-silico prediction of the splice change and experiment verification and validated the functional role of this variant. We found iDARTS could also identify the role of this variant on splicing with predicted ΔPSI = -0.091. Motivated by this, we propose utilizing iDARTS to reveal the potential functions of the splice-altering variants in diseases. With the flexibility in terms of predicting splicing from arbitrary sequences in the context of tissue-specificity, iDARTS can be applied to any transcript and overcomes the issues that using single transcript may lead to mis-classification of variants[43,44] and disease relevant tissues are unavailable.

Our model exhibits accurate, robust, and generalizable behaviors with several limitations. Owing to a potential trade-off between robustness and sensitivity to sequence changes and unaccounted splicing features, our model may underestimate the effects of splice-altering variants yet with good correlations. We expect the sensitivity of our model to sequence changes would be improved with more datasets that perturb sequences as well as the expression of RBPs in tissues via high-throughput genome editing in a massive parallel manner. Moreover, our model could be further improved by incorporating gene regulation features including chromatin marks and transcriptional factors, RBP binding profiles, and RNA modifications.

Our understanding of how genetic variants affect splicing is evolving. As splicing has become more and more appreciated in studying diseases[6,43,44,50], we anticipate that in-silico probing the effects of variants on splicing from genome-wide sequences through modeling the cis-elements and trans-acting factors determinants of splicing would be an integral part of clinical studies, thus providing splice-altering variant candidates with therapeutic potentials.

4.4 Methods

4.4.1 iDARTS framework architecture.

We propose a method called iDARTS (**Figure 4.1a**), an individualized Deep-learning Analysis of RNA Transcript Splicing which learns a deep-learning model of alternative splicing by leveraging large-scale RNA-seq datasets from 8,304 samples in 53 tissues from the GTEx project (V7)[36]. iDARTS is designed to predict an individualized PSI (percent of spliced in) of an exon skipping event. The workflow of iDARTS consists of three components. First, a personalized genome and RNA sequencing (RNA-seq) are taken as inputs to provide the mutation profiles of exon skipping events and the expressions of RNA binding proteins (RBPs). Then, a list of curated cis-features generates from the input of exon skipping with the mutation profiles and trans-features retrieved from the expressions of RBPs. Last, both cis and trans-features are fed into a deep neural network (DNN) to make tissue-specific predictions of PSI (percent of spliced in) values for every exon skipping event.

In the first component, we downloaded both the VCF file and the RNA-seq datasets from the GTEx project via dbGAP[53]. We removed the RNA-seq fastq files with inconsistent read lengths and kept the fastq files with read length 76bp. In total, 8,304 RNA-seq datasets from 53 tissues were processed with aligner STAR 2.5.3.a. Alternative splicing (AS) events were quantified by rMATS-turbo[4,54] using GENCODE[55] v26lift37 as the gene annotation file. We processed the RNA-seq fastq files with Kallisto[56] (v.43.0) to quantify gene expression levels using GENCODE (v.19) protein-coding transcripts as the index.

The second component is the feature extractions for AS events. For trans-features, we extracted the gene level transcripts per kilobase million (TPM) values of 1,498 known RBPs. We collected 997 cis-features including conservation scores, motifs, splicing strength, and RNA secondary structures. In total, 2,495 features were collected for every exon skipping event. We normalized each of the features by dividing its maximum value of all exon skipping events. To obtain a high-quality and non-redundant set of exon skipping events for training, we removed cassette exons that overlap with other exons, and exons that are very short (< 10nt) or very long (> 600nt). To avoid penetrant bias, we removed the AS events in the training datasets with heterogenic alleles within 300nt of exon-intron boundaries and exonic regions of all three exons in the AS events. We also required that the cassette exons from rMATS-turbo processed AS events in each individual have junction counts no less than 20.

The last component is a DNN model that predicts tissue-specific PSI value of exon skipping event. We designed iDARTS DNN model with three hidden layers. The model architecture was specified as follows: an input layer with 2,495 variables; three fully connected layers with 500, 250, 125 variables and the ReLU activation function; an output

layer with one variable and the Sigmoid activation function. Additionally, we added a drop-out probability between hidden layers to reduce potential over-fitting issues. The drop-out probability was set to be 0.2 for all three hidden layers. Followed by drop-out, we also added Batch Normalization to help the convergence of the DNN model. Then, we used a stochastic gradient descent (SGD) optimizer with the learning rate 1e-4 and the batch size 2,000 for the DNN model training task. The objective function for training the DNN model was defined as the sum of negative log-likelihood and mean-square-error of training samples with PSI values between 0.3 and 0.7. Specifically,

$$Objective = negative\ log\ likelihood + \ 5 \times ||Y_t - f_t(X_t)||^2 \times \mathbf{I}[0.3 \leq Y_t \leq 0.7]$$

$$negative\ log\ likelihood = \ -\sum_t \log\left(Y_t f_t(X_t) + (1 - \ Y_t)(1 - \ f_t(X_t)\right)$$

where t indicates index of training samples and Y_t indicates PSI values between 0 and1 for samples. $f_t(X_t)$ denotes as the predicted PSI values given input X_t. We designed the objective function in the way to account for the U shape distribution of the PSI values in which most PSI values are either close to 1 or close to 0. The log-likelihood function was to fit the U shape distribution. We also posed more learning weight to balance the minor fraction of training samples with PSI values between 0.3 and 0.7 by adding five times the mean-square-error of these samples in the objective function. We implemented the DNN model architecture by using Keras (https://github.com/keras-team/keras) with Tensorflow (https://www.tensorflow.org/) backend.

4.4.2 Evaluation of the performance of iDARTS

The evaluation of iDARTS was conducted in three ways. First, we evaluated the performance of iDARTS on predicting RNA-seq quantified individual PSI values from GTEx project. We randomly split the chromosomes (excluding chrY) into 5 folds with nearly equal size of exon skipping events and performed 5-fold cross-validation. For each fold as the testing data, the remaining four folds were used for training five iDARTS DNN models with five different initial random seeds. The predicted PSI values were computed as the average of PSI values predicted by the five models for each fold. The procedure was repeated five times. The final performance was obtained by aggregating all testing data together. We employed three metrics: (1) Pearson correlation; (2) Lin's concordance correlation; and (3) R-square to evaluate the performance of predicted RNA-seq PSI values vs. RNA-seq quantified PSI values. Second, we sought to evaluate whether iDARTS had learned the tissue-specific behaviors of exon skipping events. We selected whole blood tissue as controls and compared the performance of iDARTS for predicting PSI values of exon skipping events that have average PSI values at least 5% different from those in whole blood tissue. Specifically, for each tissue other than whole blood, we compared the difference of the performances predicted by using tissue-specific median RBP expressions and the median RBP expressions in whole blood. The difference in R-square would reflect the benefit of using tissue-specific RBP expression profiles. Last, we tested the performance of iDARTS on external datasets. We processed the Roadmap RNA-seq similarly as we did for the GTEx project. In total, 52 tissues with both AS quantifications and gene expressions of RBPs were obtained by using rMATS-turbo and Kallisto. We applied iDARTS to predict

the PSI values of exon skipping events with the input of gene expressions of RBPs and the curated sequence features.

4.4.3 The Splicing Quantitative Trait Loci (sQTL) analysis of GTEx

We analyzed the sQTLs in 635 people in each tissue from the GTEx project. To obtain reliable AS events, we filtered out them in each tissue individually with the following criteria: (1) The median number of splice junction reads is no less than 5 where the number of splice junction reads is counted by the number of inclusion reads / 2 + the number of skipping reads; (2) The maximum difference of PSI values of AS events across all samples is larger than 10%; (3) at least three samples in the tissue have PSI values different than the median PSI values. We implemented a linear regression model to investigate sQTLs by testing the association of PSI values with SNVs within 200kb upstream or downstream of alternative exons. For each exon skipping event, the sQTL p value was reported as the closest SNP with the most significant p value within the 200kb window.

4.4.4 Construction of exon triplets for genome-wide analysis

To enable a genome-wide analysis of the effects of single nucleotide variants (SNVs), we constructed exon triplets based on all internal exons in GENCODE v26lift37 and their two flanking exons. We removed the exon triplets with exons that are very short (<10nt) and very long (>600nt). In total, 444,248 exon triplets are generated.

4.4.5 Genome-wide analysis of splicing dysregulation in Genome Aggregation Database

Genome Aggregation Database (gnomAD v.2.1.1)[37] is composed of 15,708 genomes and 125,748 exomes from sequencing studies of unrelated individuals who generally do not have a Mendelian disease. We downloaded the gnomAD VCF file from https://gnomad.broadinstitute.org/downloads. SNVs from the VCF file within 300nt of exon-intron boundary and exon regions of all internal exons from the exon triplets were selected for downstream analysis. The minor allele frequencies for SNVs were determined using the flag "AF" in the VCF file. The annotation of intolerance and tolerance of genes were downloaded from https://gnomad.broadinstitute.org/ with the file name forweb_cleaned_exac_r03_march16_z_data_pLI_CNV-final.txt.gz.

4.4.6 Genome-wide analysis of variants induced splicing defects in COSMIC Database

The Catalogue of Somatic Mutations in Cancer (COSMIC v93)[46] is a comprehensive data resource for somatic mutations in human cancer. We downloaded the mutations from both coding and noncoding regions in VCF file formats which are CosmicCodingMuts.vcf.gz, and CosmicNonCodingVariants.vcf.gz respectively. We also downloaded a list of curated cancer consensus genes from COSMIC database. The flag "CNT" in the VCF files was utilized to determine whether the mutations were recurrent or not following the descriptions on the website. We only investigated the SNVs that are within 300nt of exon-intron boundary and exon regions of all internal exons from the exon triplets that are harbored in cancer consensus genes.

4.4.7 Genome-wide analysis of disease-related variants in ClinVar Database

ClinVar[51] is a publicly accessible and frequently updated database of disease-related variants curated from literature reviews, clinical and research studies. We downloaded ClinVar VCF files with two versions, namely, 20170516 and 20210511. We obtained the review status of variants with the flag "CLINREVSTAT" and the clinical significance of variants with the flag "CLINSIG" from the VCF files. The SNVs with clear annotations, review status, clinical significance, and within 300nt of exon-intron boundary and exon regions of all skipping exons of exon triplets were analyzed from the VCF files.

4.4.8 Construction of the cis-sequence features

The cis-sequence features were built partly based on the previous works[31,34]. We also incorporated new features to accommodate a more sequence-wise understanding of AS. For each exon skipping event, we denoted C1, A, and C2 as upstream exon, alternative exon, downstream exon, respectively. I1 represents the intronic region between upstream exon and alternative exon. I2 represents the intronic region between alternative exon and downstream exon. The whole descriptions of the cis-sequence features were listed below.

1. Log-transformed exon length of C1, A, and C2, as well as the intron length of I1 and I2.
2. Frameshift of A, where the length of A can be divided by 3
3. Transability of C1, C1C2, C1A and C1AC2. The transability is defined as the same from the previous work[31] in which it was defined whether there is no stop codon in all three possible reading frames.

4. Splice strength of donor and acceptor sites were predicted by splice strength predictor built by us. See supplemental methods for more details.
5. We weighted the splice strength of donor and acceptor sites with the average of junction conservations. The definition of the junction conservations was taken from the work[57] in which it was defined as the average conservation scores of junction sites divided by those of 100nt of nearby intronic regions. The conservation scores were downloaded from UCSC phastCons46way.
6. We ran RNAplfold[58] 2.2.10 to predict unpaired probabilities of different intronic regions described as the RNA secondary features.
7. We downloaded Alu elements from UCSC[59]. Features are computed as the counts of Alu elements on either plus strand and minus strand or both strands.
8. Hexamers with respect to Exonic and intronic splicing enhancer and silencer scores were downloaded from these works[30,60-62].
9. Short motifs and their weighted scores by conservations were incorporated from the work[31].

In total, 997 cis-sequence features were constructed for each exon skipping event.

4.5 Figures

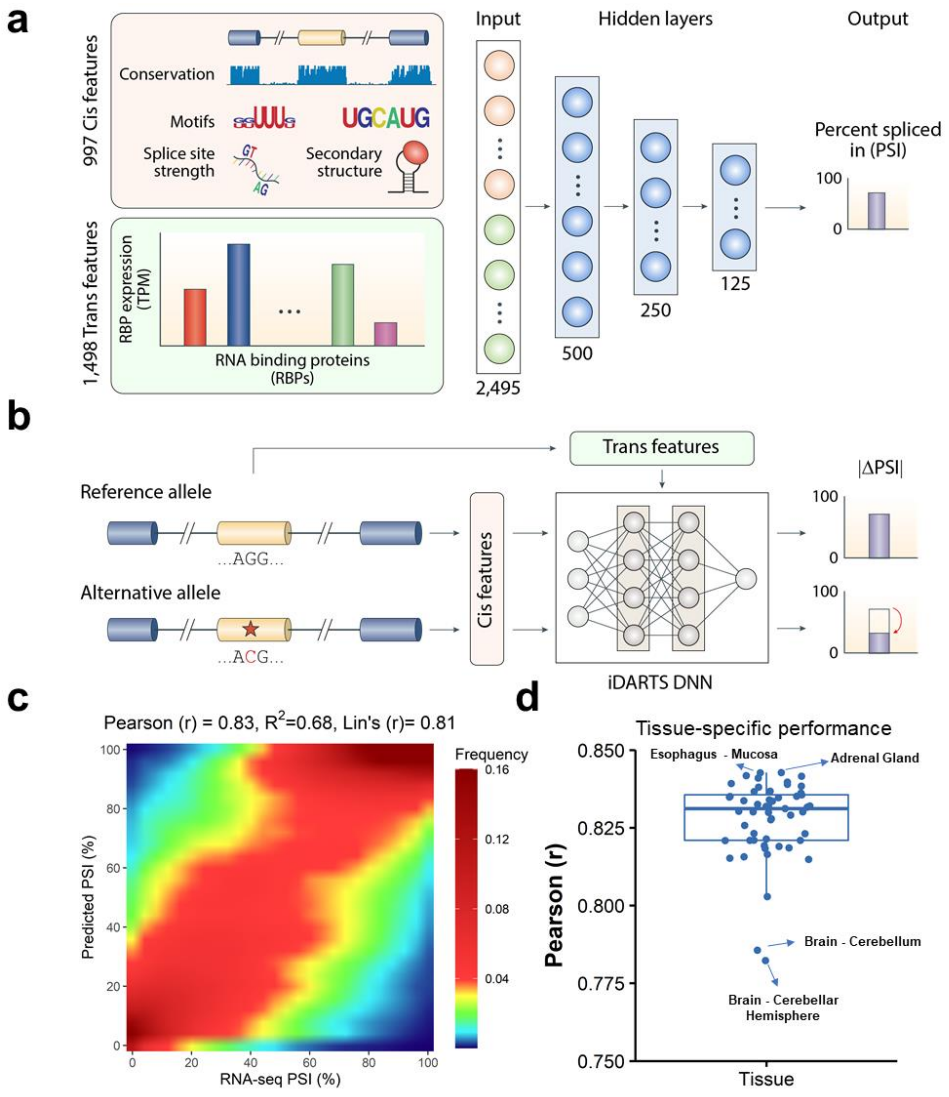


Figure 4.1 The framework and performance of iDARTS.

(a) The workflow of deep learning-based individualized alternative splicing prediction. **(b)** iDARTS DNN model extracts the cis-features from sequences with reference allele and alternative allele and makes predictions based on trans-features representing tissue-specificity for reference PSI value and alternative PSI value, respectively. The difference in predicted PSI values for alternative and reference is computed as the predicted tissue-

specific effect of alternative allele on splicing. **(c)** Performance evaluation of the predictions for 23,764 exon skipping events from 635 individuals in 53 tissues. The exon skipping events in each individual are binned based on their RNA-seq estimated PSI values. The distributions of predicted PSI values are illustrated for each bin in the plot. **(d)** Tissue-specific performance of the computational model.

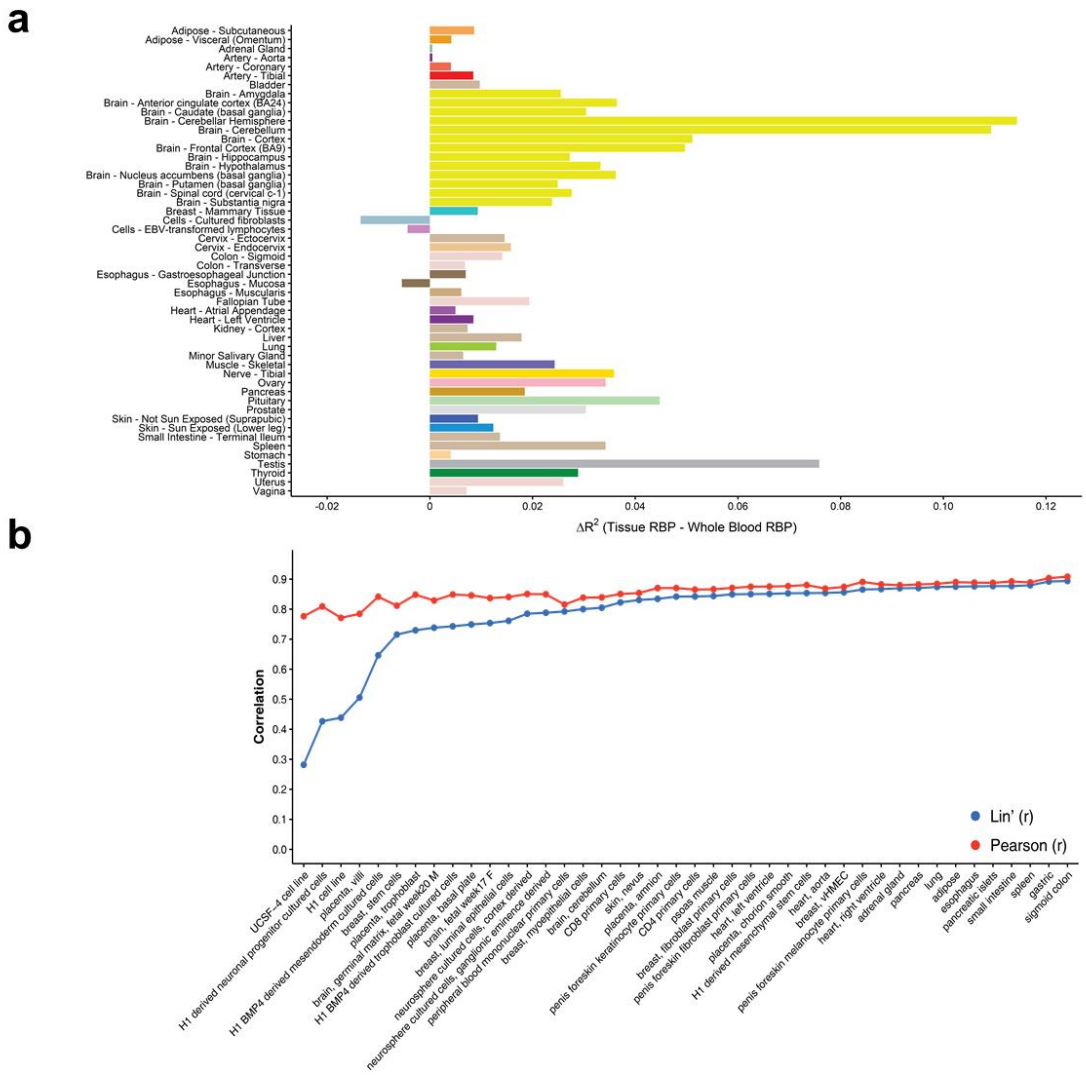


Figure 4.2 Tissue-specific evaluations of iDARTS on GTEx and Roadmap project.

(a) For each tissue other than whole blood, the performance in R-square is generally higher with tissue-specific RBP expression profiles than with whole blood RBP expression profiles except for cultured fibroblasts cells, EBV-transformed lymphocytes, and esophagus muscularis. **(b)** We evaluated the tissue-specific performances of iDARTS on exon skipping events from the Roadmap project.

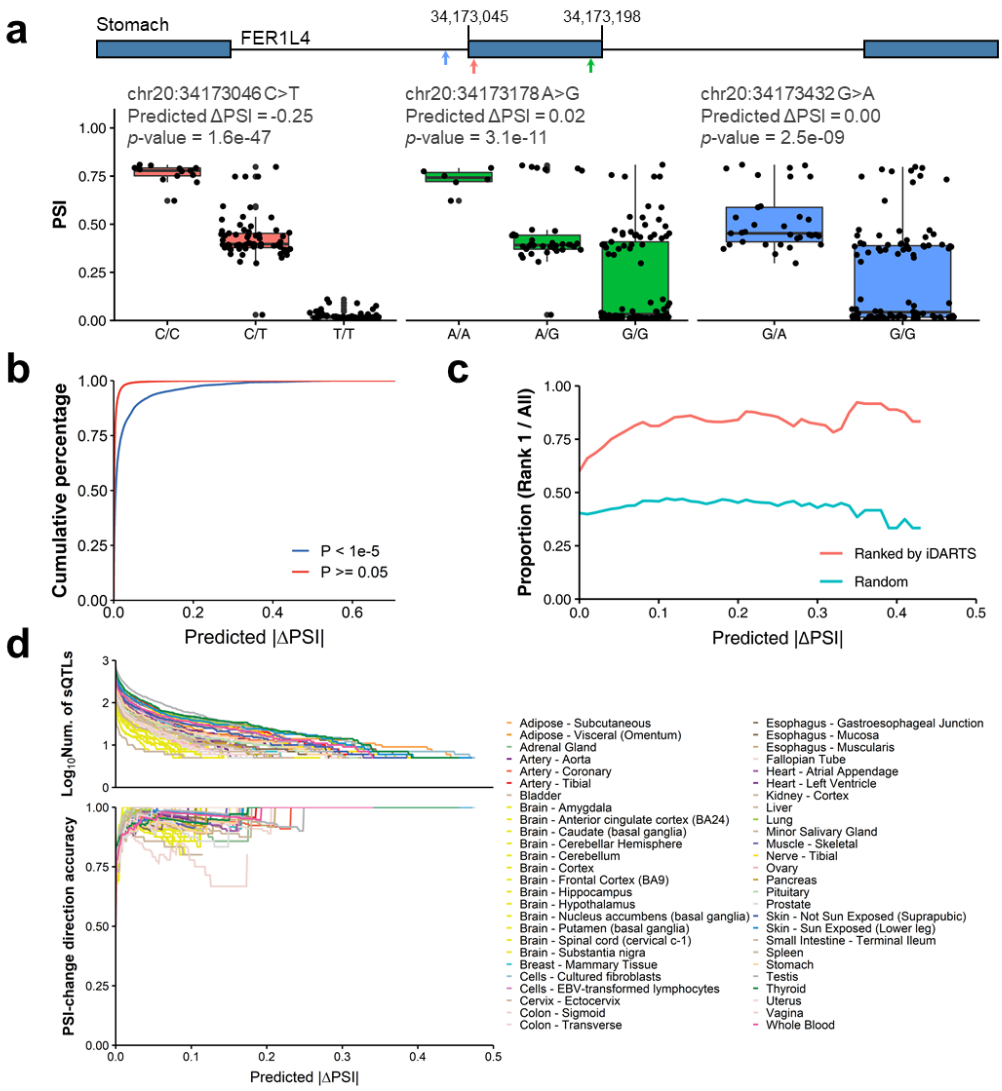


Figure 4.3 Tissue-specific predictions of the genomic effects on exon skipping.

(a) An example of three variants with significant p-values and predictions by iDARTS for the exon skipping event from gene FER1L4 in the stomach. **(b)** The cumulative percentage plot of the predicted |ΔPSI| of sQTLs with p-value < 1e-5 and variants with p-value >= 0.05 colored as blue and red, respectively. It shows that sQTLs with p-value < 1e-5 were predicted to have larger effects than variants with p-value >= 0.05. **(c)** The proportion of sQTLs being ranked first place compared to their nearby variants for all exon skipping

events. The x-axis represents the threshold of selecting the exon skipping events with the maximum predicted |ΔPSI| for all variants, and the y-axis represents the proportion of sQTLs ranked first place for all selected exon skipping events. The red line indicates the rank is made by iDARTS, which is higher than the green line for which the rank is randomly decided. **(d)** sQTL direction prediction accuracy generally increases with predicted magnitude represented as |ΔPSI|. The x-axis represents the threshold of predicted |ΔPSI| of sQTLs. The y-axis of the top figure represents the number of sQTLs given the threshold of |ΔPSI|. At least 5 sQTLs for each threshold are analyzed. The y-axis of the bottom figure represents the accuracy of predicting the directionality of splicing changes of variants with predicted |ΔPSI| >= threshold. Each line indicates each tissue.

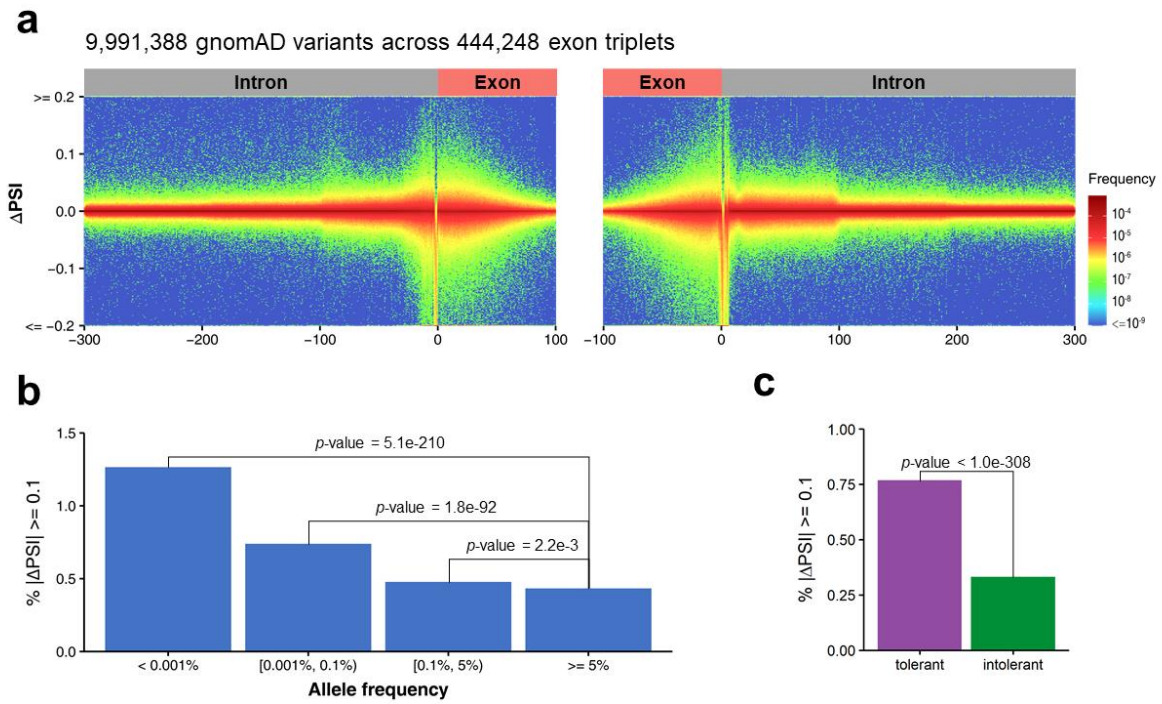


Figure 4.4 Genome-wide analysis of the effects of SNVs on splicing.

(a) 9,991,388 gnomAD variants across 444,240 exon triplets were predicted by iDARTS. The x-axis depicts the location of each variant relative to the boundaries of the middle exons of all exon triplets. The y-axis illustrates the predicted ΔPSI for each variant. The color-coded frequency represents the proportion of variants in each specific region. Most variants have no or small effects on splicing, while variants close to the splice sites impose a large effect on splicing. **(b)** The proportion of predicted large disrupted splicing variants (LDSVs) with |ΔPSI|>= 0.1 is significantly depleted from extreme variants to common variants (fisher-exact test). We divided allele frequency into four bins shown in the x-axis. The y-axis shows the proportion of large disrupted splicing variants over all variants with allele frequency in the corresponding bin. The p-values were obtained using fisher-exact test by comparing the number of LDSVs in each of the first three allele frequency bins with

the last allele frequency bin. **(c)** The proportion of LDSVs is significantly smaller in intolerant genes than in tolerant genes (fisher-exact test).

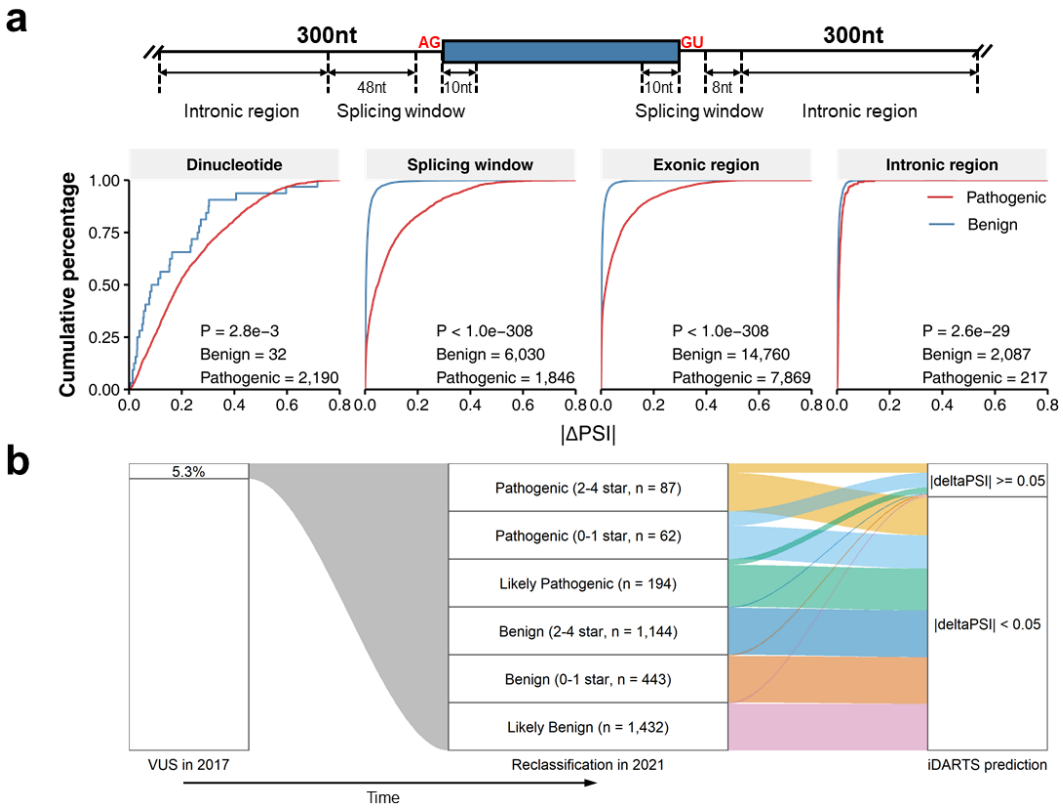
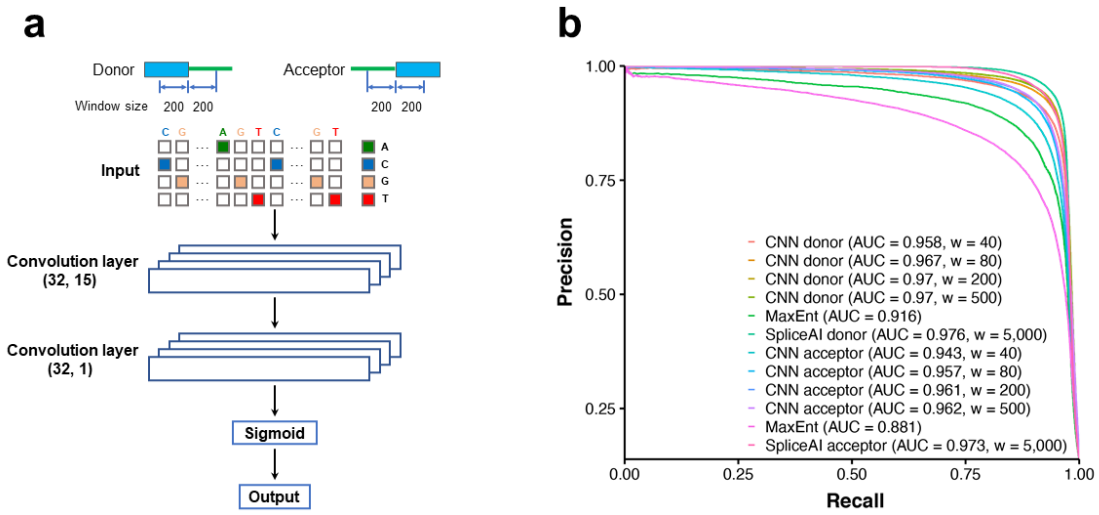


Figure 4.5 Predicting the splicing effects of disease variants helps to understand the pathogenicity of variants and reduce the number of variants of uncertain significance (VUS).

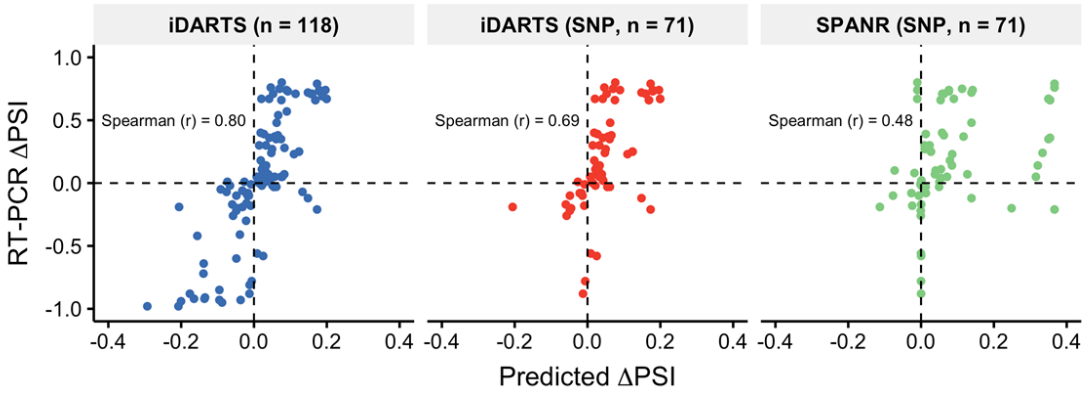
(a) pathogenic variants were predicted to have significantly larger effects on splicing than benign variants. The variants were classified into four different gene features based on their locations relative to skipping exons of exon triplets. The variants in dinucleotide are the variants found in 5' or 3' core consensus splice sites. For the variants not found in these splice sites, we characterize them as variants from the splicing window if they are within 50nt of upstream of and 10nt of downstream of the 3' splice site, or within 10nt on each side of 5' splice site. The variants located inside and 10nt away from the exon-intron boundaries of skipping exons are variants from the exonic region. For the variants in

intronic region, they are within 300nt of exon-intron boundaries of skipping exons and are not overlapping with the splicing window and the exonic region of either upstream exons and downstream exons and regions defined above. The cumulative percentage plots illustrate the effects of pathogenic and benign variants on splicing predicted by iDARTS with the x-axis being the predicted |ΔPSI|, and the y-axis being the cumulative percentage. The significance of differences in predicted |ΔPSI| for two groups was determined by Wilcoxon rank test. **(b)** Predicted splicing disrupt variants are enriched in Pathogenic and Likely Pathogenic categories when predicting the pathogenicity of VUS in 2017. The alluvial plot shows a time flow from VUS in 2017 to their reclassification in 2021 to the prediction results made by iDARTS. From the first blocks to the second blocks, we found around 5.3% of VUS could be reclassified in 2021. And the stream flows between the second and third blocks exhibit the variants being annotated as splicing disrupt variants (|ΔPSI| >= 0.05) and variants with small effects (|ΔPSI| >= 0.05). The widths of the stream flows represent the proportion of variants assigned to each of the third blocks.

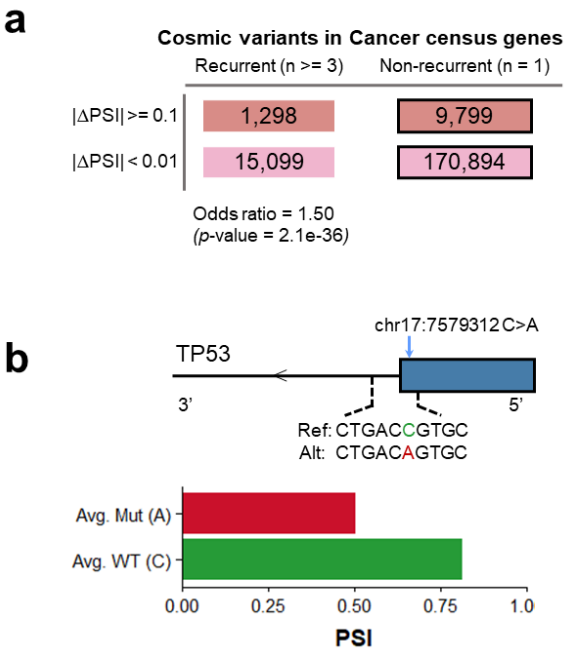


Supplementary Figure 4.6 The architecture of CNN donor and acceptor models.

(a) The schematic overview of the CNN donor and acceptor models. For both donor and acceptor model, the flanking 200nt of nucleotides on each side of the position of interest are used as input. The CNN architecture for both donor and acceptor models is composed of two convolution layers and one output layer with a sigmoid activation function. The first convolution layer is designed with 32 kernels and window size 15 and the second convolution layer is designed with 32 kernels with window size 1. **(b)** Benchmarking the performances of CNN donor and acceptor models with window size 40, 80, 200, and 500 against MaxEnt and SpliceAI.

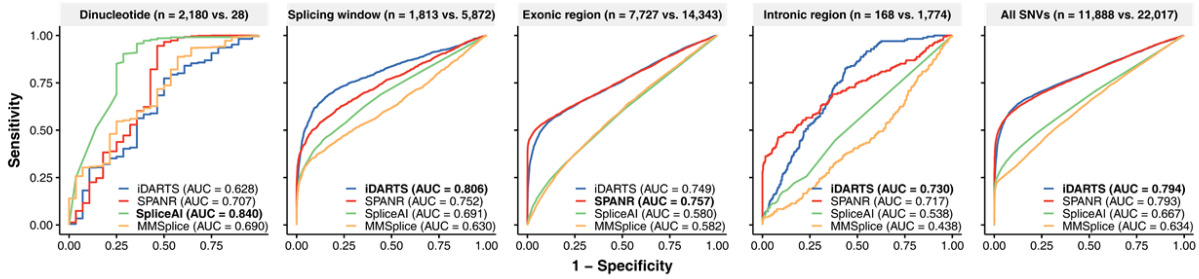


Supplementary Figure 4.7 The change in PSI for every data point containing either single substitutions or combinations of multiple substitutions for both iDARTS and SPANR.



Supplementary Figure 4.8 Genome-wide analysis of the effects of somatic mutations on splicing in cancer census genes.

(a) The predicted large splicing disruption variants (LSDVs) with |ΔPSI|>= 0.1 are strongly and significantly enriched at recurrent somatic mutations with odds ratio 1.50 (fisher-exact test). **(b)** An example of one somatic mutation in gene TP53 that appears in many cancer types is predicted to affect splicing by 31%.



Supplementary Figure 4.9 Benchmarking the performance of iDARTS, SPANR, MMSplice, and SpliceAI on predicting the pathogenicity of variants.

4.6 Table

Supplementary Table 4.1 Comparison of time requirements between CNN splice predictors and SpliceAI for the task of scoring 10,000 splice sites.

Task	Program	Time	Hardware
Scoring 10,000 donor sites	CNN donor model	32s	Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz
	SpliceAI	32719s	Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz
Scoring 10,000 acceptor sites	CNN acceptor model	33s	Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz
	SpliceAI	32686s	Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz

4.7 Appendix

4.7.1 Construction of splicing strength predictors for splicing donor and acceptor

We built splicing strength predictors for donor and acceptor separately. To construct a training dataset for both donor and acceptor models, we retrieved all splicing junctions from GENCODE v26lift37 annotated transcripts as a positive set. Within positive set, around 98% of 3' splice sites have consensus dinucleotide AG and around 97% of 5' splice sites have consensus dinucleotide GT. For the negative set, we randomly selected genomic positions that are within 300nt of but not overlapping annotated splice sites. As consensus dinucleotide AG and GT are dominant in annotated 3' splice sites and 5' splice sites, respectively, around 50% of the negative sites were chosen to have AG or GT to increase the robustness of the model. Specifically, for each annotated splice site, we randomly selected about 6 positions around 300nt of the splice site and 3 of 6 positions were required to have consensus dinucleotide GT for 5' splice sites and AG for 3' splice sites. In total, we had 6 times of negative sites compared to positive splicing sites for training.

We built a CNN model to predict the splice sites with the input of one-hot encoded sequences of window size of 200nt on each side of the position of interest. Following the model architectures from previous work[32,33], both donor and acceptor models consisted of two consecutive convolution layers and one sigmoid output layer shown in **Supplementary Figure 4a**.

4.7.2 The performance of the CNN model on identifications of donor sites and acceptor sites

We separated the training sites by chromosomes, in which chr1, chr3, chr7, and chr9 were selected for testing the performance of the model, while the remaining chromosomes for training based on the configures of training from the previous work[33].

To benchmark the performance of the model, we selected popular splice site predictors, named MaxEnt[25] and SpliceAI[33]. We also evaluated the performance of different choices of window sizes. We plotted the recall and precision rates at varying thresholds and used AUC (area under the precision-recall curve) to evaluate the performances (**Supplementary Figure 4.1b**). We found the donor and acceptor models with window size 200nt perform as well as those with window size 500nt. And both the donor and acceptor models with window size 200nt outperform maxent but performs slightly worse than SpliceAI. Furthermore, we compared the time requirements for both donor and acceptor models and SpliceAI for scoring 10,000 splice sites, they are 100 times faster than SpliceAI (**Supplementary Table 4.1**). Collectively, the donor and acceptor models built upon CNN perform similarly with SpliceAI but run much faster than SpliceAI. Therefore, we used the donor and acceptor models for scoring the strengths of splice sites.

4.8 References

1. Sharp, P.A. Split genes and RNA splicing. *Cell* **77**, 805-15 (1994).
2. Nilsen, T.W. & Graveley, B.R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457-63 (2010).
3. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-5 (2008).
4. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**, E5593-601 (2014).
5. Castle, J.C. *et al.* Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet* **40**, 1416-25 (2008).
6. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet* **102**, 11-26 (2018).
7. Kalsotra, A. & Cooper, T.A. Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet* **12**, 715-29 (2011).
8. Wang, Z. & Burge, C.B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna* **14**, 802-13 (2008).
9. Fu, X.D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**, 689-701 (2014).
10. Qu, Y.J. *et al.* A rare variant (c.863G>T) in exon 7 of SMN1 disrupts mRNA splicing and is responsible for spinal muscular atrophy. *Eur J Hum Genet* **24**, 864-70 (2016).
11. Singh, N.N., Androphy, E.J. & Singh, R.N. An extended inhibitory context causes skipping of exon 7 of SMN2 in spinal muscular atrophy. *Biochem Biophys Res Commun* **315**, 381-8 (2004).
12. Wai, H.A. *et al.* Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med* **22**, 1005-1014 (2020).
13. López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett* **579**, 1900-3 (2005).
14. Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J. & Fairbrother, W.G. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A* **108**, 11093-8 (2011).
15. Padgett, R.A. New connections between splicing and human disease. *Trends Genet* **28**, 147-54 (2012).
16. Singh, R.K. & Cooper, T.A. Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med* **18**, 472-82 (2012).

17. Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D.N. & Sanford, J.R. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res* **21**, 1563-71 (2011).

18. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324-1335 (2014).

19. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).

20. Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun* **8**, 14519 (2017).

21. Zhang, Y. *et al.* Regional Variation of Splicing QTLs in Human Brain. *Am J Hum Genet* **107**, 196-210 (2020).

22. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* **25**, 911-919 (2019).

23. Cheung, R. *et al.* A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Mol Cell* **73**, 183-194.e8 (2019).

24. Pertea, M., Lin, X. & Salzberg, S.L. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res* **29**, 1185-90 (2001).

25. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-94 (2004).

26. Desmet, F.O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* **37**, e67 (2009).

27. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53-9 (2010).

28. Ke, S. *et al.* Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* **21**, 1360-74 (2011).

29. Xiong, H.Y., Barash, Y. & Frey, B.J. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**, 2554-62 (2011).

30. Rosenberg, A.B., Patwardhan, R.P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698-711 (2015).

31. Xiong, H.Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).

32. Cheng, J. *et al.* MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol* **20**, 48 (2019).

33. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

34. Zhang, Z. *et al.* Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat Methods* **16**, 307-310 (2019).

35. Cheng, J., Celik, M.H., Kundaje, A. & Gagneur, J. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biol* **22**, 94 (2021).

36. Battle, A., Brown, C.D., Engelhardt, B.E. & Montgomery, S.B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

37. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).

38. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).

39. Singh, N.N., Androphy, E.J. & Singh, R.N. In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. *RNA* **10**, 1291-305 (2004).

40. Singh, N.N., Singh, R.N. & Androphy, E.J. Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic Acids Res* **35**, 371-89 (2007).

41. Cartegni, L., Hastings, M.L., Calarco, J.A., de Stanchina, E. & Krainer, A.R. Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. *Am J Hum Genet* **78**, 63-77 (2006).

42. Goldstein, D.B. Islands of linkage disequilibrium. *Nat Genet* **29**, 109-11 (2001).

43. Cummings, B.B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452-458 (2020).

44. Schoch, K. *et al.* Alternative transcripts in variant interpretation: the potential for missed diagnoses and misdiagnoses. *Genet Med* **22**, 1269-1275 (2020).

45. Truty, R. *et al.* Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. *Am J Hum Genet* **108**, 696-708 (2021).

46. Tate, J.G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).

47. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).

48. Shihab, H.A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536-43 (2015).

49. Varley, J.M. *et al.* Characterization of germline TP53 splicing mutations and their genetic and functional analysis. *Oncogene* **20**, 2647-54 (2001).

50. Lord, J. & Baralle, D. Splicing in the Diagnosis of Rare Disease: Advances and Challenges. *Front Genet* **12**, 689892 (2021).

51. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).

52. Merico, D. *et al.* ATP7B variant c.1934T > G p.Met645Arg causes Wilson disease by promoting exon 6 skipping. *NPJ Genom Med* **5**, 16 (2020).

53. Tryka, K.A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* **42**, D975-9 (2014).

54. Phillips, J.W. *et al.* Pathway-guided analysis identifies Myc-dependent alternative pre-mRNA splicing in aggressive prostate cancers. *Proc Natl Acad Sci U S A* **117**, 5269-5279 (2020).

55. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).

56. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-7 (2016).

57. Wainberg, M., Alipanahi, B. & Frey, B. Does conservation account for splicing patterns? *BMC Genomics* **17**, 787 (2016).

58. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).

59. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**, D853-D858 (2019).

60. Zhang, X.H. & Chasin, L.A. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18**, 1241-50 (2004).

61. Fairbrother, W.G. *et al.* RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* **32**, W187-90 (2004).

62. Rong, S. *et al.* Mutational bias and the protein code shape the evolution of splicing enhancers. *Nat Commun* **11**, 2845 (2020).

5 M6A DEPOSITION IS REGULATED BY PRMT1-MEDIATED ARGININE METHYLATION OF METTL14 IN ITS DISORDERED C-TERMINAL REGION

5.1 Introduction

N6-methyladenosine (m^6A) is the most abundant internal modification in cellular mRNA[1-3]. This chemical modification has emerged as a key regulator of mRNA metabolic processes, including transport, translation, splicing, and decay[4-6]. In mammals, m^6A is deposited by a methyltransferase complex consisting of METTL3, METTL14, and WTAP[7-9] and is actively removed from transcripts by the demethylases FTO and ALKBH5[10,11]. The biological functions of m^6A are generally carried out by its "reader" proteins, which include the YT521-B homology (YTH) domain containing proteins YTHDF1–3 and YTHDC1–2[12,13]. For

example, YTHDF1 and YTHDF3 recognize m^{6}A-modified mRNAs and promote their translation by recruiting translation initiation factors[14,15]; whereas YTHDF2 promotes the degradation of its target transcripts by recruiting the CCR4-NOT deadenylase complex [16]. m^{6}A homeostasis is crucial for normal development, and its dysregulation has been linked to the pathogenesis of many human diseases, including neurological disorders and cancer[17,18]. During early embryonic development, deposition of m^{6}A provides an "identity" to transcripts encoding pluripotency transcription factors, such as Nanog, and promotes their expedited decay[7,19,20]. In neuronal progenitor cells, METTL14 knockout causes premature differentiation and delayed specification of neuronal subtypes[21,22]. In hematologic malignancies, METTL3 and METTL14 are both highly expressed, leading to increases in m^{6}A and tumor cell proliferation[23-25]. Additionally, reduced m^{6}A has been shown to stabilize the mRNA levels of *NANOG* and *KLF4*, the key pluripotency factors required for the maintenance of breast cancer tumor-initiating cells[26]. Together, these genetic knockout studies have provided valuable information in understanding the role of m^{6}A in mRNA metabolism; however, the molecular mechanisms underlying the regulation of m^{6}A are largely unknown.

Arginine methylation is a critical post-translational modification (PTM) that regulates protein functions in mRNA metabolism[27-29]. The human genome encodes nine protein arginine methyltransferases (PRMTs), which catalyze three types of arginine methylation: monomethylation (MMA), asymmetric dimethylation (ADMA), and symmetric dimethylation (SDMA)[27,28]. Proteomic studies revealed that RNA-binding proteins form the largest group of PRMT substrates[30-32], and motif analysis uncovered a conserved arginine/glycine-rich (RGG/RG) polypeptide sequence as the preferred site for methylation[31-33]. These RGG/RG motifs are often located within the unstructured, intrinsically disordered regions (IDRs) of proteins and can mediate protein–protein and protein–nucleic acid interactions[33-36]. Arginine methylation of the RGG/RG motifs does not neutralize the cationic charge of the arginine residue but removes its potential hydrogen bond donors and imparts hydrophobicity of the protein[37]. Thus, arginine methylation has emerged as an important PTM that regulates the biochemical activity and the biological function of RGG/RG motif-containing proteins.

Here, we identify PRMT1-catalyzed arginine methylation of the RGG/RG motif-containing IDR in the C-terminus of METTL14 as a novel molecular mechanism that controls m^{6}A deposition in mammalian cells. Specifically, arginine methylation of the IDR enhances the interactions of METTL14 with RNA substrates and with RNA polymerase II (RNAPII), which are fundamental for catalyzing m^{6}A deposition *in vitro* and in cells. We performed transcriptome-wide m^{6}A analysis in mouse embryonic stem cells and identified 1,701 arginine methylation-dependent m^{6}A sites located in 1,290 genes that function in various cellular processes. We focused on the DNA interstrand crosslink (ICL) repair pathway, in which the arginine methylation-dependent m^{6}A sites are significantly enriched, and demonstrated that these m^{6}A sites are associated with enhanced translation of DNA repair genes. Consequently, mESCs expressing arginine methylation-deficient METTL14 are hypersensitive to DNA crosslinking agents. Thus, our study reveals arginine methylation of METTL14 as a novel molecular mechanism underlying the regulation of m^{6}A deposition.

5.2 Results

5.2.1 C-terminal IDR of METTL14 is arginine methylated.

RGG/RG motifs, located in the context of IDRs, are often in multiple copies[33]. We found that the C-terminus of METTL14 harbors an array of RGG/RG motifs, ranging from five in flies to ten in humans (**Figure 5.1A, Supplementary Figure 5.8**). Consistent with the low complexity of these motifs, this region of METTL14 is predicted to be highly disordered (**Supplementary Figure 5.14A and B**). To determine the extent to which the C-terminal IDR contributes to the RNA methylation activity of the METTL3/METTL14 complex, we performed *in vitro* RNA methylation assays and found that this region is essential for catalyzing m^{6}A deposition *in vitro* (**Supplementary Figure 5.14C**), consistent with a recent report[38]. As RGG/RG motifs are the preferred methylation substrates for PRMTs[33], we hypothesized that arginine methylation of the C-terminal IDR regulates the function of METTL14 in m^{6}A RNA modification. To test if METTL14 is arginine methylated, we performed *in vitro* methylation assays by incubating recombinant GST-tagged METTL14

with PRMTs (PRMT1–8) in the presence of ^{3}H-labeled *S*-adenosyl methionine (^{3}H-SAM). Among the PRMTs tested, we found that METTL14 can be methylated by PRMT1 and, to a much lesser extent, PRMT6 and PRMT3 (**Figure 5.1B, Supplementary Figure 5.9A**). To determine if arginine methylation occurs at the C-terminal IDR of METTL14, we performed *in vitro* methylation assays with full-length (FL) and C-terminal IDR-truncated mutant (1-400) METTL14. The arginine methylation of METTL14 was completely abolished after deleting the C-terminal IDR (**Figure 5.1C**), suggesting that the RGG/RG motifs are indeed the sites of methylation.

To assess METTL14 arginine methylation *in vivo*, we immunoprecipitated endogenous METTL14 from HEK293 cells and confirmed its methylation using an antibody (ASYM26) that specifically recognizes ADMA, a modification catalyzed by type I PRMTs, including PRMT1, PRMT3, and PRMT6 (**Figure 5.1D**). Next, we treated HEK293 cells with a potent type I PRMT inhibitor MS023 [39] to inhibit cellular ADMA. The level of arginine methylated METTL14 was dramatically reduced upon treatment with the inhibitor (**Figure 5.1E**), further confirming that METTL14 is arginine methylated in cells. Additionally, METTL14 arginine methylation can also be detected in various human cancer cell lines, including those derived from cervical cancer (HeLa), lung cancer (A549 and H1299), and breast cancer (MDA-MB-231 and MCF7), and MS023 treatment reduces METTL14 arginine methylation in all cell lines tested (**Supplementary Figure 5.9B–D**). Importantly, consistent with our *in vitro* methylation results, deleting the C-terminal IDR completely abolished METTL14 methylation *in vivo*, as detected using two different ADMA antibodies (**Figure 5.1F**). Altogether, these results demonstrate that the C-terminal IDR of METTL14 is arginine methylated.

5.2.2 PRMT1 catalyzes METTL14 C-terminal IDR arginine methylation.

To determine which PRMT methylates METTL14 in cells, we examined the interactions of METTL14 with PRMT1, PRMT3, and PRMT6, the PRMTs that execute METTL14 arginine methylation *in vitro* (**Figure 5.1B**). HEK293 cells were transfected with GFP-tagged PRMT1, PRMT3, and PRMT6, and co-immunoprecipitation (co-IP) assays were performed by immunoprecipitating endogenous METTL14 and detecting associated GFP-PRMTs. Our

results indicate that only PRMT1 interacts with METTL14 (**Figure 5.2A**). To further confirm this interaction, we performed reciprocal co-IP assays using endogenously expressed METTL14 and PRMT1. We were able to detect PRMT1 in the METTL14-immunoprecipitated protein complex (**Figure 5.2B**, left panel), and METTL14 in the PRMT1-immunoprecipitated protein complex (**Figure 5.2B**, right panel). To determine if METTL14 interacts with PRMT1 through its C-terminal IDR, we compared the interaction of PRMT1 with full-length (FL) and C-terminal IDR truncation mutant (1–400) METTL14 using co-IP assays. Deleting the C-terminal IDR completely abolished METTL14 interaction with PRMT1 (**Figure 5.2C**). Furthermore, we also performed GST pull-down assays and demonstrated that although full length recombinant GST-METTL14 can pull down PRMT1 from HEK293 total cell lysates, C-terminal IDR-truncated METTL14 cannot (**Figure 5.2D**), further demonstrating that the C-terminal IDR is essential for METTL14 interaction with PRMT1.

To assess if PRMT1 catalyzes METTL14 C-terminal IDR arginine methylation *in vivo*, we assessed the level of METTL14 arginine methylation in HEK293 cells with altered PRMT1 expression. Overexpressing GFP-PRMT1 increased METTL14 arginine methylation (**Figure 5.2E**), and knocking down the expression of PRMT1, but not PRMT3 or PRMT6, using siRNA reduced METTL14 arginine methylation (**Figure 5.2F, Supplementary Figure 5.10A)**, supporting the role of PRMT1 as a major PRMT catalyzing the arginine methylation of the METTL14 C-terminal IDR in cells.

5.2.3 C-terminal IDR arginine methylation enhances METTL14–RNA interaction and METTL3/METTL14 methyltransferase activity.

To identify the arginine methylation site(s), we performed *in vitro* methylation assays after introducing a series of arginine (R)-to-lysine (K) mutations in the IDR, either individually or in various combinations. However, none of the mutations tested significantly reduced METTL14 methylation (**Supplementary Figure 5.10B**), suggesting that multiple arginine residues are methylated but the combinations we selected were insufficient to cover them all. Thus, we immunoprecipitated METTL14 from HEK293 cells and performed mass spectrometry analysis to identify the methylation sites. Although five sites (Arg438,

Arg442, Arg445, Arg450, and Arg456) were found to be both mono- and dimethylated (**Figure 5.3A, Supplementary Figure 5.15A**), the mutation of all five arginine residues (5RK) only modestly reduced the methylation signal (by ~40%) *in vitro* and in cells (**Figure 5.3B and C**), suggesting that either additional arginine residues are methylated or compensatory methylation occurs when the preferred methylation sites are not available. We thus mutated all thirteen arginine residues in the IDR to lysine (13RK), which completely abolished METTL14 arginine methylation (**Figure 5.3B and C, Supplementary Figure 5.10C**), and used this RK mutant as the arginine methylation-deficient mutant METTL14 in the following studies.

RGG/RG motifs can mediate protein–protein and protein–RNA interactions and are recognized as the second most common RNA-binding domains in the human genome [33-36]. Therefore, to investigate the role of METTL14 C-terminal IDR arginine methylation, we tested the hypothesis that arginine methylation of the RGG/RG motifs of METTL14 regulates its interactions with its RNA substrates. Three independent assays were performed using recombinant METTL14 proteins purified from HEK293 cells. First, we carried out the RNA pull-down assays by incubating recombinant arginine methylated (WT), hypomethylated (MS023-treated), and RK mutant METTL14 proteins with biotin-labeled RNA oligonucleotides harboring the consensus sequence motif for m^6A modification, GGACU [8]. The loss of arginine methylation, caused by either inhibitor treatment or by R-to-K mutation, dramatically reduced the interactions of METTL14 with the RNA substrates (**Figure 5.3D**). Next, we performed an Electrophoretic Mobility Shift Assay (EMSA) by incubating the 6-carboxyfluorescein (6-FAM)-labeled RNA probe with increasing amounts of recombinant METTL14 proteins (as described in **Figure 5.3D**). Our results show that arginine methylated METTL14 exhibits much stronger binding to RNA substrates, compared to the hypomethylated and RK mutant METTL14 (**Figure 5.3E**). In addition, METTL14 protein purified from PRMT1 knockdown HEK293 cells also exhibited reduced interactions with RNA substrates, further supporting the role of PRMT1 in catalyzing METTL14 arginine methylation (**Figure 5.15B)**. Finally, to quantitatively compare the RNA-binding affinity of these recombinant METTL14 proteins, we performed fluorescence polarization assays, which measure protein binding-induced changes in the

polarization of light emitted upon excitation of a fluorescence-labeled RNA probe. Both hypomethylated and RK mutant METTL14 exhibited significantly lower RNA-binding affinities than arginine methylated METTL14 (dissociation constant [*Kd*] values of 211.2 and 227.4 nM for hypomethylated and RK mutant METTL14, respectively, compared to 49.14 nM for arginine methylated METTL14; **Figure 5.3F**). Results from these three independent experiments demonstrate that C-terminal IDR arginine methylation enhances the interactions of METTL14 with its RNA substrates.

To assess if altered METTL14–RNA interaction affect the catalytic activity of the RNA methyltransferase complex, we compared the RNA methylation activity of the recombinant METTL14 proteins *in vitro*. The arginine methylated METTL14 (in complex with METTL3) exhibited significantly (~2 fold) higher m^6A methyltransferase activity than the hypomethylated and arginine methylation-deficient enzymes (**Figure 5.3G**). Similarly, METTL14 protein purified from PRMT1 knockdown HEK293 cells also exhibited reduced RNA methylation activity (**Supplementary Figure 5.15C**). Note that the reduced RNA binding affinity (**Figure 5.3D–F**) and RNA methylation activity (**Figure 5.3G**) of RK mutant METTL14 was similar to that of the hypomethylated METTL14, suggesting that the effects observed were not artifacts caused by the R-to-K mutations. Altogether, these data demonstrate that arginine methylation of the C-terminal IDR enhances the activity of the METTL3/METTL14 complex, likely by promoting the interaction of METTL14 with RNA substrates.

5.2.4 C-terminal IDR arginine methylation enhances the METTL14–RNAPII interaction.

We next investigate the role of C-terminal IDR arginine methylation on the function of METTL14 in cells. To do so, we examined the impact of the loss of C-terminal IDR arginine methylation on the subcellular localization of METTL14 and protein–protein interactions with its known partners. Immunofluorescence assays showed that neither the removal of the C-terminal IDR nor the mutation of the arginine methylation sites affected the nuclear localization of METTL14 (**Supplementary Figure 5.16A**). Furthermore, consistent with previous crystal structure studies showing that the C-terminus of METTL14 is not involved

in its interaction with METTL3 [40-42], we observed that the interactions of METTL14 with other components of the m^{6}A methyltransferase complex, including METTL3 and WTAP, were also unaffected by C-terminal IDR truncation or RK mutation, as revealed by the GST pull-down and co-IP assays (**Figure 5.16B-D**).

Because m^{6}A deposition is co-transcriptional and the RNA methyltransferase complex has been shown to associate with RNAPII [43-45], we tested if arginine methylation regulates the interaction of METTL14 with RNAPII. First, we performed reciprocal co-IP assays of endogenous METTL14 and RNAPII and detected RNAPII in the METTL14-immunoprecipitated protein complex (**Figure 5.4A**, left panel), and METTL14 in the RNAPII-immunoprecipitated protein complex (**Figure 5.4A**, right panel), consistent with a recent report [43]. Next, to determine if C-terminal IDR arginine methylation contributes to this interaction, we transfected HA-tagged WT and arginine methylation-deficient (RK) mutant METTL14 and compared their interactions with RNAPII. Although the loss of C-terminal IDR methylation does not affect the interaction of METTL14 with METTL3, its interaction with RNAPII was dramatically impaired (**Figure 5.4B**). Note that the deletion of the C-terminal IDR also reduced the METTL14–RNAPII interaction (**Supplementary Figure 5.16E**), indicating that this region contributes to their interaction. Furthermore, we treated HEK293 cells with MS023 to inhibit METTL14 arginine methylation and observed a similarly reduced interaction between METTL14 and RNAPII (**Figure 5.4C**), suggesting that the METTL14–RNAPII interaction is regulated by C-terminal IDR arginine methylation. Because the C-terminal IDR arginine methylation enhances METTL14–RNA interaction (**Figure 5.3D–F, Supplementary Figure 5.15B**), we next tested if RNA is involved in the METTL14–RNAPII interaction. We performed the co-IP assays in the presence of RNase A, which led to a significantly reduced amount of METTL14-associated RNAPII (**Figure 5.4D**). These results suggest that METTL14–RNA interaction contributes, at least in part, to the METTL14–RNAPII interaction. Collectively, these data show that arginine methylation of the C-terminal IDR is critical for regulating the association of METTL14 with RNAPII.

5.2.5 METTL14 arginine methylation regulates m^{6}A deposition *in vivo*.

To investigate the cellular function of METTL14 arginine methylation, we used mouse embryonic stem cells (mESCs), a model system in which we confirmed the arginine methylation of METTL14 (**Supplementary Figure 5.11A**). We established three isogenic mESC lines by stably transfecting *Mettl14* knockout mESCs [20] with human WT METTL14 (KO+WT), 5RK mutant METTL14 (KO+5RK), and 13RK mutant METTL14 (KO+RK) (**Figure 5.5A, Supplementary Figure 5.17A, Supplementary Figure 5.11B**). Polyadenylated mRNA was purified from these cells and analyzed using liquid chromatography–tandem mass spectrometry (LC-MS/MS). The m^{6}A to A ratio (m^{6}A/A) was reduced by ~40% in KO+RK mESCs compared to KO+WT mESCs (**Figure 5.5B**). Importantly, the reduction in m^{6}A correlated with the degree of METTL14 methylation loss, as KO+5RK mESCs exhibited a significant but relatively modest (~10%) reduction (**Supplementary Figure 5.17B**). Similar to what has previously been reported for the *Mettl14* KO mESCs [7,20], the KO+RK mESCs exhibited reduced pluripotency and proliferation compared to WT and KO+WT mESCs (**Figure 5.5C–E**).

To identify transcriptome-wide m^{6}A sites that are regulated by METTL14 arginine methylation, we performed methylated RNA-IP (RIP) followed by high-throughput sequencing (MeRIP-seq or m^{6}A-seq) in WT, *Mettl14* KO, KO+WT, KO+5RK, and KO+RK mESCs (**Supplementary Figure 5.17A**). Crosslinking IP (CLIP)-seq analysis of multi-mapped reads (CLAM) [46] was used to identify m^{6}A peaks using different q-value cut-offs (0.05, 0.01, and 0.005). Consistent with our LC-MS/MS-based m^{6}A/A quantification (**Figure 5.5B, Supplementary Figure 5.17B**), the number of m^{6}A peaks identified by CLAM using all three cut-offs positively correlated with the degree of METTL14 arginine methylation (**Supplementary Figure 5.17C**). Principal component analysis (PCA) of the m^{6}A peaks in each cell line demonstrated strong reproducibility (low variation) among three biological replicates (**Supplementary Figure 5.17D**). We choose 0.05 as the q-value cut-off for further analysis because the number of m^{6}A peaks identified in WT mESCs using that cut-off (11,338) was similar to the number reported for mESCs in other studies [7,20,47]. *De novo* motif analysis identified the RRACU m^{6}A sequence motif as enriched at m^{6}A sites, and distribution analysis revealed that the m^{6}A sites in protein-coding genes are enriched near

the stop codon and at the beginning of the 3' UTR (**Figure 5.5F, Supplementary Figure 5.17E**), both as previously described [7,20,47]. The comparison of m^{6}A peaks in KO+WT and KO+RK mESCs revealed a significant decrease in m^{6}A peak density upon the loss of METTL14 arginine methylation (**Figure 5.5G**). We next performed differential m^{6}A analysis and identified 1,701 METTL14 arginine methylation-dependent m^{6}A sites in 1,290 genes and 10,635 methylation-independent m^{6}A sites in 5,764 genes between KO+WT and KO+RK mESCs (**Supplementary Figure 5.17F**). Although the majority of METTL14 arginine methylation dependent m^{6}A sites are found near the stop codon and 3' UTR, some of them are located in the internal exons (**Figure 5.5H**). We found that internal exons harboring these methylation-dependent m^{6}A sites are significantly longer than the exons harboring methylation-independent m^{6}A sites (**Supplementary Figure 5.17G**), indicating that m^{6}A deposition in long exons is more dependent on METTL14 methylation than m^{6}A deposition in short exons, consistent with the preference of m^{6}A enrichment in long internal exons [19,20,48]. Additionally, RNA sequences in the vicinity of these m^{6}A sites are predicted to be more likely to form secondary structures, such as Helix/Stems or multi-branched loops (**Supplementary Figure 5.17H**). Although the deposition of m^{6}A has been linked to gene expression and mRNA stability, a comparison of the expression of genes with vs. without methylation-dependent m^{6}A sites did not reveal any significant differences (**Supplementary Figure 5.17I**).

To determine the biological function of arginine methylation-dependent m^{6}A sites, we performed Gene Ontology (GO) analysis. Several known m^{6}A-regulated cellular processes are significantly enriched, such as stem cell population maintenance and regulation of the TGF-□ signaling pathway (**Figure 5.5I**), which might explain the changes in the morphology and pluripotency of KO+RK mESCs (**Figure 5.5C–E**). Unexpectedly, GO analysis also revealed a strong enrichment of METTL14 arginine methylation-dependent genes in DNA repair pathways, particularly in error-prone translesion synthesis and the Fanconi anemia pathway (**Figure 5.5I, Supplementary Figure 5.17J**). Because the role of m^{6}A-mediated RNA metabolism in DNA repair is largely unknown, we aimed to investigate how METTL14 arginine methylation-dependent m^{6}A sites function in regulating DNA repair.

5.2.6 METTL14 arginine methylation-dependent m^{6}A sites are associated with enhanced translation of DNA repair genes.

As demonstrated in the UCSC Genome Browser custom tracks (**Figure 5.6A, Supplementary Figure 5.11C**), there was a significant reduction in m^{6}A signals in the transcripts of error-prone translesion synthesis and Fanconi anemia genes, including *Atrip*, *Palb2*, *Fancm*, *Blm*, *Brca1*, and *Brca2* due to *Mettl14* KO or expression of arginine methylation-deficient (KO+RK) METTL14. Interestingly, most of these METTL14 arginine methylation-dependent m^{6}A sites are located in long internal exons, consistent with the transcriptome-wide analysis showing that internal exons harboring arginine methylation-dependent m^{6}A sites are significantly longer than exons harboring arginine methylation-independent m^{6}A sites (**Supplementary Figure 5.17G**). These results were confirmed by m^{6}A-IP, followed by quantitative reverse transcription PCR (RT-qPCR) (**Figure 5.6B, Supplementary Figure 5.11D**). Furthermore, consistent with our *in vitro* data showing an important role of arginine methylation in promoting METTL14–RNA interaction (**Figure 5.3D–F**), METTL14 RIP-qPCR revealed that arginine methylation deficiency dramatically reduced the interactions of METTL14 with m^{6}A-positive regions of the target transcripts *in vivo* (**Figure 5.6C, Supplementary Figure 5.11E**). Knocking out *Prmt1* [49] or treating mESCs with the type I PRMT inhibitor MS023, which inhibits METTL14 arginine methylation (**Fig 1E, Fig 2F, Appendix Fig S2B–D**), also reduced m^{6}A deposition (**Fig EV5A**) and the interactions of METTL14 with target transcripts (**Fig EV5B**), further supporting the role of PRMT1-catalyzed METTL14 arginine methylation in regulating m^{6}A deposition.

The deposition of m^{6}A is tightly linked to gene expression [12,13]. Therefore, we next investigated how changes in m^{6}A deposition due to loss of METTL14 arginine methylation affect the expression of DNA repair genes. Western blot analysis revealed that the expression of *Atrip*, *Palb2*, and *Fancm* was reduced upon *Mettl14* KO or expression of arginine methylation-deficient (KO+RK) METTL14 (**Figure 5.6D**). A similar reduction in protein expression was also detected in *Prmt1* KO and MS023-treated mESCs (**Supplementary Figure 5.18C and D**). This reduction in protein expression was not due to reduced mRNA production (**Supplementary Figure 5.12A and B**) or increased mRNA

degradation (**Supplementary Figure 5.12C**). Instead, polysome profiling analysis revealed a significant reduction in the association of these DNA repair transcripts with polyribosomes in KO+RK mESCs compared to KO+WT mESCs, suggesting a reduction in protein translation upon the loss of METTL14 arginine methylation (**Figure 5.6E and F**). Together, these results reveal that METTL14 arginine methylation is important for the efficient translation of DNA repair genes, likely through an m^{6}A-dependent mechanism.

5.2.7 Loss of METTL14 arginine methylation sensitizes mESCs to DNA damage.

The expression of genes involved in error-prone translesion synthesis and the Fanconi anemia pathway is essential for the repair of DNA interstrand crosslinks (ICLs), which form when both strands of DNA are covalently linked [50]. ICLs prevent DNA strand separation, blocking DNA replication and transcription, and thus exerting potent biological effects. We next examined if METTL14 arginine methylation loss impairs cellular responses to ICLs. Consistent with our observation that *Mettl14* KO and KO+RK mESCs exhibited reduced expression of ICL repair genes (**Figure 5.6D**), we found that these mESCs were significantly more sensitive than WT and KO+WT mESCs to treatment with mitomycin C (MMC) and cisplatin, two chemotherapeutic agents that kill cancer cells by inducing ICLs (**Figure 5.7A and B**). Similar sensitivity was also observed in *Prmt1* KO and MS023-treated mESCs (**Supplementary Figure 5.18E and F**). However, these cells were not differentially sensitive to ionizing radiation, which causes DNA double-strand breaks (**Supplementary Figure 5.12D**). Importantly, increasing the expression of Fanconi anemia pathway genes, such as *Palb2*, by transient transfection can partially reduce the sensitivity of KO+RK mESCs to MMC (**Supplementary Figure 5.12E**).

5.3 Discussion

This study identified a unique functional role of arginine methylation in RNA m^{6}A modification and gene expression through the regulation of IDR-mediated protein–RNA and protein–protein interactions (**Figure 5.7C**), expanding our current knowledge about the role of arginine methylation in RNA metabolism. Considering the widespread impact of

m^{6}A-mediated regulation across the human genome, the discovery of m^{6}A regulation by METTL14 arginine methylation has broad implications in normal development and human diseases.

5.3.1 Arginine methylation as a regulator of RGG/RG motif-containing IDRs

Although most protein domains must adopt a well-defined structure to function, a large fraction of the proteome consists of IDRs that do not form defined three-dimensional structures yet exhibit biological activity [51,52]. Specifically, IDR-mediated liquid-liquid phase separation (LLPS) has emerged as a fundamental biophysical process governing the organization of high-order chromatin architecture [53-55], transcription [56-58], and DNA repair [59,60], as well as many other membraneless organelles, such as stress granules and P-bodies [61,62]. IDRs exhibit a marked bias in their amino acid composition, including a high proportion of charged residues, such as arginine and lysine, and are predicted to be enriched for methylation [51]. Arginine can mediate multivalent interactions with nucleotides and proteins through hydrogen bonding and π-stacking. The RGG/RG motif-containing IDR of the METTL14 C-terminus is conserved from flies to humans (**Supplementary Figure 5.8**) and is crucial for METTL3/METTL14 RNA methyltransferase activity by contributing to RNA substrate binding [38]. Although it is yet to be determined if the C-terminal IDR can promote METTL14 LLPS, our study reveals, for the first time, that PRMT1 can catalyze the arginine methylation of this IDR and regulate METTL14 protein function and m^{6}A deposition.

Arginine methylation imparts bulkiness and hydrophobicity of a protein and can either positively or negatively affect protein–RNA and protein–protein interactions. We show that arginine methylation of the C-terminal IDR of METTL14 enhances its interactions with RNA substrates and RNAPII (**Figure 5.3 and Figure 5.4**). Although our data suggest that RNA is involved in mediating the METTL14–RNAPII interaction, arginine methylation may regulate this interaction through other mechanisms. For example, arginine methylation may promote the interaction with methylarginine “reader” proteins, such as the Tudor-domain containing protein 3 (TDRD3) and the survival motor neuron (SMN) protein, both of which have been reported to interact with RNAPII [63-65].

Alternatively, this modification could enhance the interaction of the RGG/RG motif with RNAPII, as a recent report demonstrated that hnRNPG can directly bind to the phosphorylated carboxy-terminal domain (CTD) of RNAPII through its RGG/RG motifs [66]. While beyond the scope of this study, testing these two hypotheses will provide additional insights into the molecular mechanisms by which arginine methylation regulates METTL14 function.

5.3.2 METTL14 arginine methylation and co-transcriptional m^{6}A deposition

m^{6}A has been identified in chromatin-associated pre-mRNA [45,67], suggesting that its deposition is co-transcriptional. However, it is still unclear how transcription machinery modulates the activity and specificity of the METTL3/METTL14 methyltransferase complex to control m^{6}A deposition. The interaction of the METTL3/METTL14 complex with RNAPII, as shown in this study (**Figure 5.4**) and reported by others [43,44], provides a molecular basis for this co-transcriptional RNA modification. Surprisingly, this interaction is dramatically reduced upon loss of METTL14 methylation (**Figure 5.4B and C**), indicating that arginine methylation of METTL14 could be an important molecular mechanism regulating co-transcriptional m^{6}A deposition. Indeed, mESCs expressing arginine methylation-deficient METTL14 exhibited a significant reduction in global m^{6}A levels (~40%, **Figure 5.5B**), particularly near the stop codon and at the beginning of the 3' UTR of protein-coding genes (**Figure 5.5H**). Although the loss of METTL14 arginine methylation reduces its interaction with all forms of RNAPII (**Figure 5.4B**), it is possible that elongating RNAPII (S2-p) prefers to interact with hypermethylated METTL14 for m^{6}A deposition in the coding sequence (CDS) and the 3' UTR, because PRMT1, the enzyme that catalyzes METTL14 arginine methylation, has been found in the RNAPII elongation complex through interacting with the transcription elongating factor SPT5 [68]. Furthermore, histone H3 trimethylation at lysine 36 (H3K36me3), a histone mark that is tightly associated with transcription elongation, was recently shown to guide co-transcriptional m^{6}A deposition [43]. H3K36me3 recruits METTL14 through a direct interaction, thus enriches the METTL3/METTL14 methyltransferase complex at this histone mark. Although the H3K36me3-interacting region of METTL14 was mapped to its N-terminal α-helical motif [43], it remains possible that

arginine methylation of the C-terminal RGG/RG motif-containing IDR could enhance METTL14–H3K36me3 engagement through processes such as LLPS *in vivo*. Alternatively, H3K36me3 could directly or indirectly promote PRMT1-catalyzed METTL14 arginine methylation, thus enabling the enrichment of hypermethylated METTL14 in the vicinity of this elongation-associated histone mark for enhanced m^{6}A deposition.

5.3.3 m^{6}A RNA methylation in the regulation of DNA repair

Loss of METTL14 arginine methylation leads to ~40% m^{6}A reduction on cellular mRNAs (**Figure 5.5B**), an effect likely caused by overall reduced METTL14/METTL3 methyltransferase activity (**Figure 5.3G**) and/or uncoupling of co-transcriptional m^{6}A deposition (**Figure 5.4B**). Our GO analysis of METTL14 arginine methylation-dependent m^{6}A sites not only identified known m^{6}A-regulated cellular processes, such as stem cell population maintenance, but also revealed a previously underappreciated role of m^{6}A in regulating DNA repair gene expression (**Figure 5.5I**). Interestingly, these m^{6}A sites are mainly located in the long internal exons of DNA repair genes (**Figure 5.6A and Supplementary Figure 5.11C**). Using polysome profiling analysis, we demonstrate that the METTL14 arginine methylation-dependent m^{6}A modification of these transcripts is essential for promoting their efficient protein translation (**Figure 5.6D–F**). It was recently reported that m^{6}A in mRNA coding regions can promote translation by recruiting m^{6}A reader YTHDC2 [69]. Therefore, it is possible that YTHDC2 is involved in the translation of these DNA repair genes.

Although our study uncovered a novel function of m^{6}A in promoting DNA repair gene expression, a recent study by Xiang and colleagues reported the rapid, reversible accumulation of m^{6}A RNA at the sites of UV irradiation, which recruits DNA polymerase κ (POLK) as an early response for DNA repair [70]. Interestingly, we found that the POLK transcript is also decorated with m^{6}A, and loss of METTL14 arginine methylation caused a ~50% reduction in its m^{6}A levels, indicating that POLK expression may also be subjected to arginine methylation-dependent m^{6}A regulation. Recently, METTL3 was reported to be recruited to DNA damage sites through ATM-mediated phosphorylation at S43, which enhances m^{6}A deposition on DNA damage-associated RNAs to facilitate DNA repair [71].

Together, these studies highlight a crucial function of the m^{6}A RNA modification in the regulation of DNA repair through the direct recruitment of DNA repair machinery as an early response to DNA damage and the enhancement of DNA repair gene expression as a sustained, long-term response.

Consistent with these findings, genetic knockout and inhibition of PRMT1, which dampens METTL14 arginine methylation, also sensitized mESCs to MMC- and cisplatin-induced cell death (**Supplementary Figure 5.18E-F**). Of particular relevance to this observation, Musiani and colleagues reported that, in response to cisplatin treatment, PRMT1 is recruited to chromatin to activate the transcription of genes involved in the senescence-associated secretory phenotype by methylating histone H4 [72]. This finding suggests that PRMT1 functions through multiple pathways to promote cell survival in response to DNA damage. Recently, PRMTs have emerged as promising therapeutic targets for treating human malignancies, including solid tumors and blood cancers [28,73]. Our work reveals that deficiencies in the repair of ICLs could be a specific vulnerability of PRMT inhibitor-treated cells, suggesting that PRMT inhibition may be a promising strategy to sensitize cancer cells to existing chemotherapy drugs.

5.4 Methods

5.4.1 Plasmids and antibodies

Flag-METTL3 (#53739), Flag-METTL14 (#53740), Flag-RNA Pol II (#35175), pMD2.G (#12259), and pSPAX2 (#12260) were purchased from Addgene. GST-tagged PRMT1, PRMT2, PRMT3, CARM1, PRMT6, PRMT7, and PRMT8, as well as Myc-PRMT5, plasmids were used to purify recombinant enzymes and have been described before [74]. GFP-tagged PRMT1, PRMT3, and PRMT6 were used for mammalian expression and have been described before [75]. Human METTL14 cDNA was cloned into pGEX-6P-1, pCMV-HA (Clontech), p3xFlag-CMV-7.1 (Sigma), and pLV-EF1a-IRES-Blast (Addgene, #85133) vectors. All R-to-K mutants of METTL14 were generated using a site-directed mutagenesis kit (Agilent Technologies). The sequences of all primers used in this study are listed in **Supplementary Table 5.1**.

The following antibodies were used for either IP or Western blot analysis: anti-METTL14 (HPA038002, Sigma), anti-METTL3 (A301-567A, Bethyl), anti-PRMT1 (A300-722A, Bethyl), anti-PRMT6 (IMG-506, IMGENEX), anti-Atrip (A7139, ABClonal), anti-Fancm (12954-1-AP, Proteintech), anti-Palb2 (14340-1-AP, Proteintech), anti-Flag (F3165, Sigma), rabbit anti-GFP (A6455), mouse anti-GFP (sc9996, Santa Cruz Biotechnology), anti-β-ACTIN (A5441, Sigma), anti-RNAPII (39097, Active motif), anti-RNAPII S2p (91115, Active motif), anti-RNAPII S5p (sc-47701, Santa Cruz Biotechnology), mouse anti-HA (901501, Biolegend), rabbit anti-HA (3724S, Cell Signaling Technology), and anti-ADMA (13522S, Cell Signaling Technology). The ASYM26 antibody was kindly provided by Dr. Stéphane Richard (McGill University). The PRMT3 antibody was kindly provided by Dr. Mark T. Bedford (MD Anderson Cancer Center).

5.4.2 *In vitro* methylation assays

For the *in vitro* protein methylation assay, the reactions were carried out in 30 µl of phosphate-buffered saline (PBS; pH 7.4) containing 0.5–1.0 µg substrate, 3 µg recombinant enzymes, and 0.42 µM ^{3}H-SAM (79 Ci/mmol from 7.5 µM stock solution; PerkinElmer Life Sciences). Each reaction was incubated at 30°C for 1 h, separated by SDS-PAGE, transferred to a PVDF membrane, and exposed to film for 1 day at -80°C. After exposure, the membrane was washed with methanol and stained with Ponceau S to visualize total protein loaded.

For the *in vitro* RNA methylation assay, reactions were carried out in a 96-well Streptavidin FlashPlate (#SM9103001PK, PerkinElmer). In each well, the 20-µl reaction mixture contained 200 nM biotin-labeled RNA oligonucleotides (5'UACACUCGAUCU**GGACU**AAAGCUGCUC3'), 20 mM Tris (pH 7.5), 0.01% Triton X-100, 1 mM DTT, 0.2 U/mL RNasin, 1% glycerol, 420 nM ^{3}H-SAM, and the indicated amounts of recombinant Flag-METTL3 and Flag-METTL14. Each *in vitro* methylation reaction was incubated at room temperature for 2 h. Enzymatic activity was measured in counts per min using a scintillation counter (PerkinElmer).

5.4.3 Immunoprecipitation of arginine methylated proteins

To detected arginine methylated proteins, cells were either left untreated or treated with the methylation inhibitors AdOx (20 μM) or MS023 (10 μM) for 2 days. Cell pellets were lysed in 1x RIPA buffer (20 mM Tris-HCl, [pH 7.5], 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, and protease inhibitor) for 1 h at 4°C. The lysates were sonicated on ice and clarified by centrifugation, followed by pre-clearing with protein G agarose. The lysates were subsequently immunoprecipitated with specific antibodies, as indicated. Immunoprecipitated proteins were analyzed by Western blot using arginine methylation-specific antibodies.

5.4.4 Recombinant protein purification

GST-tagged proteins were purified from Escherichia coli strain BL21(DE3). A single colony of indicated plasmids was picked and cultured in 10ml LB Broth with 100 μg/ml ampicillin overnight. 40 ml fresh LB Broth with 100 μg/ml ampicillin was added the next day. The protein expression was induced with 1 mM IPTG at 30°C for 4 h. The cells were sonicated in PBS on ice and clarified by centrifugation. The lysates were subsequently incubated with Glutathione Sepharose 4B resin (GE Healthcare Life Sciences) overnight at 4°C. The GST-tagged proteins were eluted with 10 mg/ml reduced L-Glutathione in elution buffer (100 mM Tris-HCl, pH 7.4, with 150 mM NaCl) after washing three times with PBS buffer.

For the purification of Flag-tagged recombinant proteins, HEK293 cells were transfected with indicated plasmids for 48 h and lysed in Co-IP buffer (20 mM Tris-HCl, [pH 7.5], 150 mM NaCl, 1% NP-40, and protease inhibitor) at 4°C for 1 h. The lysates were briefly sonicated on ice and clarified by centrifugation. The lysates were subsequently incubated anti-Flag M2 magnetic beads overnight at 4°C. The Flag-tagged proteins were eluted with 200 μg/ml 3xFlag peptide in TBS buffer (50 mM Tris-HCl, pH 7.4, with 150 mM NaCl) after washing three times with Co-IP buffer.

5.4.5 GST pull-down

All GST-tagged proteins used in this study were purified from *Escherichia coli* strain BL21(DE3). Cells were lysed in lysis buffer containing 20 mM Tris-HCl (pH 7.4), 150 mM NaCl, 0.1% NP-40, and protease inhibitors, and the cell lysates were incubated with purified GST-tagged recombinant proteins with gentle rocking overnight at 4°C. Glutathione Sepharose beads (GE Healthcare Life Sciences) were added to the protein and lysate mixture and incubated with gentle rocking for 2 h at 4°C. The mixture was centrifuged, the supernatant was discarded, and the beads were washed three times with the cell lysis buffer. After centrifuging again, the pellet was resuspended in 30 μl 2X SDS sample buffer and heated at 95°C for 5 min. The samples were loaded on SDS-PAGE gels and analyzed by Western blot using the indicated antibodies.

5.4.6 Co-IP assay

Cells were lysed in Co-IP buffer (20 mM Tris-HCl [pH 7.4], 150 mM NaCl, 0.1% NP-40, and protease inhibitors). After brief sonication, the lysate was centrifuged at 12,000 rpm for 10 min at 4°C. For each IP, the supernatant was incubated with 2 μg of the indicated antibody with gentle rocking overnight at 4°C. The next day, protein A/G beads (Thermo Scientific) were added to the antibody–cell lysate mixture and incubated with gentle rocking for 2 h at 4°C. The immunocomplex was precipitated by centrifugation and washed three times with the cell lysis buffer. The samples were loaded on SDS-PAGE gels and analyzed by Western blot using the indicated antibodies.

5.4.7 Immunofluorescence

The HeLa cells transfected with the indicated plasmids were grown on glass coverslips to the desired confluence (85%) before fixation. First, the cells were rinsed with PBS and were fixed with ice-cold methanol for 20 min at room temperature. After blocking with 20% newborn calf serum for 1 h, the cells were incubated with the indicated antibodies at 4°C overnight. The cells were then stained with a fluorescence-labeled secondary antibody

and stained with 4',6-diamidino-2-phenylindole (DAPI). The coverslips were then sealed and examined using an Olympus BX50 fluorescence microscope.

5.4.8 Electrophoretic mobility shift assay (EMSA)

The 5' 6-FAM labeled ssRNA oligonucleotide (5'UACACUCGAUCUGGACUAAAGCUGCUC3') was incubated with increasing amounts of indicated proteins at 4°C in 10 ul reaction buffer containing 50 mM Tris (pH 7.9), 250 mM KCl, 50 mM $MgCl_2$, 0.5mM EDTA, and 0.2 U/mL RNasin for 1 h. The reactions were then resolved on 6% native acrylamide gels (37.5:1 acrylamide:bis-acrylamide) in 0.5xTBE buffer. The mobility shift of oligonucleotides was detected using Bio-Rad ChemiDoc Imaging System.

5.4.9 Identification of METTL14 arginine methylation sites by LC-MS/MS

Flag-tagged recombinant METTL14 protein purified from HEK293 cells was resolved on an 8% SDS-PAGE gel and stained with SimplyBlue™ SafeStain (Invitrogen™, cat. no. LC6065). The protein band was excised and de-stained, followed by in-gel digestion using Trypsin/Lys-C Mix (Promega, cat. no. V5073), according to the manufacturer's instructions. After overnight digestion, the peptides were extracted three times by adding 50% ACN/0.1% TFA solution, 60% ACN/0.1% TFA solution, and 80% ACN/0.1% TFA solution to the gel pieces. The combined peptide extracts were evaporated using a Savant SpeedVac SVC 100H Centrifugal Evaporator. The peptides were dissolved in 1% formic acid (Fisher Chemical, cat. no. A11750) and analyzed by reversed-phase LC/MS. The mass spectrometric analysis was carried out using a Thermo Scientific Orbitrap Fusion Mass Spectrometer equipped with an Easy Spray source and an Easy-nLC1000 system. The raw spectra files were searched using both Proteome Discoverer Software with Sequest (Version 2.0) and the Mascot algorithm (Mascot 2.5.1).

5.4.10 Fluorescence polarization assay

Fluorescence polarization assays were performed in black, low-flange, flat-bottom 384-well microplates with a nonbinding surface (Corning, MA). Various amounts of recombinant WT

and mutant METTL14 proteins were incubated with 1 nM of a 5' 6-FAM-labeled RNA probe (5'UACACUCGAUCU**GGACU**AAAGCUGCUC3') in the binding buffer containing 20 mM Tris (pH 7.5), 0.01% Triton X-100, 1 mM DTT, 0.2 U/mL RNasin, and 1% glycerol. Binding was performed for 15 min at 37°C. Fluorescence polarization was measured using a Biotech Synergy H4 plate reader (excitation 485 nm, emission 528 nm), and dissociation constants were calculated using GraphPad Prism 8.0.

5.4.11 Lentivirus packaging and stable mESC line generation

Lentiviruses were made by co-transfecting each pLV-EF1a-IRES-Blast METTL14 overexpression vector (WT and RK) with pSPAX2 and pMD2.G at a 4:3:1 ratio into 293T cells. The supernatant was harvested 48 h after transfection and filtered through a 0.45 μm filter. The virus was concentrated using PEG-it Virus Precipitation Solution (#LV810A-1, System Biosciences). *Mettl14* KO mESCs, kindly provided by Dr. Jacob H. Hanna (Weizmann Institute of Science, Israel), were seeded in 6-well plates and infected with the packaged lentiviruses in the presence of 5 μg/ml polybrene (Sigma). 24 h after infection, the mESCs were treated with 5 μg/ml blasticidin for one week to select those expressing WT or RK mutant METTL14 (KO+WT and KO+RK, respectively). mESCs were cultured under feeder-free conditions supplemented with mouse leukemia inhibitory factor (GeminiBio).

5.4.12 RNA m^6A quantification by LC-MS/MS

Total RNA was isolated from the indicated cell lines using TRIzol reagent (Invitrogen). The polyadenylated RNA from these cells was isolated using two rounds of purification on oligo $d(T)_{25}$ magnetic beads (Thermo Fisher). 25 ng of poly(A)+ RNA was digested using nuclease P1 (1 U, Sigma) in 20 μl of buffer containing 20 mM NH_4OAc, (pH 5.5) at 42°C for 2 h, followed by the addition of FastAP buffer (2.3 μL) and alkaline phosphatase (1 U, Thermo Fisher) and incubation at 37°C for 4 h. The sample was then filtered (0.22 μm pore size, 4 mm diameter, Millipore), and 5 μl of the solution was injected into a SCIEX Triple Quad 6500+ LC-MS/MS system. The nucleosides were separated by reverse-phase ultra-performance liquid chromatography on a C18 column (Agilent) with online mass spectrometry detection performed in positive electrospray ionization mode. The

nucleosides were quantified using the nucleoside-to-base ion mass transitions of 282 to 150 (m^6A) and 268 to 136 (A). Nucleoside concentrations were determined by comparison to a standard curve obtained from pure nucleoside standards run with the same batch of samples. The m^6A/A ratio was calculated based on the calibrated concentrations.

5.4.13 Colony formation and alkaline phosphatase staining assay

The mESCs were seeded at a 500 cells/well concentration in a 6-well plate for 7 days. The cell culture medium was aspirated, and the cells were washed once with 1 ml of 1x PBST (1xPBS containing 0.05% Tween-20). Subsequently, the cells were stained with an Alkaline Phosphatase Staining Kit (Biopioneer), according to the manufacturer's instructions.

5.4.14 Proliferation and viability assay

To assess cell proliferation and viability, cells were cultured in 96-well plates and counted at the indicated times using a CCK-8 Cell Counting Kit-8 (Dojindo), measuring the absorbance on a microplate reader using a 450-nm filter.

5.4.15 Protein sequence alignment using ClustalW

The parameters for the alignment using ClustalW were the following: Gap Penalty: 10, Gap Length Penalty: 0.2, Delay Divergent Sequence: 30%, Protein Weight Matrix: Gonnet Series for multiple alignment parameters. For pairwise alignment: Gap Penalty: 10, Gap Length: 0.1, Protein Weight Matrix: Gonnet 250.

5.4.16 MeRIP-seq (m^6A-seq)

Using a modified m^6A-seq protocol [76], we profiled the genome-wide m^6A methylomes of WT, *Mettl14* KO, KO+WT, and KO+RK mESCs. For each sample, we analyzed three biological replicates using both RIP-seq with an m^6A-specific antibody and standard RNA-seq of the input control. Total RNA was extracted from the mESCs using TRIzol reagent. Enrichment of mRNA from total RNA was performed using a Dynabeads mRNA Purification Kit (Invitrogen), according to the manufacturer's instructions. mRNA samples were chemically

fragmented into ~100-nucleotide-long fragments by incubation with 10X RNA Fragmentation Reagent (Invitrogen) at 90°C for 90 s. The fragmentation reaction was stopped by adding 0.5 M EDTA, followed by standard ethanol precipitation. The fragmented RNA samples were resuspended in 10 mM Tris-HCl (pH 7.4). 20 ng of fragmented RNA to be used as input control was stored in -80°C in a final volume of 8.5 µl in FPF (Fragment, Prime, Finish) mix from the TruSeq Stranded Total RNA Kit (Illumina). The remaining RNA was subjected to m^{6}A-seq. Specifically, 2 µg of fragmented mRNA was incubated for 2 h at 4°C with 4 µg of affinity-purified anti-m^{6}A polyclonal antibody (Synaptic Systems) in m^{6}A-IP buffer (150 mM NaCl, 0.1% NP-40, 10 mM Tris-HCl, [pH 7.4], and 0.2 U/µl of RNasin). The RNA–antibody mixture was immunoprecipitated by incubation with protein A beads (Millipore) at 4°C for another 2 h. The beads were extensively washed with m^{6}A-IP buffer. The bound RNA was eluted with m^{6}A elution buffer (150 mM NaCl, 6.7 mM m^{6}A nucleotides, 0.1% NP-40, 10 mM Tris-HCl, [pH 7.4], and 0.2 U/µl of RNasin) and elution wash buffer (150 mM NaCl, 0.1% NP-40, 10 mM Tris-HCl, [pH 7.4], 0.2 U/µl of RNasin). The m^{6}A eluate was purified using Agencourt AMPure XP beads (Beckman Coulter). The immunopurified RNA and non-IP input control samples were used for library generation using the TruSeq Stranded Total RNA Kit (Illumina). Single-end, 50-nucleotide sequencing was performed on an Illumina HiSeq 4000 platform, according to the manufacturer's instructions.

5.4.17 Gene expression quantification

The RNA-seq reads of the input control RNA from WT, *Mettl14* KO, KO+WT, and KO+RK mESCs were aligned to the mouse genome (GRCm38) with GENCODE annotation (release M13) using Kallisto (version 0.43.0) [77]. Gene expression was reported in TPM (transcripts per million).

5.4.18 m^{6}A peaks and differential m^{6}A peaks calling procedure

We mapped the m^{6}A-seq reads to the mouse genome (GRCm38) with GENCODE annotation (release M13) using STAR (version 2.5.3a) [78]. Only uniquely mapped reads were used for m^{6}A peak calling. Then we ran CLAM (version 1.2.0) with local window size w = 100, p-

value correction using the Bonferroni correction, and m^{6}A peaks were called as significant 100-bp windows. To call METTL14 arginine methylation-dependent (differential) m^{6}A peaks between any two samples, we selected 100-bp windows called as m^{6}A peaks (peak window) in at least one replicate of one sample and compared the average peak window intensity signals for each sample (the ratio of reads per kilobase per million mapped reads [RPKM] in IP over input control). First, we calculated the fold change between samples for each peak window. Then, if the fold change was greater than 1.5 and the average RPKM of the peak window in the input controls was greater than 1 for both samples, t-tests were performed to compare the peak intensities. Peak windows with $p < 0.05$ were considered as differential m^{6}A sites. To call METTL14 arginine methylation-independent (common) m^{6}A peaks between any two samples, we followed a similar procedure as we did for differential m^{6}A sites, expect that we required that $1/1.1 <$ Fold Change < 1.1, and that the p-value ≥ 0.1.

5.4.19 PCA analysis of biological replicates based on m^{6}A peaks

To investigate the reproducibility (variation) of the m^{6}A peaks among biological replicates in WT, KO, KO+WT, and KO+RK mESCs, we performed unsupervised principal component analysis (PCA) of m^{6}A peaks. First, we selected the peaks that were called by CLAM in at least one sample. Then, for each peak, we assigned 1 to the sample if the peaks were present and 0 if not. PCA was conducted via sklearn function in Python. The top two principle components that explained the highest percentage of the variance were chosen to visualize the m^{6}A profiles of the twelve samples.

5.4.20 m^{6}A motif finding, topological distribution, and composition analysis

Motifs were identified in the m^{6}A peak windows for each sample using HOMER [79] with the following parameters: motif_len= 5,6,7; size = 100; motif_num = 10. We performed a genome-wide analysis to determine the topological distribution of m^{6}A in the 5'UTR, coding sequence (CDS), and 3'UTR by splitting each transcript region into 50 bins with equal size. The frequency of m^{6}A peaks in each bin was calculated as the number of m^{6}A peaks per transcript. We then analyzed the composition of m^{6}A peaks by looking at the proportions of

m^{6}A peaks in the 3'UTR, 5'UTR, CDS, other exons, and intron regions. Other exons were defined as the exons that could not be mapped to 3'UTR, 5'UTR, or CDS.

5.4.21 Differential topological distribution analysis of m^{6}A peaks

Transcripts were first binned into 50 parts for the 5'UTR, CDS, and 3'UTR, respectively. To determine which sample of each comparison had more transcripts with m^{6}A, the number of transcripts with or without m^{6}A peaks were counted for each bin. Fisher's exact tests were performed to evaluate statistical significance. A stringent FDR threshold of 0.01 was used to correct for multiple hypothesis testing.

5.4.22 RNA secondary structure analysis

To analyze the likelihood of RNA sequences to form secondary structures, we first obtained 200-nt RNA sequences covering 100 nt of each putative m^{6}A peak (including the RRACU motif), 50 nt upstream, and 50 nt downstream. The RNA secondary structures of differential and common m^{6}A sites between the KO+WT and KO+RK mESCs were analyzed by using RNAfold (https://github.com/ViennaRNA/ViennaRNA) [80], which predicts RNA secondary structures and forgi (https://viennarna.github.io/forgi/graph_tutorial.html), a package developed by the RNAfold group to define the structure results predicted by RNAfold. Four RNA secondary structures (Helix/Stem, Hairpin loop, Bulge loop + Interior loop, and Multi-branched loop) can be identified using the forgi software.

5.4.23 GO analysis of genes with differential m^{6}A peaks

The GO annotation file was downloaded from ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/MOUSE/. We analyzed the enrichment of GO terms among genes with differential m^{6}A peaks and background genes (genes with expression levels no less than 1 in both samples) for KO+RK vs. KO+WT. The hypergeometric test was performed to identify the significantly enriched GO terms for each comparison.

5.4.24 Reverse transcription-quantitative PCR (RT-qPCR)

Total cellular RNA was extracted using TRIzol reagent and analyzed for integrity using the Agilent 2100 Bioanalyzer (Agilent Technologies). Total RNA (1 μg) was then used as a template to synthesize cDNA using the High-Capacity cDNA Archive Kit (Applied Biosystems), according to the manufacturer's instructions, and qPCR was subsequently performed on a CFX96 Real-time System C1000 Touch Thermal Cycler (Bio-Rad). RNA levels were normalized to the endogenous control gene *Actb* (ACTIN). Data analysis was performed using the Bio-Rad CFX Manager 3.1. The experimental cycle threshold (Ct) was calibrated against the ACTIN control product. All amplifications were performed in triplicate.

5.4.25 RNA immunoprecipitation (RIP)-qPCR

Cells were crosslinked with 1% formaldehyde for 10 min, and crosslinking was stopped by the addition of glycine to a final concentration of 0.25 M for 5 min. Cells were washed twice with cold PBS and lysed in RIP buffer (50 mM Tris-HCl [pH 7.5], 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 1 mM PMSF, 2 mM VRC, and protease inhibitors) with sonication. The cell lysates were centrifuged, and the supernatant was transferred to a clean tube. 2 μg METTL14 antibody was added to the supernatant and incubated overnight at 4°C with shaking. Dynabeads were added to the supernatant the next day and incubated at 4°C for 4 h with shaking. The beads were washed three times for 5 min with washing buffer I (50 mM Tris-HCl [pH 7.5], 1 M NaCl; 1% NP-40, 1% sodium deoxycholate, and 2 mM VRC) and three times for 5 min with washing buffer II (50 mM Tris-HCl [pH 7.5], 1 M NaCl, 1% NP-40, 1% sodium deoxycholate, 2 mM VRC, and 1 M urea). After washing, the beads were incubated in 100 μL of elution buffer (100 mM Tris-HCl [pH 8.0], 200 mM NaCl, 10mM EDTA, 1% SDS, and 0.2 mg/mL Proteinase K) for 1 h at 42°C, followed by 1 h at 65°C. RNA was then extracted using TRIzol reagent and reverse transcribed into cDNA using the High-Capacity cDNA Reverse Transcription Kit. The primers for RIP-qPCR are listed in **Supplementary Table 5.1**.

5.4.26 mRNA half-life

Cells were treated with 5 μg/ml actinomycin D at different time points (0 h, 3 h, 6 h, and 9 h) before harvest. RNA was purified using TRIzol reagent and reverse transcribed into cDNA using the High-Capacity cDNA Reverse Transcription Kit. The primers for the mRNA half-life assay are listed in **Supplementary Table 5.1**.

5.4.27 Polysome profiling

Cells were pre-treated with 100 μg/ml cycloheximide for 5 min at 37°C, followed by washing using ice-cold PBS containing 100 μg/ml cycloheximide. Cells were pelleted, lysed on ice in lysis buffer, then centrifuged. The supernatant was collected and loaded onto a 10/50% (w/v) sucrose gradient, followed by centrifugation at 39,000 rpm in an SW40 rotor (Beckman) for 3 h at 4°C. Sucrose solutions were freshly prepared in cell lysis buffer (20 mM HEPES [pH 7.6], 100 mM KCl, 5 mM $MgCl_2$, 1% Triton X-100, and 100 μg/ml cycloheximide, supplemented with protease inhibitor and RNase inhibitor). Separated samples were fractionated, and OD254 values were measured. An aliquot of the ribosome fraction was used to extract total RNA using TRIzol reagent for real-time PCR analysis.

5.4.28 Statistical analysis

All experiments were performed at least three times. Statistical comparisons were performed using Student's t-tests. $p < 0.05$ was considered statistically significant.

5.4.29 Data availability

The MeRIP-seq (m6A-seq) datasets produced in this study are available in the following database:

- MeRIP-seq (m6A-seq) data: Gene Expression Omnibus GSE160108 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160108

Acknowledgements

We thank Dr. Yi Xing (Children's Hospital of Philadelphia) for discussion on the bioinformatics analysis. We thank Ross Tomaino (Taplin Mass Spectrometry Facility) for mass spectrometry analysis of METTL14 arginine methylation sites. We thank Dr. Jacob Hanna (Weizmann Institute of Science) for sharing WT and *Mettl14* KO mESCs. We thank Dr. Stéphane Richard (McGill University) for sharing *Prmt1* KO mESCs. B.T. Harada is supported by an NIH fellowship (CA221007); C. He is supported by an NIH grant (HG008935); Y.Z. Yang is supported by an NIH grant (GM133850); and L. Lin is supported by an NIH grant (GM121827). C. He. is an investigator of the Howard Hughes Medical Institute. We thank Dr. Kerin Higa for providing helpful comments on scientific writing.

5.5 Figures

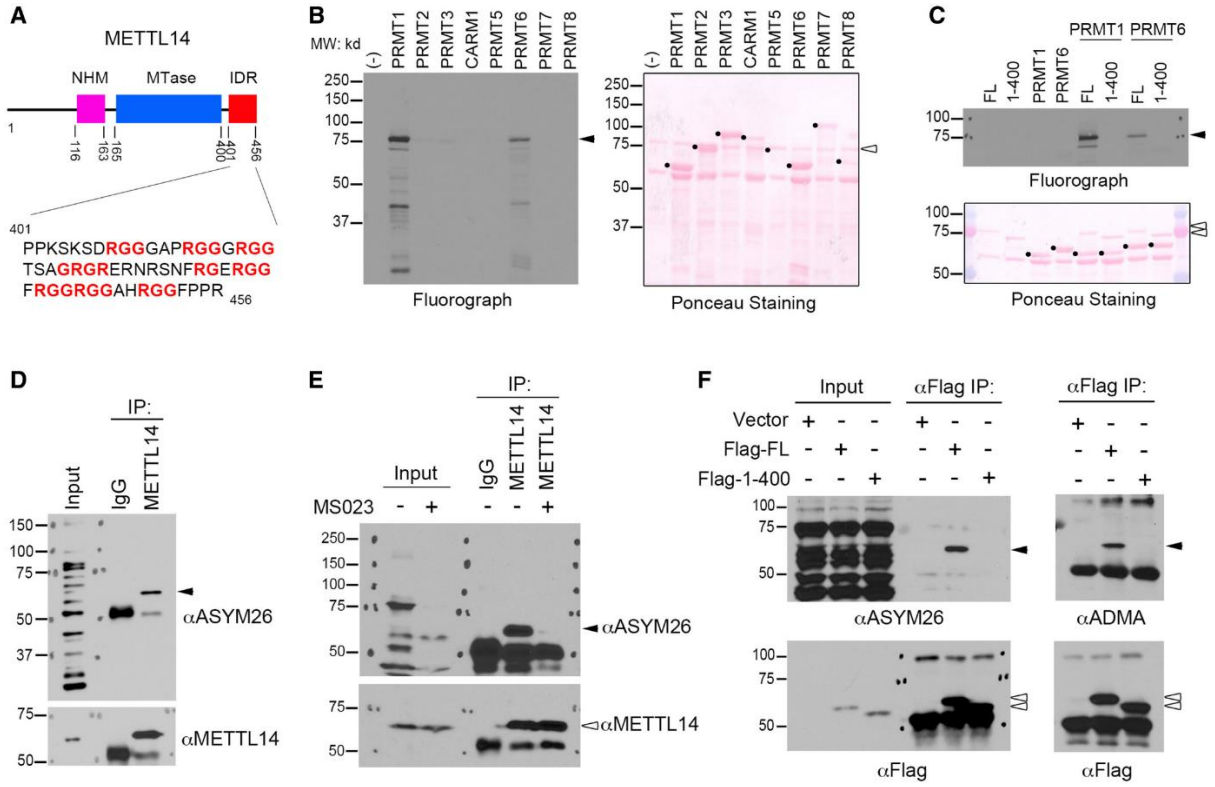


Figure 5.1 METTL14 C-terminal IDR is arginine methylated in vitro and in cells.

(A) Schematic representation of the domain structure of METTL14. The C-terminal IDR, containing an array of RGG motifs, is highlighted. NHM: N-terminal α-helical motif; MTase: Methyltransferase domain; IDR: Intrinsically disordered region. **(B)** METTL14 is arginine methylated *in vitro*. *In vitro* methylation assays were performed by incubating recombinant PRMTs (1–8) with purified GST-tagged METTL14. **(C)** METTL14 is arginine methylated at its C-terminal IDR *in vitro*. GST-tagged full-length (FL) and C-terminal IDR-truncated (1–400) METTL14 were incubated with recombinant PRMT1 and PRMT6. **(D)** METTL14 is arginine methylated in cells. Endogenous METTL14 was immunoprecipitated from HEK293 cells under denaturing conditions and detected using the ADMA antibody ASYM26. **(E)**

Inhibiting type I PRMT activity reduces METTL14 arginine methylation. HEK293 cells were treated with the type I PRMT inhibitor MS023 (1 μM, 48 h). METTL14 was immunoprecipitated from the cells and detected by Western blot analysis using anti-METTL14 and anti-ASYM26 antibodies. **(F)** METTL14 is arginine methylated at its C-terminal IDR in cells. HEK293 cells expressing Flag-tagged FL or C-terminal IDR-truncated (1–400) METTL14 were lysed and immunoprecipitated with an anti-Flag antibody. Arginine methylation of immunoprecipitated METTL14 was analyzed by Western blot using two different antibodies that recognize ADMA.

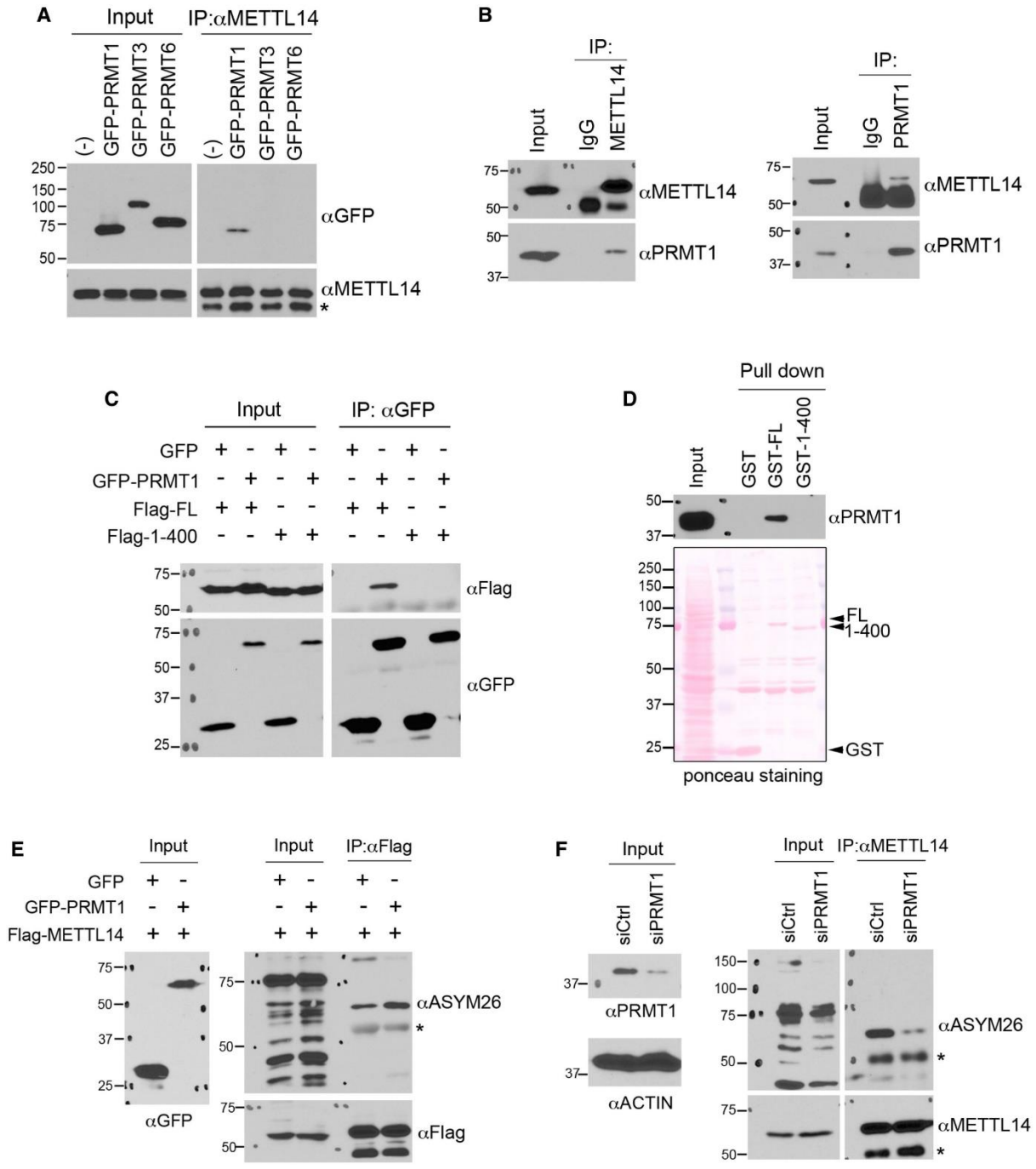


Figure 5.2 PRMT1 catalyzes METTL14 C-terminal IDR arginine methylation.

(A) PRMT1 interacts with METTL14 in cells. HEK293 cells expressing GFP-tagged PRMT1, PRMT3, or PRMT6 were lysed and immunoprecipitated with an anti-METTL14 antibody, followed by Western blot analysis using an anti-GFP antibody. * indicates the location of

the IgG heavy chain. **(B)** Endogenous METTL14 interacts with PRMT1. The reciprocal endogenous co-immunoprecipitation (IP) assays were performed using the METTL14 antibody for IP and the PRMT1 antibody for Western blot detection (left panel) and using the PRMT1 antibody for IP and the METTL14 antibody for Western blot detection (right panel). **(C)** The C-terminal IDR of METTL14 is essential for its interaction with PRMT1. HEK293 cells were transfected with GFP-tagged PRMT1 and Flag-tagged FL or C-terminal IDR-truncated (1–400) METTL14. IP was performed using an anti-GFP antibody, and Western blot analysis was performed using anti-GFP and anti-Flag antibodies. **(D)** GST pull-down detection of the interactions of PRMT1 with GST-tagged FL and truncated (1–400) recombinant METTL14. The black triangles indicate recombinant METTL14 proteins. **(E)** Overexpression of PRMT1 enhances METTL14 arginine methylation. HEK293 cells were transfected with either GFP vector or GFP-PRMT1, together with Flag-METTL14. The level of METTL14 methylation was detected by IP using an anti-Flag antibody, followed by Western blot analysis using an anti-ASYM26 antibody. * indicates the location of the IgG heavy chain. **(F)** Knockdown of PRMT1 expression reduces METTL14 arginine methylation. HEK293 cells were transfected with control siRNA (siCtrl) and the siRNA targeting PRMT1 (siPRMT1). METTL14 was immunoprecipitated from these cells, and its methylation level was detected by Western blot analysis using an anti-ASYM26 antibody. * indicates the location of the IgG heavy chain.

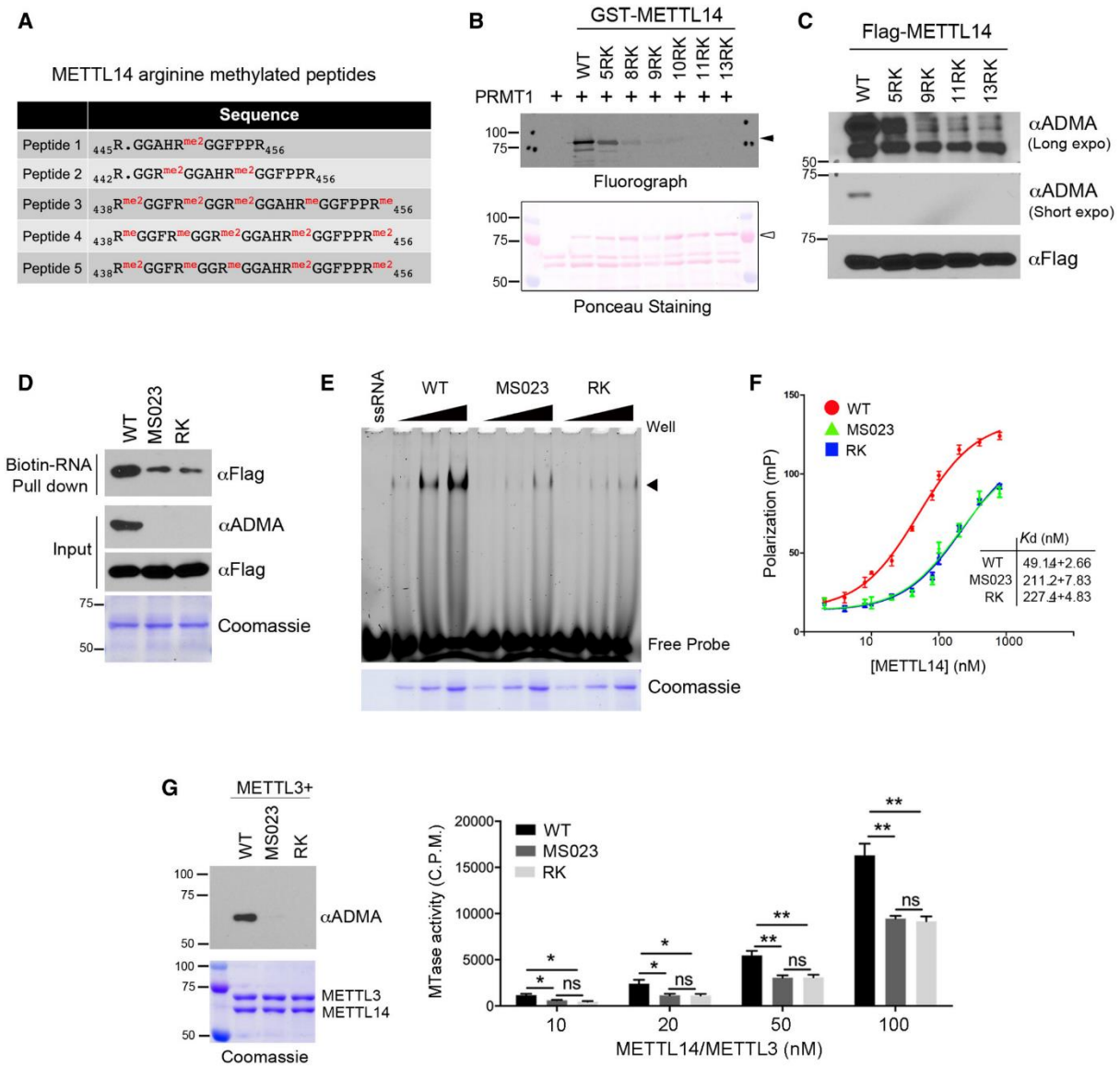


Figure 5.3 C-terminal IDR arginine methylation enhances METTL14–RNA interactions and METTL3/METTL14 RNA methylation activity.

(A) Summary of METTL14 arginine-methylated peptides identified by LC-MS/MS. **(B)** METTL14 IDR arginine methylation occurs at multiple arginine residues within RGG/RG motifs. Mutation of five arginine sites identified from mass spectrometry reduces METTL14 arginine methylation, but only mutation of all arginine residues to lysine completed blocked METTL14 methylation. Ponceau S staining shows the loading of the recombinant

proteins. The black triangles indicate arginine methylated-METTL14; open triangles indicate recombinant METTL14 proteins. **(C)** METTL14 is methylated at multiple arginine residues in cells. HEK293 cells expressing Flag-tagged WT or various R-to-K METTL14 mutants were lysed and immunoprecipitated with an anti-Flag antibody. Arginine methylation of immunoprecipitated METTL14 was detected by Western blot analysis using an anti-ADMA antibody. Both short and long exposures of the chemiluminescence signals are shown. **(D)** Arginine methylation of the METTL14 IDR enhances its interaction with RNA substrates. RNA pull-down assay was performed by incubating biotin-labeled RNA with WT, hypomethylated (MS023), and arginine methylation-deficient (RK) mutant METTL14. The pull-down samples were detected by Western blot analysis using an anti-Flag antibody. The methylation status of the recombinant proteins was confirmed by Western blot analysis using an anti-ADMA antibody. **(E)** EMSA was performed to compare the interactions of WT, hypomethylated (MS023), and arginine methylation-deficient (RK) mutant METTL14 with 6-FAM-labeled RNA. Arrow indicates the shift of the RNA probe caused by the protein–RNA interaction. Coomassie staining shows the increasing amounts of recombinant proteins used in the assay. **(F)** Fluorescence polarization assays were performed by incubating 6-FAM-labeled RNA with WT, hypomethylated (MS023), and arginine methylation-deficient (RK) mutant METTL14. Each point represents the average of three independent replicates and error bars represent standard deviation (SD). The dissociation constant values (Kd) were listed as mean ± SD. **(G)** Arginine methylation of the C-terminal IDR enhances the RNA methylation activity of the METTL14/METTL3 complex in vitro. In vitro RNA methylation assays were performed by incubating biotin-labeled RNA substrates with METTL3/METTL14 methyltransferase complexes containing WT,

hypomethylated (MS023), and arginine methylation-deficient (RK) mutant METTL14 in various concentrations (10–100 nM). The methylation status of the METTL3/METTL14 complex was confirmed by Western blot analysis using an anti-ADMA antibody. Coomassie staining shows the purification of the enzyme complex. Enzymatic activity was measured in counts per minute (c.p.m.) using a scintillation counter. Data from three replicates were analyzed by Student's t-test and shown as mean ± SD. *P < 0.05; **P < 0.01; ns, not significant.

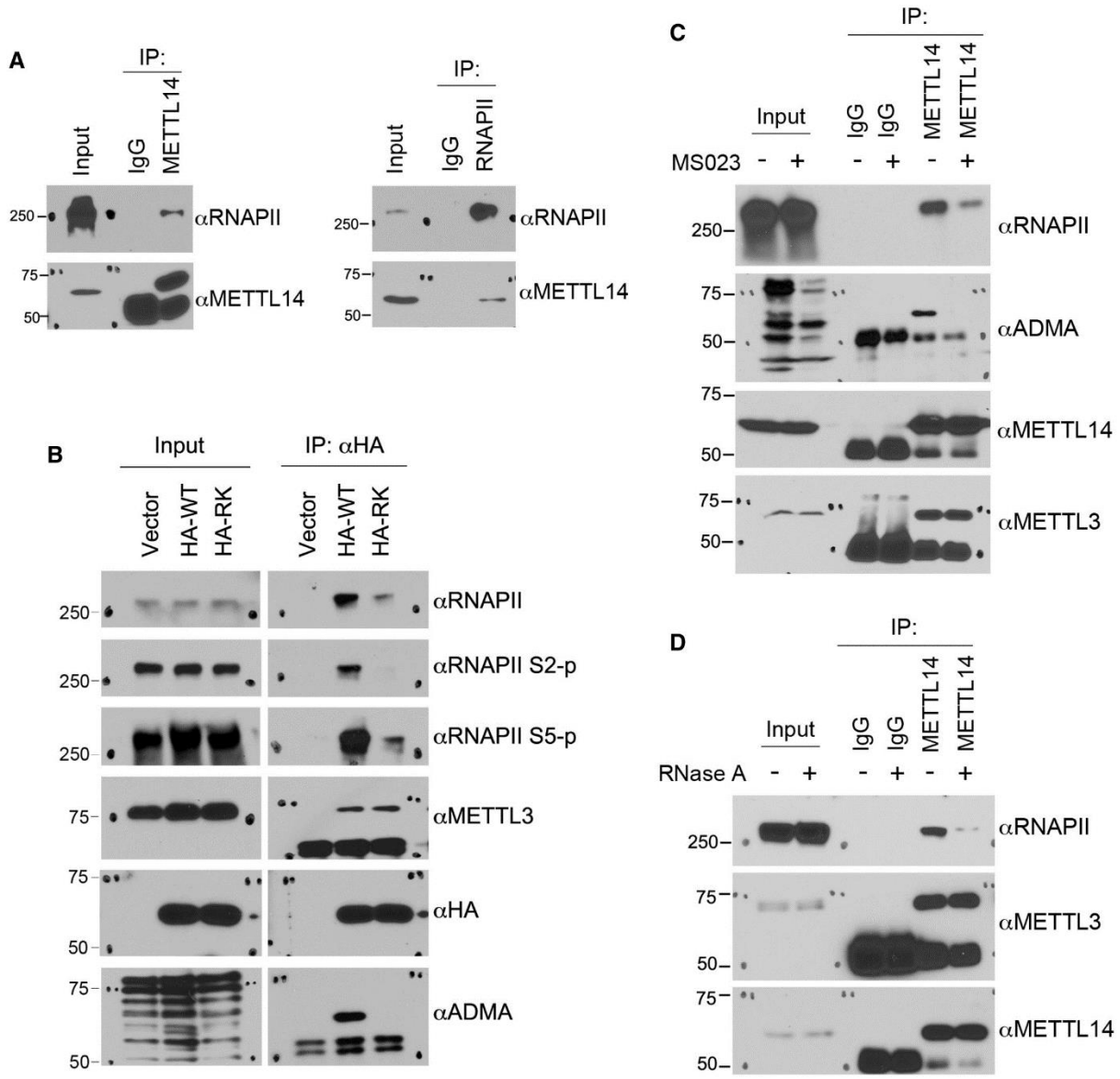


Figure 5.4 Arginine methylation of the C-terminal IDR enhances the interaction of METTL14 with RNAPII in cells.

(A) Endogenous METTL14 interacts with RNAPII. Endogenous co-immunoprecipitation (IP) was performed using the METTL14 antibody for IP and the RNAPII antibody for Western blot detection (left panel) and using the RNAPII antibody for IP and the METTL14 antibody for Western blot detection (right panel). **(B)** Arginine methylation of the

METTL14 C-terminal IDR enhances its interaction with RNAPII. HEK293 cells were transfected with HA-tagged WT or arginine methylation-deficient (RK) mutant METTL14. IP was performed using an anti-HA antibody, and Western blot analysis was performed using the indicated antibodies. **(C)** Co-IP assays were performed to compare the interactions between METTL14 and RNAPII in control and MS023-treated HEK293 cells. Cells were treated with either DMSO or MS023 (1 μM) for 48 h before they were lysed. IP was performed using control IgG and METTL14 antibodies, respectively. Western blot analysis was performed using anti-RNAPII, anti-ADMA, anti-METTL14, and anti-METTL3 antibodies. **(D)** Co-IP assays were performed to examine the involvement of RNA in the METTL14–RNAPII interaction. Total cell lysates were either left untreated or treated with RNase A to remove the RNA component before IP. Western blot analysis was performed using anti-RNAPII, anti-METTL3, and anti-METTL14 antibodies.

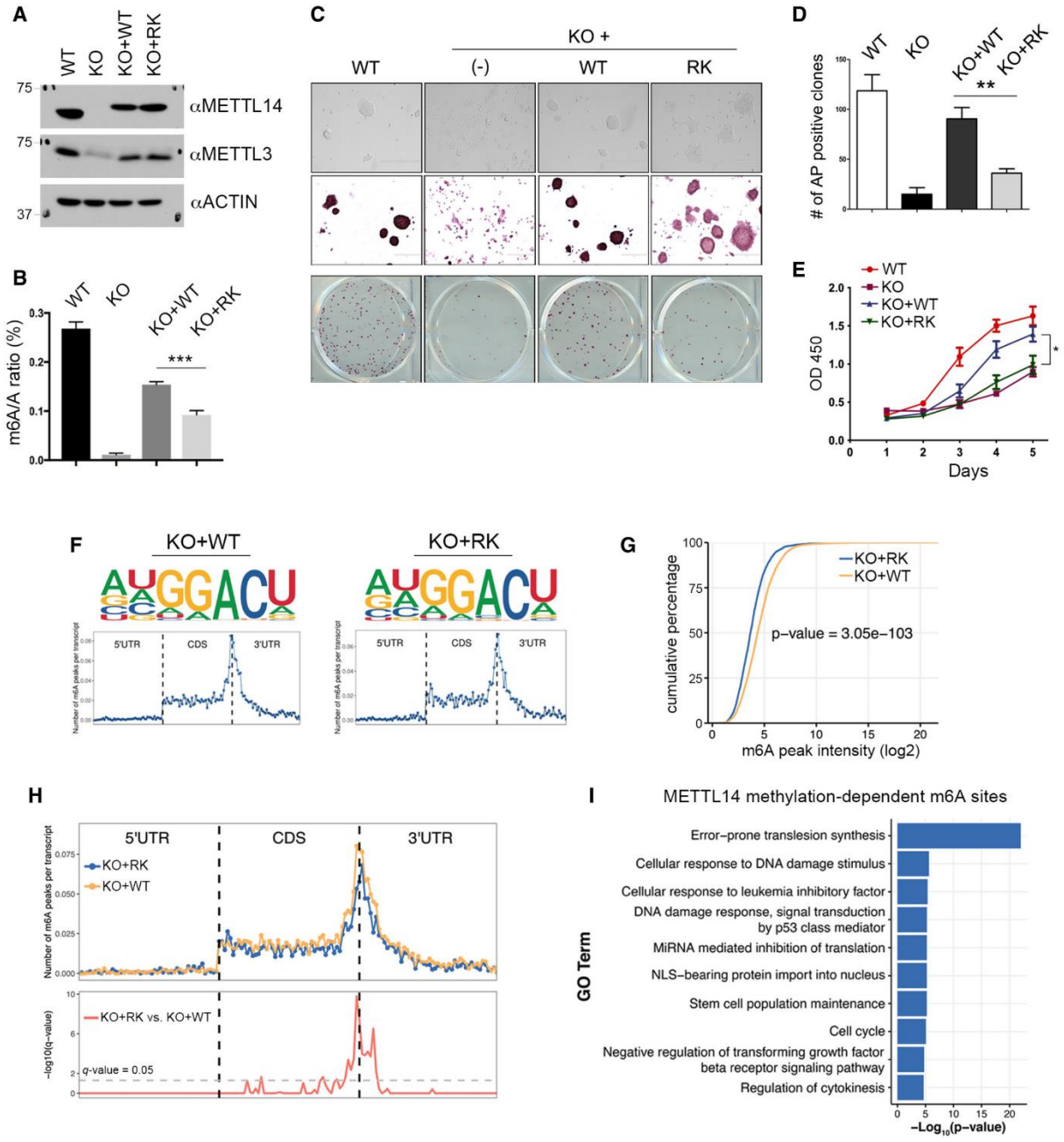


Figure 5.5 Analysis of METTL14 arginine methylation-dependent m6A sites

(A) Generation of isogenic mESC lines expressing WT and arginine methylation-deficient mutant (RK) METTL14. Mettl14 KO mESCs were transfected with Flag-tagged WT or RK mutant METTL14 using a lentivirus expression system. The expression of METTL14 and METTL3 in these cells was detected by Western blot analysis using anti-METTL14 and anti-

METTL3 antibodies. ACTIN was used as a loading control. **(B)** m6A levels are reduced in mESCs expressing arginine methylation-deficient (RK) mutant METTL14. The mRNA purified from WT, Mettl14 KO, KO + WT, and KO + RK mESCs was subjected to LC-MS/MS analysis to quantify m6A levels (presented as the m6A/A ratio). **(C)** Morphology and alkaline phosphatase (AP) staining of mESCs expressing WT, Mettl14 KO, KO + WT, and KO + RK METTL14. Scale bar: 400 μm. **(D)** Quantification of AP-positive clones in (C). **(E)** Proliferation of mESCs expressing WT, Mettl14 KO, KO + WT, and KO + RK METTL14 over a 5-day period. Each point represents the average of three independent replicates, and error bars represent standard deviation (SD). **(F)** Sequence motifs of m6A-enriched regions in KO + WT and KO + RK mESCs (upper panels). Topological distribution of normalized m6A peaks across the 5′UTR, CDS, and 3′UTR of mRNAs (lower panels). **(G)** Cumulative distribution of log2 m6A peak intensity in KO + WT and KO + RK mESCs. Statistical analysis was performed using the Wilcoxon test to measure the median difference of peak intensities between the two groups. **(H)** Overlay of m6A distributions across the 5′UTR, CDS, and 3′UTR of mRNAs in KO + WT and KO + RK mESCs (upper panel). Statistical analysis of differential m6A peaks in KO + RK versus KO + WT mESCs (lower panel). The y-axis represents the q-value (−log10). The dashed gray line indicates q-value = 0.05. **(I)** Gene Ontology (GO) analysis of genes harboring METTL14 arginine methylation-dependent m6A sites. Statistical analysis was performed using Hypergeometric test. The P-value for the enrichment of each biological process (GO term) is shown.

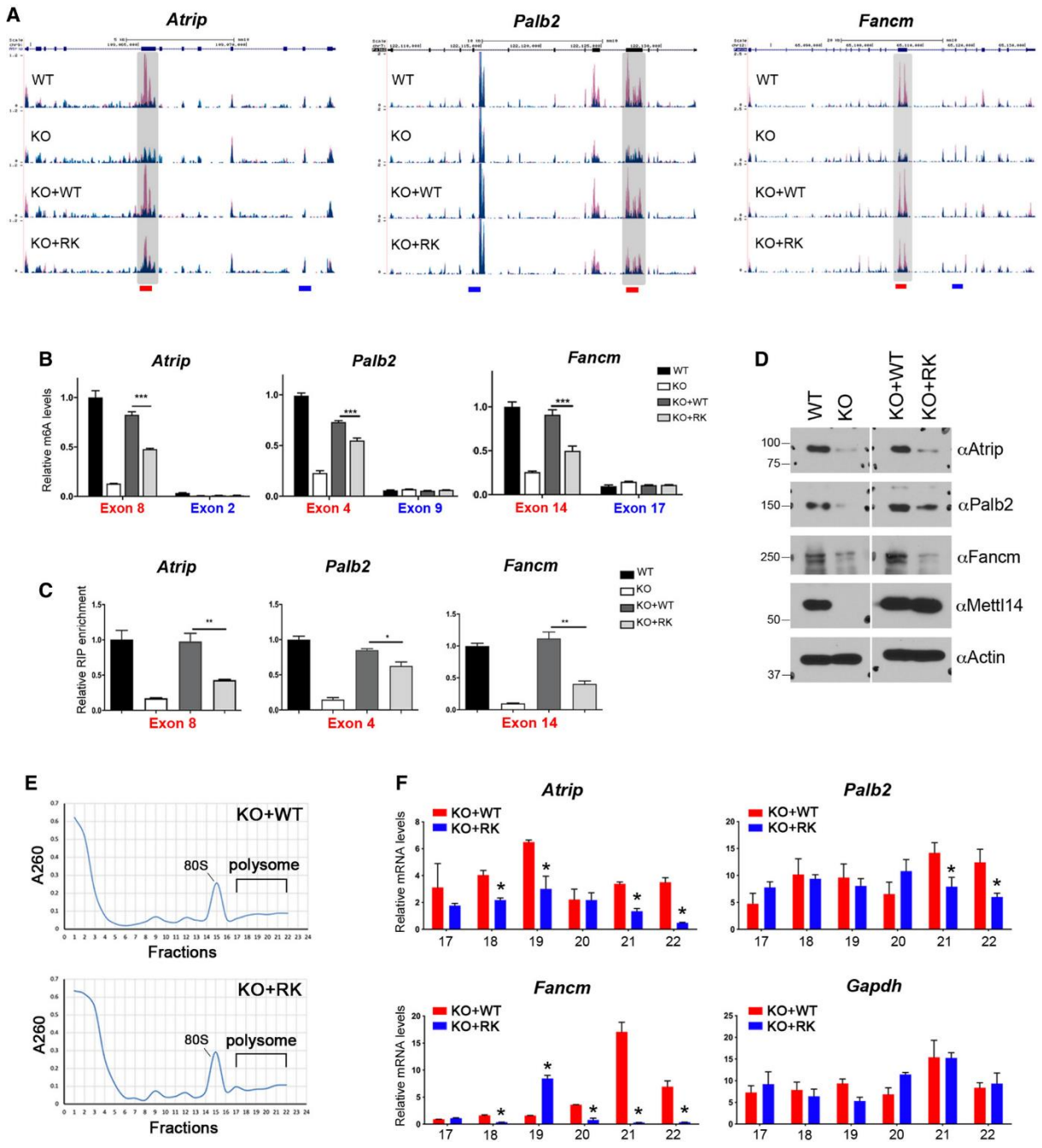


Figure 5.6 METTL14 arginine methylation-dependent m6A sites are associated with enhanced translation of DNA repair genes

(A) UCSC Genome Browser custom tracks of m6A-seq reads along the indicated mRNAs in WT, Mettl14 KO, KO + WT, and KO + RK mESCs. The y-axis represents the normalized

number of reads. Blue reads are from non-immunoprecipitated input libraries, and red reads are from m6A-IP libraries. Above the custom tracks, the thick blue boxes represent the protein-coding regions (CDSs), the thin blue boxes represent the untranslated regions (UTRs), and the blue lines represent introns. The bars at the bottom of the custom tracks indicate the amplicon locations for MeRIP (m6A-IP)-qPCR assays (B) and METTL14 RIP-qPCR assays (C) to detect m6A-positive (red) and negative (blue) regions. **(B)** MeRIP (m6A-IP)-qPCR assays were performed for WT, Mettl14 KO, KO + WT, and KO + RK mESCs to validate the MeRIP-seq results. m6A-negative regions of the transcripts (blue) were included as negative controls. **(C)** METTL14 RIP-qPCR assays were performed for WT, Mettl14 KO, KO + WT, and KO + RK mESCs to compare the binding of WT and RK mutant METTL14 to mRNA targets. Primers (red color) that amplify m6A positive regions of the transcripts were used. **(D)** The expression of ICL repair genes is reduced in mESCs expressing arginine methylation-deficient mutant (RK) METTL14. Total cell lysates from WT, Mettl14 KO, KO + WT, and KO + RK mESCs were subjected to Western blot analysis using the indicated antibodies. **(E)** Polysome profiling was performed for KO + WT and KO + RK mESCs. Whole-cell extracts were fractionated through centrifugation in a sucrose density gradient. Optical scans (OD260) of the collected fractions are shown. **(F)** Quantification of ribosome-bound mRNA for the indicated genes from individual fractions (as in (E)), relative to the amount of the total mRNA in all fractions. Gapdh was included as a negative control.

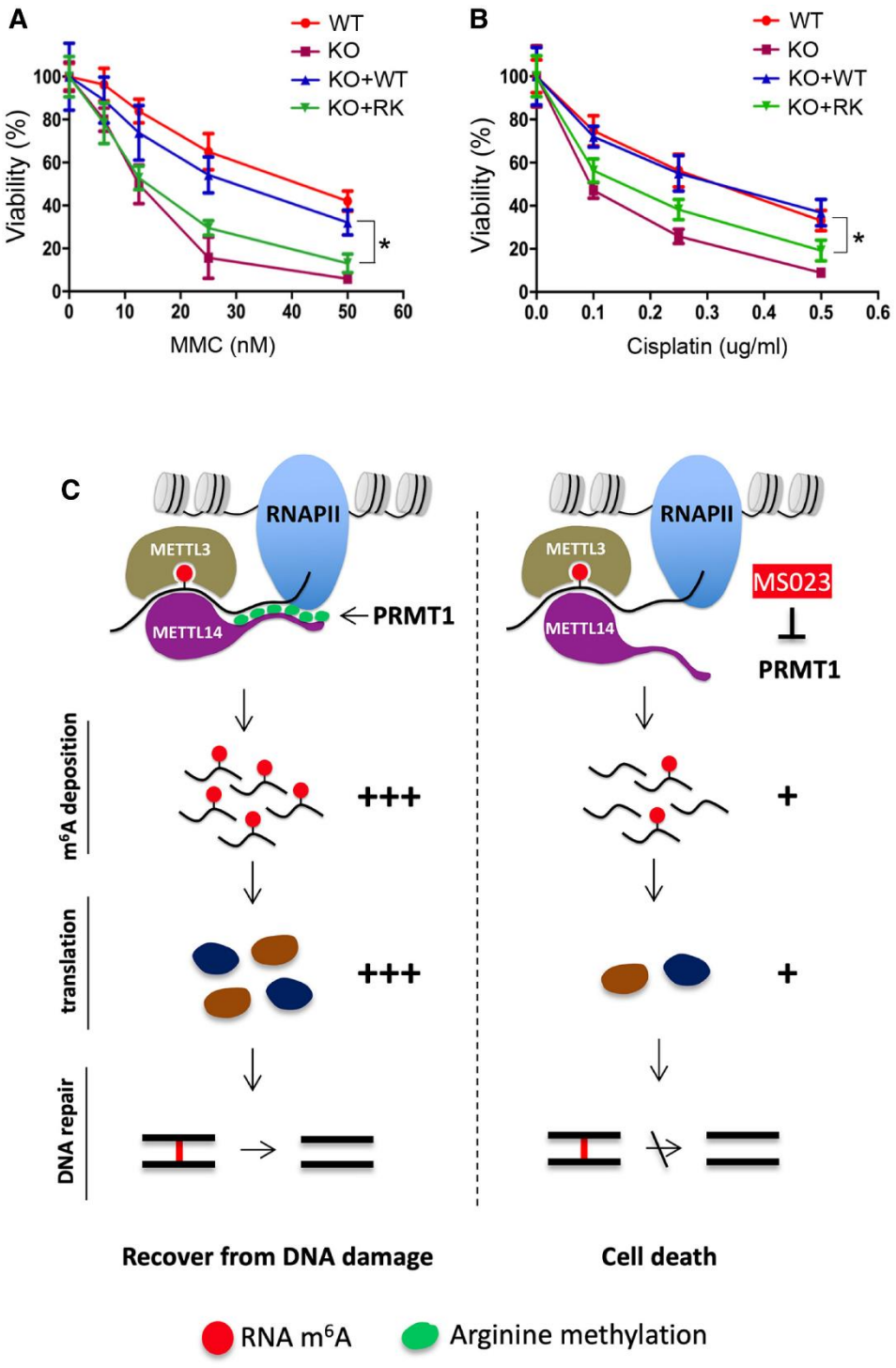


Figure 5.7 Loss of METTL14 arginine methylation sensitizes mESCs to DNA damage

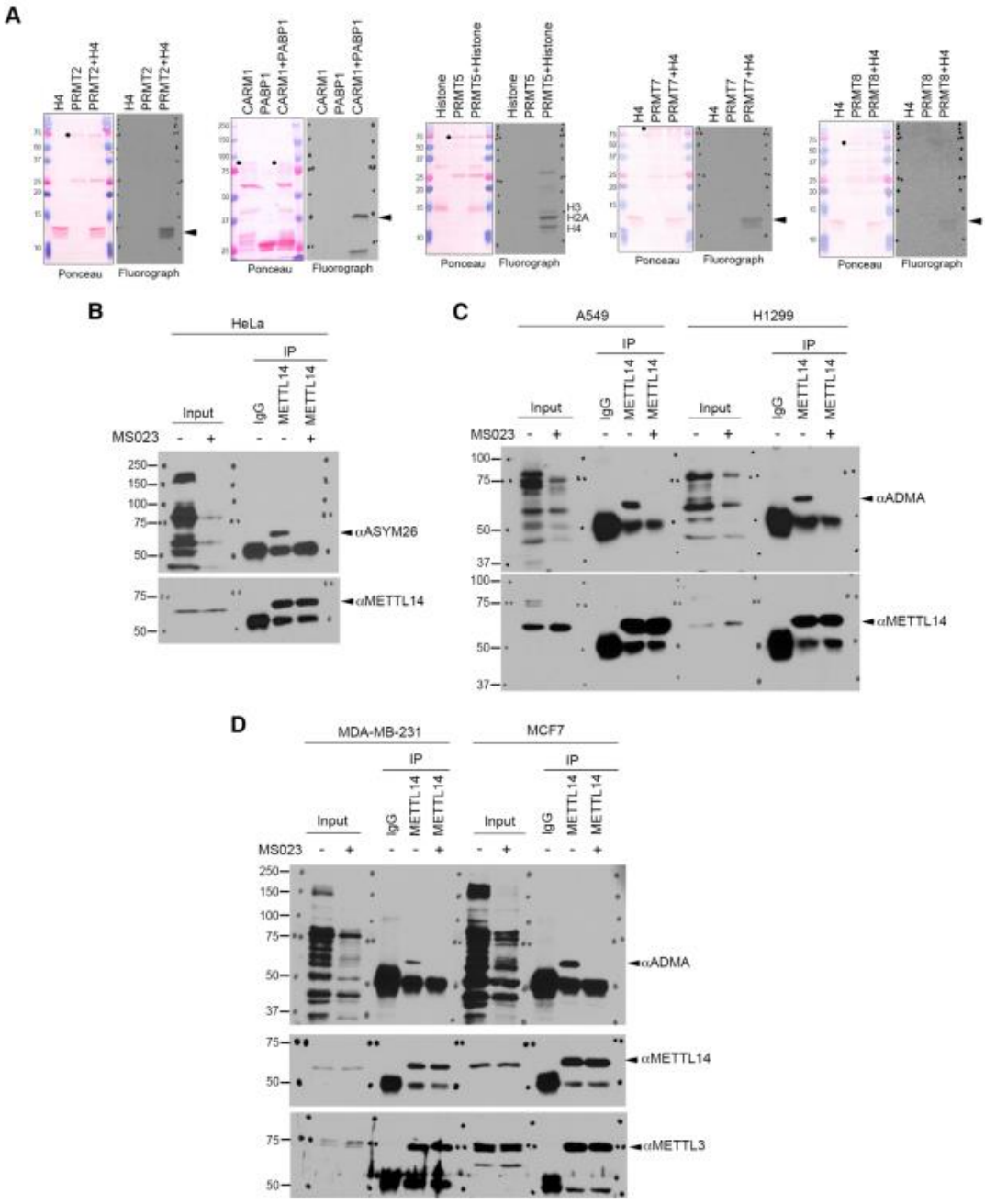
(A) mESCs expressing arginine methylation-deficient mutant (RK) METTL14 are sensitive to ICL damage induced by MMC. WT, Mettl14 KO, KO + WT, and KO + RK mESCs were

treated with various concentrations of MMC for 4 days before cell viability was measured. **(B)** Similar to (A), except that mESCs were treated with cisplatin, another ICL-inducing chemical, at various concentrations. **(C)** Proposed model for METTL14 C-terminal IDR arginine methylation-mediated regulation of m6A RNA modification and its effects on ICL DNA repair. PRMT1-mediated arginine methylation of the C-terminal IDR of METTL14 promotes its interactions with RNA substrates and RNAPII, which enables efficient m6A deposition on transcripts involved in ICL repair. The deposition of m6A enhances the translation efficiency of these DNA repair genes, promoting the recovery of mESCs from DNA damage. Inhibiting METTL14 arginine methylation using the type I PRMT inhibitor MS023 reduces m6A deposition and the protein synthesis of ICL repair genes, thus sensitizing mESCs to DNA damage-induced cell death.

Supplementary Figure 5.8 METTL14 harbors a conserved arginine/glycine (RGG/RG)-rich C terminus.

The C-terminal intrinsically disordered region (IDR) of METTL14 is conserved among species. METTL14 protein sequences from Drosophila melanogaster, Danio rerio, Xenopus tropicalis, Gallus gallus, Mus musculus, and Homo sapiens were aligned using ClustalW software. Colored strips indicate the conserved protein domains/motifs. The C-terminal

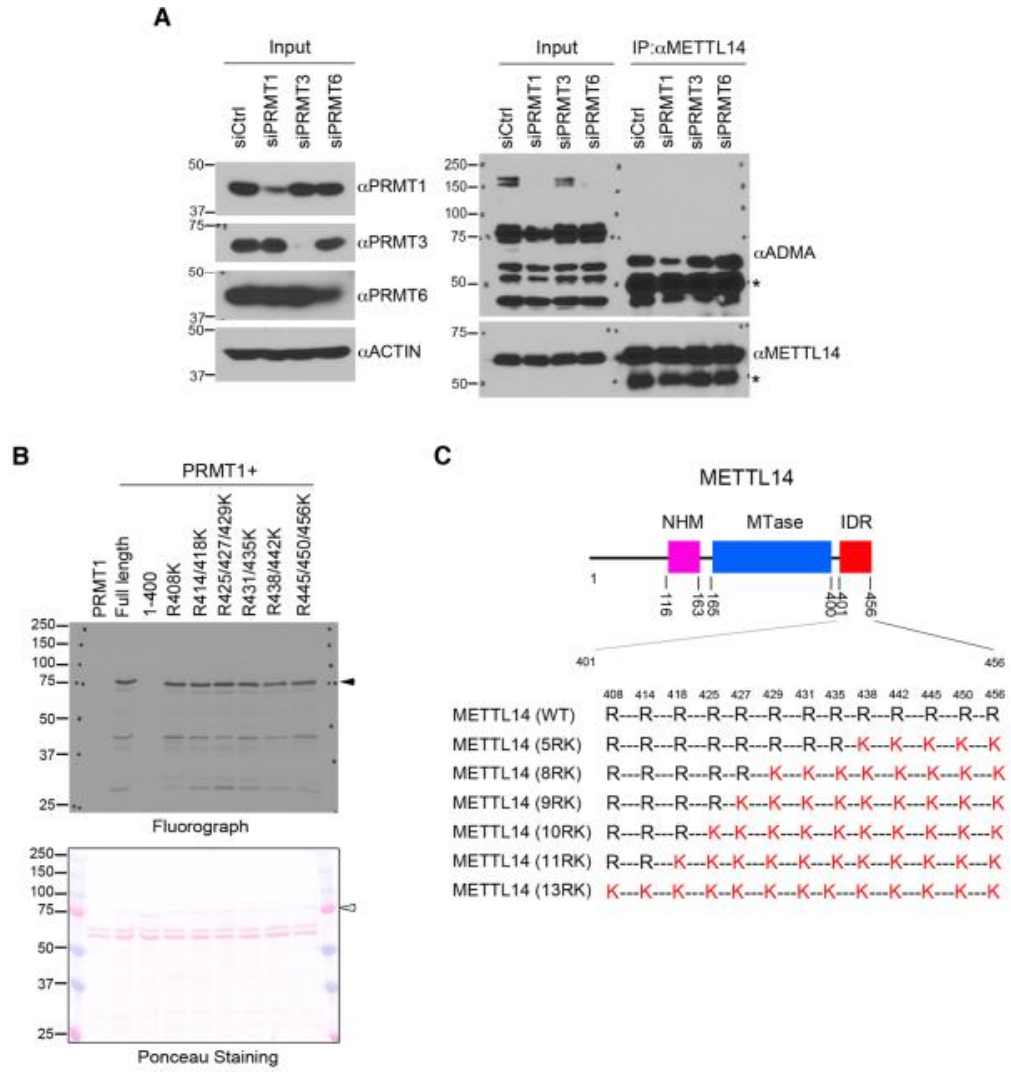
IDR, which contains multiple RGG motifs, is conserved in vertebrates, whereas the same region in Drosophila harbors a much shorter RGG motif.



Supplementary Figure 5.9 Characterization of METTL14 arginine methylation in vitro and in vivo.

(A) In vitro methylation assays were performed to confirm the activities of PRMTs used in **Figure 5.1B**. Recombinant proteins of PRMTs were incubated with their respective

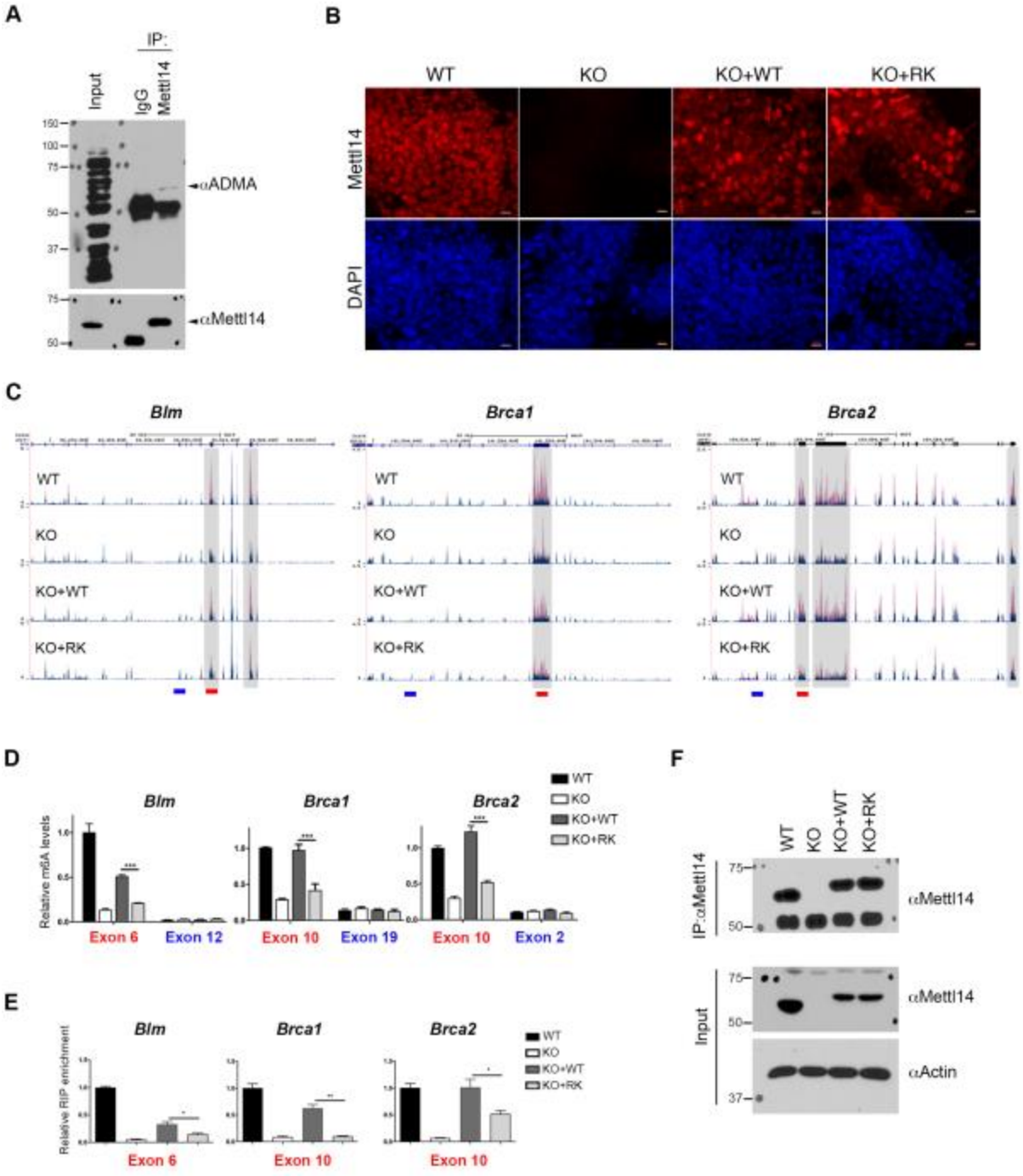
substrates, including histone H4 (H4), Polyadenylate-binding protein 1 (PABP1), and core histones. The Ponceau staining shows the loading of the recombinant proteins. Black dots indicate PRMT enzymes; triangles indicate fluorograph signals from substrate methylation. Human cervical cancer cell line HeLa **(B)**, Lung cancer cell line A549 and H1299 **(C)**, and breast cancer cell line MDA-MB231 and MCF7 **(D)** were either left untreated or treated with Type I PRMT inhibitor MS023 (1 μM, 48 h). The level of METTL14 arginine methylation was detected by IP/Western blot analysis using indicated antibodies.



Supplementary Figure 5.10 Identification of PRMT1-catalyzed methylation sites on METTL14.

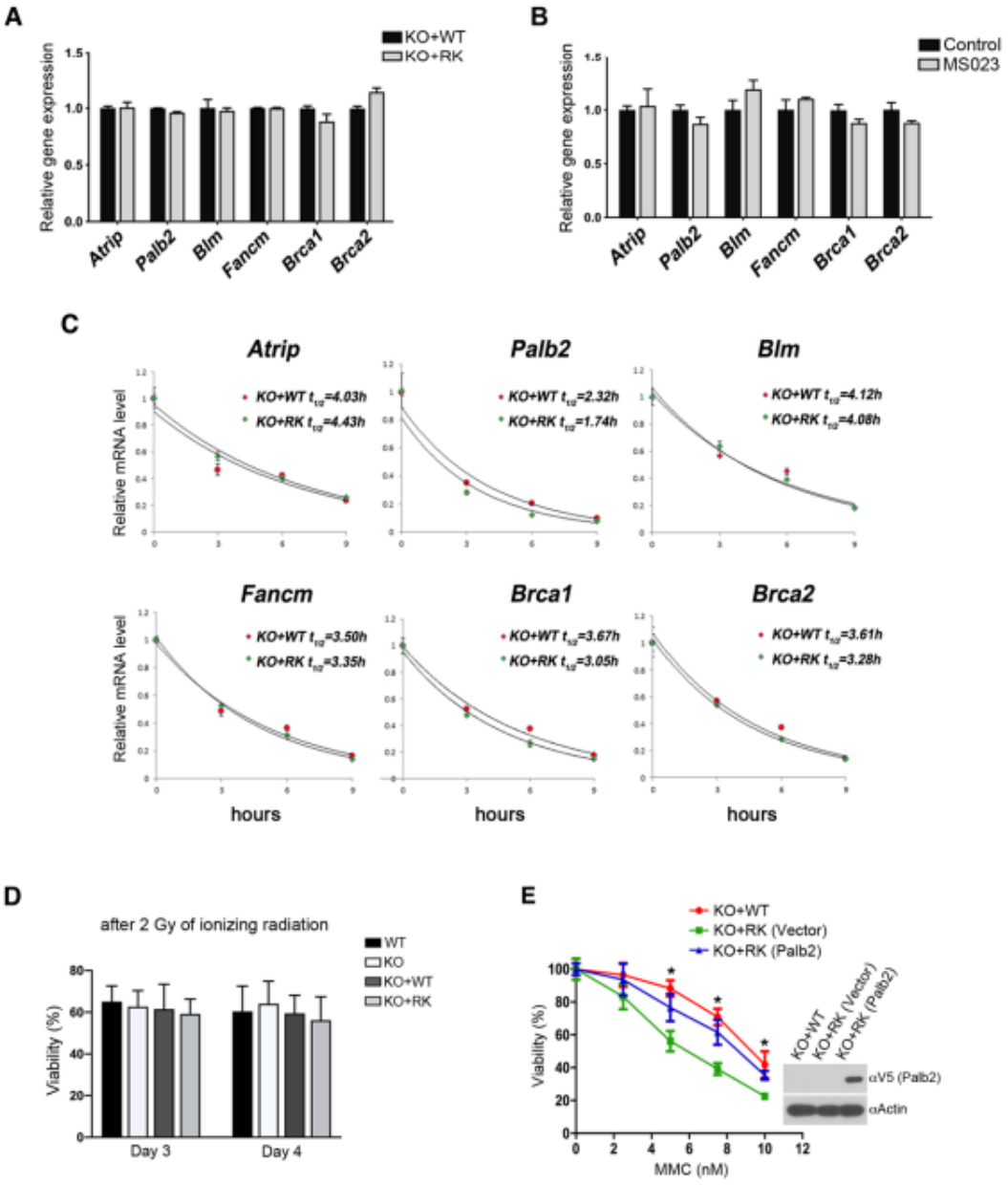
(A) PRMT1, but not PRMT3 and PRMT6, is responsible for METTL14 arginine methylation in vivo. The levels of METTL14 arginine methylation were compared in cells transfected with control siRNA (siCtrl), PRMT1-specific siRNA (siPRMT1), PRMT3-specific siRNA (siPRMT3), and PRMT6-specific siRNA (siPRMT6). The knockdown efficiency was confirmed by Western blot analysis of total cell lysates using indicated antibodies. The levels of METTL14 arginine methylation were detected by IP/WB analysis. **(B)** Selective

mutation analysis of single, double, or triple arginine sites does not impair METTL14 methylation in vitro. The in vitro methylation assays were performed by incubating recombinant PRMT1 with purified GST-tagged WT, 1-400 truncation, and various arginine to lysine (R-to-K) METTL14 mutants. The Ponceau S staining shows the loading of the recombinant proteins used in the exact methylation assay. **(C)** Schematic representation of the mutated arginine residues in each METTL14 mutant constructs used in **Figure 5.3B** and **Figure 5.3C**.



Supplementary Figure 5.11 Characterization of METTL14 arginine methylation-dependent m6 A sites in mESCs.

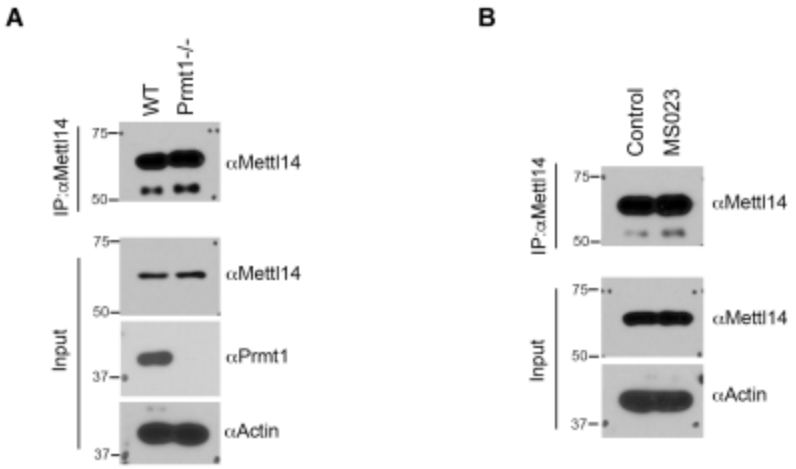
(A) Detection of METTL14 arginine methylation in mESCs. METTL14 was immunoprecipitated from mESCs, and Western blot analysis was performed to detect its methylation using anti-ADMA and anti-METTL14 antibodies. **(B)** Detection of METT1L4 expression in WT, Mettl14 KO, KO+WT, KO+RK mESCs by immunofluorescence using an anti-METTL14 antibody. DAPI staining indicates the cell nucleus. Scale bar: 20 μM **(C)** UCSC Genome Browser custom tracks of m6 A-seq reads along the indicated mRNAs in WT, Mettl14 KO, KO+WT, and KO+RK mESCs. The y-axis represents the normalized number of reads. Blue reads are from non-immunoprecipitated input libraries, and red reads are from m6 A-IP libraries. Above the custom tracks, the thick blue boxes represent the protein coding regions (CDSs), the thin blue boxes represent the untranslated regions (UTRs), and the blue lines represent introns. The bars at the bottom of the custom tracks indicate the amplicon locations for MeRIP (m6 A-IP)-qPCR assays (D) and METTL14 RIP-qPCR assays (E) to detect m6 A-positive (red) and negative (blue) regions. **(D)** MeRIP (m6 A-IP)-qPCR assays were performed for WT, Mettl14 KO, KO+WT, and KO+RK mESCs to validate the MeRIP-seq results. Four target mRNAs encoded by genes in the Fanconi anemia pathway were analyzed. m6 A-negative regions of the transcripts (blue) were included as negative controls. Data are shown as mean ± SD from three biological replicates. ***, $p < 0.001$. **(E)** METTL14 RIP-qPCR assays were performed for WT, Mettl14 KO, KO+WT, and KO+RK mESCs to compare the binding of WT and RK mutant METTL14 to the indicated mRNA targets. Primers (red color) that amplify m6 A positive regions of the transcripts were used. Data are shown as mean ± SD from three biological replicates. *, $p < 0.05$; **, $p < 0.01$. **(F)** The amount of METTL14 protein immunoprecipitated in the RIP experiments described in **Figure 5.6C** and **Supplementary Figure 5.11E** was detected by Western blot analysis.



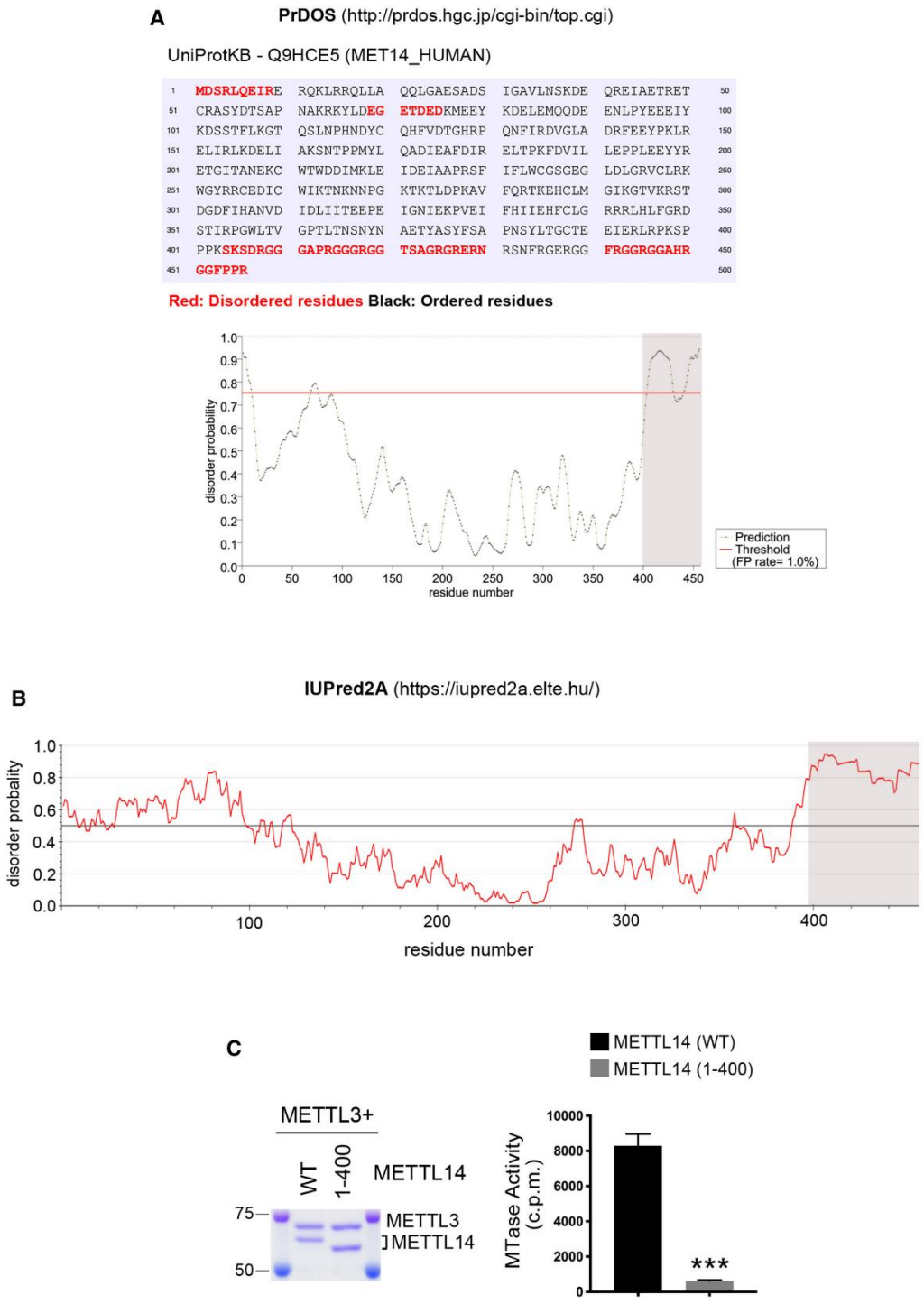
Supplementary Figure 5.12 Examine the impact of METTL14 arginine methylation loss on mRNA expression, stability, and cellular response to DNA damage.

(A) The mRNA levels of Fanconi anemia pathway genes were analyzed by RT-qPCR for mESCs expressing WT and RK mutant METTL14. Data are shown as mean ± SD from three biological replicates. **(B)** The mRNA levels of Fanconi anemia pathway genes were analyzed

by RT-qPCR for mESCs treated with DMSO (control) or type I PRMT inhibitor (MS023). Data are shown as mean ± SD from three biological replicates. **(C)** mRNA half-life assays were performed to compare the mRNA stability of genes involved in the Fanconi anemia pathway for mESCs expressing WT and RK mutant METTL14. **(D)** The viability of WT, Mettl14 KO, KO+WT, and KO+RK mESCs was measured on days 3 and 4 after ionizing radiation (2 Gy). **(E)** The KO+RK mESCs transfected with V5-tagged Palb2, as well as KO+WT and KO+RK mESCs, were treated with various amounts of MMC. Cell viability was measured on day 4. The expression of transfected Palb2 was confirmed by Western blot analysis using an anti-V5 antibody. Data are shown as mean ± SD from three biological replicates. *, $p < 0.05$.

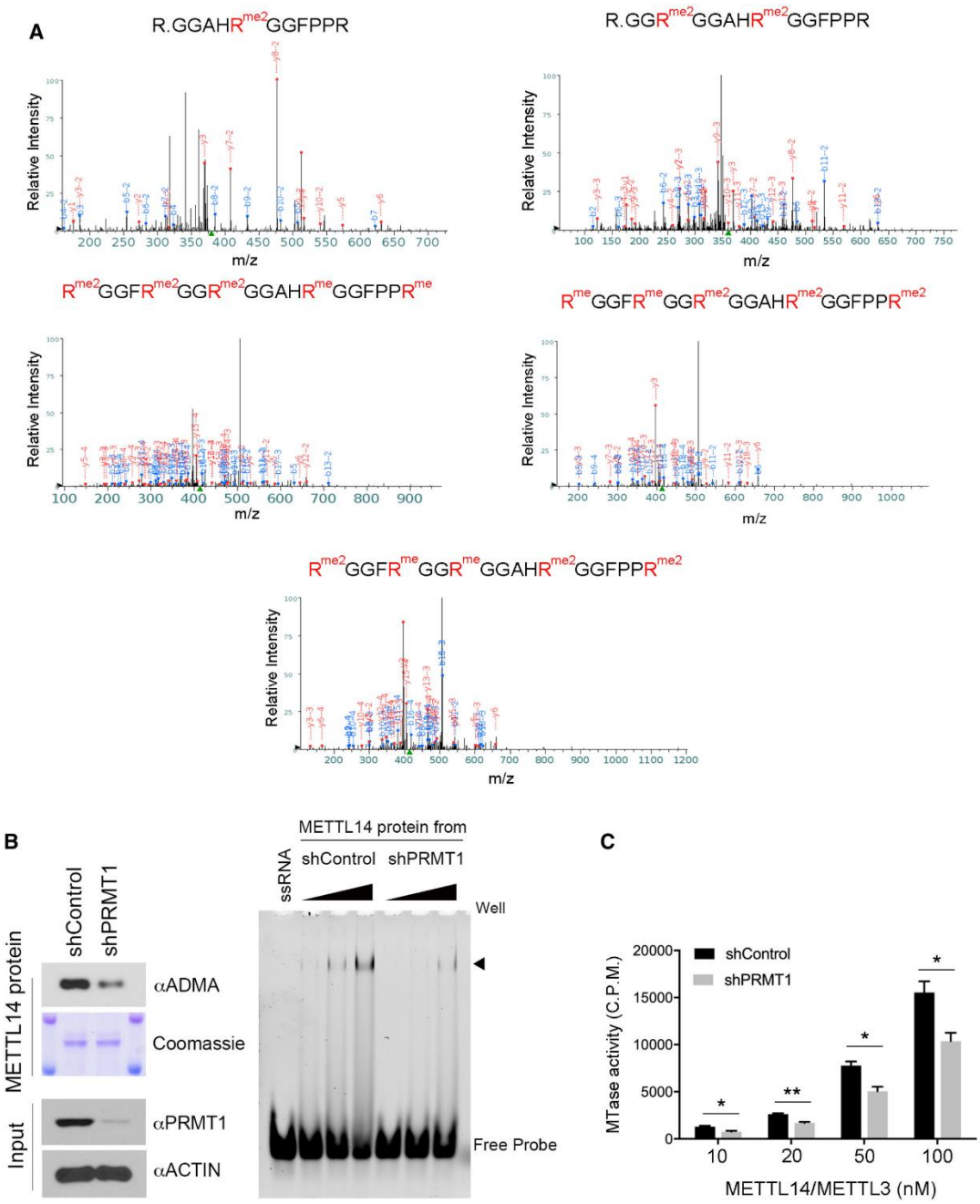


Supplementary Figure 5.13 The amount of METTL14 protein immunoprecipitated in the RIP experiments performed in Prmt1 KO (A) and MS023-treated (B) mESCs, as described in Supplementary Figure 5.18B, was detected by Western blot analysis.



Supplementary Figure 5.14 The RG-rich C terminus of METTL14 is intrinsically disordered and is essential for the RNA methyltransferase activity of METTL3/METTL14 complex.

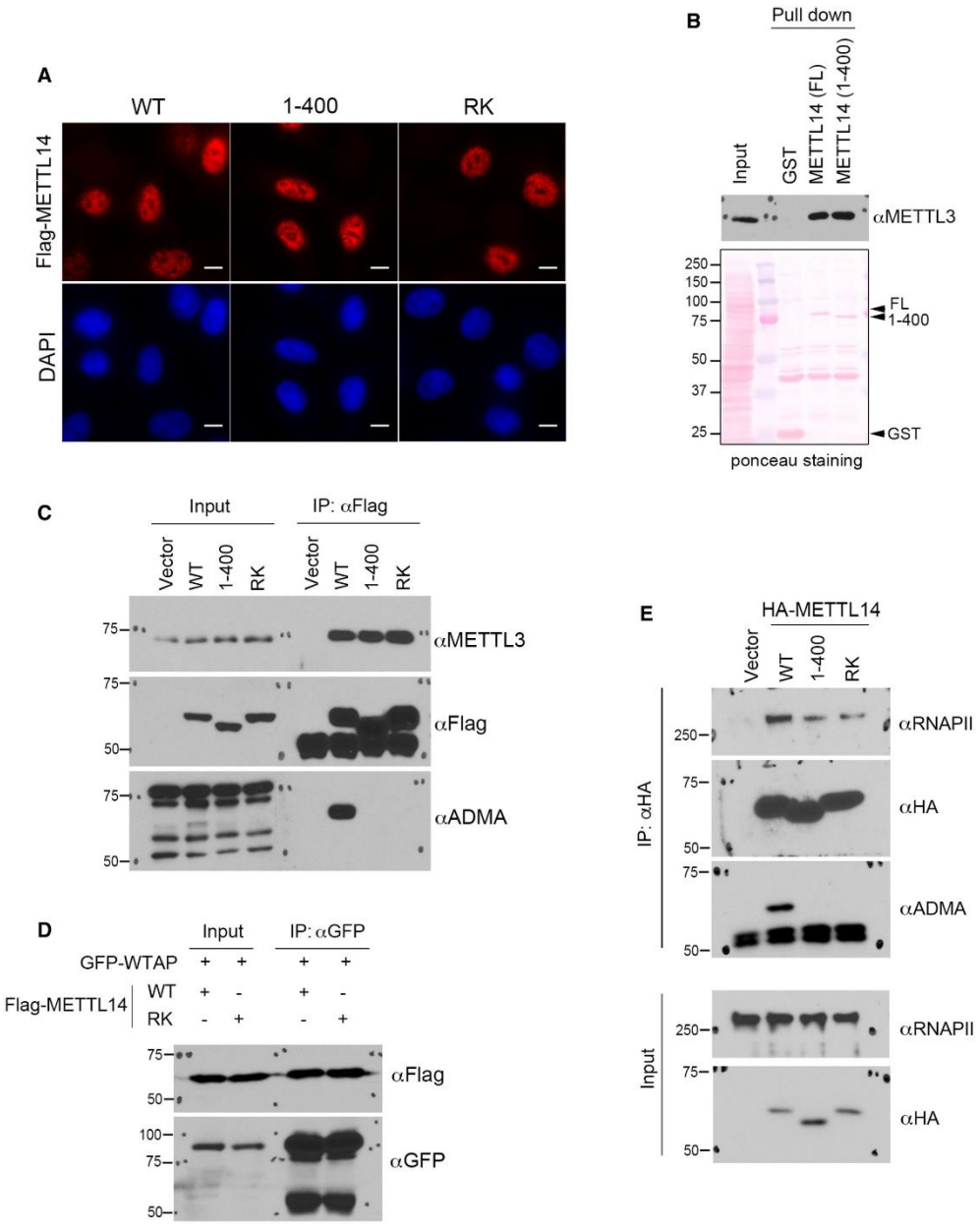
(A) Prediction of METTL14 IDRs and disorder probability using PrDOS. Disordered amino acids are highlighted in red. FP: false positive. **(B)** Prediction of METTL14 disorder probability using IUPred2A. **(C)** The C-terminal IDR of METTL14 is essential for the RNA methylation activity of the METTL3/METTL14 complex. In vitro RNA methylation assays were performed by incubating biotin-labeled RNA substrates with METTL3/METTL14 methyltransferase complexes containing WT and C-terminal IDR-truncated mutant (1–400) METTL14. The Coomassie staining shows the purification of the enzyme complex. The enzymatic activity was measured in counts per minute (c.p.m.) using a scintillation counter. Data from three independent replicates were analyzed by Student's t-test and shown as mean ± SD. ***P < 0.001.



Supplementary Figure 5.15 Characterization of the impacts of arginine methylation on METTL14–RNA interactions and RNA methylation activity.

(A) Identification of METTL14 arginine methylation sites by mass spectrometry. LC-MS/MS was performed on METTL14 proteins purified from HEK293 cells. Five peptides that are

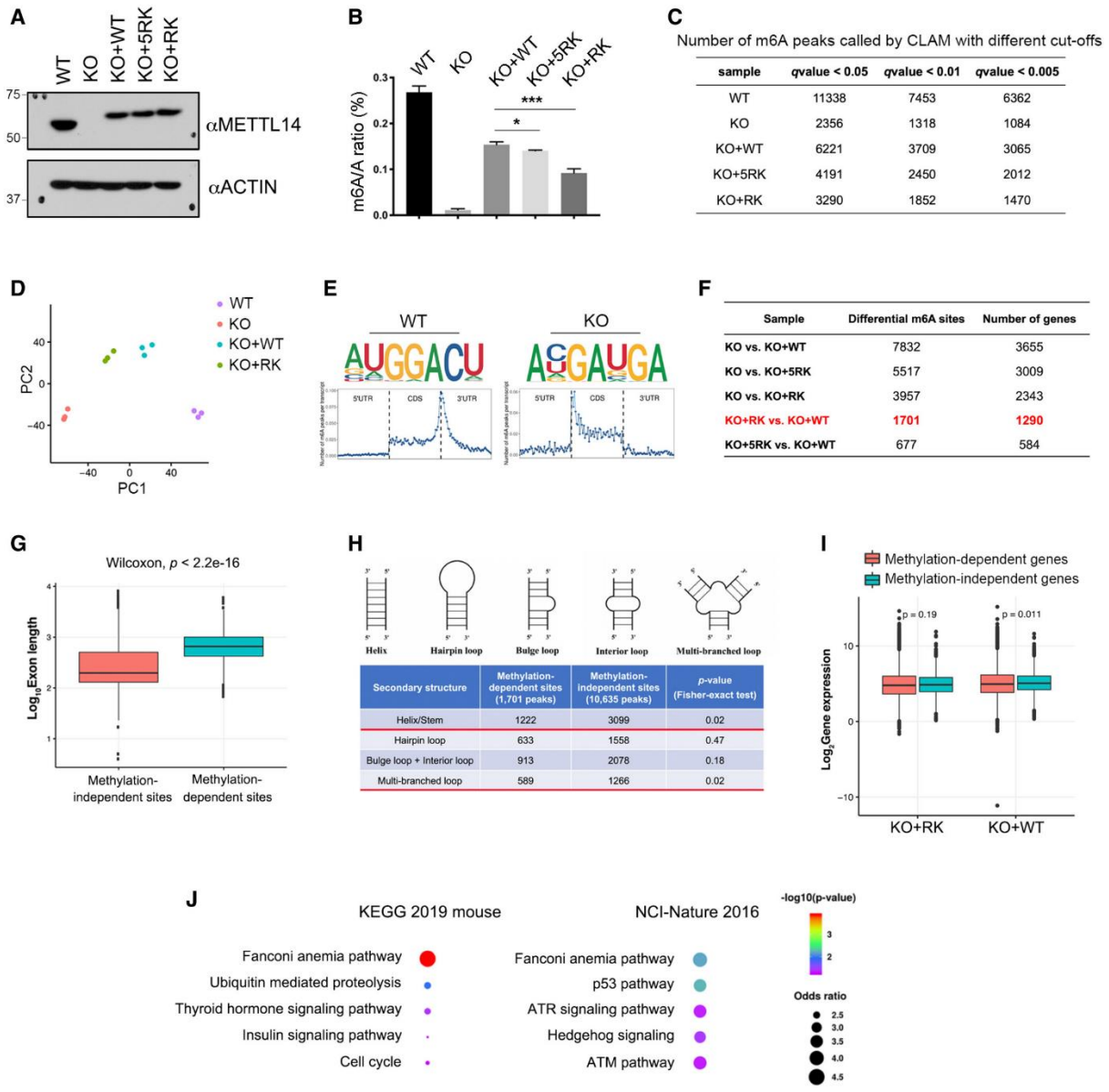
mono- or dimethylated were identified (R438, R442, R445, R450, and R456). **(B)** Recombinant METTL14 proteins purified from PRMT1 knockdown HEK293 cells exhibits reduced RNA interactions. Flag-METTL14 was expressed and purified from control (shControl) and PRMT1 knockdown (shPRMT1) HEK293 cells. The methylation level of METTL14 was detected by Western blot using an anti-ADMA antibody. The amount of protein was visualized by Coomassie staining (left panel). EMSA was performed to compare the interaction of recombinant METTL14 purified from shControl and shPRMT1 HEK293 cells with 6-FAM-labeled RNA. Arrow indicates the shift of the RNA probe caused by the protein–RNA interaction (right panel). **(C)** Recombinant METTL14 purified from PRMT1 knockdown HEK293 cells exhibits reduced RNA methylation activity. In vitro RNA methylation assays were performed by incubating biotin-labeled RNA substrates with METTL3/METTL14 methyltransferase complexes purified from control (shControl) and PRMT1 knockdown (shPRMT1) HEK293 cells. The enzymatic activity was measured in counts per minute (c.p.m.) using a scintillation counter. Data from three independent replicates were analyzed by Student's t-test and shown as mean ± SD. *P < 0.05; **P < 0.01.



Supplementary Figure 5.16 Localization and interaction analysis of arginine methylation-deficient METTL14 in cells.

(A) Immunofluorescence assays were performed to examine the subcellular localization of WT, C-terminal IDR-truncated (1–400), and arginine methylation-deficient (RK) mutant METTL14 in HeLa cells. The localizations of all three proteins were detected by using an

anti-Flag antibody. DAPI staining was performed to mark the cell nucleus. Scale bar: 10 μm. **(B)** GST pull-down assays were performed by incubating GST-tagged full-length (FL) and C-terminal IDR-truncated (1–400) METTL14 with HEK293 cell lysate. Western blot analysis was performed using an anti-METTL3 antibody. Ponceau S staining shows the loading of the recombinant proteins in the pull-down samples. **(C)** Co-IP assays were performed to detect the interactions of WT, C-terminal IDR-truncated (1–400), and arginine methylation-deficient (RK) mutant METTL14 with METTL3. The methylation of METTL14 protein was confirmed by Western blot analysis using an anti-ADMA antibody. **(D)** Co-IP assays were performed to detect the interactions of WT and arginine methylation-deficient (RK) mutant METTL14 with GFP-WTAP. The HEK293 cells were transiently transfected with the indicated plasmids 48 h before the assays were performed. **(E)** Co-IP assays were performed to detect the interactions of WT, C-terminal IDR-truncated (1–400), and arginine methylation-deficient (RK) mutant METTL14 with RNAPII. The methylation of METTL14 protein was confirmed by Western blot analysis using an anti-ADMA antibody.

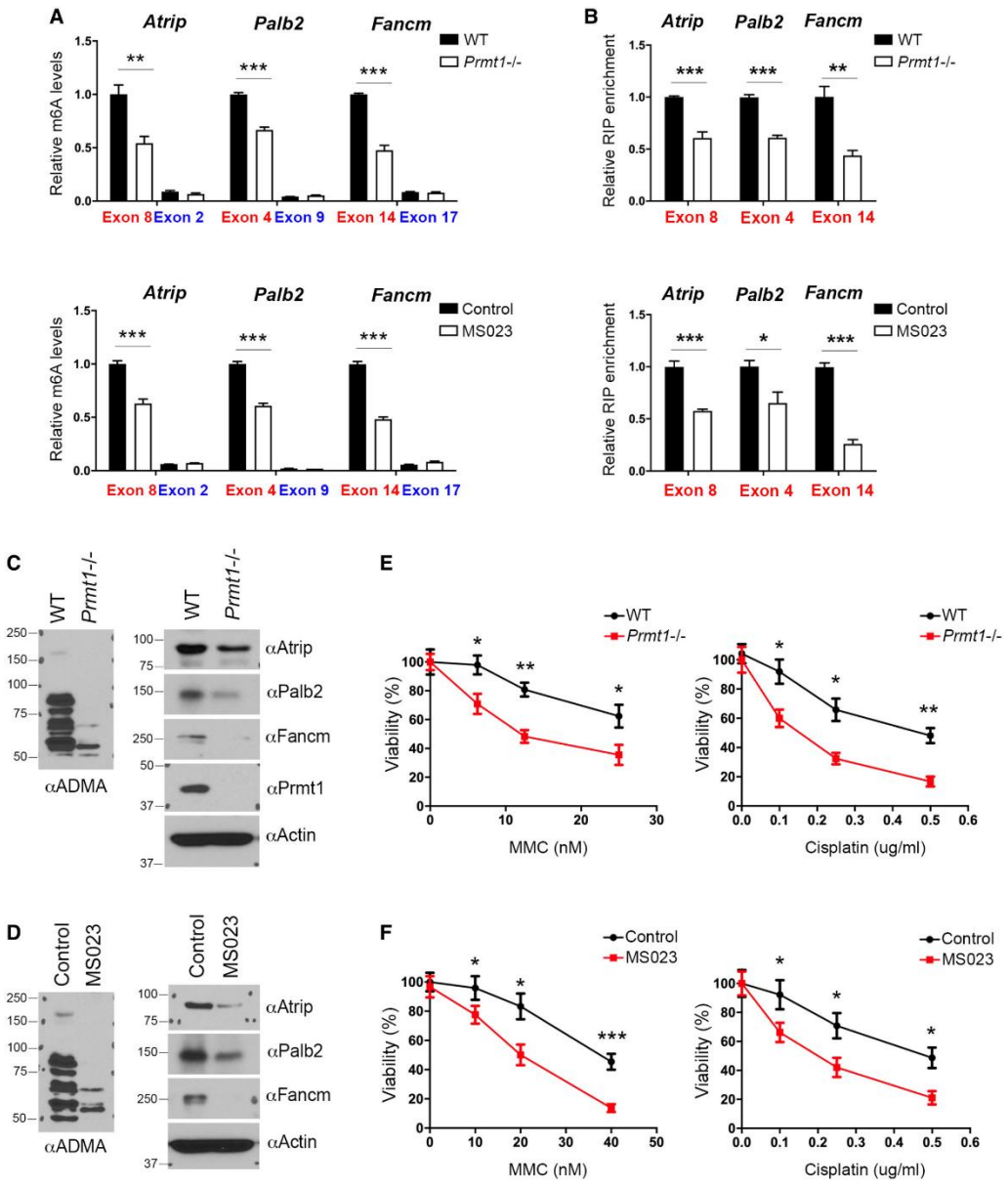


Supplementary Figure 5.17 MeRIP-seq (m6A-seq) analysis of mESCs expressing WT and arginine methylation-deficient METTL14.

(A) Three isogenic mESC lines were established by re-expressing WT (KO + WT), 5RK (KO + 5RK), and 13RK (KO + RK) mutant METTL14 in Mettl14 KO mESCs through lentivirus transduction. The expression of METTL14 in these cell lines was detected by Western blot analysis using an anti-METTL14 antibody. ACTIN was used as a loading control. **(B)** LC-

MS/MS was performed to quantify m6A levels (presented as the m6A/A ratio) in WT, Mettl14 KO, KO + WT, KO + 5RK, and KO + RK mESCs. The total RNA was extracted using the TRIzol reagent, and the poly(A) mRNA was purified for LC-MS/MS analysis. Data from three biological replicates were analyzed by Student's t-test and shown as mean ± SD. *P < 0.05; ***P < 0.001. **(C)** Summary of the numbers of m6A peaks in WT, Mettl14 KO, KO + WT, KO + 5RK, and KO + RK mESCs using different q-value cutoffs. **(D)** Principal component analysis (PCA) plot of m6A peaks in WT, Mettl14 KO, KO + WT, and KO + RK mESCs, each with three biological replicates. PC1 and PC2 are the top two principle components that explained the highest percentage of the variance. **(E)** Sequence motifs of m6A-enriched regions in WT and Mettl14 KO mESCs (upper panels). Topological distribution of normalized m6A peaks across the 5′UTR, CDS, and 3′UTR of mRNAs (lower panels). **(F)** Summary of the numbers of differential m6A peaks and corresponding numbers of genes for each pair of comparison among all established mESC cell lines. The differential m6A sites and the number of genes harboring these sites compared between KO + RK and KO + WT mESCs were highlighted in red. **(G)** Length comparison between internal exons harboring METTL14 arginine methylation-independent and -dependent m6A sites. Statistical analysis was performed using the Wilcoxon test to measure the median difference between the two groups. **(H)** Secondary structure prediction of RNA sequences harboring METTL14 arginine methylation-dependent and -independent m6A sites. Statistical analysis was performed using Fisher's exact test. **(I)** Gene expression level comparison of genes harboring METTL14 arginine methylation-independent and -dependent m6A sites in KO + RK and KO + WT mESCs. Statistical analysis was performed using the Wilcoxon test. **(J)** GO pathway analysis using EnrichR (Kuleshov et al, 2016)

reveals that genes harboring METTL14 arginine methylation-dependent m6A peaks are enriched for the Fanconi anemia pathway. Examples of analysis using two-pathway interaction annotation databases (KEGG 2019 mouse and NCI-Nature 2016) are shown. Statistical analysis was performed using Fisher exact test, as defined in EnrichR.



Supplementary Figure 5.18 METTL14 C-terminal IDR arginine methylation regulates m6A deposition on DNA repair genes.

(A) MeRIP (m6A-IP)-qPCR assays were performed for WT and Prmt1 KO mESCs (upper panel), as well as for control and MS023-treated mESCs (lower panel), to detect the impact of PRMT1 loss or inhibition on m6A deposition at targeted transcripts. m6A-negative regions of the transcripts (blue) were included as negative controls. **(B)** METTL14 RIP-

qPCR assays were performed for WT and Prmt1 KO mESCs (upper panel), as well as for control and MS023-treated mESCs (lower panel), to detect the impact of PRMT1 loss or inhibition on the interactions of METTL14 with targeted transcripts. **(C)** The expression of ICL repair genes is reduced in Prmt1 KO mESCs. Total cell lysates from WT and Prmt1 KO mESCs were subjected to Western blot analysis using the indicated antibodies. **(D)** The expression of ICL repair genes is reduced in MS023-treated mESCs. Total cell lysates from control and MS023-treated mESCs were subjected to Western blot analysis using the indicated antibodies. **(E)** Knockout of Prmt1 sensitizes mESCs to ICL damage. WT and Prmt1 KO mESCs were treated with various concentrations of MMC (left panel) or cisplatin (right panel) for 4 days before cell viability was measured. **(F)** Similar to (E), except that mESCs were treated with MS023, to inhibit type I PRMT activity, while they were treated with MMC (left panel) or cisplatin (right panel).

5.6 Tables

Supplementary Table 5.1 Primers used in this study

Primer Name	Primer sequence (5'-3')
Cloning primers	
GST-METTL14 Forward	CGGGATCCATGGATAGCCGCTTGC
GST-METTL14 Reverse	CCGCTCGAGTTATCGAGGTGGAAAG
GST-METTL14 (1-400) Forward	CGGGATCCATGGATAGCCGCTTGC
GST-METTL14 (1-400) Reverse	CCGCTCGAGTTAAGGCGATTTTGGTCG
3xFlag-METTL14 Forward	CCCAAGCTTATGGATAGCCGCTTGC
3xFlag-METTL14 Reverse	GGGGTACCTTATCGAGGTGGAAAG
3xFlag-METTL14 (1-400) Forward	CCCAAGCTTATGGATAGCCGCTTGC
3Flag-METTL14 (1-400) Reverse	GGGGTACCTTAAGGCGATTTTGGTCG
GFP-WTAP Forward	CCGCTCGAGCTATGACCAACGAAGAAC
GFP-WTAP Reverse	CGGGATCCTTACAAAACTGAACC
pLV-EF1a-IRES-Blast METTL14 Forward	CGGGATCCATGGACTACAAAGACCATGA
pLV-EF1a-IRES-Blast METTL14 Reverse	CGGAATTCTTATCGAGGTGGAAAG
HA-METTL14 Forward	CGGAATTCGGATGGATAGCCGCTTGC
HA-METTL14 Reverse	CCGCTCGAGTTATCGAGGTGGAAAG
METTL14 site mutagenesis primers	
METTL14 R408K-Forwrad	CAAATCTAAATCTGAC**AAA**GGAGGTGGAGCTCCC
METTL14 R408K-Reverse	GGGAGCTCCACCTCC**TTT**GTCAGATTTAGATTTG
METTL14 R414K/R418K-Forward	GGAGGTGGAGCTCCC**AAA**GGTGGAGGA**AAA**GGTGG AACTTCTGC
METTL14 R414K/R418K-Reverse	GCAGAAGTTCCACC**TTT**TCCTCCACC**TTT**GGGAGCTC CACCTCC
METTL14 R425K/R427K/R429K-Forward	GGAACTTCTGCTGGC**AAA**GGA**AAA**GAA**AAA**AATAGAT CTAACTTC
METTL14 R425K/R427K/R429K-Reverse	GAAGTTAGATCTATT**TTT**TTC**TTT**TCC**TTT**GCCAGCA GAAGTTCC
METTL14 R431K/R435K-Forward	GGACGAGAAAGAAAT**AAA**TCTAACTTC**AAA**GGAGAAA GAGGTGGC
METTL14 R431K/R435K-Reverse	GCCACCTCTTTCTCC**TTT**GAAGTTAGA**TTT**ATTTCTTT CTCGTCC
METTL14 R438K/R442K-	CTAACTTCCGAGGAGAA**AAA**GGTGGCTTT**AAA**GGGG GCCGTGGAGGAG

Forward	
METTL14 R438K/R442K-Reverse	CTCCTCCACGGCCCCC**TTT**AAAGCCACC**TTT**TTCTCCT CGGAAGTTAG
METTL14 R445K-Forward	GGCTTTAGAGGGGGC**AAA**GGAGGAGCACACAG
METTL14 R445K-Reverse	CTGTGTGCTCCTCC**TTT**GCCCCCTCTAAAGCC
METTL14 R450K-Forward	GTGGAGGAGCACAC**AAA**GGTGGCTTTCCACCTC
METTL14 R450K-Reverse	GAGGTGGAAAGCCACC**TTT**GTGTGCTCCTCCAC
METTL14 R456K-Forward	GGTGGCTTTCCACCT**AAA**TAAGGTACCAGTCG
METTL14 R456K-Reverse	CGACTGGTACCTTA**TTT**AGGTGGAAAGCCACC
RT-qPCR primers	
Atrip-Forward	CTCATAAGGTCCGCCGATTAG
Atrip-Reverse	CTGCTCAGAAGGTGACAAAGA
Blm-Forward	TGTGATTCATGCATCTCTTCCTAAA
Blm-Reverse	CAGCTCGGCCGGATTCT
Brca1-Forward	GGAGATGTTGTGACTGGAAGAA
Brca1-Reverse	GTGAAGGGCTCACAACAATAGA
Brca2-Forward	TCCCCCCTACCATCAGTTTG
Brca2-Reverse	CAGTGGTAGAGTTTGACTTCGTTCTT
Fancm-Forward	GGCAGAACGTGTCCAAGATTG
Fancm-Reverse	GCGGAGCCTTTTCTGATGTT
Palb2-Forward	CTGGTGATGACAGTGAAAAGCAA
Palb2-Reverse	CAGGCCAAGCATAGCTTTTATATCT
RIP-qPCR primers	
Atrip-Forward	ATCTTTAGCAGTGGGTGCTG
Atrip-Reverse	GGTCCAGACTTGTGCAGATAC
Blm-Forward	GGAAGATTTGCTGGCTGGAA
Blm-Reverse	ACGGCCAGGCTTCCTAT
Brca1-Forward	GCTAACTGTGTGCACTGTACT
Brca1-Reverse	GAGGGACGATTTGAGAGACATAC
Brca2-Forward	CAGTGAAACAAGAACTGATGAA
Brca2-Reverse	GATCACTCTCTCTTAGTTCCATTT
Fancm-Forward	TGTGTCTGGAAGGCATTCTG
Fancm-Reverse	GGGATTGGTGATATGGCTCTAC
Palb2-Forward	GAGGTGCGGGCTGATTT
Palb2-Reverse	CCAGGACCTGCTGGAAAG

5.7 References

1. Desrosiers, R., Friderici, K. & Rottman, F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc Natl Acad Sci U S A* **71**, 3971-5 (1974).
2. Adams, J.M. & Cory, S. Modified nucleosides and bizarre 5'-termini in mouse myeloma mRNA. *Nature* **255**, 28-33 (1975).
3. Dubin, D.T. & Taylor, R.H. The methylation state of poly A-containing messenger RNA from cultured hamster cells. *Nucleic Acids Res* **2**, 1653-68 (1975).
4. Yue, Y., Liu, J. & He, C. RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes Dev* **29**, 1343-55 (2015).
5. Roignant, J.Y. & Soller, M. m(6)A in mRNA: An Ancient Mechanism for Fine-Tuning Gene Expression. *Trends Genet* **33**, 380-390 (2017).
6. Roundtree, I.A., Evans, M.E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187-1200 (2017).
7. Wang, Y. *et al.* N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat Cell Biol* **16**, 191-8 (2014).
8. Liu, J. *et al.* A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat Chem Biol* **10**, 93-5 (2014).
9. Ping, X.L. *et al.* Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Res* **24**, 177-89 (2014).
10. Jia, G. *et al.* N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol* **7**, 885-7 (2011).
11. Zheng, G. *et al.* ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol Cell* **49**, 18-29 (2013).
12. Shi, H., Wei, J. & He, C. Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. *Mol Cell* **74**, 640-650 (2019).
13. Zaccara, S., Ries, R.J. & Jaffrey, S.R. Reading, writing and erasing mRNA methylation. *Nat Rev Mol Cell Biol* **20**, 608-624 (2019).
14. Wang, X. *et al.* N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* **161**, 1388-99 (2015).
15. Li, A. *et al.* Cytoplasmic m(6)A reader YTHDF3 promotes mRNA translation. *Cell Res* **27**, 444-447 (2017).
16. Wang, X. *et al.* N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117-20 (2014).
17. Jaffrey, S.R. & Kharas, M.G. Emerging links between m(6)A and misregulated mRNA methylation in cancer. *Genome Med* **9**, 2 (2017).

18. Noack, F. & Calegari, F. Epitranscriptomics: A New Regulatory Mechanism of Brain Development and Function. *Front Neurosci* **12**, 85 (2018).

19. Batista, P.J. *et al.* m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* **15**, 707-19 (2014).

20. Geula, S. *et al.* Stem cells. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. *Science* **347**, 1002-6 (2015).

21. Yoon, K.J. *et al.* Temporal Control of Mammalian Cortical Neurogenesis by m(6)A Methylation. *Cell* **171**, 877-889 e17 (2017).

22. Wang, Y. *et al.* N(6)-methyladenosine RNA modification regulates embryonic neural stem cell self-renewal through histone modifications. *Nat Neurosci* **21**, 195-206 (2018).

23. Vu, L.P. *et al.* The N(6)-methyladenosine (m(6)A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nat Med* **23**, 1369-1376 (2017).

24. Barbieri, I. *et al.* Promoter-bound METTL3 maintains myeloid leukaemia by m(6)A-dependent translation control. *Nature* **552**, 126-131 (2017).

25. Weng, H. *et al.* METTL14 Inhibits Hematopoietic Stem/Progenitor Differentiation and Promotes Leukemogenesis via mRNA m(6)A Modification. *Cell Stem Cell* **22**, 191-205 e9 (2018).

26. Zhang, C. *et al.* Hypoxia induces the breast cancer stem cell phenotype by HIF-dependent and ALKBH5-mediated m(6)A-demethylation of NANOG mRNA. *Proc Natl Acad Sci U S A* **113**, E2047-56 (2016).

27. Bedford, M.T. & Clarke, S.G. Protein arginine methylation in mammals: who, what, and why. *Mol Cell* **33**, 1-13 (2009).

28. Yang, Y. & Bedford, M.T. Protein arginine methyltransferases and cancer. *Nat Rev Cancer* **13**, 37-50 (2013).

29. Blanc, R.S. & Richard, S. Arginine Methylation: The Coming of Age. *Mol Cell* **65**, 8-24 (2017).

30. Boisvert, F.M., Cote, J., Boulanger, M.C. & Richard, S. A proteomic analysis of arginine-methylated protein complexes. *Mol Cell Proteomics* **2**, 1319-30 (2003).

31. Guo, A. *et al.* Immunoaffinity enrichment and mass spectrometry analysis of protein methylation. *Mol Cell Proteomics* **13**, 372-87 (2014).

32. Geoghegan, V., Guo, A., Trudgian, D., Thomas, B. & Acuto, O. Comprehensive identification of arginine methylation in primary T cells reveals regulatory roles in cell signalling. *Nat Commun* **6**, 6758 (2015).

33. Thandapani, P., O'Connor, T.R., Bailey, T.L. & Richard, S. Defining the RGG/RG motif. *Mol Cell* **50**, 613-23 (2013).

34. Rajyaguru, P. & Parker, R. RGG motif proteins: modulators of mRNA functional states. *Cell Cycle* **11**, 2594-9 (2012).

35. Ozdilek, B.A. *et al.* Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. *Nucleic Acids Res* **45**, 7984-7996 (2017).

36. Chong, P.A., Vernon, R.M. & Forman-Kay, J.D. RGG/RG Motif Regions in RNA Binding and Phase Separation. *J Mol Biol* **430**, 4650-4665 (2018).

37. Tripsianes, K. *et al.* Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins. *Nat Struct Mol Biol* **18**, 1414-20 (2011).

38. Scholler, E. *et al.* Interactions, localization, and phosphorylation of the m(6)A generating METTL3-METTL14-WTAP complex. *RNA* **24**, 499-512 (2018).

39. Eram, M.S. *et al.* A Potent, Selective, and Cell-Active Inhibitor of Human Type I Protein Arginine Methyltransferases. *ACS Chem Biol* **11**, 772-781 (2016).

40. Sledz, P. & Jinek, M. Structural insights into the molecular mechanism of the m(6)A writer complex. *Elife* **5**(2016).

41. Wang, P., Doxtader, K.A. & Nam, Y. Structural Basis for Cooperative Function of Mettl3 and Mettl14 Methyltransferases. *Mol Cell* **63**, 306-317 (2016).

42. Wang, X. *et al.* Structural basis of N(6)-adenosine methylation by the METTL3-METTL14 complex. *Nature* **534**, 575-8 (2016).

43. Huang, H. *et al.* Histone H3 trimethylation at lysine 36 guides m(6)A RNA modification co-transcriptionally. *Nature* **567**, 414-419 (2019).

44. Slobodin, B. *et al.* Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation. *Cell* **169**, 326-337 e12 (2017).

45. Ke, S. *et al.* m(6)A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes Dev* **31**, 990-1006 (2017).

46. Zhang, Z. & Xing, Y. CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res* **45**, 9260-9271 (2017).

47. Aguilo, F. *et al.* Coordination of m(6)A mRNA Methylation and Gene Transcription by ZFP217 Regulates Pluripotency and Reprogramming. *Cell Stem Cell* **17**, 689-704 (2015).

48. Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201-6 (2012).

49. Pawlak, M.R., Scherer, C.A., Chen, J., Roshon, M.J. & Ruley, H.E. Arginine N-methyltransferase 1 is required for early postimplantation mouse development, but cells deficient in the enzyme are viable. *Mol Cell Biol* **20**, 4859-69 (2000).

50. Deans, A.J. & West, S.C. DNA interstrand crosslink repair and cancer. *Nat Rev Cancer* **11**, 467-80 (2011).

51. Oldfield, C.J. & Dunker, A.K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83**, 553-84 (2014).

52. Protter, D.S.W. *et al.* Intrinsically Disordered Regions Can Contribute Promiscuous Interactions to RNP Granule Assembly. *Cell Rep* **22**, 1401-1412 (2018).

53. Gibson, B.A. *et al.* Organization of Chromatin by Intrinsic and Regulated Phase Separation. *Cell* **179**, 470-484 e21 (2019).

54. Zhang, Y. & Kutateladze, T.G. Liquid-liquid phase separation is an intrinsic physicochemical property of chromatin. *Nat Struct Mol Biol* **26**, 1085-1086 (2019).

55. Strom, A.R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547**, 241-245 (2017).

56. Sabari, B.R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**(2018).

57. Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855 e16 (2018).

58. Cho, W.K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412-415 (2018).

59. Pessina, F. *et al.* Functional transcription promoters at DNA double-strand breaks mediate RNA-driven phase separation of damage-response factors. *Nat Cell Biol* **21**, 1286-1299 (2019).

60. Kilic, S. *et al.* Phase separation of 53BP1 determines liquid-like behavior of DNA repair compartments. *EMBO J* **38**, e101379 (2019).

61. Boeynaems, S. *et al.* Phase Separation of C9orf72 Dipeptide Repeats Perturbs Stress Granule Dynamics. *Mol Cell* **65**, 1044-1055 e5 (2017).

62. Molliex, A. *et al.* Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123-33 (2015).

63. Sims, R.J., 3rd *et al.* The C-terminal domain of RNA polymerase II is modified by site-specific methylation. *Science* **332**, 99-103 (2011).

64. Yang, Y. *et al.* Arginine methylation facilitates the recruitment of TOP3B to chromatin to prevent R loop accumulation. *Mol Cell* **53**, 484-97 (2014).

65. Zhao, D.Y. *et al.* SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination. *Nature* **529**, 48-53 (2016).

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

67. Louloupi, A., Ntini, E., Conrad, T. & Orom, U.A.V. Transient N-6-Methyladenosine Transcriptome Sequencing Reveals a Regulatory Role of m6A in Splicing Efficiency. *Cell Rep* **23**, 3429-3437 (2018).

68. Kwak, Y.T. *et al.* Methylation of SPT5 regulates its interaction with RNA polymerase II and transcriptional elongation properties. *Mol Cell* **11**, 1055-66 (2003).

69. Mao, Y. *et al.* m(6)A in mRNA coding regions promotes translation via the RNA helicase-containing YTHDC2. *Nat Commun* **10**, 5332 (2019).

70. Xiang, Y. *et al.* RNA m(6)A methylation regulates the ultraviolet-induced DNA damage response. *Nature* **543**, 573-576 (2017).

71. Zhang, C. *et al.* METTL3 and N6-Methyladenosine Promote Homologous Recombination-Mediated Repair of DSBs by Modulating DNA-RNA Hybrid Accumulation. *Mol Cell* **79**, 425-442 e7 (2020).

72. Musiani, D. *et al.* PRMT1 Is Recruited via DNA-PK to Chromatin Where It Sustains the Senescence-Associated Secretory Phenotype in Response to Cisplatin. *Cell Rep* **30**, 1208-1222 e9 (2020).

73. Guccione, E. & Richard, S. The regulation, functions and clinical relevance of arginine methylation. *Nat Rev Mol Cell Biol* **20**, 642-657 (2019).

74. Huang, L., Wang, Z., Narayanan, N. & Yang, Y. Arginine methylation of the C-terminus RGG motif promotes TOP3B topoisomerase activity and stress granule localization. *Nucleic Acids Res* **46**, 3061-3074 (2018).

75. Yang, Y. *et al.* PRMT9 is a type II methyltransferase that methylates the splicing factor SAP145. *Nat Commun* **6**, 6428 (2015).

76. Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N. & Rechavi, G. Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing. *Nat Protoc* **8**, 176-89 (2013).

77. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-7 (2016).

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

79. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).

80. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).

6 CONCLUDING REMARKS

Since the initiation of the Human Genome Project (HGP), our knowledge of how genes are regulated has been drastically expanded. From studying a limited number of protein-coding genes to deciphering regulatory elements in noncoding regions over the past decade, researchers have made tremendous findings and achievements related to the in-depth understanding of the dynamic and complex regulations of genes. With the advent of high-throughput sequencing technologies, multi-omics data including epigenomics, transcriptomics, proteomics, and metabolomics has proved to be invaluable for gaining biological insights into molecular phenotypes of regulatory elements. Immense efforts from consortia provide enormous datasets covering many aspects of biological processes at an unprecedented scale and resolution.

Continuous and rapid accumulation of sequencing datasets has brought both opportunities and challenges to researchers for studying multicellular complexity. Developing computational approaches to interrogate the large amount of data has become an urgent need. Machine learning shows a promising power for extracting biological

knowledge from big data compendia. This dissertation has been focused on developing machine learning strategies to study alternative splicing, a crucial gene regulatory mechanism. Alternative splicing is a major source of transcriptome diversity. The defects of alternative splicing are frequently observed and overlooked in human diseases and cancers. Our incomplete understanding of regulatory elements that govern alternative splicing limits the ability to interpret the functional consequences of splice-altering variants and support precision medicine. To understand the regulatory mechanisms of and the effects of genetic variants on alternative splicing, machine learning strategies were developed by leveraging large-scale RNA sequencing datasets across diverse biological conditions in the following Chapters.

In Chapter 2, we developed a Deep-learning Augmented RNA-seq analysis of Transcript Splicing (DARTS) framework by utilizing large-scale publicly available RNA-seq datasets to infer differential splicing between biological conditions. A key feature for DARTS is that it allows the discoveries of differential splicing when sequencing data is shallow or the target gene is lowly expressed. Unlike previous computational tools that only use RNA sequence features to predict splicing, DARTS adds an additional layer by introducing trans RBP levels in the framework. The trans RBP levels inherently characterize biological condition-specificity. With the trans RBP features, DARTS can be easily extended and generalized to diverse biological systems.

In Chapter 3, we developed a computational tool, Systematic Investigation of Retained Introns (SIRI), that reliably quantifies intron retention levels as well as a deep-learning-based computational approach that predicts intron retention regulatory patterns at the subcellular level. Conventional transcriptome sequencing captures RNA molecules in

whole cells, thus ignoring subcellular distributions of processed and unprocessed transcripts. We generated extensive RNA-seq datasets at subcellular level and demonstrated that polyadenylated RNA abundance does not indicate functional gene expression from the analysis of switching intron regulatory patterns across cell development. These findings recommend future directions of designing subcellular transcriptome analyses towards more profound biological discoveries. We expect that SIRI coupled with the deep-learning-based computational approach will contribute to new discoveries of functional elements that determine subcellular-specific regulations of introns under various biological environments.

In Chapter 4, motivated by the success of DARTS, we developed a deep-learning-based framework, individualized Deep-learning Analysis of RNA Transcript Splicing (iDARTS), for predicting tissue-specific splicing levels. An inherent limitation of DARTS is that it could not make quantitative predictions of alternative splicing in biological samples. Therefore, we extended the framework of DARTS to iDARTS that directly models the cis elements and trans RBPs determinants of alternative splicing in tissues. iDARTS shows accurate, robust, and generalizable behaviours in predicting splicing levels in tissues. The unidirectional flow from genomic sequence and trans RBP levels to splicing makes iDARTS capable of inferring causality by measuring the effects of variants on splicing, enabling a broader application in genetic and clinical studies. A potential future improvement will be to incorporate additional co-transcriptional regulations including chromatin marks, transcriptional factors, RBP binding profiles, and RNA modifications in the iDARTS framework.

In Chapter 5, we studied the regulation of N6-methyladenosine (m^6A) modification through investigating the functionalities of arginine-methylation of METTL14 on m^6A. We found that methylation deficient METTL14 negatively impact m^6A levels globally. These arginine methylation-dependent m^6A sites are predicted to show preferences of RNA secondary structures such as helix/stem or multi-branched loops. As m^6A is reported to occur co-transcriptional and impact on splicing, future works will be expected to focus on utilizing machine learning strategies to dissect the regulatory elements underlying m^6A, thereby helping to further understand the regulation of alternative splicing via RNA modifications.

In the long run, we expect leveraging machine learning strategies to extract biological knowledge from multi-omics datasets will be routinely conducted. Emerging experimental approaches, such as the third generation long-reads sequencing, and single-cell sequencing have opened new perspectives on the dynamic and complex regulations of genes at single-cell resolution. The ever-growing size of the datasets that profile various aspects of biological regulations necessitate the development of machine learning strategies to analyse, interrogate, and integrate these large-scale datasets for comprehensively and systematically characterizing biological mechanisms. With more and more diverse datasets generated, machine learning approaches will become more and more accurate in modelling biological complexities. We anticipate the transformed biological knowledge from large-scale datasets via machine learning will eventually benefit clinical studies and precision medicine.