UNIVERSITY OF CALIFORNIA

Los Angeles

# COMPUTATIONAL APPROACHES TO STUDY SPLICING REGULATION IN DEVELOPMENT AND DISEASE

A dissertation submitted in partial satisfaction

of the requirement for the degree

Doctor of Philosophy in Bioinformatics

by

Yuanyuan Wang

2021

ABSTRACT OF THE DISSERTATION

# COMPUTATIONAL APPROACHES TO STUDY SPLICING REGULATION IN DEVELOPMENT AND DISEASE

by

Yuanyuan Wang

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2021

Professor Yi Xing, Chair

Alternative splicing is an elaborately regulated co-/post-transcriptional process that dramatically expands the diversity and complexity of the eukaryotic transcriptome and proteome. A coordinated cell type-specific alternative splicing network is essential for cell-fate determination and tissue-identity acquisition. Defects in splicing machinery, including the cis-acting elements and trans-acting factors, can result in extensive aberrant splicing, which has been implicated in a wide range of human diseases, especially cancers and neurological disorders. Large-scale RNA sequencing (RNA-seq) data accumulated in public repositories or generated by consortium projects provide an unprecedented resource for more comprehensive elucidation on splicing regulation in development and disease. In the

meantime, it also posed new challenges for the development of computational tools on faster profiling and more precise interpretation of the alternative splicing.

The first part of the dissertation presents rMATS-turbo, referring to rMATS 4.0.1 or above, an ultra-fast computational tool for alternative splicing analysis in a time- and memory-efficient manner. We provide two major application scenarios of rMATS-turbo to demonstrate its capability for straightforward and fast splicing analysis. Firstly, we described a single-command differential splicing analysis between two cell lines, yielding a robust identification of splicing alterations, including those derived from novel splice sites. Secondly, we demonstrated the workflow for comprehensive profiling of splicing landscape using 1,019 RNA-seq datasets (18.58 T base) from the Cancer Cell Line Encyclopedia. Benchmarks of time and memory consumption revealed that rMATS-turbo still performs well even with increasing read depth or sample size. These results illustrated the ultra-fast nature of rMATS-turbo, which makes it a useful tool for splicing analysis on large-scale RNA-seq data.

In the second and third parts of the dissertation, we exploited rMATS-turbo and other computational approaches to study the dynamics and regulation of splicing in tissue development and disease. In the second part, we sought to evaluate how alternative splicing, under the control of RNA binding proteins (RBPs), affects cell fate commitment during induced osteogenic differentiation of human bone marrow-derived multipotent stem/stromal progenitor cells (MSPCs). Our analysis revealed temporal coordination between widespread alternative splicing changes and RBP expression alterations. We also developed a new computational platform to screen key RBPs during development using time-course RNA-seq data. Nine RBPs were identified as potential key splicing regulators

during osteogenic differentiation. Perturbation of two candidate RBPs, KHDRBS3 and CPEB2 inhibited MSPC osteogenesis in vitro, validating our computational prediction of "driver" RBPs.

In the third part of the dissertation, inspired by previous studies implying a linkage of PRMT9 with splicing and brain development, we aimed to unravel the direct molecular, cellular, and pathological contributions of PRMT9 on neurological disorders. First, we showed that the autosomal recessive intellectual disability-associated variant, PRMT9 G189R, cannot catalyze SF3B2 methylation on R508 (R508me2s) and is extremely unstable. We also demonstrated that Prmt9 conditional KO in excitatory neurons resulted in impairment of learning, memory, and maturation of functional synapses in mice. The transcriptomic analysis discovered widespread splicing alterations, but no steady-state gene expression changes in KO mice, which indicates that alternative splicing independently defines the brain-specific transcriptome in Prmt9 KO mice. Moreover, genes with splicing changes were enriched in neuron- and synapse-related pathways. All of those findings indicated a PRMT9-SF3B2-splicing-synapse regulatory cascade linking PRMT9 with brain development. Finally, a working model was proposed that PRMT9-mediated SF3B2 R508me2s regulates splicing through 3' splice site competition by altering SF3B2/pre-mRNA interaction. Overall, this work clarified the molecular, cellular, and functional contributions of PRMT9 and also deepened our insights into the splicing regulations in the pathogenesis of intellectual disability and related disorders.

The dissertation of Yuanyuan Wang is approved.

John S. Adams

Douglas L. Black

Linda M. Liau

Yi Xing, Committee Chair

University of California, Los Angeles

2021

To Mom, Dad, Grandfather and Grandmother

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES AND BOXES

# ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my thesis advisor, Yi Xing, who has provided me with enormous support and guidance during my Ph.D. training. His expertise and broad knowledge have inspired me to view, conduct and present my research more comprehensively and rigorously. I joined the lab with a background in molecular biology and learned from him the main topics of the field, the analytical skills, writing, and everything I need to know to become a computational biologist.

I would also like to thank all my committee members, John S. Adams, Douglas L. Black, and Linda M. Liau for their support, encouragement, and insightful advice. I am very lucky and grateful to have them on my committee.

I would specifically like to convey my gratitude to my collaborators, which made tremendous contribution to the projects included in my dissertation, including Rene F. Chun, Mason Henrich, and John S. Adams from UCLA, Lei Shen and Yanzhong (Frankie) Yang from the City of Hope, Shenfeng Qiu from the University of Arizona, as well as Hao Chen and Yang Shi from Harvard University.

I would also like to thank my friends and colleagues in Xing lab for making this journey full of joy. My Ph.D. life would be less colorful without them.

Last but not least, I would like to thank my parents and grandparents for their unconditional love and support during my entire life. It is them who teach me that I am the owner of my life, and it is me who can decide what kind of person I want to be. I love them forever.

# 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, LIFESCI 4 - GENETICS
University of California, Los Angeles 2017

Exchange Student, Molecular & Medical Pharmacology
University of California, Los Angeles 2014-2015

B.S. in Biological Sciences, College of Life Sciences
Nanjing University 2011-2015

## HONORS AND AWARDS

Biomedical Big Data (BBD) Training Grant, University of California, Los Angeles 2017

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

National Encouragement Scholarship, China Ministry of Education 2013-2014

Asia Regional Bronze Medal, The iGEM competition 2013

National Scholarship, China Ministry of Education 2012

## SELECTED PUBLICATIONS/MANUSCRIPTS

* co-first author

1. **Wang Y**, et al. rMATS-turbo: an ultra-fast computational platform for comprehensive alternative splicing analysis of large-scale RNA-seq data. In preparation

2. **Lei S*, Wang Y***, et al. PRMT9 affects neuron development by regulating splicing through SF3B2 R508 methylation. In preparation

3. Henrich M*, Ha P*, **Wang Y***, Adams JS, Soo C, Ting K, Stodeick L, Chun R. (2021) Alternative splicing regulates the physiological adaptation of the mouse hind limb postural and phasic muscles to microgravity. In submission

4. Chen H, Gu L, Orellana EA, **Wang Y**, Guo J, Liu Q, Wang L, Shen Z, Wu H, Gregory RI, Xing Y. (2020) METTL4 is an snRNA m6Am methyltransferase that regulates RNA splicing. Cell Research, 30(6):544-547

5. **Wang Y**, Chun RF, Adhikari S, Lopez CM, Henrich M, Yacoubian V, Lin L, Adams JS, Xing Y. (2020) Elucidating dynamics and regulation of alternative splicing in osteogenic differentiation. bioRxiv, 843268

6. Pan Y, Lee AH, Yang HT, **Wang Y**, Xu Y, Kadash-Edmondson, K, Phillips J, Champhekar A, Puig C, Ribas A, Witte ON, Prins RM, Xing Y. (2019) IRIS: Big data-informed discovery of

cancer immunotherapy targets arising from pre-mRNA alternative splicing. bioRxiv, 843268

7. Shen S, **Wang Y**, Wang C, Wu YN, Xing Y. (2016) SURVIV: Survival analysis of mRNA isoform variation. Nature Communications, 7:11548

8. Wu NC, Du Y, Le S, Young AP, Zhang TH, **Wang Y**, Zhou J, Yoshizawa JM, Dong L, Li X, Wu TT, Sun R. (2016) Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza A virus M segment. BMC genomics. Dec;17(1):1-5

# 1 INTRODUCTION

The generic information coded by DNAs can be passed to RNA molecules though transcription, which acts as the molecular templates for the synthesis of proteins. However, the number of protein-coding genes is extremely low compared to the diversity of proteome in eukaryotic cells. The proteomic complexity is achieved by both co-/post-transcriptional processing of the precursor mRNA (pre-mRNA), and the post-translational modifications of protein products. A ubiquitous and pivotal step of nascent RNA processing is alternative splicing, an essential biological process that generates a vast and diverse RNA population by selective intron removal and exon joining [1]. Indeed, genome-wide studies estimated that 90-95% of human genes undergo alternative splicing.

RNA splicing is elaborately coordinated by context-dependent interaction between cis-acting elements and trans-acting factors [2-5]. The core cis-acting elements strictly required by the spliceosome consist of 5' and 3' splice site, branch point site, and polypyrimidine track upstream of the 3' splice site [2,6] on the pre-mRNA. Other cis-acting elements comprise splicing regulatory elements (SREs) on pre-mRNA, including exonic

splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs). Trans-acting factors can be categorized into two major types: 1) RNA components of spliceosome (small nuclear RNAs, snRNAs) that base-pairing with cis-acting elements. For example, U1 snRNA and U2 snRNA can be anchored to 5' splice site and branch point site adjacent 3' splice site; 2) protein components of spliceosome complex and other RNA binding proteins (RBPs) that selectively binds to the SREs.

It is now well established that temporal control of alternative splicing is critical for cell-fate determination and tissue-identity acquisition [7]. Dysregulation of the splicing networks has also been reported to contributes to the pathogenesis of diseases [8,9], especially cancers [10] and neurological disorders [11,12], yielding a massive inventory for potential therapeutic targets [13-15]. Splicing alterations underlying disease progression often involve changes in splicing regulatory machinery, including both cis-acting elements and trans-acting factors. In fact, population-scale transcriptomic studies have shown that about one-third of disease-causing genetic mutations are related to mis-splicing, for example, the creation of cryptic splice sites [16]. The physiological and pathological contributions of mutations in cis-acting elements are usually mediated by splicing changes in individual gene [17], while defects in trans-acting factors usually induce transcriptome-wide alternations in splicing. Trans-acting snRNAs has been reported to be recurrently mutated in multiple cancer types [18,19] as well as neurodegenerative disease [20], which lead to splicing alterations with an excessive usage of cryptic splice sites [18,19]. Besides, internal m6Am modifications of U2 snRNA can also induce splicing changes [21,22]. As for the protein component of trans-acting factors, cancer-associated hotspot mutations in spliceosomal

proteins or RBPs has been identified and proved to generate aberrant splicing [23,24], which can be targeted by pharmacological modulation of the spliceosome [25]. Variation of protein abundance of splicing factors also plays an important role in development [26] or disease [27]. Last but not least, post-translational modifications of splicing factors become an active research topic for splicing [28,29] and splicing-related clinical investigations [30].

To date, there are a number of computational tools developed for the detection, quantification, and differential analysis of alternative splicing from RNA sequencing (RNA-seq) data [31-38]. A common challenge for most of those computational tools is that they are computationally intensive, which limits their application on large-/population-scale dataset. In the meantime, due to the advances in sequencing technologies and reduction of sequencing cost, enormous RNA-seq dataset have been generated, leading to a rapid accumulation of RNA-seq data in public repositories, such as Sequence Read Archive (SRA) [39]. Moreover, the emerge of large-scale consortium-based studies also provides unprecedented resources for population-scale alternative splicing analysis (for example, TCGA [40], ENCODE [41], GTEx [42] and CCLE [43]). This further poses computational burdens on alternative splicing analysis, and urges the development of computationally efficient tools for comprehensive profiling of splicing landscape from massive RNA-seq datasets.

As an effort to resolve this challenge, in Chapter 2 of this dissertation, we describe the rMATS-turbo, referring to ultra-fast version rMATS 4.0.1 or above, which exhibits dramatic improvement in processing speed and storage efficiency for alternative splicing analysis. We provide two major application scenarios of rMATS-turbo to demonstrate its capabilities as well as benchmark its time and memory usage. Firstly, we performed a general single-command two-group comparison between two cell lines, which identified

robust differential splicing changes, including events derived from cryptic splice sites. Secondly, we demonstrated the workflow for parallel processing of large-scale datasets in a time- and memory-efficient manner using 1,019 RNA-seq datasets (18.58 T base) from the Cancer Cell Line Encyclopedia. Both examples were performed on a shared high-performance computing cluster with a Unix-based operating system using a single thread. The first example takes ~6 h with 11 GB peak RAM. The second example takes ~ 3 days when running the prep steps in parallel (24 GB RAM), which illustrated the ultra-fast processing speed of rMATS-turbo to perform splicing analysis using large-scale RNA-seq data.

In Chapter 3 and Chapter 4 of the dissertation, we applied rMATS-turbo on various datasets to elucidate the dynamics and regulation of alternative splicing under normal tissue development or disease scenario. In these two sections, we demonstrated the regulation of splicing by temporal expression of trans-acting RBPs and post-translational modification of trans-acting spliceosomal proteins, respectively. In Chapter 3, We sought to evaluate how AS, under the control of RBPs, affects cell fate commitment during induced osteogenic differentiation of human bone marrow-derived multipotent stem/stromal progenitor cells (MSPCs). We generated a time-course RNA sequencing (RNA-seq) dataset representative of induced MSPC differentiation to osteoblasts. Our analysis utilizing rMATS-turbo revealed widespread AS changes coordinated with differential RBP expression at multiple time points, including many AS changes in non-differentially expressed genes. We also developed a computational platform to identify key splicing regulators of alternative splicing during osteogenic differentiation using time-course RNA-seq data, which takes into account the temporal patterns of exon skipping and RBP

expression as well as RBP binding in the vicinity of regulated exons. In total, we identified nine RBPs as potential key splicing regulators during MSPC osteogenic differentiation. Perturbation of two candidate RBP genes, KHDRBS3 and CPEB2, by siRNA knockdown, inhibited MSPC osteogenesis in vitro, validating our computational prediction of "driver" RBPs. Overall, this work highlighted a high degree of complexity in the splicing regulation during osteogenic differentiation. Our computational approach may be applied to other time-course RNA-seq data to explore dynamic regulation of alternative splicing by RBPs in other biological processes or disease trajectories.

In Chapter 4, we aimed to measure the enzymatic activity of PRMT9 G189R mutant, which is previously proved causal in autosomal recessive intellectual disability disease, and to decipher the linkage between PRMT9 and brain-related functions. Both in vitro and in vivo methylation assays showed that wild-type but not the G189R mutant PRMT9 can catalyze the symmetric dimethylarginine of SF3B2 at R508. G189R mutant PRMT9 has also been proved unstable, as demonstrated by significantly shortened protein half-life. To dissect the behavior and cellular consequences of Prmt9 depletion in excitatory neurons, Prmt9 conditional knockout (cKO) mice were bred, which exhibited impaired learning, memory and formation of functional synapses. RNA-seq samples were extracted from hippocampus tissue of two-week-old wild-type or whole body Prmt9 KO mice and subject to RNA-seq. Our transcriptomic analysis discovered wide-spread splicing alterations, but no steady-state gene expression changes in KO samples. Also, we revealed that genes with splicing changes were enriched in neuron- and synapse-related pathways. All of those findings indicated a PRMT9-SF3B2-splicing-synapse regulatory cascade linking PRMT9 with brain development. Finally, sequence feature comparisons suggested that

PRMT9-mediated SF3B2 R508me2s regulates splicing through 3' splice site competition by altering SF3B2/pre-mRNA interaction. The CLIP-qPCR measurement of SF3B2 interaction with 3' splice site sequences in upstream and downstream of differentially spliced exons supported this model. Overall, this work clarified the molecular, cellular and functional contributions of PRMT9, which deepened our insights into the pathogenesis of intellectual disability and related disorders.

# References


1 Sharp, P. A. Split genes and RNA splicing. Cell **77**, 805-815, doi:10.1016/0092-8674(94)90130-9 (1994).

2 Fu, X. D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet **15**, 689-701, doi:10.1038/nrg3778 (2014).

3 Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. Nature **583**, 711-719, doi:10.1038/s41586-020-2077-3 (2020).

4 Matera, A. G. & Wang, Z. A day in the life of the spliceosome. Nat Rev Mol Cell Biol **15**, 108-121, doi:10.1038/nrm3742 (2014).

5 Shi, Y. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nat Rev Mol Cell Biol **18**, 655-670, doi:10.1038/nrm.2017.86 (2017).

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, doi:10.1016/j.ajhg.2017.11.002 (2018).

7 Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol **18**, 437-451, doi:10.1038/nrm.2017.27 (2017).

8 Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. Nat Rev Genet **17**, 19-32, doi:10.1038/nrg.2015.3 (2016).

9 Montes, M., Sanford, B. L., Comiskey, D. F. & Chandler, D. S. RNA Splicing and Disease: Animal Models to Therapies. Trends Genet **35**, 68-87, doi:10.1016/j.tig.2018.10.002 (2019).

10 Rahman, M. A., Krainer, A. R. & Abdel-Wahab, O. SnapShot: Splicing Alterations in Cancer. Cell **180**, 208-208 e201, doi:10.1016/j.cell.2019.12.011 (2020).

11 Raj, B. & Blencowe, B. J. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron **87**, 14-27, doi:10.1016/j.neuron.2015.05.004 (2015).

12 Vuong, C. K., Black, D. L. & Zheng, S. The neurogenetics of alternative splicing. Nat Rev Neurosci **17**, 265-281, doi:10.1038/nrn.2016.27 (2016).

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

14 Bonnal, S. C., Lopez-Oreja, I. & Valcarcel, J. Roles and mechanisms of alternative splicing in cancer - implications for care. Nat Rev Clin Oncol **17**, 457-474, doi:10.1038/s41571-020-0350-x (2020).

15 Pan, Y. et al. RNA Dysregulation: An Expanding Source of Cancer Immunotherapy Targets. Trends Pharmacol Sci **42**, 268-282, doi:10.1016/j.tips.2021.01.006 (2021).

16 Vaz-Drago, R. et al. Transcription-coupled RNA surveillance in human genetic diseases caused by splice site mutations. Hum Mol Genet **24**, 2784-2795, doi:10.1093/hmg/ddv039 (2015).

17 Valentonyte, R. et al. Sarcoidosis is associated with a truncating splice site mutation in BTNL2. Nat Genet **37**, 357-364, doi:10.1038/ng1519 (2005).

18 Suzuki, H. et al. Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. Nature **574**, 707-711, doi:10.1038/s41586-019-1650-0 (2019).

19 Shuai, S. et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. Nature **574**, 712-716, doi:10.1038/s41586-019-1651-z (2019).

20 Jia, Y., Mu, J. C. & Ackerman, S. L. Mutation of a U2 snRNA gene causes global disruption of alternative splicing and neurodegeneration. Cell **148**, 296-308, doi:10.1016/j.cell.2011.11.057 (2012).

21 Goh, Y. T., Koh, C. W. Q., Sim, D. Y., Roca, X. & Goh, W. S. S. METTL4 catalyzes m6Am methylation in U2 snRNA to regulate pre-mRNA splicing. Nucleic Acids Res **48**, 9250-9261, doi:10.1093/nar/gkaa684 (2020).

22 Chen, H. et al. METTL4 is an snRNA m(6)Am methyltransferase that regulates RNA splicing. Cell Res **30**, 544-547, doi:10.1038/s41422-019-0270-4 (2020).

23 Darman, R. B. et al. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. Cell Rep **13**, 1033-1045, doi:10.1016/j.celrep.2015.09.053 (2015).

24 Zhang, J. et al. Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. Mol Cell **76**, 82-95 e87, doi:10.1016/j.molcel.2019.07.017 (2019).

25 Shirai, C. L. et al. Mutant U2AF1-expressing cells are sensitive to pharmacological modulation of the spliceosome. Nat Commun **8**, 14060, doi:10.1038/ncomms14060 (2017).

26 Cieply, B. et al. Multiphasic and Dynamic Changes in Alternative Splicing during Induction of Pluripotency Are Coordinated by Numerous RNA-Binding Proteins. Cell Rep **15**, 247-255, doi:10.1016/j.celrep.2016.03.025 (2016).

27 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, doi:10.1073/pnas.1915975117 (2020).

28 Yang, Y. et al. PRMT9 is a type II methyltransferase that methylates the splicing factor SAP145. Nat Commun **6**, 6428, doi:10.1038/ncomms7428 (2015).

29 Sachamitr, P. et al. PRMT5 inhibition disrupts splicing and stemness in glioblastoma. Nat Commun **12**, 979, doi:10.1038/s41467-021-21204-5 (2021).

30 Fong, J. Y. et al. Therapeutic Targeting of RNA Splicing Catalysis through Inhibition of Protein Arginine Methylation. Cancer Cell **36**, 194-209 e199, doi:10.1016/j.ccell.2019.07.003 (2019).

31 Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol **31**, 46-53, doi:10.1038/nbt.2450 (2013).

32 Hu, Y. et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res **41**, e39, doi:10.1093/nar/gks1026 (2013).

33 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. Genome Res **22**, 2008-2017, doi:10.1101/gr.133744.111 (2012).

34 Hartley, S. W. & Mullikin, J. C. Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. Nucleic Acids Res **44**, e127, doi:10.1093/nar/gkw501 (2016).

35 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**, E5593-5601, doi:10.1073/pnas.1419161111 (2014).

36 Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet **50**, 151-158, doi:10.1038/s41588-017-0004-9 (2018).

37 Norton, S. S., Vaquero-Garcia, J., Lahens, N. F., Grant, G. R. & Barash, Y. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. Bioinformatics **34**, 1488-1497, doi:10.1093/bioinformatics/btx790 (2018).

38 Trincado, J. L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol **19**, 40, doi:10.1186/s13059-018-1417-1 (2018).

39 Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database, C. The sequence read archive. Nucleic Acids Res **39**, D19-21, doi:10.1093/nar/gkq1019 (2011).

40 Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature **455**, 1061-1068, doi:10.1038/nature07385 (2008).

41 Consortium, E. P. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science **306**, 636-640, doi:10.1126/science.1105136 (2004).

42 Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank **13**, 307-308, doi:10.1089/bio.2015.29031.hmm (2015).

43 Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature **569**, 503-508, doi:10.1038/s41586-019-1186-3 (2019).

# 2 rMATS-turbo: an ultra-fast computational platform for profiling of splicing landscape using large-scale RNA-seq data

## 2.1 Introduction

Enormous RNA sequencing (RNA-seq) data have been produced due to the advances in sequencing technologies and reduction of sequencing cost, resulting in rapid accumulation of RNA-seq data in public repositories, such as Sequence Read Archive (SRA) [39]. Moreover, the emergence of large-scale consortium-based studies also provides unprecedented

resources for population-scale alternative splicing analysis (for example, TCGA [40], ENCODE [41], GTEx [42] and CCLE [43]). One key computational challenge underlying the accumulation of data is the time and memory limitation required for alternative splicing analysis on large-scale datasets.

A plethora of computational approaches have been developed to facilitate the identification and quantification of differential alternative splicing from RNA-seq data. There are two major quantification strategies: isoform-based models (e.g. Cuffdiff2[31], DiffSplice[32]) and count-based models. The latter can be further divided into exon/junction-based models (e.g. DEXSeq[33], JunctionSeq[34]) and event-based models (e.g. rMATS[35], LeafCutter[36], MAJIQ[37], SUPPA2[38]). Isoform-based methods aims to reconstruct full-length transcripts and perform statistical test based on isoform-resolution abundances. However, they highly depend on accurate transcript quantification and have decreased resolution on the genome compared to count-based models. Exon/junction-based methods have hyper-focused resolutions on the exons or junctions, but cannot infer the types of the splicing events. Based on a comprehensive evaluation of differential splicing tools [44], rMATS and MAJIQ scored better than other event-based methods in terms of number of detected events, as well as precision and recall rate. However, no single tool outperformed the others across all measurement. For example, MAJIQ does not support paired sample comparisons between groups, although it requires the least maximum memory. Another major common limitation shared by isoform-based methods and event-based methods is that the time and memory consumption increased dramatically with the number of samples, which restricts their application on large-scale and multi-consortium datasets.

Among those methods, rMATS [35], a software that enables the robust and flexible detection of differential alternative splicing from replicate RNA-seq, is developed and actively maintained by our lab. It has been widely used for alternative splicing analysis in various species under different circumstances [27,45-50], including the reveal of functional map of human RNA binding proteins by ENCODE consortium [3]. Over the years we have made substantial updates and added numerous features to the rMATS software, especially the rMATS-turbo re-implementation, which refers to version rMATS 4.0.1 or above. rMATS-turbo inherits the statistical model and robustness of rMATS while incorporating new splicing graph data structure and data processing pipeline, allowing for more sensitive differential alternative splicing detection and ultra-fast processing speed, making it a more powerful tool on large-scale RNA-seq datasets. While widely used by the research community [27,47-56], the rMATS-turbo software has never been systematically introduced. Here we will describe the workflow and features of rMATS-turbo, and demonstrate how to use it to process massive RNA-seq datasets in a parallel manner on a computer cluster.

# 2.2 Results

## 2.2.1 Overview of the rMATS-turbo framework

The rMATS-turbo computational program is designed to identify and quantify alternative splicing events in large-scale RNA-seq datasets, as well as to analyze the statistical significance of splicing changes between user-defined groups. **Figure 2.1** provides an overview of the rMATS-turbo software workflow. The whole rMATS-turbo framework can be separated into two steps: 1) the prep step and 2) the post step. In the prep step, rMATS-turbo takes FASTQ or BAM files as input and transforms them into individual splicing

graphs, with exons represented by nodes and exon-exon junctions represented by edges. The splicing graphs are expanded by RNA-seq reads on the backbone constructed from the reference annotation file. In the post step, the splicing graphs from all inputs are merged together into a single splicing graph. This merged splicing graph can then be used 1) to identify different types of splicing events based on the graph structure, and 2) to quantify exon usage based on the edge weights. The post step also implements a statistical model to facilitate the detection of significant differential splicing changes between user-defined groups. **Box 2.1** provides detailed descriptions of how to use the prep and post steps, as well as other arguments of rMATS-turbo.

The rMATS-turbo software can identify and quantify the five basic patterns of alternative splicing, including skipped exon (SE, exon skipping), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), mutually exclusive exon (MXE), and retained intron (RI, intron retention) events. Exon inclusion levels are assessed by using the percent spliced in (PSI) value, which represents the percentage of mRNA transcripts that include a specific exon or splice junction. PSI can be calculated from the following formula:

where I and S represent read counts for the inclusion isoform and skipping isoform, respectively; and  and  represent the effective lengths for the inclusion isoform and skipping isoform, respectively. **Supplementary Figure 2.6** presents a detailed illustration of the classification of supporting reads and calculation of effective lengths. Each type of alternative splicing event has a corresponding set of output files, which are described in **Box 2**.

### 2.2.2 rMATS-turbo facilitate easy and robust analysis of alternative splicing in a user-friendly manner

Compared to rMATS, rMATS-turbo retains the statistical robustness in identifying differential splicing changes while introducing new options and modules to enhance its performance, as itemized below.

- Most alternative splicing analysis tools underperform on large-scale datasets because of long computational time and high memory usage. Thus, the most important improvement offered by rMATS-turbo is that the computational time and memory usage are dramatically decreased compared to rMATS, especially when running on large-scale datasets ($n \geq 200$). This feature makes rMATS-turbo an extremely powerful tool for alternative splicing profiling on large-scale datasets from various consortium projects. This improvement is enabled by decoupling of the prep and post steps (**Figure 2.1**), which allows for the parallel processing of input files in individual prep steps. Specifically, the graph-construction prep step is separated from the splicing event-detection post step by using the '--task prep' and the '--task post' options. Separation of the prep and post steps also allows for the addition of additional samples after the initial run. Rather than running the entire tool on all samples whenever a new sample arrives, users can process new samples in a separate prep step. Then, the splicing graphs from previous runs can be merged together by running the post steps together.
- rMATS-turbo incorporates a new feature, '--novelSS', which allows the detection of splicing events that involve unannotated (or cryptic) splice sites. Although an exon definition mechanism is present to maintain the splicing fidelity and prevent

uncontrolled splicing at cryptic sites, non-annotated splicing events are increasingly being discovered through the analysis of mouse and human RNA-seq data [57]. Moreover, mutations in the RNA [18,19] or protein [23] components of the spliceosome have been reported in multiple cancers, which can result in the global selection of cryptic splice sites. The '--novelSS' feature could be useful for deciphering aberrant splicing in cancers that exhibit excess usage of cryptic splice sites.

- The statistical model in rMATS was built into the entire pipeline as a single module. In rMATS-turbo, the user has the option (using '--task stat') to decouple the statistical model from the event identification and quantification procedure, to fulfill the demand for multiple comparisons when more than two groups are involved. Taking the event definition file and count file (**Box 2.1**) from the post step of rMATS-turbo as input, the '--task stat' option only executes the statistical comparison between defined groups. When applying this option, users should 1) process all samples together using the typical rMATS pipeline (prep and post steps) with the '--statoff' tag to disable the statistical model. 2) The output directory will contain all of the necessary event definition files and count files for all five types of splicing events. The count files will contain information for all samples. Count information for a specific comparison can be extracted from the count files by using the code provided in rMATS_P/prepare_stat_inputs.py. 3) Finally, users can use the event definition files and extracted count files as input to conduct statistical comparisons between desired groups.

- Whereas the default statistical model considers the samples to be unpaired, we now provide an option ('--paired-stat') to perform paired statistical comparisons, using the model described in our PAIRADISE software [58].
- We have included an option to allow users to focus on a splicing event of interest, and to enable simpler multi-lab or multi-project comparisons. Specifically, the '--fixed-event-set' option allows rMATS-turbo to read a user-defined event set and perform the quantification.
- One-sample or one-group quantification is now permitted. To enable this, the user should only provide '--b1' or '--s1' options and should omit '--b2' or '--s2' options.
- Other minor improvements include user-defined permission of variable read length by using the '--variable-read-length' flag and permission of hard/soft-clipping of RNA-seq reads by using the '--allow-clipping' flag.
- Splicing events detected by rMATS-turbo can be visualized by using the rmats2sashimiplot software (https://github.com/Xinglab/rmats2sashimiplot) designed specifically for rMATS-turbo output.

### 2.2.3 Applications of the protocol

The original rMATS implementation and the current ultra-fast rMATS-turbo version [27,47-56] are widely used to perform transcriptome-wide alternative splicing analysis. Especially, we actively maintained the user interface of the software, which helped users quickly and easily adopt the new versions of rMATS, including rMATS-turbo. To date, various versions of rMATS have been downloaded > 70k times from different resources, including SourceForge (http://rnaseq-mats.sourceforge.net), Bioconda

(https://anaconda.org/bioconda/rmats), Github (https://github.com/Xinglab/rmats-turbo) and Docker (https://hub.docker.com/r/xinglab/rmats). Although mostly applied to RNA-seq data of human and mouse [51,53], the rMATS software has proven useful in various non-mammalian organisms, including Drosophila [48], birds [49], as well as in plants, including soybean [46], grape [45], and Arabidopsis thaliana [47,50]. The software has been successfully exploited to reveal wide-spread splicing changes in normal tissue development [55,59,60] and in different diseases [27,52,54,56]. It has also recently been used to reveal a large-scale binding and functional map of RNA binding proteins in ENCODE consortium [3].

The general analytical framework of rMATS-turbo can be utilized to infer the statistical significance of splicing changes between user-defined groups, for example, normal tissue versus disease tissue [27], control group versus treatment group [53], or step-wise time series of samples during development [26,61]. This analysis can aid in the prioritization of splicing changes that are of high biological relevance or diagnostic/prognostic/therapeutic value [52,54]. Moreover, through the parallel processing of input FASTQ/BAM files by separate prep steps and a summarizing post step, rMATS-turbo is remarkably competent for comprehensive splicing quantification on large- or population-scale RNA-seq studies. Finally, the resulting PSI value matrix can be further adapted for clustering to uncover the potential relationships between samples. The PSI matrix can also be correlated to score-based values by linear regression for other purposes, such as when using gene expression values to indicate potential regulation of splicing by certain RNA binding proteins.

To illustrate the application of rMATS-turbo, we describe in detail how to use rMATS-turbo in two representative application scenarios. Note, however, that this protocol

can be generalized to many other scenarios and datasets. The datasets from Examples 1 and 2 are described in the 'Materials - Required data' section, with detailed information given in the **Supplementary Table 2.1** and **2.2**.

2.2.3.1 Example 1: sinlge-command general two-group differential splicing analysis

For this example, we performed rMATS-turbo with a one-line command with prep and post steps running together, using RNA-seq data from two prostate cancer cell lines (PC3E and GS689). In addition, the '--novelSS' flag was used to demonstrate the cryptic splice site detection feature. Epithelial-mesenchymal transition (EMT) is a reversible and dynamic process with changes in cellular organization from epithelial to mesenchymal phenotypes [62], leading to functional consequences in cell migration, invasion and metastasis. The PC3E cell line has epithelial cell-like characteristics, whereas the GC689 cell line exhibits mesenchymal and invasive properties.

The final output files (n = 36) will be generated to the output folder specified by the '--od' parameter. There will be 7 files for each of the five types of alternative splicing events and 1 file summarizing the total number of events detected by rMATS-turbo. A more detailed description of the output files can be found in the public rMATS-turbo GitHub repository (https://github.com/Xinglab/rmats-turbo) and in **Box 2.2**.

Using junction reads from RNA-seq with rMATS-turbo, we identified a total of 690,094 alternative splicing events, including all 5 basic types (**Figure 2.1**). After we filtered out events supported by fewer than 10 junction reads and events with extreme PSI values (average PSI value < 0.05 or > 0.95 in both groups), the number of alternative splicing events decreased to 155,056 (**Figure 2.2a**). SE events were the most prevalent

type, accounting for ~40% of the total number of filtered alternative splicing events (**Figure 2.2a**).

rMATS-turbo also tests for statistical differences of alternative splicing between the PC3E and GS689 cell lines. rMATS-turbo reports p-values and delta PSI in the final output file. After filtering by FDR (≤0.01) and delta PSI value (≥0.05), rMATS-turbo identified 14,220 events (~ 9.16%) that were significantly different between the PC3E and GS689 cell lines (**Figure 2.2b, 2.2c**). For example, the USO1 (General Vesicular Transport Factor P115) gene demonstrated a dramatic isoform switch on cassette exon 14 (chr4:76716488-76716509), from the exon-inclusion isoform (91% inclusion level) in the epithelial-like PC3E cell line to the exon-skipping isoform (5% inclusion level) in the mesenchymal-like GS689 cell line (**Figure 2.2d**). This is consistent with previous observation that USO1 showed a reduction in exon 14 (chr4:76716488-76716509) inclusion in Hepatocellular carcinoma [63].

Alternative splicing events derived from novel splice sites were discovered by turning on the novel splice site detection feature ('--novelSS' flag). Most of these novel splice sites were of the canonical 'GT-AG' dinucleotide (data not shown), which verifies the reliability and validity of novel splice site detection. For example, a novel exon (chr19:18232916-18232937) in the MAST3 (Microtubule Associated Serine/Threonine Kinase 3) gene is exclusively included/expressed in GS689 cells but not in PC3E cells (**Figure 2.2e**). This novel exon is a 21-nucleotide 'micro-exon' that is highly evolutionarily conserved, which has been reported to be misregulated in human brain [64,65].

2.2.3.2 Example 2: multi-command ultra-fast profiling of alternative splicing in a large RNA-seq dataset

In this example, the rMATS-turbo process is split into parallel processing of input files via multiple separate prep steps, followed by a single post step. RNA-seq data of 1,019 human cell lines from the Cancer Cell Line Encyclopedia (CCLE) are leveraged to illustrate the ability of rMATS-turbo to parallelly process a large amount of input files. We predicted the EMT phenotype of CCLE cell lines by adapting the quantitative measurement of and EMT score given by a two-sample Kolmogorov–Smirnov test (2KS), which is based on the expression of specific gene signatures [66,67]. The SE events highly correlated with the EMT scores were visualized by a heatmap.

After running the prep steps (Experiment 2, step 4) separately and in parallel for each BAM file, the splicing graphs will be generated and saved to '.rmats' files in the '--tmp' directory. The '.rmats' files track info from each BAM file separately according to the path of the BAM files specified in the input configuration files from '--b1' and/or '--b2' parameters. Then, in the post step (Experiment 2, step 5), the '.rmats' splicing graph files will be merged together to identify and quantify alternative splicing events. Similar to Example 1, the final output files of rMATS turbo (n = 36) will be generated to the output folder specified by the '--od' parameter. PSI values for each event are indicated in the 'IncLevel1' and/or 'IncLevel2' columns in the [AS type].MATS.JC.txt output file. In contrast to Example 1, the p-value and delta PSI columns are marked as NA because all BAM files are assigned to group 1 ('--b1') and statistical inference was not employed.

Using rMATS-turbo, we identified a total of 1,428,168 alternative splicing events by junction reads from the RNA-seq data of 1,019 CCLE cell lines. PSI value quantifications

can be further utilized to characterize the cell lines and can be correlated to any other score-based values to select events of interest after post-processing by stringent filters. For example, of the 1,078,334 SE events identified by junction reads, 52,797 SE events remained after filtering by read count (average junction reads ≥ 20) and PSI value range ($PSI_{5\%\ percentile} \leq 0.95$ and $PSI_{95\%\ percentile} \geq 0.05$). Of these, 162 were highly correlated (Pearson correlation $R^2 > 0.4$) with EMT scores calculated from the two-sample KS test based on expression levels of signature genes [66,67]. We were able to match 207 signature genes (164 epithelial signature genes; 43 mesenchymal signature genes) from the cell line signature genes described by Tan et al. to genes in our expression data. As shown in **Figure 2.3**, the epithelial or mesenchymal status classification was supported by previous studies [67,68]. The selected SE events showed a clear transition pattern in cell lines originated from tissues with hybrid epithelial/mesenchymal states (e.g. breast and lung) [62]. By contrast, the events showed a more unified pattern in cell lines marked as exclusively mesenchymal (e.g. central nervous system, kidney, liver, and skin) or mostly epithelial (e.g. large intestine) [68].

### 2.2.4 rMATS-turbo enables ultra-fast splicing analysis on large-scale dataset

To evaluate the runtime and memory requirement for both prep and post steps of rMATS-turbo, we documented the wall clock time and maximum memory usage by analyzing the 1019 CCLE cell lines from CCLE dataset. The benchmarking was performed on a shared high-performance computing (HPC) cluster maintained by the Children's Hospital of Philadelphia (CHOP). Jobs were run on a system with 24 cores and 128 GB of memory, using only one core for each job.

Since prep steps can be run in parallel, the benchmarks are more relevant to the size of input FASTQ/BAM files. As shown in **Figure 2.4a** and **2.4b**, the prep step takes ~ 1 h and < 1 G memory to generate the splicing graph for a single BAM file under default settings, even for samples with deep sequencing depth (>200M). To conserve time, prep steps can be run separately and in parallel for different samples. Ultimately, the total time required for prep steps on all 1019 will depend on how many resources can be allocated to run individual jobs. For the summarizing post step, which merges the splicing graphs generated by individual prep steps, the benchmarks are more relevant to the number of samples. Under default settings, both runtime and maximum memory usage increased linearly along with the number of samples. And it takes ~ 5 h and < 25 G to detect and quantify alternative splicing from splicing graphs generated from 1,000 samples (**Figure 2.5a, 2.5b**).

We also tested the influence of user's selected options (eg. Novel splice site detection feature) on runtime and memory usage of both prep and post steps. For prep steps, the time usage for a single sample remains unchanged upon novel splice site detection, and the memory usage increased very mildly, with most jobs completed using < 3 G memory (**Figure 2.4c, 2.4d**). However, for post step, both time and memory usage increased exponentially along with the number of samples when utilizing the novel splice site detection feature (**Figure 2.5c, 2.5d**), which is expected because of the dramatic increase of novel nodes and edges added to the merged splicing graphs from different samples. One recommended solution for detection of novel splicing events from large scale dataset (> 500 samples) is to 1) run post steps on separate batches (< 200 samples), 2) merge the detected events from different batches manually as a user-defined event list, and

then 3) run the post step on all samples to only perform splicing quantification with '--fixed-event-set' option.

In total, using the resources described in above, it takes ~ 3 d to run both the splicing graph-generating prep steps and the events-detecting post step on 1,019 CCLE cell lines under default setting. This demonstrated the ultra-fast nature of rMATS-turbo and its competence of comprehensive splicing profiling on large-scale or cross-consortium datasets.

## 2.3 Discussion

Overall, rMATS-turbo has been proved to be computationally efficient. With decoupled splicing graph-generating prep steps and the events-detecting post step, it is able to process large-scale datasets with limited memory and dramatically increased speed. Although statistical test is designed for two-group comparison, the decoupling of statistical test module to the whole analysis pipeline makes it more convenient to perform pairwise differential splicing comparisons when multiple conditions are present, which is extremely helpful for time-course data, such as RNA-seq data of cell differentiation [48]. For two-group comparison of relatively small dataset, users can use a straightforward one-line command to finish the splicing quantification and differential splicing detection simultaneously (application example 1). For splicing profiling of large-scale datasets in a time- and memory-efficient manner, users can perform the splicing graph-generation prep steps parallelly for each individual sample, and then run a single summarizing post step to detect and quantify splicing changes (application example 2). Both procedures help to identify splicing changes that are biologically relevant.

Nonetheless, one big drawback of rMATS-turbo is that the time and memory usage are sensitive to the number of input samples when the novel splice site detection feature is turned on ('--novelSS' flag). This makes sense because the number of novel junctions (novel nodes and novel edges) would increase dramatically with the number of input reads/samples, resulting in exponential expansion of the splicing graph. To make use of this novel splice site detection feature on large-scale dataset, one recommended solution is to run the post steps on smaller batches (< 200 samples), then merge the detected splicing events, including the novel ones, into one single repository, and finally run a single post step to only perform the quantification using the '--fixed-event-set' option.

In addition, the numbers of reads for either 'inclusion isoform' and 'skipping isoform' (**Supplementary Figure 2.6**) are the sum of all involved junctions. This will create less reliable events if the read counts from one junction is in imbalance with another junction, especially in regions with complex splicing patterns. Reliable results will require appropriate filtering criteria. Further updates will incorporate reports of read counts for individual junctions.

# 2.4 Methods

Since this is a protocols paper, the Method section will describe in details a list of the essential materials (including equipment, required data and equipment setup) as well as the procedures as a numbered list of direct instructions on how we performed the analysis. Critical steps will be highlighted to emphasize those that should be performed in a precise manner to maximize the likelihood of success or ensure the best performance.

## 2.4.1 Materials

### 2.4.1.1 Equipment

- Hardware: Computer with Unix-based operating system with ≥25 GB of RAM (32 GB is needed for read alignment using STAR if FASTQ files are used as input).
- Software: Listed below are the software and versions used in this analysis.
    - rMATS-turbo 4.1.1 (https://github.com/Xinglab/rmats-turbo)
    - rmats2sashimiplot 2.0.4 (https://github.com/Xinglab/rmats2sashimiplot)
    - rMATS-turbo dependencies
    - Python 2.7 (Python 3 is also supported)
    - Python libraries
    - Cython 0.27.3
    - numpy 1.16.6
    - BLAS and LAPACK 0.3.7
    - gcc 4.8.5
    - gfortran 4.8.5
    - cmake 3.14.0
    - PAIRADISE (optional) (https://github.com/Xinglab/PAIRADISE)
    - samtools 1.10 (optional)
    - rmast2sashimiplot dependencies
    - Python 2.7 (Python 3 can be used after running 2to3.sh)
    - Python libraries
    - scipy 1.2.1
    - matplotlib 2.2.3

- pysam 0.15.4
- bedtools 2.29.2
- sratoolkit 2.9.2: for downloading the FASTQ files from SRA archive.
- STAR 2.7.1a: for alignment of downloaded FASTQ files.
- R 3.6.1 (optional)
- Conda 4.8.3 (optional)
- wget 1.14 (optional)

#### 2.4.1.2 Required data

For application example 1, RNA-seq data for PC3E and GS689 cell lines (151.64 GB bases) can be downloaded from the SRA archive under accession BioProject PRJNA438990 (**Supplementary Table 2.1**). For application example 2, RNA-seq data for the 1,019 cancer cell lines from CCLE (18.58 TB bases) can be downloaded from the SRA archive under accession BioProject PRJNA523380 (**Supplementary Table 2.2**). In addition, other required input data include:

- Human hg19 reference genome (6.96 GB) (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/GRCh37_mapping/GRCh37.primary_assembly.genome.fa.gz)
- Human hg19 GTF annotation file (65 MB) (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/GRCh37_mapping/gencode.v31lift37.annotation.gtf.gz)
- Human hg19 GFF3 annotation file (77 MB) (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/GRCh37_mappin

g/gencode.v31lift37.annotation.gff3.gz). This input file is needed by the rmats2sashimiplot software when generating the sashimi plot based on genome coordinates.

#### 2.4.1.3 Equipment setup

**Downloading and installing rMATS-turbo**

After the required dependencies are installed, rMATS-turbo can be downloaded and installed through the GitHub repository:

```
git clone https://github.com/Xinglab/rmats-turbo
cd rmats-turbo
./build_rmats
```

Alternatively, rMATS-turbo and all required dependencies can be installed at once through Conda by using the following command:

```
conda install -c bioconda rmats
```

**Downloading and installing rmats2sashimiplot**

The rmats2sashimiplot package can be downloaded from the Github repository:

```
git clone https://github.com/Xinglab/rmats-turbo
cd rmats-turbo
./build_rmats
```

rmats2sashimiplot is written in Python 2. If using Python 3, the following command must first be run to convert the package to Python 3 scripts:

```
bash 2to3.sh
```

Next, the rmats2sashimiplot package can be installed by running the setup.py file:

```
python ./setup.py install
```

The package can also be used without installation by providing the path to the script:

```
python ./src/rmats2sashimiplot/rmats2sashimiplot.py
```

**Downloading and preparing the required data**

The human hg19 reference genome and related GTF and GFF3 annotation files can be downloaded by using 'wget' with the following commands:

```
wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/GRCh37_mapping/GRCh
37.primary_assembly.genome.fa.gz
wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/GRCh37_mapping/genc
ode.v31lift37.annotation.gtf.gz
wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/GRCh37_mapping/genc
ode.v31lift37.annotation.gff3.gz
```

rMATS-turbo can take either FASTQ or BAM files as input. If FASTQ files are used, rMATS-turbo will first call the STAR program to perform read mapping. We recommend that users perform the alignment separately and use the pre-aligned BAM files as input for rMATS-turbo. FASTQ files for the datasets used for Example 1 (PC3E and GS689 cell lines) and Example 2 (1,019 CCLE cell lines) can be downloaded directly from the SRA archive with the sratoolkit software, and then mapped to the human hg19 genome by STAR [69] as follows:

1. Download the .sra files to the workspace of sratoolkit (sra_workspace, specified while installing sratoolkit) and then convert the files to FASTQ files.

```
./sratoolkit.2.9.2-ubuntu64/bin/prefetch.2.9.2 $SRA_RUN && ./sratoolkit.2.9.2-
ubuntu64/bin/fastq-dump.2.9.2 --outdir ./ ---split-files
$sra_workspace/$SRA_RUN.sra
```

$SRA_RUN represents the SRA run accession number for each cell line. A full list of SRA run accession numbers and related sample information for the PC3E and GS689 cell lines (BioProject PRJNA438990) and CCLE cell lines (BioProject PRJNA523380) are available in the **Supplementary Table 2.1** and **2.2**, respectively.

2. Generate a reference genome index for STAR using the downloaded hg19 reference genome and the related GTF annotation file.

```
STAR ---runThreadN 4 --runMode genomeGenerate --genomeDir
STAR_index/gencode.v31lift37 ---genomeFastaFiles GRCh37.primary_assembly.genome.fa
--sjdbGTFfile gencode.v31lift37.annotation.gtf
```

The index will be generated to the STAR_index/gencode.v31lift37 folder indicated by the '--genomeDir' parameter.

3. Align the RNA-seq reads to the hg19 genome with STAR.

```
STAR ---genomeDir STAR_index/gencode.v31lift37 ---readFilesIn $sample_name_1.fastq
$sample_name_2.fastq --outFileNamePrefix ./$sample_name/ --outSAMunmapped Within -
-outSAMattributes NH HI AS NM MD XS --twopassMode Basic --alignSJDBoverhangMin 1 -
-alignSJoverhangMin 8 ---alignEndsType EndToEnd --runThreadN 6 --outSAMtype BAM
SortedByCoordinate --outSAMstrandField intronMotif
```

$sample_name_1.fastq and $sample_name_2.fastq represent the paths of two unzipped FASTQ read pairs from paired-end RNA-seq.

▲ **CRITICAL STEP** To facilitate detection of novel/cryptic splice site-derived events, we recommend that users use the two-pass alignment mode (--twopassMode Basic) from STAR [69].

▲ **CRITICAL STEP** STAR usually requires ~32 GB of RAM for mapping to the human genome.

## 2.4.2 Procedures

Here we describe all steps for two representative application scenarios of rMATS-turbo. Example 1 is a general two-group differential splicing analysis using RNA-seq data from the PC3E and GS689 prostate cancer cell lines (BioProject PRJNA438990). Example 2 is the profiling of alternative splicing in a large-scale dataset, using RNA-seq data of 1,019 human cell lines from CCLE (BioProject PRJNA523380). Supplementary Table 2.1 provides troubleshooting advices for running rMATS-turbo, on steps marked with 'Troubleshooting' marks.

### 2.4.2.1 Example 1: sinlge-command general two-group differential splicing analysis

**Set up working directory and input files for rMATS-turbo analysis • Timing ~ 5 min**

1. Set up working directory where all outputs will be written.

```
mkdir -p PC3E-GS689/rmats
cd PC3E-GS689/rmats
```

2. Generate configuration text files (b1.txt and b2.txt) as input file for rMATS-turbo. These two files contain comma-separated lists of FASTQ/BAM files for group1 and group2, respectively.

   ▲ **CRITICAL STEP** Either FASTQ files or BAM files can be used by rMATS-turbo. If FASTQ files are provided, rMATS-turbo will perform sequencing alignment using STAR software.

```
ls $prefix_dir_group1 | tr '\n' ',' | sed 's/,$/\n/' > ./b1.txt
ls $prefix_dir_group2 | tr '\n' ',' | sed 's/,$/\n/' > ./b2.txt
```

$prefix_dir_group1: Folder containing all FASTQ/BAM files for group 1.

$prefix_dir_group2: Folder containing all FASTQ/BAM files for group 2.

**Run rMATS-turbo to perform differential splicing analysis using <--task both> parameter • Timing ~ 4 h 45 min**

3. Run rmats.py with specified parameters.

   ▲ **CRITICAL STEP** If BAM files are used, the configuration file(s) should be specified by '--b1' and/or '--b2'; if FASTQ files are used, the configuration files should be specified by '--s1' and/or '--s2'.

   ▲ **CRITICAL STEP** To facilitate detection of novel/cryptic splice site-derived events, we recommend using the two-pass alignment mode (--twopassMode Basic) from STAR [69].

   ▲ **CRITICAL STEP** The '--paired-stats' flag can be used if each entry in '--b1' is matched with its pair in '--b2' (paired replicates). If this flag is used, the PAIRADISE [58] software will be utilized to identify differential alternative splicing events based on a paired statistical model.

   ▲ **CRITICAL STEP** rMATS-turbo can be executed on multiple threads (specified by the '--nthread' parameter) to significantly shorten the runtime.

   ```
   python $rmats_dir/rmats.py —gtf $GTF --tmp prep —od post --readLength 101 --b1
   b1.txt --b2 b2.txt -t paired --anchorLength 1 --nthread 1 —libType fr-unstranded —
   task both —variable-read-length --novelSS
   ```

The '--novelSS' flag enables detection of splicing events derived from novel splice sites.

? TROUBLESHOOTING

**Perform downstream analysis and visulization of rMATS-turbo results** • Timing ~ 1 h

4. Filter significant events from the two-group comparisons. The following criteria were used and are recommended for Example 1:
    - Read coverage filter: average read count ≥ 10 for both groups.
    - PSI range filter: filter out events with average PSI value < 0.05 or > 0.95 for both groups.
    - FDR filter: FDR ≤ 0.01
    - PSI value difference filter: |ΔPSI| ≥ 0.05
5. Run rmats2sashimiplot software to generate sashimi plots for selected splicing events.

```
mkdir -p PC3E-GS689/rmats2sashimi
cd PC3E-GS689/rmats2sashimi
$rmats2sashimiplot --b1 $bam1,$bam2,$bam3 --b2 $bam4,$bam5,$bam6 -t SE -e
sashimi_events.txt --l1 PC3E_rep --l2 GS689_rep —exon_s 1 --intron_s 5 -o ./output —
group-info sashimi_groupInfo.txt
```

where:

-e refers to the path to a file containing events for which the sashimi plot will be generated. The format of this file should be the same as the rMATS-turbo final output (e.g. SE.MATS.JC.txt);

--l1 and –l2 are labels for group 1 and group 2, respectively;

--group-info refers to the path to a file that groups the replicates. One sashimi plot will be generated for each group (in contrast to the default behavior of one plot per replicate). Each line of the file defines a group and is formatted as "group name: indices of mapping files". The content of the example sashimi_groupInfo.txt file used in Example 1 is as follows:

```
PC3E:1-3
GS689:4-6
```

? TROUBLESHOOTING

2.4.2.2 Example 2: multi-command ultra-fast profiling of alternative splicing in a large RNA-seq dataset

**Set up input files for rMATS-turbo analysis with prep and post steps separated** •

**Timing ~ 5 min**

1. Set up working directory where all outputs will be written.

```
mkdir -p CCLE/rmats
cd CCLE/rmats
```

2. Generate configuration text files for prep step of rMATS-turbo.

```
mkdir -p bamConfiguration_prep
ls $bam1 > ./bamConfiguration_prep/bam1.txt
ls $bam2 > ./bamConfiguration_prep/bam2.txt
...
ls $bam1019 > ./bamConfiguration_prep/bam1019.txt
```

$bam1: File path of the input BAM file.

3. Generate configuration file for post step of rMATS-turbo. This file contains comma-separated paths of all FASTQ/BAM files.

▲ **CRITICAL STEP** Regardless of whether they are absolute paths or relative paths, the paths of the FASTQ/BAM files must be the same in the prep configuration file and the post configuration file.

```
mkdir -p bamConfiguration_post
ls $bam1 $bam2 [...] $bam1019 | tr '\n' ',' | sed
's/,$/\n/' > ./bamConfiguration_post/b1_1019.txt
```

**Run rMATS-turbo with prep and post steps separated** • **Timing ~ 3 d**

4. Run the prep step with the '--task prep' parameter on each sample (BAM files) separately. Each prep step takes ~ 1 h to generate the splicing graph for a single BAM file (**Figure 2.4a**).

   ▲ **CRITICAL STEP** For Example 2, in this step, the splicing graph-generating prep steps were performed separately and in parallel for each BAM file. This approach is extremely helpful for large-scale data analysis because it dramatically decreases time and memory usage. The use of '--b2' and '--s2' can be skipped because only one FASTQ/BAM configuration file is used.

```
python $rmats_dir/rmats.py —gtf $GTF --tmp prep --od post_1019 --readLength 101 --b1
bamConfiguration_prep/bam1.txt -t paired —anchorLength 1 —nthread 1 —libType fr-
unstranded --task prep --variable-read-length
python $rmats_dir/rmats.py —gtf $GTF --tmp prep --od post_1019 --readLength 101 --b1
bamConfiguration_prep/bam2.txt -t paired —anchorLength 1 —nthread 1 —libType fr-
unstranded --task prep --variable-read-length
...
python $rmats_dir/rmats.py —gtf $GTF --tmp prep --od post_1019 --readLength 101 --b1
bamConfiguration_prep/bam3.txt -t paired —anchorLength 1 —nthread 1 —libType fr-
unstranded --task prep --variable-read-length
```

5. Run the post step with the '--task post' parameter on all samples. In this step, the splicing graphs generated by the prep steps are merged together to enable detection of alternative splicing events.

   ▲ **CRITICAL STEP** The command uses the '--statoff' flag to disable the statistical test because only one group is provided. If this flag is not added and only one group is provided, the statistical test would be automatically disabled (with a warning message).

```
python $rmats_dir/rmats.py ——gtf $GTF --tmp prep --od post_1019 --readLength 101 --b1
bamConfiguration_post/b1_1019.txt  -t paired --anchorLength 1 --nthread 1 --libType
fr-unstranded --task post --variable-read-length --statoff
```

   ? TROUBLESHOOTING

**Perform downstream analysis of rMATS-turbo results** • **Timing ~ 30 min**

6. Filter high-confidence alternative splicing events detected by rMATS-turbo. The following criteria were used and are recommended for Example 2:
   - Read coverage filter: average read count ≥ 20 across all 1,019 samples.
   - PSI range filter: 5% quantile of PSI values ≤ 0.95; 95% quantile of PSI values ≥ 0.05
7. Calculate the EMT score matrix of the 1,019 CCLE cell lines using the two-sample KS test based on expression levels of EMT signature genes, as described in the literature [66,67]. A total of 207 signature genes (164 epithelial signature genes; 43 mesenchymal signature genes) from the cell line signature genes described by Tan et al. are mapped to genes in our expression data.

8. Generate heatmap visualization of alternative splicing events detected by rMATS-turbo. Skipped exon events showing a high correlation (Pearson correlation R2 > 0.4) with EMT scores were selected to be displayed in the heatmap.

### 2.4.3 Code availability

rMATS-turbo is freely available on Github (https://github.com/Xinglab/rmats-turbo), Bioconda (https://anaconda.org/bioconda/rmats), and SourceForge (http://rnaseq-mats.sourceforge.net). Source code for rmats2sashimiplot is publicly available at the following GitHub repository (https://github.com/Xinglab/rmats2sashimiplot).

### 2.4.4 Data availability

The FASTQ files of PC3E and GS689 cell lines for Example 1 (BioProject PRJNA438990) and CCLE cell lines for Example 2 (BioProject PRJNA523380) can be downloaded freely at the SRA archive (https://www.ncbi.nlm.nih.gov/sra). The demonstration output files of rMATS-turbo and rmats2sashimiplot, as well as the code required to generate the plots are available at the GitHub repository (https://github.com/ywang1993/nature_protocols). EMT scores calculated from the two-sample KS test based on 207 signature genes for the 1,019 CCLE cell lines can be retrieved from the folder CCLE/EMT_score in this repository.

# 2.5 Figures

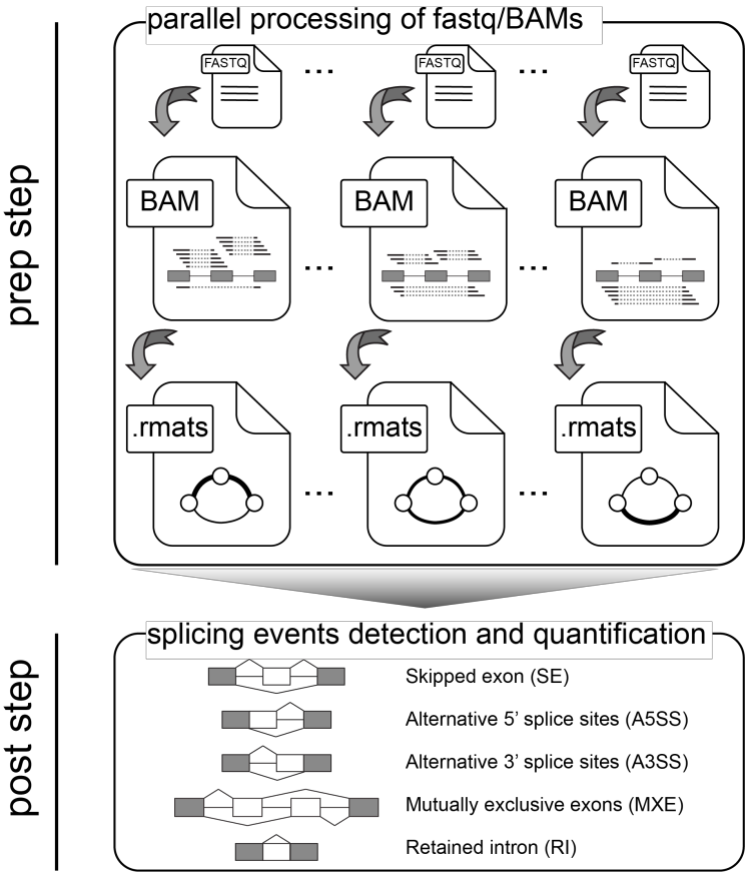



**Figure 2.1 Overview of the rMATS-turbo workflow to identify, quantify, and analyze alternative splicing events in large-scale RNA-seq datasets.**

The entire workflow comprises two major steps: 1) the prep step, and 2) the post step. The prep step enables parallel processing of either FASTQ or BAM files as input to generate corresponding splicing graphs. When FASTQ files are used, RNA-seq reads are first aligned to the reference genome and converted to standard BAM format by calling STAR software. One splicing graph per BAM file is generated and stored in an .rmats file, with exons as nodes and junction reads as edges. The post step merges all splicing graphs generated by the prep step(s), and then detects and quantifies alternative splicing events. The post step

also implements a statistical model that permits the sensitive and robust detection of differential alternative splicing events when two groups are provided.

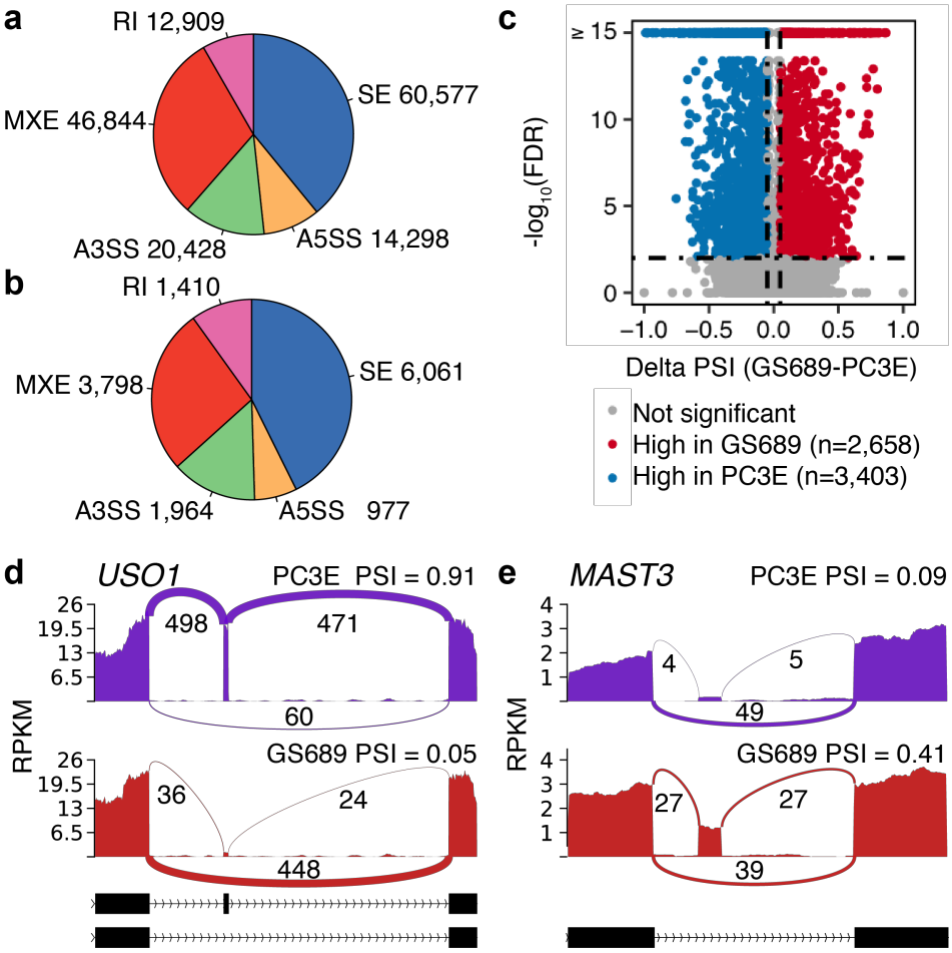



**Figure 2.2 rMATS-turbo enables identification of significant differential alternative splicing events, including those derived from novel splice sites, between PC3E and GS689 cell lines.**

**(a)** Summary pie chart of total alternative splicing events identified by rMATS-turbo in PC3E-GS689 dataset after filtering by read count (≥10 in both groups) and PSI value range (filter out events with average PSI value < 0.05 or > 0.95 in both groups). **(b)** Summary pie chart of alternative splicing events with significant changes between the two groups (FDR ≤ 0.01 and |ΔPSI| ≥ 0.05). **(c)** Volcano plot of skipped exon (SE) events. Each dot represents one SE event. The horizontal and vertical dashed lines mark the threshold of significance

level (FDR ≤ 0.01) and PSI value change (|ΔPSI| ≥ 0.05), respectively. **(d)** Sashimi plot showing the change in usage of a target exon in USO1. **(e)** Sashimi plot showing the change in usage of a novel exon in MAST3. Black bars and dashed lines on the bottom represent exons and introns annotated in the reference, respectively. Solid peaks represent reads per kilobase per million mapped (RPKM) mapped to each region. Arches represent splice junctions. Numbers represent numbers of reads mapped to each splice junction. PSI values are indicated on top of the sashimi plot. Number and PSI value of events shown in this figure are calculated by junction reads only.

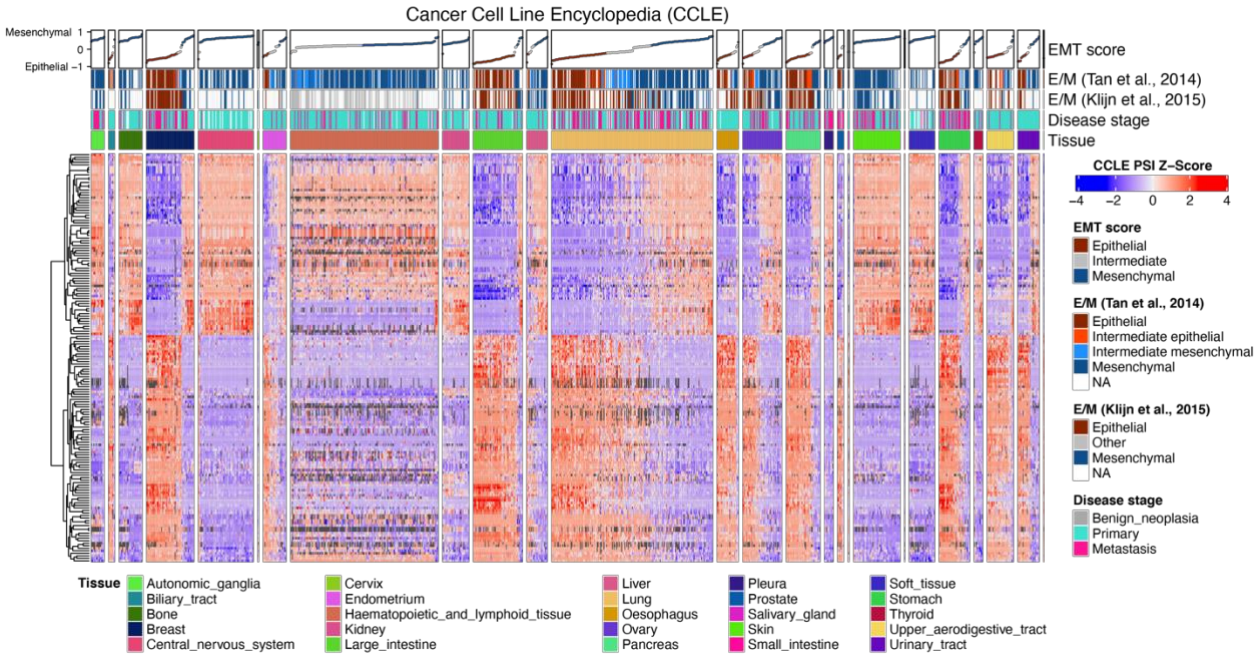



**Figure 2.3 Global alternative splicing profiling of the CCLE dataset (n = 1,019).**

The 1,019 cell lines (represented by columns) are grouped by their tissue of origin and then ordered by EMT scores calculated from the two-sample KS test based on expression level of signature genes (depicted on the top of the heatmap). Red and blue colors represent epithelial and mesenchymal status, respectively, classified based on the EMT score (red < 0; blue > 0) and KS test p-values ($p < 0.05$). Heatmap colors represent the Z-score-transformed PSI values for 162 skipped exon events (represented by rows) that are highly correlated ($R^2 > 0.4$) with EMT scores. E/M classification of cell lines from two other papers are also shown [67,68]. E, epithelial; M, mesenchymal; EMT, epithelial to mesenchymal transition.

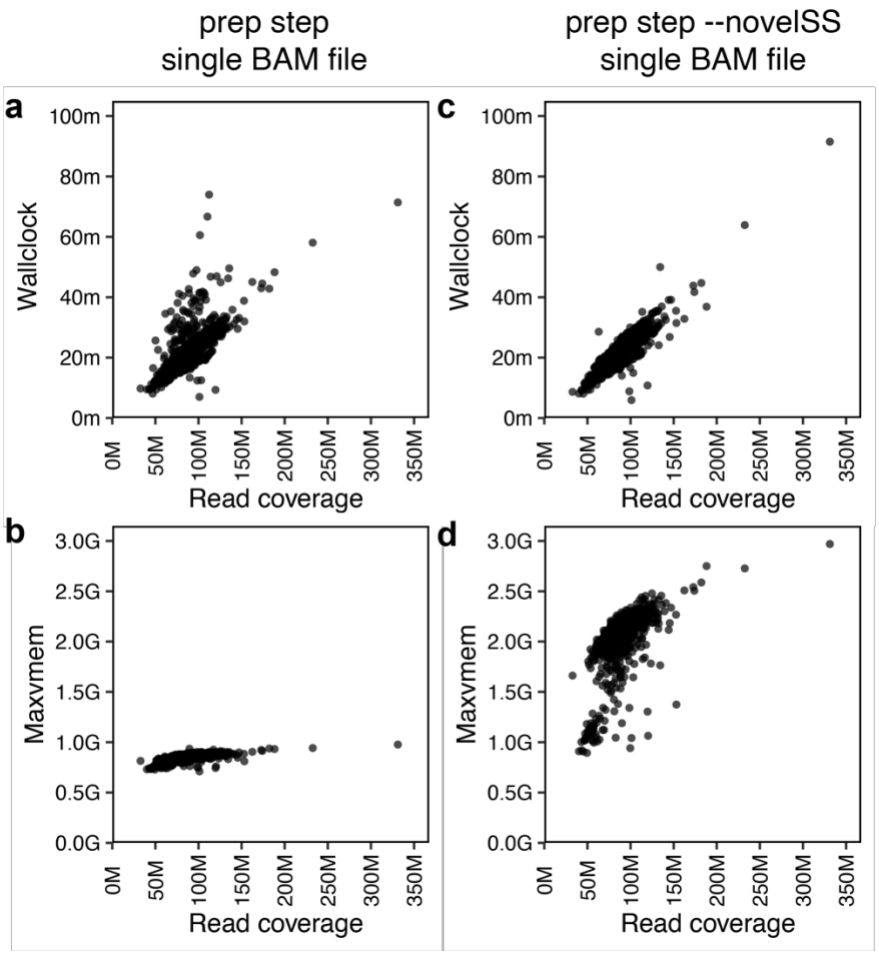



**Figure 2.4 Benchmarks of runtime and memory usage for the rMATS-turbo prep step based on the read coverage when running on a single BAM file.**

Panel **a** and **b** represent the prep steps running without the novel splice site detection feature. Panel c and d represent the prep steps running with the novel splice site detection feature turned on by adding the '--novelSS' flag. Panel **a** and **c** represent the wall clock runtime for the job to finish. Panel b and d represent the maxvmem, which is the maximum amount of RAM used by the jobs when running.

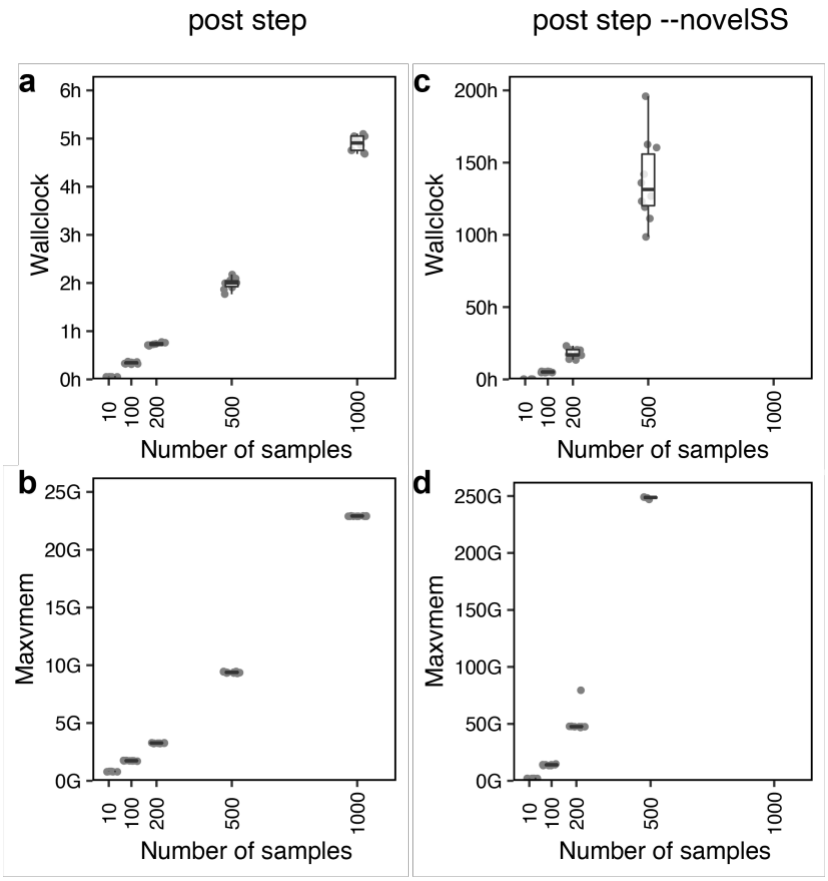
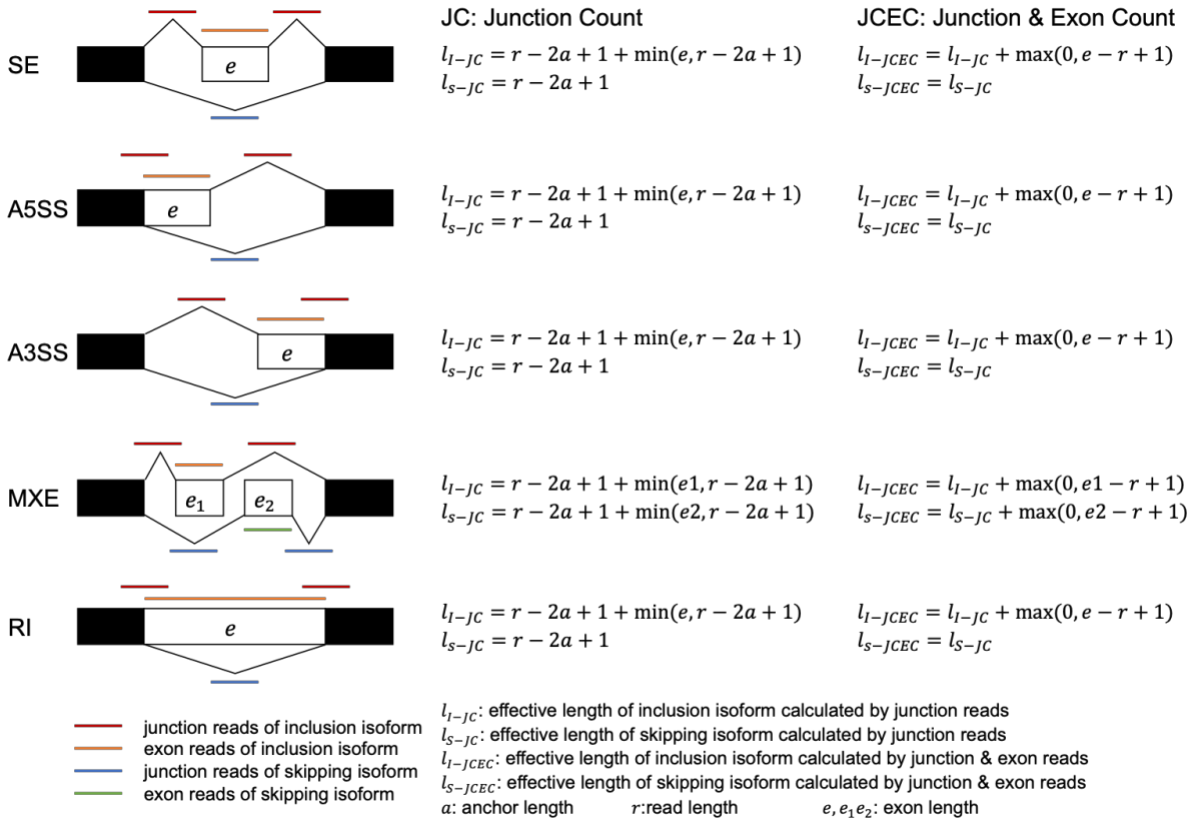



**Figure 2.5 Benchmarks of runtime and memory usage for the rMATS-turbo post step based on the number of input splicing graphs generated by the prep steps.**

Panel a and b represent the post steps running without novel splice site detection feature. Panel c and d represent the post steps running with the novel splice site detection feature turned on by adding the '--novelSS' flag. Panel a and c represent the wall clock runtime for the job to finish. Panel b and d represent the maxvmem, which is the maximum amount of RAM used by the jobs when running.




**Supplementary Figure 2.6 Schematic illustration of the classification of supporting reads and calculation of effective lengths for alternative splicing events.**

Diagrams on the left demonstrate the structures of different alternative splicing events that rMATS-turbo recognizes from the splicing graphs. The number of supporting reads can be counted by using either only the junction reads (JC) or both the junction and exon body reads (JCEC). Depending on whether JC or JCEC reads are used, the effective length calculation is adjusted as shown in the formulas on the right side of the diagram. Red line: position of junction read of inclusion isoform; orange line: position of exon read of inclusion isoform; blue line: position of junction read of skipping isoform; green line: position of exon read of skipping isoform.

# 2.6 Tables

**Supplementary Table 2.1 Summary of PC3E and GS689 cell lines used in application example 1.**

| RUN | ASSAY TYPE | BIOPROJECT | CELL LINE | INSTRUMENT | LIBRARY LAYOUT | LIBRARY SELECTION | TISSUE |
|---|---|---|---|---|---|---|---|
| **SRR6862379** | RNA-Seq | PRJNA438990 | PC3E | Illumina HiSeq 4000 | PAIRED | cDNA | prostate cancer |
| **SRR6862380** | RNA-Seq | PRJNA438990 | PC3E | Illumina HiSeq 4000 | PAIRED | cDNA | prostate cancer |
| **SRR6862381** | RNA-Seq | PRJNA438990 | PC3E | Illumina HiSeq 4000 | PAIRED | cDNA | prostate cancer |
| **SRR6862382** | RNA-Seq | PRJNA438990 | GS689 | Illumina HiSeq 4000 | PAIRED | cDNA | prostate cancer |
| **SRR6862383** | RNA-Seq | PRJNA438990 | GS689 | Illumina HiSeq 4000 | PAIRED | cDNA | prostate cancer |
| **SRR6862384** | RNA-Seq | PRJNA438990 | GS689 | Illumina HiSeq 4000 | PAIRED | cDNA | prostate cancer |

**Supplementary Table 2.2 Summary of the 1,019 CCLE cell lines used in application example 2.**

Available at Github repository https://github.com/ywang1993/nature_protocols

## Supplementary Table 2.3 Troubleshooting table for rMATS-turbo.

| Step | Problem | Possible reason | Solution |
|---|---|---|---|
| 3 (Example 1) | Output files are empty or only contain headers | 1. '--readLength' parameters are not provided correctly<br>2. Input reads differ in read length from that of the '--readLength' parameter, but the '--variable-read-length' flag is not added<br>3. Input reads have hard/soft clipping, but the '--allow-clipping' flag is not added | 1. Provide a precise read length after the '--readLength' flag<br>2. Add '--variable-read-length' flag if lengths of input reads differ significantly<br>3. Add '--allow-clipping' flag if there are many hard/soft clipped reads in the input BAM files |
| 5 (Example 1) | rmats2sashimi job takes too long to finish | Input file contains too many events to plot | Only select a limited number of events to plot, or split input events into multiple files and run rmats2sashimiplot on each file separately |
| 5 (Example 2) | ValueError: invalid literal for int() with base 10 | Statistical test is performed but only one group of samples is provided | Add '--statoff' flag to disable the statistical test; or ensure you have updated to the latest version, in which the statistical test is automatically disabled with a warning message when only one group is provided |
| 5 (Example 2) | XX.bam not found in .rmats files;<br>XX.bam found X times in .rmats files | Splicing graphs (in .rmats file) for FASTQ/BAM files in --b1 or --b2 file are not found or found multiple times in the --tmp folder. | 1. Ensure that the FASTQ/BAM file paths are exactly the same between the prep and post steps. Assignment of FASTQ/BAM files to prep steps does not restrict the choice/order of '--b1'/'--s1' and/or '--b2'/'--s2' for a post step<br>2. Ensure that only one splicing graph (in .rmats file) is present in the --tmp folder (and its subfolders) for one input FASTQ/BAM file |

# 2.7 Boxes

### Box 2.1 Description of arguments of rMATS-turbo

```
-h, --help            show this help message and exit
--version             show program's version number and exit
--gtf GTF             An annotation of genes and transcripts in GTF format
--b1 B1               A text file containing a comma separated list of the
                      BAM files for sample_1. (Only if using BAM)
--b2 B2               A text file containing a comma separated list of the
                      BAM files for sample_2. (Only if using BAM)
--s1 S1               A text file containing a comma separated list of the
                      FASTQ files for sample_1. If using paired reads the
                      format is ":" to separate pairs and "," to separate
                      replicates. (Only if using fastq)
--s2 S2               A text file containing a comma separated list of the
                      FASTQ files for sample_2. If using paired reads the
                      format is ":" to separate pairs and "," to separate
                      replicates. (Only if using fastq)
--od OD               The directory for final output from the post step
--tmp TMP             The directory for intermediate output such as ".rmats"
                      files from the prep step
-t {paired,single}    Type of read used in the analysis: either "paired" for
                      paired-end data or "single" for single-end data.
                      Default: paired
--libType {fr-unstranded,fr-firststrand,fr-secondstrand}
                      Library type. Use fr-firststrand or fr-secondstrand
                      for strand-specific data. Default: fr-unstranded
--readLength READLENGTH
                      The length of each read
--variable-read-length
                      Allow reads with lengths that differ from --readLength
                      to be processed. --readLength will still be used to
                      determine IncFormLen and SkipFormLen
--anchorLength ANCHORLENGTH
                      The anchor length. Default is 1
--tophatAnchor TOPHATANCHOR
                      The "anchor length" or "overhang length" used in the
                      aligner. At least "anchor length" NT must be mapped to
                      each end of a given junction. The default is 6. (Only
                      if using fastq)
--bi BINDEX           The directory name of the STAR binary indices (name of
                      the directory that contains the SA file). (Only if
                      using fastq)
--nthread NTHREAD     The number of threads. The optimal number of threads
                      should be equal to the number of CPU cores. Default: 1
--tstat TSTAT         The number of threads for the statistical model. If
                      not set then the value of --nthread is used
```

```
--cstat CSTAT         The cutoff splicing difference. The cutoff used in the
                      null hypothesis test for differential splicing. The
                      default is 0.0001 for 0.01% difference. Valid: 0 <=
                      cutoff < 1. Does not apply to the paired stats model
--task {prep,post,both,inte,stat}
                      Specify which step(s) of rMATS to run. Default: both.
                      prep: preprocess BAMs and generate a .rmats file.
                      post: load .rmats file(s) into memory, detect and
                      count alternative splicing events, and calculate P
                      value (if not --statoff). both: prep + post. inte
                      (integrity): check that the BAM filenames recorded by
                      the prep task(s) match the BAM filenames for the
                      current command line. stat: run statistical test on
                      existing output files
--statoff             Skip the statistical analysis
--paired-stats        Use the paired stats model
--novelSS             Enable detection of novel splice sites (unannotated
                      splice sites). Default is no detection of novel splice
                      sites
--mil MIL             Minimum Intron Length. Only impacts --novelSS
                      behavior. Default: 50
--mel MEL             Maximum Exon Length. Only impacts --novelSS behavior.
                      Default: 500
--allow-clipping      Allow alignments with soft or hard clipping to be used
--fixed-event-set     A directory containing fromGTF.[AS].txt files to be used
                      Instead of detecting a new set of events.
```

## Box 2.2 Output files of rMATS-turbo

Each type of alternative splicing event has a corresponding set of output files. In the filename templates below, `[AS_Event]` is replaced by one of the five alternative splicing patterns: skipped exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), mutually exclusive exon (MXE), or retained intron (RI). As shown in **Supplementary Figure 1**, the number of supporting reads can be counted by only the junction reads (JC) or by both junction and exon body reads (JCEC). The output file from different counting methods is also indicated in the file name.

`--od` contains the final output files from the post step:

- `[AS_Event].MATS.JC.txt`: Final output including only reads that span junctions defined by rMATS.
- `[AS_Event].MATS.JCEC.txt`: Final output including both reads that span junctions defined by rMATS and reads that do not cross an exon boundary.
- `fromGTF.[AS_Event].txt`: All identified alternative splicing (AS) events derived from GTF and RNA.
- `fromGTF.novelJunction.[AS_Event].txt`: Alternative splicing (AS) events that were identified only after considering the RNA (as opposed to analyzing the GTF in isolation). Does not include events with an unannotated splice site.
- `fromGTF.novelSpliceSite.[AS_Event].txt`: This file contains only events that include an unannotated splice site. Only relevant if `--novelSS` is enabled.
- `JC.raw.input.[AS_Event].txt`: Event counts including only reads that span junctions defined by rMATS.
- `JCEC.raw.input.[AS_Event].txt`: Event counts including both reads that span junctions defined by rMATS and reads that do not cross an exon boundary.
- Shared columns:
  - `ID`: rMATS event id
  - `GeneID`: Gene id
  - `geneSymbol`: Gene name
  - `chr`: Chromosome
  - `strand`: Strand of the gene
  - `IJC_SAMPLE_1`: Inclusion counts for sample 1. Replicates are comma separated

- SJC_SAMPLE_1: Skipping counts for sample 1. Replicates are comma separated
- IJC_SAMPLE_2: Inclusion counts for sample 2. Replicates are comma separated
- SJC_SAMPLE_2: Skipping counts for sample 2. Replicates are comma separated
- IncFormLen: Length of inclusion form, used for normalization
- SkipFormLen: Length of skipping form, used for normalization
- PValue: Significance of splicing difference between the two sample groups. (Only available if the statistical model is on)
- FDR: False Discovery Rate calculated from p-value. (Only available if statistical model is on)
- IncLevel1: Inclusion level for sample 1. Replicates are comma separated. Calculated from normalized counts
- IncLevel2: Inclusion level for sample 2. Replicates are comma separated. Calculated from normalized counts
- IncLevelDifference: average(IncLevel1) - average(IncLevel2)

- Event specific columns (event coordinates):
  - SE: exonStart_0base exonEnd upstreamES upstreamEE downstreamES downstreamEE
    - The inclusion form includes the target exon (exonStart_0base, exonEnd)
  - MXE: 1stExonStart_0base 1stExonEnd 2ndExonStart_0base 2ndExonEnd upstreamES upstreamEE downstreamES downstreamEE
    - If the strand is + then the inclusion form includes the 1st exon (1stExonStart_0base, 1stExonEnd) and skips the 2nd exon
    - If the strand is − then the inclusion form includes the 2nd exon (2ndExonStart_0base, 2ndExonEnd) and skips the 1st exon
  - A3SS, A5SS: longExonStart_0base longExonEnd shortES shortEE flankingES flankingEE
    - The inclusion form includes the long exon (longExonStart_0base, longExonEnd) instead of the short exon (shortES, shortEE)
  - RI: riExonStart_0base riExonEnd upstreamES upstreamEE downstreamES downstreamEE

        - The inclusion form includes (retains) the intron (`upstreamEE`, `downstreamES`)
- `summary.txt`: Brief summary of all AS event types. Includes the total event counts and significant event counts. By default, events are counted as significant if FDR <= 0.05. The summary can be regenerated with different criteria by running `rMATS_P/summary.py`

`--tmp` contains the intermediate files generated by the prep step:

- `[datetime]_[id].rmats`: `Summary` generated from processing a BAM
- `[datetime]_bam[sample_num]_[replicate_num]/Aligned.sortedByCoord.out.bam`: result of mapping input FASTQ files
- `[datetime]_read_outcomes_by_bam.txt`: Counts of the reads used from each BAM along with counts of the reasons that reads were not able to be used

# 2.8 References


3 Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. Nature **583**, 711-719, doi:10.1038/s41586-020-2077-3 (2020).

18 Suzuki, H. et al. Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. Nature **574**, 707-711, doi:10.1038/s41586-019-1650-0 (2019).

19 Shuai, S. et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. Nature **574**, 712-716, doi:10.1038/s41586-019-1651-z (2019).

23 Darman, R. B. et al. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. Cell Rep **13**, 1033-1045, doi:10.1016/j.celrep.2015.09.053 (2015).

26 Cieply, B. et al. Multiphasic and Dynamic Changes in Alternative Splicing during Induction of Pluripotency Are Coordinated by Numerous RNA-Binding Proteins. Cell Rep **15**, 247-255, doi:10.1016/j.celrep.2016.03.025 (2016).

27 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, doi:10.1073/pnas.1915975117 (2020).

31 Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol **31**, 46-53, doi:10.1038/nbt.2450 (2013).

32 Hu, Y. et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res **41**, e39, doi:10.1093/nar/gks1026 (2013).

33 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. Genome Res **22**, 2008-2017, doi:10.1101/gr.133744.111 (2012).

34 Hartley, S. W. & Mullikin, J. C. Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. Nucleic Acids Res **44**, e127, doi:10.1093/nar/gkw501 (2016).

35 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**, E5593-5601, doi:10.1073/pnas.1419161111 (2014).

36 Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet **50**, 151-158, doi:10.1038/s41588-017-0004-9 (2018).

37 Norton, S. S., Vaquero-Garcia, J., Lahens, N. F., Grant, G. R. & Barash, Y. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. Bioinformatics **34**, 1488-1497, doi:10.1093/bioinformatics/btx790 (2018).

38 Trincado, J. L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol **19**, 40, doi:10.1186/s13059-018-1417-1 (2018).

39 Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database, C. The sequence read archive. Nucleic Acids Res **39**, D19-21, doi:10.1093/nar/gkq1019 (2011).

40 Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature **455**, 1061-1068, doi:10.1038/nature07385 (2008).

41 Consortium, E. P. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science **306**, 636-640, doi:10.1126/science.1105136 (2004).

42 Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank **13**, 307-308, doi:10.1089/bio.2015.29031.hmm (2015).

43 Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature **569**, 503-508, doi:10.1038/s41586-019-1186-3 (2019).

44 Mehmood, A. et al. Systematic evaluation of differential splicing tools for RNA-seq studies. Brief Bioinform **21**, 2052-2065, doi:10.1093/bib/bbz126 (2020).

45 Jiang, J. et al. Integrating Omics and Alternative Splicing Reveals Insights into Grape Response to High Temperature. Plant Physiol **173**, 1502-1518, doi:10.1104/pp.16.01305 (2017).

46 Huang, J. et al. An oomycete plant pathogen reprograms host pre-mRNA splicing to subvert immunity. Nat Commun **8**, 2051, doi:10.1038/s41467-017-02233-5 (2017).

47 Wang, L. et al. RALF1-FERONIA complex affects splicing dynamics to modulate stress responses and growth in plants. Sci Adv **6**, eaaz1622, doi:10.1126/sciadv.aaz1622 (2020).

48 Wang, Y. et al. Role of Hakai in m(6)A modification pathway in Drosophila. Nat Commun **12**, 2159, doi:10.1038/s41467-021-22424-5 (2021).

49 Rogers, T. F., Palmer, D. H. & Wright, A. E. Sex-Specific Selection Drives the Evolution of Alternative Splicing in Birds. Mol Biol Evol **38**, 519-530, doi:10.1093/molbev/msaa242 (2021).

50 Liu, W. et al. Ectopic targeting of CG DNA methylation in Arabidopsis with the bacterial SssI methyltransferase. Nat Commun **12**, 3130, doi:10.1038/s41467-021-23346-y (2021).

51 Begg, B. E., Jens, M., Wang, P. Y., Minor, C. M. & Burge, C. B. Concentration-dependent splicing is enabled by Rbfox motifs of intermediate affinity. Nat Struct Mol Biol **27**, 901-912, doi:10.1038/s41594-020-0475-8 (2020).

52 Hu, X. et al. The RNA-binding protein AKAP8 suppresses tumor metastasis by antagonizing EMT-associated alternative splicing. Nat Commun **11**, 486, doi:10.1038/s41467-020-14304-1 (2020).

53 Jourdain, A. A. et al. Loss of LUC7L2 and U1 snRNP subunits shifts energy metabolism from glycolysis to OXPHOS. Mol Cell **81**, 1905-1919 e1912, doi:10.1016/j.molcel.2021.02.033 (2021).

54 Leclair, N. K. et al. Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. Mol Cell **80**, 648-665 e649, doi:10.1016/j.molcel.2020.10.019 (2020).

55 Lau, E. et al. Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. Cell Rep **29**, 3751-3765 e3755, doi:10.1016/j.celrep.2019.11.026 (2019).

56 Daniels, N. J. et al. Functional analyses of human LUC7-like proteins involved in splicing regulation and myeloid neoplasms. Cell Rep **35**, 108989, doi:10.1016/j.celrep.2021.108989 (2021).

57 Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. Nat Rev Genet **17**, 407-421, doi:10.1038/nrg.2016.46 (2016).

58 Demirdjian, L. et al. Detecting Allele-Specific Alternative Splicing from Population-Scale RNA-Seq Data. Am J Hum Genet **107**, 461-472, doi:10.1016/j.ajhg.2020.07.005 (2020).

59 Jin, L. et al. STRAP regulates alternative splicing fidelity during lineage commitment of mouse embryonic stem cells. Nat Commun **11**, 5941, doi:10.1038/s41467-020-19698-6 (2020).

60 Litzler, L. C. et al. PRMT5 is essential for B cell development and germinal center dynamics. Nat Commun **10**, 22, doi:10.1038/s41467-018-07884-6 (2019).

61 Yuanyuan Wang, R. F. C., Samir Adhikari, Christopher M Lopez, Mason Henrich, Vahe Yacoubian, Lan Lin, John S. Adams, Yi Xing. Elucidating dynamics and regulation of alternative splicing in osteogenic differentiation. BioRxiv, doi:https://doi.org/10.1101/2020.10.30.362384 (2020).

62 Yang, J. et al. Guidelines and definitions for research on epithelial-mesenchymal transition. Nat Rev Mol Cell Biol **21**, 341-352, doi:10.1038/s41580-020-0237-9 (2020).

63 Lin, K. T., Ma, W. K., Scharner, J., Liu, Y. R. & Krainer, A. R. A human-specific switch of alternatively spliced AFMID isoforms contributes to TP53 mutations and tumor

recurrence in hepatocellular carcinoma. Genome Res, doi:10.1101/gr.227181.117 (2018).

64 Irimia, M. et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell **159**, 1511-1523, doi:10.1016/j.cell.2014.11.035 (2014).

65 Li, Y. I., Sanchez-Pulido, L., Haerty, W. & Ponting, C. P. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res **25**, 1-13, doi:10.1101/gr.181990.114 (2015).

66 Chakraborty, P., George, J. T., Tripathi, S., Levine, H. & Jolly, M. K. Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. Front Bioeng Biotechnol **8**, 220, doi:10.3389/fbioe.2020.00220 (2020).

67 Tan, T. Z. et al. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol Med **6**, 1279-1293, doi:10.15252/emmm.201404208 (2014).

68 Klijn, C. et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol **33**, 306-312, doi:10.1038/nbt.3080 (2015).

69 Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

# 3 ELUCIDATING DYNAMICS AND REGULATION OF ALTERNATIVE SPLICING DURING OSTEOGENIC DIFFERENTIATION

## 3.1 Introduction

As humans age the balance of bone and fat content in the skeleton shifts toward increased fat with reciprocal reduction in bone content; this eventually leads to enhanced skeletal fragility [70,71]. Age-related bone loss resulting in osteoporosis and increased fractures represents a major health problem with increased morbidity and decreased quality of life [72,73]. Human bone-forming cells or osteoblasts are derived from bone marrow-derived multipotent stem/stromal progenitor cells (MSPCs). As their name implies, when provoked MSPCs are capable of undergoing osteogenesis, adipogenesis, chondrogenesis and

myogenesis [74,75] with a major differentiation branch point between osteogenesis and adipogenesis. This branch point represents a binary fate choice for MSPC. The signals that specify MSPC to specific cell differentiation pathways are not fully understood but transcription factors are considered by most to be the dominant signal. RUNX2 [76,77] and PPARγ [78,79] are well known transcription factors that have central roles in osteogenesis and adipogenesis, respectively. Our understanding of osteogenic differentiation from MSPC has made tremendous progress in the last several decades. This includes: 1) establishing criteria for MSPC characterization [80-82]; 2) optimizing procedures for isolation [83,84] and culture of MSPC [85,86]; and 3) determining some of the key regulatory molecules, particularly those that control transcription, that modulate MSPC differentiation patterns [87,88].

In the mouse, recent studies in transcription and epigenetic mechanisms regulating MSPC differentiation using single-cell RNA sequencing (scRNA-seq) approaches have yielded important findings with regard to the fate decisions in MSPC differentiation [89-91]. For example, Zhong et al identified a lineage commitment progenitor population that could differentiate to adipocytes or osteocytes [91]. Additionally, two groups [89,90] found that the expression profiles of transcription factors revealed that adipogenesis required a much larger number of expressed transcription factors compared to osteogenesis, and that there existed a subset of transcription factors that acted in both pro-osteogenic and anti-adipogenic manner, supporting a net gain in osteoblasts over adipocytes. One report demonstrated that adipogenesis invoked more chromatin remodeling relative to osteogenesis [89]. In each report noted above, selection of either the osteogenesis or adipogenesis differentiation path appeared to be binary.

In this work, we studied the role of AS during induced osteogenic differentiation of MSPC. We conducted a time course study with primary MSPCs derived from the marrow space of human femurs cultured in osteogenic media. At selected time points, RNA was obtained for RNA-seq. Extensive and dynamic transcriptomic alterations were found to be orchestrated during induced osteogenic differentiation, including remodeling of the gene expression and splicing programs. We developed a multi-step bioinformatic strategy to identify significantly differentially-expressed RBPs that may be key regulators of AS during the process of osteogenesis.

# 3.2 Results

### 3.2.1 Extensive transcriptomic alterations characterize the induced, stepwise differentiation of primary MSPCs to osteoblasts.

In order to evaluate gene expression and AS regulation during osteogenesis, we cultured primary bone marrow-derived MSPCs in osteogenic differentiation media over 12 days to obtain temporal MSPC osteogenic differentiation datasets. Cells were stained for alkaline phosphatase activity and mineralization by alizarin red every two days (**Figure 3.1A**). The anticipated increase in staining intensity was observed qualitatively (**Figure 3.1A**) and quantitatively (**Figure 3.1B**). RNAs were isolated in triplicate on day 0, day 2, day 4, day 6, day 8 and day 12, followed by high-throughput RNA-seq (**Supplementary Figure 3.6A**). The RNA-seq-derived gene expression data showed that markers of immature osteoblasts (MYC and SOX9) were gradually downregulated as markers of osteoblast differentiation (RUNX2 and ALPL) were upregulated during the time course (**Figure 3.1C**), consistent with osteogenic differentiation of the MSPCs. These sequential changes in histological staining

and RNA-seq-derived marker gene expression confirmed MSPC progression toward a mature osteoblast.

A substantial remodeling of the gene expression and AS profiles occurred in osteogenic media-induced osteogenic differentiation of MSPCs. Using rMATS-turbo (rMATS 4.0.2) [35], we uncovered ~23,000 AS events across all 18 samples (**Supplementary Figure 3.6D**), including exon skipping, alternative 5' splice sites, alternative 3' splice sites and intron retention. Because exon skipping is the most prevalent and most well-characterized type of AS events in human transcriptomes [6] and because it represented 72% of the total AS events in this study, we focused on exon skipping events (**Supplementary Figure 3.6D**). Temporal profiling of induced changes in the MSPC transcriptome from RNA-seq was assessed globally by principal component analysis (PCA). The PCA plots of total gene expression, RBP gene expression, and exon skipping showed similar temporal patterns (**Figure 3.1D**), suggesting coordination between total gene expression, RBP gene expression and exon skipping. Moreover, the genes contributing most to the separation of samples on PC1 in the total gene expression PCA plot were highly enriched in RNA splicing-related gene ontology terms (**Supplementary Figure 3.6B**, **3.6C**), further indicating the coupling of gene expression and splicing programs during induced MSPC osteogenic differentiation. Interestingly, genes contributing most to the separation of samples on PC2 in the exon skipping PCA plot were enriched in regulation of transmembrane calcium ion transport terms as well as regulation of RNA splicing terms (**Supplementary Figure 3.6E**, **3.6F**). This is consistent with the observation that the net flux of calcium to the extracellular environment in osteoblasts is closely related to mineralization of the collagenous extracellular matrix [92], again highlighting the importance of AS in osteogenic

differentiation. Together, this transcriptome-wide analysis of osteogenic differentiation identifies interplay of gene expression, splicing and bone development related biological processes.

### 3.2.2 Pair-wise differential analysis identifies temporal patterns of gene expression and exon skipping during MSPC-to-osteoblast differentiation

It has been well established that coordinated splicing networks play a vital role in cell fate determination and result in physiological consequences in various developmental and tissue remodeling processes in humans [7]. In the field of skeletal modeling and remodeling the focus has been on changes in transcription factor gene expression [89]. On the other hand, there is a lack of systematic assessment of splicing networks in skeletal development and maintenance. To fill this knowledge gap and elucidate the dynamics and regulation of AS during osteogenic differentiation, we performed pair-wise comparisons on both gene expression and exon skipping in MSPC over 12 days of induced osteogenesis (**Figure 3.2A**). Beginning in the early stages of osteogenesis (day 0-2), there was a robust and dynamic alteration in the splicing program of MSPCs as they matured to an osteoblast phenotype (**Figure 3.2A**).

To further decipher the relationship between gene expression and exon skipping in the process of induced MSPC osteogenic differentiation, we investigated the expression patterns of splicing-regulated genes. Interestingly, most of those splicing-regulated genes were not differentially expressed (**Figure 3.2B**), suggesting that splicing, in and of itself, represents another layer of transcriptome remodeling during osteogenesis. For example, the filamin A gene (FLNA), of which missense point mutations are associated with a range

of X-linked skeletal dysplasias [93], was persistently highly expressed in both the MSPC and mature osteoblast population of cells with no change in gene expression over the 12-days of induced osteogenesis (**Figure 3.2B, 3.2C**). Despite no change in expression of FLNA, the inclusion level (percent spliced in, PSI) of the alternatively spliced exon 30 (ENST00000369850), which maps to the functional filamin domain repeat 15, steadily decreased from 0.72 to 0.52 over 12 days of osteogenic induction (**Figure 3.2C**).

Notably, of 604 exon skipping events (from 488 genes) with significant PSI value changes in at least 3 pair-wise comparisons, 53 were located in 47 genes encoding transcription regulators (**Supplementary Table 3.1**). Those exon skipping-regulated transcription regulators were often affected by 1) loss of a portion or the entirety of a functional domain encoded by the alternatively spliced exon, 2) a frame shift resulting in disruption or presence of downstream functional domains of the translated protein or 3) nonsense-mediated decay (**Supplementary Table 3.1**); either of these scenarios can lead to loss/gain of transcriptional action of the protein with a corresponding global reconstruction of the gene expression network. Exon skipping events residing in transcription regulator genes without frameshift, NMD induction or functional domain encoding exons, can also exert functional consequences. An example is differential inclusion/exclusion of exon 4 in the transcriptional co-activator gene, paired mesoderm homeobox protein 1 gene (PRRX1; also known as PRX1) (**Supplementary Figure 3.7**). Two isoforms were produced by alternative splicing of the cassette exon 4 in PRRX1: the exon 4 skipping isoform, PRRX1A and exon 4 inclusion isoform, PRRX1B. Interestingly, the inclusion of exon 4 in the PRRX1 gene introduced an early stop codon, shifting exon 5 to the 3' untranslated region encoding a shorter protein isoform (PRRX1b) lacking the functional

OAR (otp, aristaless, and rax) domain (**Supplementary Figure 3.7A, 3.7B**). The OAR domain is involved in DNA binding and inhibition of transcription activation [94,95]. Previous studies have established that PRRX1a and PRRX1b act differentially to regulate progenitor cell proliferation and differentiation [96,97]. Indeed, the sashimi plot in **Supplementary Figure 3.7C** shows that the inclusion level of exon 4 decreased from 63% on day 0 to 41% on day 12, resulting in an isoform switch from the OAR-absent PRRX1B to the OAR-containing PRRX1A during induced MSPC osteogenesis. This suggests that OAR-containing PRRX1A favors an osteogenic differentiation fate more than does the OAR-absent PRRX1B. This result parallels with previous observations in the mouse system where overexpression of Prrx1b, but not Prxx1a, interferes with Osx- and Runx2-directed mRNA expression and inhibits pre-osteoblast-to-osteoblast differentiation [98].

Many AS events are coordinately regulated by trans-acting RBPs in a developmental stage-specific manner [2,99]. The global changes in RBP gene expression (**Figure 3.2D**) and exon skipping PSI values (**Figure 3.2F**) were displayed as heatmaps; each defined by 5 clusters with distinct inter-cluster temporal patterns determined by unsupervised hierarchical clustering. Overall, the most pronounced changes for both RBP gene expression and exon skipping were observed in the early stages of induced osteogenesis (from day 0 to day 2, **Figure 3.2D, 3.2F**). The temporal RBP gene expression patterns and temporal exon skipping PSI value patterns were further delineated into lines (**Figure 3.2E, 3.2G**). Distinct patterns of RBP gene expression emerged from different clusters (**Figure 3.2E**): steady upregulation (e.g., cluster 3); steady downregulation (e.g., cluster 2); early stage change (e.g., cluster 1) or late stage change (e.g., cluster 5). Distinct patterns of exon skipping PSI values were also observed (**Figure 3.2G**). This included

clusters of exons with: progressive inclusion (e.g., cluster 5); progressive exclusion (e.g., clusters 3) and early stage change (e.g., clusters 1 and 2).

### 3.2.3 Computational screening identifies RBP candidates for regulation of exon skipping during osteogenic differentiation

To further identify RBPs that could control the splicing program in induced MSPC osteogenic differentiation, we developed a computational screening method searching for potential 'key splicing regulators' among 129 RBPs with known motif position weight matrix [100,101]. Trans-acting regulatory RBPs usually bind to cis-acting RNA elements in the precursor mRNA (pre-mRNA) in a sequence-specific manner with the resulting RBP-directed splicing behavior frequently dependent on the location of the RBP binding site relative to the regulated exon [2,102]. Taking this cis-trans locational information into consideration, the method described here combineed both correlation analysis and region-specific motif enrichment analysis to estimate the region-specific regulatory potential of RBPs to bind to and influence pre-mRNA splicing (**Figure 3.3A**). Detected exon skipping events not significantly changed in any of the 5 pair-wise comparisons with day 0 were classified as background events, while exon skipping events that were significantly changed in at least three of the five comparisons were retrieved as foreground events. For each differentially expressed RBP, foreground events were assigned into a 'positive' or 'negative' foreground event set based on the correlation coefficient between its PSI value and the RBP gene expression if highly correlated ($r^2$>0.5) (**Figure 3.3A, top**). Sequences extracted from each exon skipping event were subsequently scanned for the presence of an RBP binding site based on the motif scores calculated from matching the position weight matrix of the RBP motifs to possible binding positions (**Figure 3.3A, middle**). Finally, one-tailed Fisher's

exact tests were performed to estimate the region-specific enrichment of motifs in the foreground event sets compared to the background event set (**Figure 3.3A, bottom**). Eleven RNA binding motifs for nine RBPs were determined to be significantly enriched in at least one region for at least one foreground event set (**Figure 3.3B, 3.3C**).

KH RNA Binding Domain Containing, Signal Transduction Associated 3B, (KHDRBS3, SLM2, T-STAR) was one of the identified RBP splicing regulators with its RNA binding motif enriched in exon body region from negatively-correlated foreground event set and showed a temporal increase in expression during induced osteogenic differentiation (**Figure 3.3B, 3.3C**). Previous studies reported that KHDRBS3 and another member from the STAR family, KHDRBS1 (SAM68), are involved in splicing regulation in a variety of developmental and disease processes [7,103-107]. Examples of two significantly changing exon skipping events, unaccompanied by a difference in gene expression, are shown in **Supplementary Figure 3.8**. During induced osteogenesis two potential KHDRBS3 pre-mRNA targets, the transcription factor coregulator gene AIRD4B (RBP1L1, RBBP1L1) and the periostin (POSTN) gene were alternatively spliced but not differentially expressed. In both instances a KHDRBS3 RNA binding motif was present in the body of the regulated exon with the PSI value of the target exon being significantly negatively correlated with KHDRBS3 gene expression (**Supplementary Figure 3.10, 3.8A, 3.8C**).

ARID4B and ARID4A (RBP1, RBBP1) are two homologous members of the AT-rich interaction domain (ARID) gene superfamily [108]; they encode subunits of the SIN3 Transcription Regulator Family Member A (SIN3A)/HDAC (histone deacetylase) transcriptional corepressor complex which functions in various cellular processes including proliferation, differentiation and cell fate decision [109,110]. As depicted in

**Supplementary Figure 3.8B**, the alternatively spliced cassette exon 16 in ARID4B, which encodes part of the Tudor-knot domain, is significantly more often skipped (PSI diminished) in day-12 compared to day-0 differentiating MSPC. Compared to the traditional Tudor domain, which is involved in protein-protein interaction [111], the Tudor-knot contains crucial configurations needed for RNA binding activity [112]. Therefore, partial disruption of the Tudor-knot during induced osteogenesis might result in changes in the assembly or stability of ARID4B involved supramolecular complexes. For example, in the mouse Arid4a is reported to be a Runx2 coactivator that promotes osteoblastic differentiation [113]. Considering the fact that ARID4A and ARID4B can also physically interact with each other [114], the alteration of their protein-protein or protein-RNA interaction might squelch ARID4A's function as a RUNX2 coactivator.

Periostin is a secreted extracellular matrix protein that was originally identified in cells from the mesenchymal lineage in the skeleton (e.g., osteoblasts and osteoblast-derived cells). Periostin expression promotes bone anabolism partially though its ability to regulate osteoblast differentiation from mesenchymal progenitors [115]; deletion of Postn gene impairs fracture consolidation in mice [116]. As was the case with ARID4B, POSTN expression did not change significantly over the course of induced osteogenesis but exon skipping in POSTN pre-mRNA did, indicating that POSTN has the potential to contribute to osteogenic differentiation without a change in expression level. There are multiple isoforms of periostin reported, all differing in their C-terminal sequences [117]. This is consistent with our observation that the inclusion level of exon 18 which resides in the C-terminal region of POSTN decreased from 76% to 50% over the 12-day course of MSPC osteogenesis (**Supplementary Figure 3.8D**).

Cytoplasmic Polyadenylation Element Binding Protein 2 (CPEB2) was another of the identified RBP splicing regulators. CPEB2 showed a temporal increase in expression during induced osteogenic differentiation (**Figure 3.3B, 3.3C**) like KHDRBS3. The CPEB2's RNA binding motif was found enriched in the downstream intron region of some skipped exons in the negatively-correlated foreground event set. CPEB family proteins function as regulators of cytoplasmatic polyadenylation and translation of target mRNAs (Hagele et al., 2009; Turimella et al., 2014); however, recent studies report the involvement of CPEB family proteins in splicing regulation in a variety of developmental and disease processes [7,103-107].

One significant differentially skipped exon event unaccompanied by a difference in gene expression during induced osteogenesis was found in the transcription factor FOXM1 (HFH-11, HNF-3, MPP-2, INS-1) (**Supplementary Figure 3.9**). In the downstream intron of this FOXM1 skipped exon XX a CPEB2 RNA binding motif was present and PSI value change of this exon was negatively correlated with CPEB2 gene expression (**Supplementary Figure 3.10, 3.9A**). FOXM1 is a member of the forkhead-box family of transcription factors, which function in various cellular processes, principally proliferation (Costa et al., 2003; Laoukili et al., 2007), with dysfunction associated with a number of human diseases (Banayoun et al., 2011). With regard to cell fate decisions and MSPC differentiation specifically, FOXM1 has been shown in both human embryonic kidney (Zhang et al., 2011) and human osteosarcoma (Zhang et al., 2017) cell lines to interact with β-catenin, promote β-catenin nuclear localization, and stimulate osteogenesis via expression of WNT target genes. Four FOXM1 splice variants arise from AS of exons 6 and 9 (Zhang et al., 2016). Of

most interest has been the inclusion of exon 9, which inactivates FOXM1 via disruption of its trans-activation domain (Kelly et al., 1997). As depicted in **Supplementary Figure 3.9B**, the alternatively spliced exon 9 in FOXM1 was significantly more often skipped (PSI diminished) in day-12 compared to day-0 differentiating MSPC. Considering the disruptive properties of exon 9, skipping this inhibitory exon during induced osteogenesis may enhance FOXM1's pro-osteoblastic function by enriching its interaction with β-catenin.

### 3.2.4 siRNA knockdown of KHDRBS3 and CPEB2 reduce osteogenesis in vitro

As noted above, KHDRBS3 and CPEB2 were two of the nine RBPs that emerged from our computational screening pipeline (**Figure 3.3** and **Supplementary Figure 3.10**). Both RBPs showed a robust increase in expression over time during induced osteogenesis, making them ideal candidates for siRNA knockdown to test the biological significance of KHDRBS3 and CPEB2 during osteogenesis. KHDRBS3 and CPEB2 specific, and non-targeting negative control siRNA knockdown was performed in MSPC effects observed over seven days of induced osteogenic differentiation (**Figure 3.4, 3.5**). As assessed by qPCR, statistically significant knockdowns of KHDRBS3 (**Figure 3.4A**) and CPEB2 (**Figure 3.5A**) relative to non-targeting negative control in all time points and two of three time points respectively was obtained. Additionally, KHDRBS3 and CPEB2 reduction at protein level (**Figure 3.4E, 3.5E**) was also observed in western blots at all post-transfection time points. Diminished osteogenic differentiation was confirmed by significant reduction of the osteogenic maturation markers RUNX2 (**Figure 3.4B, 3.5B**), BGLAP (bone gla protein; **Figure 3.4C, 3.5C**), and ALPL (alkaline phospahatase activity; **Figure 3.4D, 3.5D**) in at least two out of three and in most cases three out of three time points. The knockdown of

these CPEB2 yielded significant reductions in these osteogenic maturation markers as did siRNA knockdown of RUNX2, a known transcription factor crucial to osteogenesis (**Figure 3.5B, 3.5C, 3.5D**). These data indicate that KHDRBS3 and CPEB2 expression normally supports osteogenesis perhaps via action upon downstream target genes (e.g. ARID4B, POSTN, and FOXM1) promoting maturation of MSPC down the osteogenic pathway.

# 3.3 Discussion

While there have been some studies of AS events that occur during tissue-specific adipogenesis [118-121], there is much less known about AS during the differentiation of mesenchymal progenitors down the osteogenic pathway in the human bone marrow niche. Studies of AS during osteogenesis have focused primarily on exon skipping in RUNX2, the master regulator gene for osteogenesis [122]. Skipping exon 5 and/or 7 in RUNX2 pre-mRNA produces isoforms of RUNX2 incapable of DNA binding and downstream transactivation of genes required for normal bone formation, including osterix (OSX), OCN (osteocalcin), OPN (osteopontin) and COL1A1 (Park, et al., 2020). Some of these downstream genes can also be regulated by AS. For example, COL1A1 encodes alpha-1 type I collagen, which is the most plentiful collagen in bone. Aberrant AS of the COL1A1 gives rise to a form of the human skeletal disease osteogenesis imperfecta [123-125].

In this study, we generated a 12-day time-course RNA-seq dataset from primary cultures of MSPCs harvested from a human femur after they were induced to differentiate to bone-forming osteoblasts in vitro. This dataset not only permited a comprehensive examination of stepwise changes for both gene expression and AS, but also opened for the first time the opportunity to examine, in a completely unbiased mode, the potential

regulatory role of differentially expressed RBP-directed changes in AS in MSPC differentiation. We found a high degree of similarity between the temporal patterns of overall gene expression, RBP gene expression and exon skipping (**Figure 3.1D**), suggesting that these events are mechanistically linked. We also observed that genes with the greatest variance in expression were significantly enriched for splicing-related gene ontology terms (**Supplementary Figure 3.6B, 3.6C**) and that a large proportion of the differentially spliced genes encode transcription regulators, defined as transcription factors and co-regulators of transcription (**Supplementary Table 3.1**).

By combining temporal correlation of exon skipping and RBP expression with RBP binding site enrichment in the vicinity of regulated exons (**Figure 3.3**), we present a computational approach to identify key RBPs that drive AS changes in osteogenic differentiation. Considering the fact that AS is often regulated by binding of trans-acting RBPs to cis-acting RNA elements in a position-dependent manner [2,6,126], it is important to understand how the region-specific RBP binding influences nearby AS events. Instead of using the whole set of differential AS events for each candidate RBP, two sets of AS events were determined based on positive or negative temporal correlation with RBP expression in osteogenic differentiation. RBP motif enrichment was assessed in three regions (upstream intron, exon body, downstream intron) as a proxy for region-specific RBP association. It should be noted that this computational strategy is generic and can be applied to any time-course RNA-seq dataset and to any type of AS patterns to elucidate RBP regulation of AS.

In this work, nine RBPs were identified as potential key splicing regulators in the process of MSPC to osteoblast differentiation. Among those were three RBP genes, two of which, KHDRBS3 and CPEB2, exhibited a two-fold increased expression during osteogenesis (**Figure 3.3C**). KHDRBS3 has been found to control cell fate in development or disease [7,106,107], and CPEB2 to impact splicing regulation in a variety of developmental and disease processes [7,103-107]. As depicted in **Figure 3.4 and 3.5**, siRNA knockdown of KHDRBS3 and CPEB2 resulted in a commensurate decrease in osteogenic differentiation markers RUNX2, BGLAP, and RUNX2. This result indicated that an increase in KHDRBS3 and CPEB2 expression influenced normal MSPC osteogenesis and that our computational strategy was successful in identifying key splicing regulators in the dynamic setting of MSPC-to-osteoblast differentiation. Although suggestive, further evidence from RNA-seq identification of altered alternative splicing after KHDRBS3 and CPEB2 siRNA knockdown and KHDRBS3 and CPEB2 CLIP-seq would be needed to prove KHDRBS3 and CPEB2's role in regulating the splicing of the targeted genes identified here.

# 3.4 Methods

## 3.4.1 MSPC culture

Primary cultures of MSPCs were obtained from PromoCell (C-12974, Heidelberg, Germany). Cells were characterized by the vendor according to criteria proposed by the International Society for Cellular Therapy [81]. Lot 402Z027 (47, male, Caucasian) was used in the RNA-seq study reported here. Lot 429Z013.1 (56, male, Caucasian) was used in the siRNA knockdown experiments. MSPCs were initially cultured in recommended growth media (PromoCell, C28009) and differentiated in MSPC osteogenic differentiation medium

(PromoCell, C-28013) on plates coated with human fibronectin (PromoCell, C-43060). Media was changed every two or three days. Samples for osteogenic differentiation RNA-seq data were obtained at day 0, 2, 4, 6, 8, and 12 and daily samples from day 0 to 7 for siRNA knockdown.

### 3.4.2 Cell staining for biomarkers of osteogenic differentiation

Alkaline phosphatase staining reagent (5-Bromo-4-chloro-3-indolyl phosphate/Nitro blue tetrazolium) was prepared from BCIP/NBT tablet (Sigma B-5655, St. Louis, MO) in 10 ml water and incubated on cell monolayers after PBS wash for 10 minutes. BCIP/NBT reagent was removed by washing with PBS-Tween 0.05% followed by a PBS wash. Alkaline phosphatase staining was quantified by spectrophotometry at 620 nm. Alizarin Red S (ARS; Sigma A-5533) was prepared at 2% in water and adjusted to pH 4.1 and filtered before usage. Cells were fixed with 10% buffered formalin (Fisher) and washed with water prior to addition of 2% ARS, pH 4.1 for 20 minutes. Excess ARS stain was washed from cells by water four times. Staining was quantified by spectrophotometry at 405 nm.

### 3.4.3 siRNA knockdown

KHDRBS3, CPEB2, and RUNX2 SMART pool On-Target Plus siRNA (Dharmacon) and negative control On-Target Plus non-targeting pool siRNA (Dharmacon) was used for experiments. Transfection of 7000 cells per well (96-well plates) was conducted with X-treme GENE siRNA transfection reagent (Sigma) at 160 ng siRNA to 1 ul regent ratio. After eight hours, siRNA and transfection reagent containing media was removed and replaced with MSPC osteogenic differentiation medium. Media was changed every two or three days

for a total of seven days. RNA was isolated from 96-well plates using RNeasy 96 (Qiagen). For qPCR gene expression analysis, cDNA was synthesized by SuperScript IV reverse transcriptase (Thermofisher) and qPCR performed with TaqMan Fast Advanced Master Mix (Thermofisher) with eukaryotic 18S rRNA endogenous control probe/primer (ThermoFisher) and gene specific probe/primers: RUNX2 (Hs01047973_m1), BGLAP (Hs01587814_g1), ALPL (Hs01029144_m1), CPEB2 (Hs01039673_m1) and KHDRBS3 (Hs00938827_m1).

### 3.4.4 Western Blot Analysis

Osteogenesis induced hMSC (PromoCell, C-12974, lot 445Z012.1, white male 61 years old) cultured in two wells of a 12-well plate were used at each time-point (day 0, 3, 5, and 7) and treatment condition (non-targeting negative control siRNA, KHDRBS3 or CPEB2 siRNA). Protein samples lysed with RIPA and denatured and reduced with 6x Laemmli sample buffer with incubation at 95C for 5 minutes. Proteins were separated on Bis-Tris 4-20% precast PAGE gel (GenScript, M0065) in Tris-MOPS-SDS running buffer (GenScript, M00138) and transferred to PVDF (Millipore, IPFL10100). After blocking for 1 hour at room temperature, primary antibodies rabbit anti-KHDRBS3 (1:250; Sigma-Aldrich, HPA000981) or rabbit anti-CPEB2 (1:500; Genetex GTX117457) and mouse anti-actin (1:5000; Abcam, ab8226) were incubated with membrane overnight at 4°C. Blots were washed 2x rapidly followed by 3x5 minute washes with PBS/0.1% Tween-20. Secondary antibodies 680RD anti-mouse (Licor, 926-68070) and 800CW anti-rabbit-800 (Licor, 926-32211) were incubated at 1:20,000 each for 1 hour at RT. Blots were washed 2x rapidly followed by 3x5 minute washes with PBS/0.1% Tween-20. Blots were imaged and

quantitated on Licor Odyssey CLx imaging system. KHDRBS3 and CPEB2 signals were normalized against actin and expressed relative to day 0 pre-transfection cells.

### 3.4.5 RNA isolation and sequencing library preparation

For RNA-seq, RNA was extracted from 24-well plate MSPC cultures at 0, 2, 4 6, 8 and 12 days of induced osteogenesis with Trizol (ThermoFisher) and purified with Direct-zol RNA microprep columns (Zymo Research). Three biological replicates were isolated at each time point. RNA-seq libraries were prepared with TruSeq Stranded mRNA Library Kit (lllumina) after which RNA was assessed for quality by Tape Station (Agilent) and quantified by Qubit (ThermoFisher). RNA-seq libraries were pooled, quantified by Qubit 3.0, diluted accordingly and committed to Illumina Paired End 101 base sequencing at the UCLA Broad Stem Cell Research Center High Throughput Sequencing Facility.

### 3.4.6 RNA-seq read alignment

High-quality raw sequencing reads were obtained and assigned to a corresponding sample by demultiplexing with a maximum of 1 mismatch allowed in the barcode sequence (barcode sequence length 7). Alignment was done using Hisat2 (v2.0.3-beta) [127] with default parameters and a pre-built index for reference plus transcripts based on genome assembly GRCh37 (hg19) annotation (grch37_tran, ftp://ftp.ccb.jhu.edu/pub/infphilo/hisat2/data/grch37_tran.tar.gz).

### 3.4.7 Gene expression quantification and differential gene expression analysis

Gene expression/transcript abundance were measured in both raw counts and TPM (Transcripts Per Million) using the alignment tool kallisto (v0.43.1) [128]. Ensemble v75 GRCh37 (hg19) cDNA annotation was used as the guiding reference for kallisto. Transcript-level estimates from kallisto were summarized into gene expression matrices by tximport (v1.6.0, R package) [129] for downstream gene-level analysis. Differential expression analysis was conducted with the count-based tool DeSeq2 (v1.18.1, R package) [130]. Technical replicates were collapsed, and lowly expressed genes (TPM ≤ 5 in all samples) were filtered out before performing differential expression analysis. For each comparison, genes with an absolute log2 fold change > log2(1.5) and an FDR (false discovery rate)-adjusted p-value < 0.01 were assumed to be differentially expressed genes. The differentially expressed gene list for the entire osteogenic differentiation pathway was defined as genes differentially expressed in all comparisons between time point day 0 and other time points (day 2, 4, 6, 8, 12).

### 3.4.8 Alternative splicing analysis to identify significantly changing foreground events and background events

AS events were detected and quantified by rMATS-turbo [35], with Ensemble v75 GRCH37 (hg19) GTF annotation. Exon inclusion levels, measured as PSI values, were calculated by junction reads (reads spanning the splicing junctions) normalized by effective junction length. AS events with low junction read support (≤ 10 average junction reads, ≤ 10 total inclusion junction reads or ≤ 10 total skipping junction reads over all 18 samples), or with extreme PSI value ranges (PSI ≤ 0.05 or ≥ 0.95 in all 18 samples) were excluded from

downstream analysis. Differential exon skipping analysis was then performed using rMATS-turbo (with default parameter -c 0.0001) for five pair-wise comparisons between time point day 0 and other available time points (day 2, 4, 6, 8, 12). Exon skipping events for each comparison were considered differential if they met the following criteria: 1) >10 average junction reads (inclusion and skipping junction reads) in both groups; 2) do not have extreme PSI values (PSI ≤ 0.05 or PSI ≥ 0.95 for all 6 samples in the comparison); 3) FDR < 0.01; and 4) absolute change in PSI (|ΔPSI|) > 0.05. The significant event set for the whole osteogenic differentiation pathway was composed of events that were identified to be differentially spliced in at least 3 of the 5 comparisons. The background event set for the whole osteogenic differentiation pathway was defined as events with no significant change during MSPC osteogenic differentiation which meet the following cutoffs in all 5 comparisons: 1) > 10 average junction reads in both groups; 2) do not have extreme PSI values (PSI ≤ 0.05 or PSI ≥ 0.95 for all 6 samples in the comparison; and 3) FDR > 0.5.

### 3.4.9 Principle component analysis (PCA)

A total of 129 RBPs (including many well-characterized splicing factors) were curated from two different sources [100,101] and included in PCA analysis of RBP expression. For total gene or RBP expression, a pseudo-count of 1 was added to each TPM value before log2 transformation to avoid arithmetic error and large negative values. PCA was then performed after removing genes/RBPs/exon skipping events with no variance among samples. Samples were projected to their PC1-PC2 space by PCA score. LOESS (locally estimated scatterplot smoothing) regression lines with 95% confidence intervals were added to the PC1-PC2 plot using R package ggplot2 (v3.1.0).

### 3.4.10 Gene set enrichment analysis (GSEA)

GSEA (v3.0) software [131] was utilized on pre-ranked gene lists based on absolute values of principle component loadings from PCA. Ranked lists from total gene expression PCA included only the top 10,000 genes; for ranked lists from exon skipping, duplicated genes with lower rank were removed. All gene ontology gene sets (c5, v7.0, https://www.gsea-msigdb.org/gsea/msigdb/download_file.jsp?filePath=/msigdb/release/7.0/c5.all.v7.0.symbols.gmt ) were used as a gene sets database with 1000 permutations to calculate the enrichment score and p-values. Top gene ontology terms from the GSEA analysis served as input for REViGO webserver (http://revigo.irb.hr/) to account for the semantic similarities and dispersibilities of gene sets. Representative gene ontology terms were visualized in semantic similarity-based scatter plots after removing redundant terms.

### 3.4.11 Hierarchical clustering of time course datasets and heatmaps

Hierarchical clustering was performed on Z-score transformed TPMs of differentially expressed RBPs (from a total of 1542 RBPs) [132] or PSI values of differentially spliced exons detected in the whole osteogenesis pathway as described above (hclust function in R package stats, v3.4.4).

### 3.4.12 Protein family domain analysis

Pfam domain scanning was conducted to search for potential functional domains affected by exon skipping events. Preprocessed Pfam annotation data, which maps HMM predicted high-quality Pfam-A domains to UCSC hg19 coordinates, were downloaded from the UCSC hg19 annotation database

(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/ucscGenePfam.txt.gz). Sequences of target exons or frameshifted downstream exons from differential exon skipping events were extracted and then scanned against the Pfam annotation data using bedtools (v2.25.0) [133].

### 3.4.13 RBP candidate screening method

189 RBP binding motifs with position weight matrix information for 129 RBPs (including many well-characterized splicing factors) were curated from two different sources and screened in this analysis. This includes 78 6-mer motifs for 78 RBPs from RNA Bind-n-Seq (RBNS) [100] and 111 7-mer motifs for 82 RBPs from RNAcompete [101].

Significant/background event lists and differentially expressed RBPs were identified as described before (see AS analysis section and gene expression analysis section). For each motif of the differentially expressed RBPs, significant exon skipping events were further assigned to two foreground event sets. They were composed of events ($n > 50$) whose PSI value was positively or negatively correlated ($R^2 > 0.5$, rlm function in R package MASS, v7.3-49) with differential RBP gene expression across different time points.

To identify region-specific RBP regulatory patterns for exon skipping events, we evaluated three regions around the alternatively spliced exons: 1) 300 nt of intronic sequence upstream of the target exon; 2) the exon body sequences; and 3) 300 nt of intronic sequence downstream the target exon. Scores for each motif were calculated by sliding window scanning of the position weight matrix at each possible binding position. Region-specific motif occurrence was then determined by comparing the calculated motif scores with a threshold score (80% of the maximum PWM score). If there was any position with a

calculated motif score ≥ the threshold score for a particular exon skipping event, then the motif occurrence was marked as "True" for this event in the corresponding region; otherwise it was marked "False".

To determine whether a motif occurred in a specific region more often in foreground event sets than in the background event set, a one-tailed Fisher's exact test was used to test the null hypothesis that the number of events with motif occurrence at a specific region was not different between the foreground (either positive or negative set) and the background event set. If an RBP motif was of significantly enriched occurrence ($p < 0.05$) in any region for any foreground event set, it was considered a key RBP for exon skipping regulation in the MSPC osteogenic differentiation process.

# 3.5 Figures

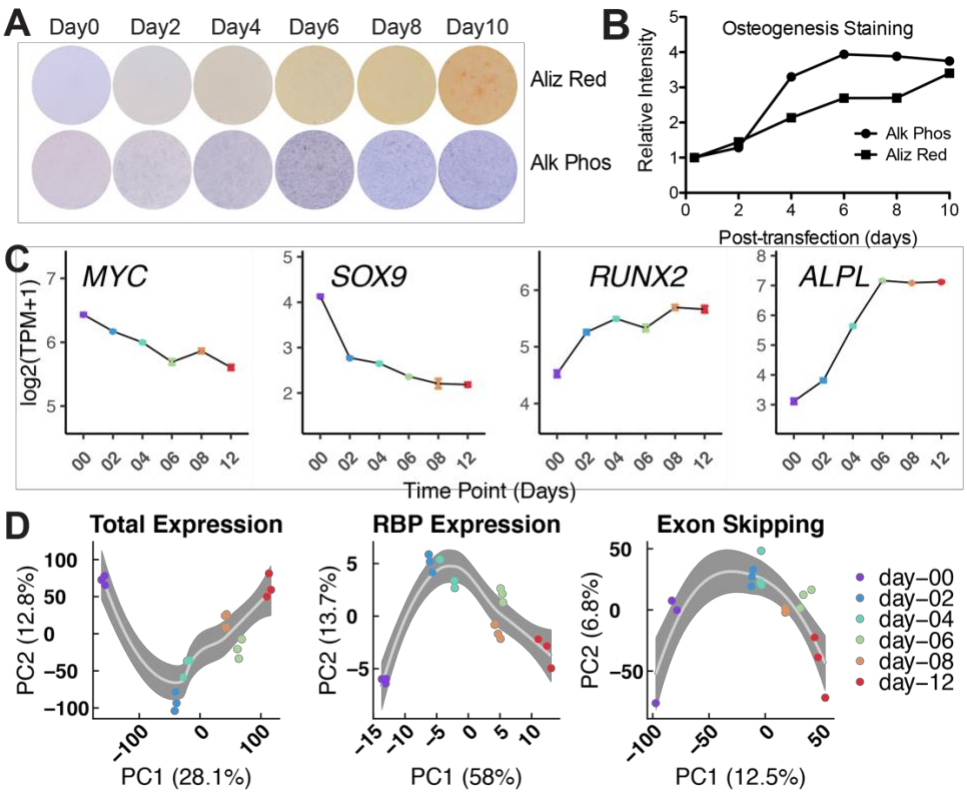



**Figure 3.1 Extensive transcriptomic alterations characterize the induced stepwise differentiation of MSPC to osteoblasts.**

**(A)** Photographs of blue alkaline phosphatase and red hydroxyapatite (bone calcification) staining of MSPCs cultured in osteoblast differentiation medium over 10 days.

**(B)** Measurement of optical density in photographed wells in panel A.

**(C)** From-left-to-right are normalized expression measures from RNA-seq at indicated time points, showing a temporal decrease in immature marker (MYC and SOX9) expression and increase in osteoblast marker (RUNX2 and ALPL) expression. Error bars represent the MEAN±SEM (n=3).

**(D)** From left-to-right, principal component analysis (PCA) plots of total gene expression (left), RBP gene expression (middle), and exon skipping (right). Samples were projected to the space of the first two principal components (PCs) with percentage of variation explained shown in x- and y-axis labels. Local regression lines are added by LOESS (locally estimated scatterplot smoothing) method, with 95% confidence intervals shown in grey.

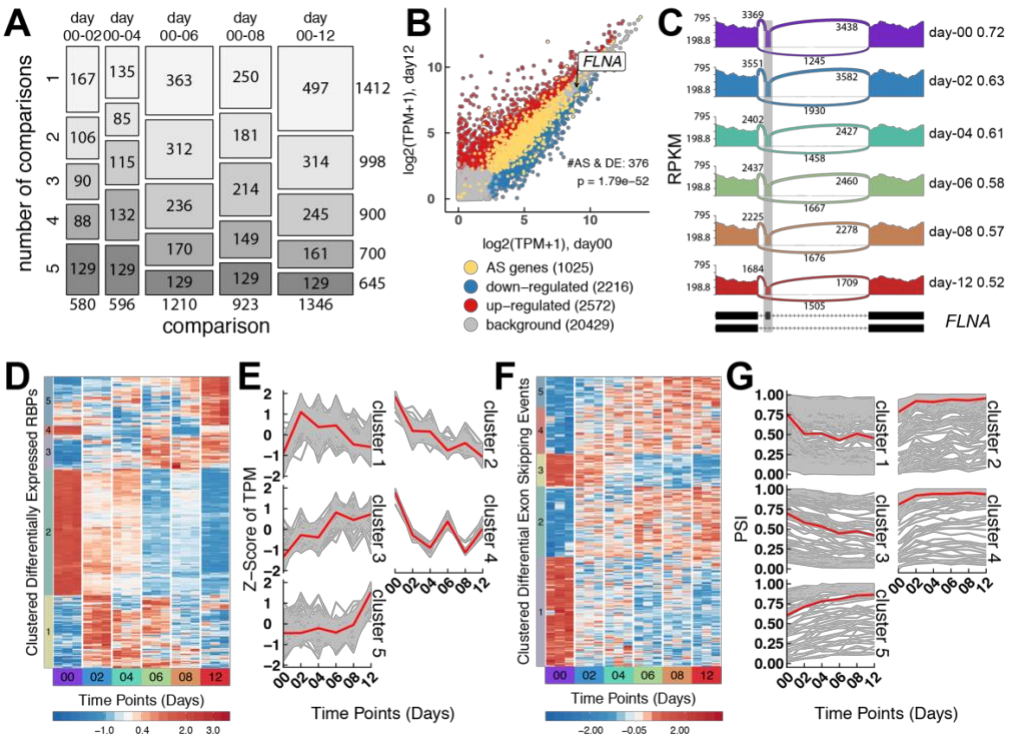



**Figure 3.2 Pair-wise differential analysis identifies temporal patterns of gene expression and exon skipping during MSPC-to-osteoblast differentiation.**

**(A)** Mosaic plot display of significant exon skipping events in pair-wise comparisons with day 0. Each box is constructed by two factors, the specific comparison (x-axis) and total number of comparisons (y-axis) where events were identified as significant. The size of the box and number shown for each box represent the number of exon skipping events in each category.

**(B)** Scatter plot comparing gene expression on day 0 vs. day 12. Red and blue dots represent up-regulated and down-regulated genes, respectively. Genes with significant exon skipping changes (denoted as AS genes) are depicted in yellow. Background genes, e.g., those with no differential gene expression and no exon skipping changes, are shown in grey.

**(C)** Sashimi plot of FLNA gene from panel B is an example of a gene whose expression is not significantly changed but does harbor a significant change in exon skipping over time. The black bars and dashed lines on the bottom represent exons and introns, respectively. Solid peaks represent reads per kilobase per million mapped (RPKM) mapped to each region. Arches represent splice junctions and the numbers represents number of reads mapped to each splice junction. PSI values are indicated on the right side of the plot for each time point.

**(D-G)** Heatmaps and line graphs illustrate the temporal coordination among RBP gene expression and exon skipping during induced osteogenic differentiation. Panel **D**, heatmap showing Z-score transformed transcripts per million (TPM) for 523 differentially expressed RBP genes; panel **E**, line plots revealing the patterns of change for corresponding clusters in panel D; panel **F**, heatmap showing Z-score transformed PSI values for 604 significantly changed exon skipping events; panel **G**, line plots revealing the patterns of change in PSI value for corresponding clusters in panel F. Red line represents the median value for each cluster.

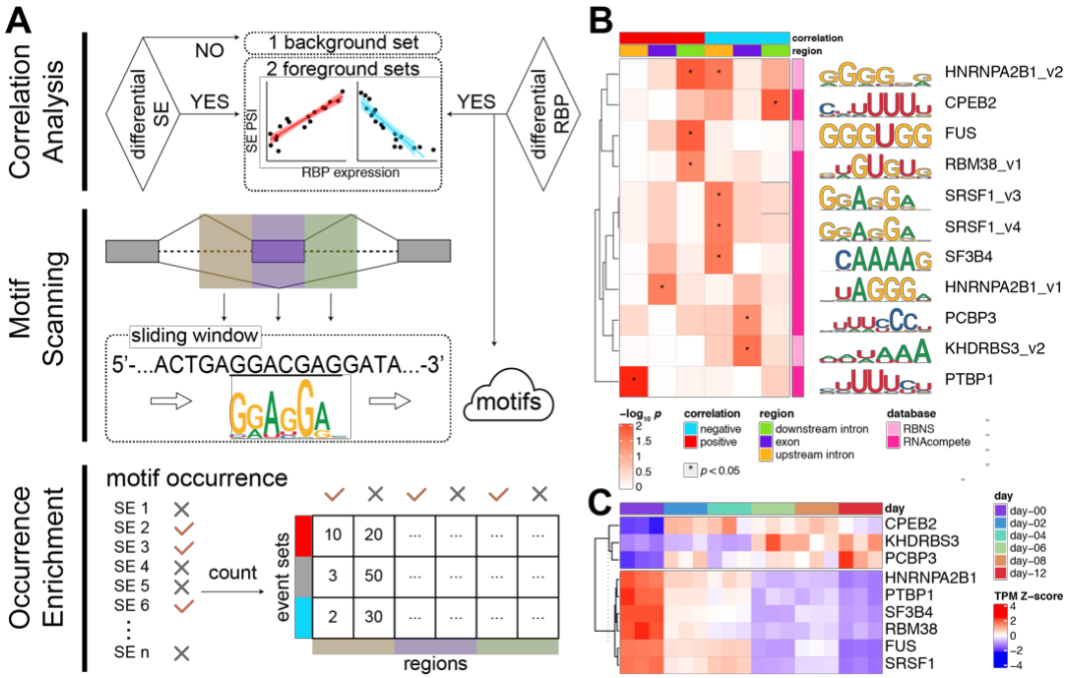



**Figure 3.3 Computational screening identifies nine RBP candidates that may regulate exon skipping in osteogenic differentiation.**

**(A)** Schematic diagram describing the computational workflow for key splicing regulator screening. As depicted in the top panel, differential exon skipping analysis was performed to determine a background event set. For each differentially expressed RBP, significantly changing exon skipping events during osteogenesis were further split into two foreground sets, whose PSI change was positively (red, $r^2 > 0.5$) or negatively (blue, $r^2 > 0.5$) correlated with the corresponding change in RBP gene expression. As shown in the middle panel, a sliding window scanning tool was used to detect putative RBP RNA binding motifs in three regions in and around the skipped exon: 300 nt into the upstream intron (orange); the exon body (purple); and 300 nt into the downstream intron (green). The bottom panel shows a tabular registration of detected motif occurrence. One-tailed Fisher's exact test was performed to ascertain the significance of motif occurrence enrichment in one of the three

regions (upstream intron, exon body, or downstream intron) on the positive (red) or the negative (blue) foreground compared to background (grey). SE, exon skipping.

**(B)** Heatmap depicting the log-transformed p values of 11 candidate RBP motifs (y-axis) exhibiting statistically significant enrichment in at least one of the foreground-region combinations; p values < 0.05 are marked by asterisk. Sequence logos are shown on right side of the heatmap. The two foreground event sets and the three regions are indicated on the top of the heatmap.

**(C)** Heatmap depicting Z-score of transformed TPM values of 9 candidate RBPs identified by the computational screening method.

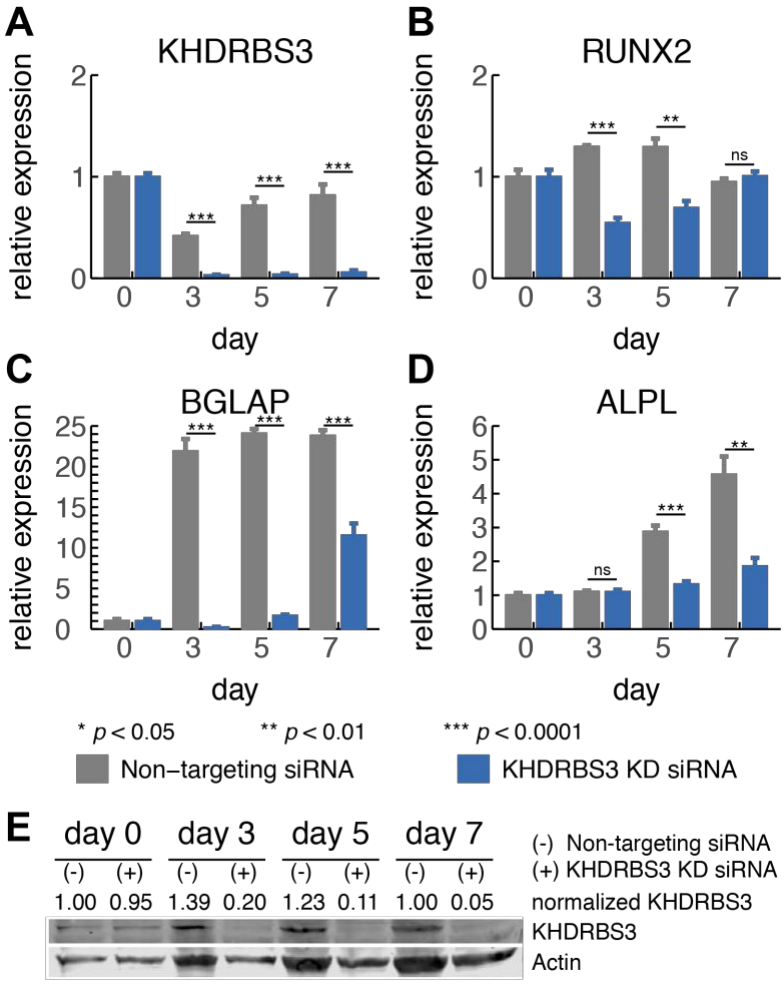



**Figure 3.4 KHDRBS3 knockdown reduced osteogenic differentiation of MSPC.**

KHDRBS3-specific siRNA knockdown (n=4) compared to non-targeting siRNA (n=4) at time-points post-transfection and exposure to osteogenic differentiation media was assayed by qPCR for effects upon: (A) KHDRBS3, (B) RUNX2, (C) BGLAP, and (D) ALPL expression; and (E) KHDRBS3 protein expression by western blot and normalized to actin expression. For qPCR graphs (panel ABCD) error bars represent the MEAN±SD (n=4). Asterisks indicate significance by two-tailed t-test. For the western blot (panel E), pt = pre-transfection.

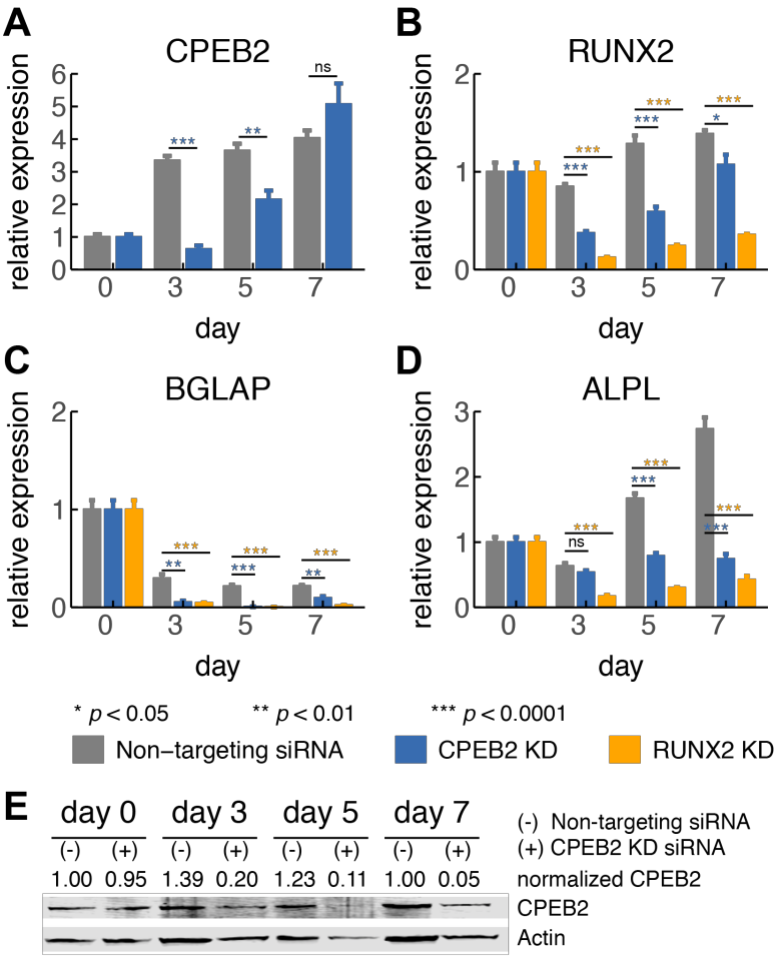
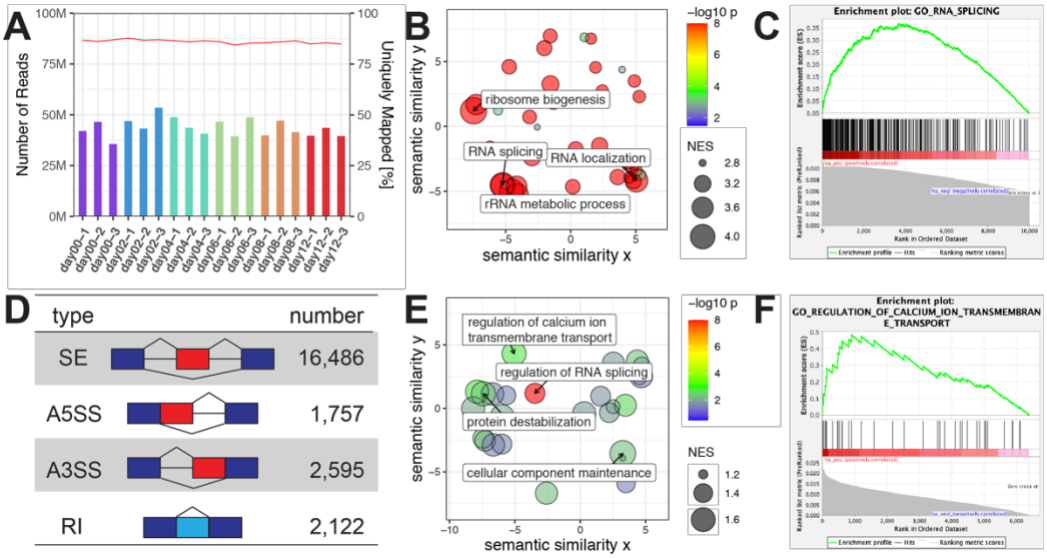



**Figure 3.5 CPEB2 knockdown reduced osteogenic differentiation of MSPC.**

CPEB2 specific siRNA knockdown (n=4) compared to non-targeting siRNA (n=4) at time-points post-transfection and exposure to osteogenic differentiation media was assayed by qPCR for effects upon: (A) CPEB2, (B) RUNX2, (C) BGLAP, and (D) ALPL expression; and (E) CPEB2 protein expression by western blot and normalized to actin expression. For qPCR graphs (panel ABCD), error bars represent the MEAN±SD (n=4). RUNX2, a known transcription factor crucial to osteogenesis, specific siRNA knockdown (n=4) is shown by the orange bars in graphs BCD. Asterisks indicate significance by two-tailed t-test. For western blot (panel E), pt = pre-transfection.




**Supplementary Figure 3.6 Transcriptome-wide analysis of osteogenic differentiation identifies interplay of gene expression, splicing and bone development related biological processes.**

**(A)** Summary of read depth and mapping statistics from RNA-Seq dataset.

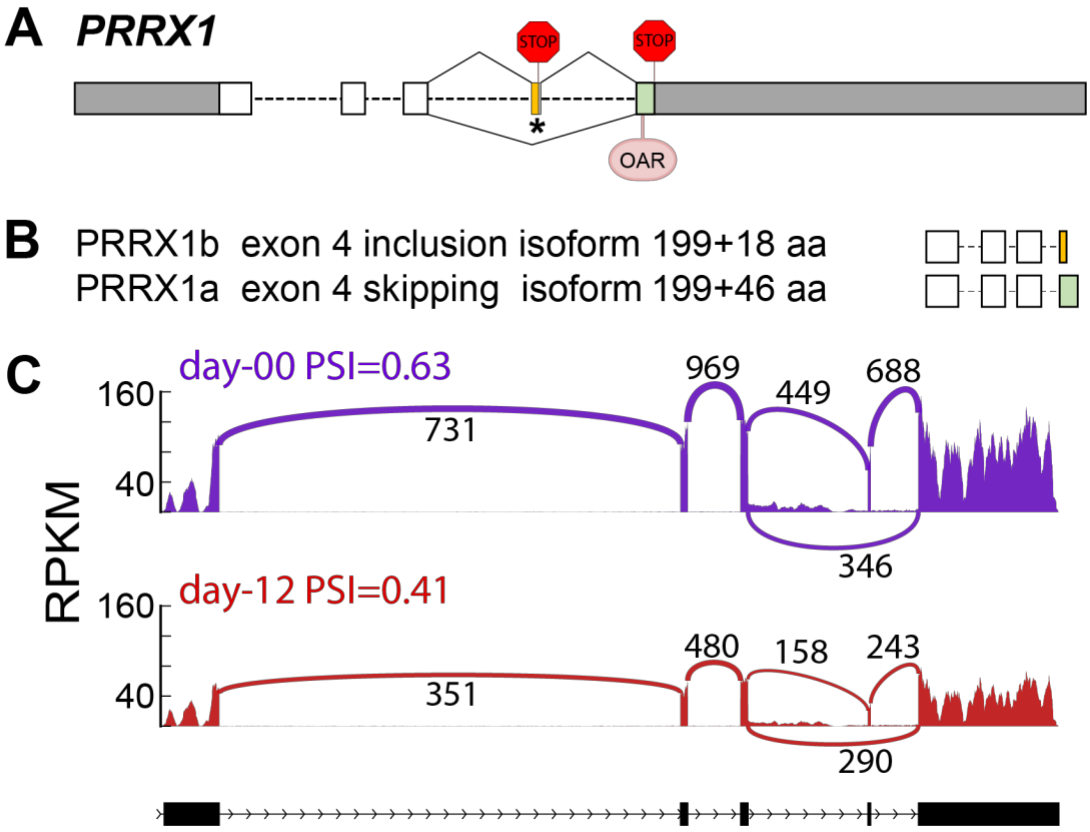
**(B)** REVIGO scatter plot depicting gene ontology terms enriched among genes with high PC1 loading from total gene expression PCA (see left panel, **Figure 3.1D**). Gene ontology enrichment was evaluated by gene set enrichment analysis (GSEA). GSEA calculated enrichment scores and p-values are indicated by size of circles and color scale, respectively; x- and y-axis represent the semantic similarities between terms. Representative gene ontology terms were labeled.

**(C)** Representative GSEA enrichment plot (GO_RNA_SPLICING) from panel A.

**(D)** Summary table of AS events detected by rMATS-turbo after filtering by read coverage and PSI value range. SE, exon skipping; A5SS, alternative 5' splice sites; A3SS, alternative 3' splice sites; RI, intron retention.

**(E)** REVIGO scatter plot depicting GO terms enriched among genes with high PC2 loading from exon skipping PCA (see right panel, **Figure 3.1D**).

**(F)** Representative GSEA enrichment plot (GO_REGULATION_OF_CALCIUM_ION_TRANSMEMBRANE_TRANSPORT) from panel E.

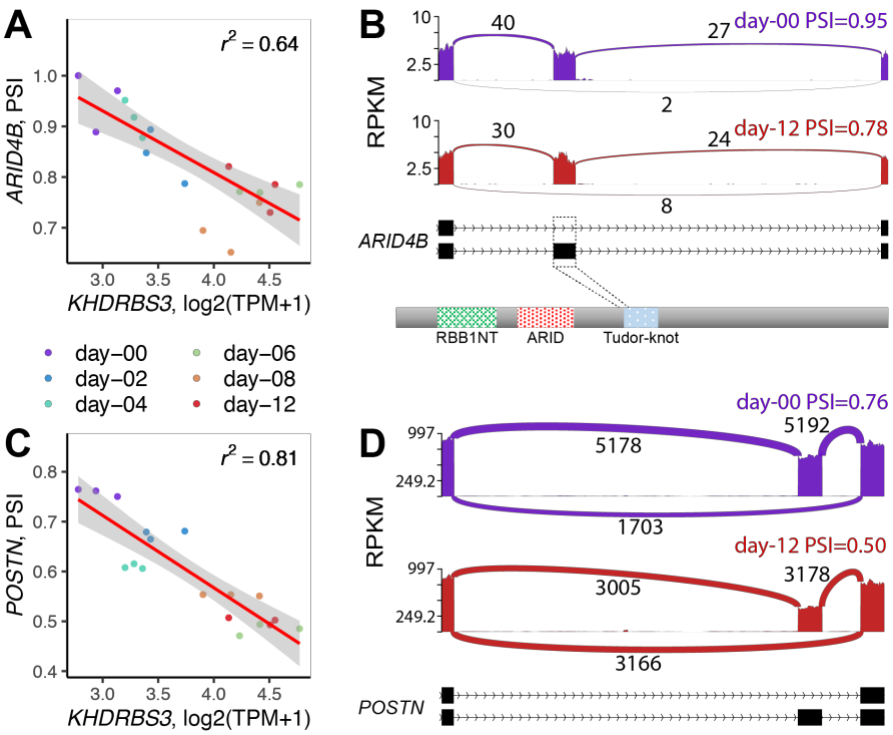



**Supplementary Figure 3.7 Exon 4 inclusion/exclusion of PRRX1 results in an isoform switch.**

**(A)** Gene structure representation of the PRRX1 gene. Bars and dashed lines represent exons and introns, respectively. Untranslated regions (UTR) are denoted in grey. Two stop codons residing in exon 4 and 5 are depicted as STOP signs. The OAR (otp, aristaless, and rax) domain is located in the green region in exon 5.

**(B)** Protein structure representation of the two PRRX1 isoforms. The top PRRX1b isoform has exon 4 incorporated and the bottom PRRX1a isoform has exon 4 excluded. The stop codon in exon 4 of the inclusion isoform renders exon 5 and OAR domain untranslated.

**(C)** Sashimi plot showing AS changes of the PRRX1 gene on day 0 (upper panel) and day12 (lower panel). The black bars and dashed lines on the bottom represent exons and introns, respectively. Solid peaks represent RPKM of reads mapped to each region. Arches represent splice junctions and the numbers represent number of reads mapped to each splice junction.



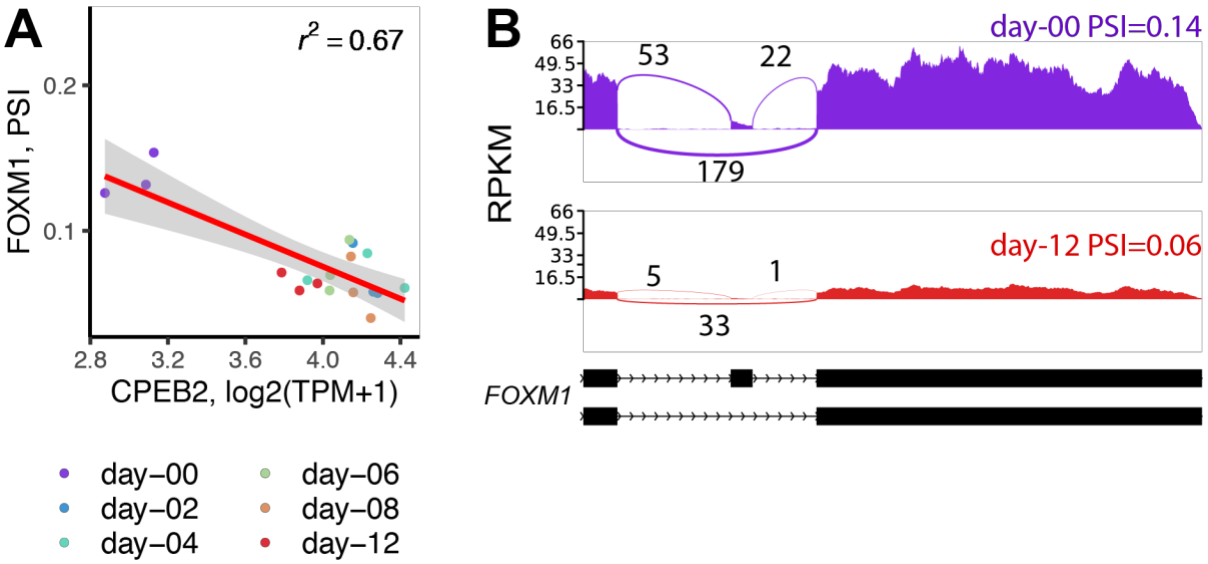

**Supplementary Figure 3.8 Examples of exon skipping events in putative targets of KHDRBS3.**

**(A, C)** Correlation between exon skipping (PSI) and KHDRBS3 gene expression changes over 12 days of induced osteogenesis in MSPC. Genes harboring the significantly changing exon skipping event are indicated on the y-axis. Linear regression lines (red) and confidence intervals (grey) are shown for correlated KHDRBS3 expression and ARID4B PSI (A) and POSTN PSI (C) during induced osteogenesis and $r^2$ values of the correlation are shown in the upper right corner.

**(B)** Sashimi plot of an exon skipping event in the transcription factor gene ARID4B. The black bars and dashed lines on the bottom represent exons and introns, respectively. Solid peaks represent RPKM mapped to each region. Arches represent splice junctions and the

numbers of reads mapped to each splice junction. On the bottom is the depiction of protein family (Pfam) domains in ARID4B protein.

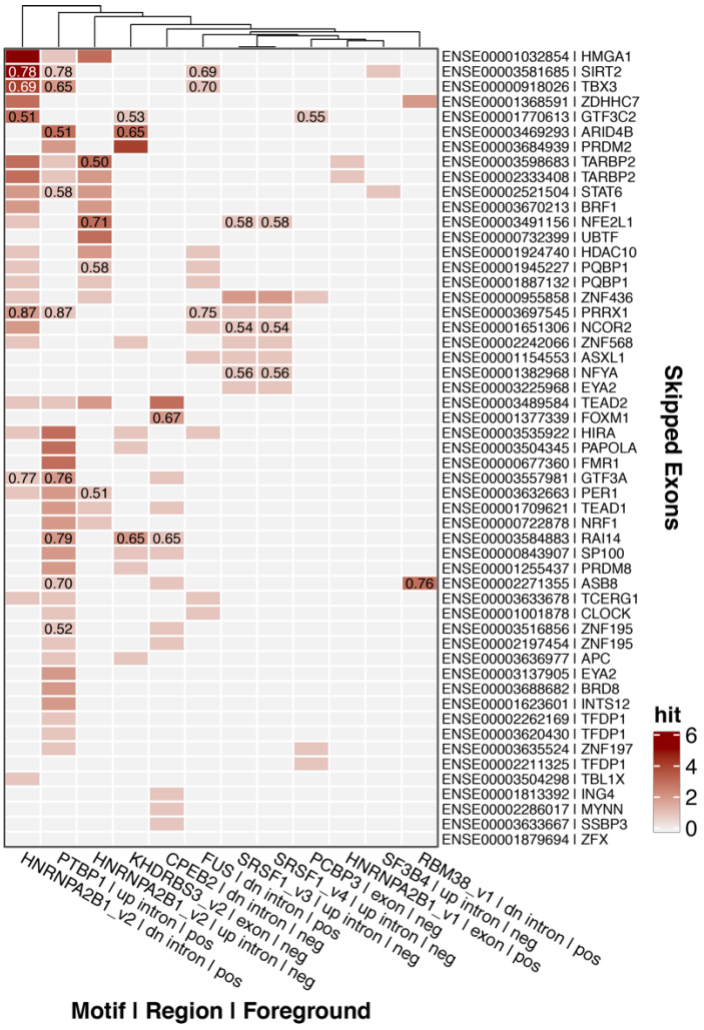
**(D)** Sashimi plot of the exon skipping event in gene POSTN. The target exon 18 is located in the C-terminal sequence of POSTN, where extensive AS changes occur.




**Supplementary Figure 3.9 Examples of exon skipping events in a putative target of CPEB2.**

**(A)** Correlation between FOXM1 exon skipping (PSI) and CPEB2 gene expression changes over 12 days of induced osteogenesis in MSPC. FOXM1 significant exon skipping event PSI value are indicated on the y-axis. Linear regression lines (red), confidence intervals (grey) are shown, and $r^2$ value of the correlation is shown in the upper right corner.

**(B)** Sashimi plot of an exon skipping event in the transcription factor gene FOXM1. The black bars and dashed lines on the bottom represent exons and introns, respectively. Solid peaks represent RPKM mapped to each region. Arches represent splice junctions and the numbers of reads mapped to each splice junction. On the bottom is the depiction of exons 8, 9, and 10 of FOXM1.




**Supplementary Figure 3.10 Heatmap of candidate RBP motif hits in exon skipping events occurring in a transcription factor.**

Candidate RBP motifs (from **Figure 3.3B**) are indicated along the x-axis; additional information includes the region of significant RBP binding motif enrichment (up intron=upstream intron, exon=exon body, dn intron=downstream intron) and direction of correlation (pos=positive or neg=negative). Posted along the y-axis are the gene symbols of the transcription factor and exon identification number in that transcription factor where the differential exon skipping event was detected. The colors of heatmap represent motif

occurrence in the designated region. $r^2$ values > 0.5 for the correlation between RBP gene expression and PSI value of the exon skipping event are indicated numerically in the appropriate boxes.

# 3.6 Tables

| GeneID | geneSymbol | PFAM | frame shift | NMD | $PSI_{00}$-$PSI_{12}$ | FDR(day00-day12) | chrom | strand | exon start | exon end | uptreamES | upstreamEE | downstreamES | downstreamEE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ENSG00000039560 | RAI14 | - | NO | - | 0.168606812 | 0 | chr5 | + | 34813678 | 34813765 | 34812284 | 34812313 | 34814687 | 34814774 |
| ENSG00000116132 | PRRX1 | - | NO | - | 0.21618986 | 1.19E-13 | chr1 | + | 170699417 | 170699489 | 170695360 | 170695542 | 170705188 | 170705374 |
| ENSG00000122034 | GTF3A | zf-C2H2 | YES | - | 0.097629224 | 2.21E-13 | chr13 | + | 28006867 | 28006941 | 28004669 | 28004758 | 28008275 | 28008287 |
| ENSG00000001167 | NFYA | - | NO | - | -0.267664243 | 3.48E-11 | chr6 | + | 41048549 | 41048636 | 41046767 | 41046903 | 41051784 | 41051931 |
| ENSG00000068903 | SIRT2 | - | YES | - | 0.154433978 | 6.01E-08 | chr19 | - | 39389018 | 39389065 | 39384458 | 39384507 | 39390145 | 39390351 |
| ENSG00000108312 | UBTF | HMG_box | NO | - | 0.171796905 | 1.37E-07 | chr17 | - | 42289711 | 42289822 | 42289240 | 42289374 | 42290186 | 42290307 |
| ENSG00000157216 | SSBP3 | SSDP | NO | - | 0.092628716 | 2.68E-07 | chr1 | - | 54723741 | 54723822 | 54722799 | 54722859 | 54747110 | 54747200 |
| ENSG00000135111 | TBX3 | T-box | NO | - | 0.27274122 | 6.55E-07 | chr12 | - | 115117717 | 115117777 | 115117309 | 115117456 | 115118683 | 115118951 |
| ENSG00000125945 | ZNF436 | - | NO | - | 0.350190291 | 1.01E-06 | chr1 | - | 23694465 | 23694558 | 23693534 | 23693661 | 23695858 | 23695935 |
| ENSG00000064655 | EYA2 | - | YES | - | 0.218708723 | 1.97E-06 | chr20 | + | 45700823 | 45700891 | 45644819 | 45644936 | 45717877 | 45718020 |
| ENSG00000166888 | STAT6 | STAT_int* | YES | NMD* | 0.09131425 | 2.06E-06 | chr12 | - | 57503825 | 57503893 | 57501945 | 57502082 | 57504911 | 57505157 |
| ENSG00000153786 | ZDHHC7 | - | NO | - | -0.192762388 | 3.03E-06 | chr16 | - | 85022368 | 85022479 | 85015475 | 85015600 | 85023909 | 85024241 |
| ENSG00000082641 | NFE2L1 | - | NO | - | -0.136495916 | 3.44E-06 | chr17 | + | 46134393 | 46134483 | 46133747 | 46133960 | 46134705 | 46134864 |
| ENSG00000113649 | TCERG1 | FF | YES | NMD | -0.16991098 | 6.02E-06 | chr5 | + | 145889629 | 145889723 | 145888707 | 145888808 | 145890003 | 145890328 |
| ENSG00000005801 | ZNF195 | KRAB* | YES | - | 0.276500981 | 6.31E-06 | chr11 | - | 3394806 | 3394886 | 3392806 | 3392933 | 3400267 | 3400355 |
| ENSG00000102103 | PQBP1 | - | YES | - | -0.127281594 | 1.30E-05 | chrX | + | 48759495 | 48759794 | 48759206 | 48759319 | 48760008 | 48760072 |
| ENSG00000198453 | ZNF568 | - | NO | - | -0.263769264 | 1.58E-05 | chr19 | + | 37413487 | 37413748 | 37408480 | 37408550 | 37416101 | 37416160 |
| ENSG00000074219 | TEAD2 | TEA | NO | - | -0.237218087 | 1.65E-05 | chr19 | - | 49859215 | 49859227 | 49858559 | 49858676 | 49860508 | 49860571 |
| ENSG00000196498 | NCOR2 | - | NO | - | -0.093681883 | 2.67E-05 | chr12 | - | 124825147 | 124825240 | 124824839 | 124824989 | 124826368 | 124826601 |
| ENSG00000139546 | TARBP2 | dsrm | YES | NMD* | -0.119554016 | 3.19E-05 | chr12 | + | 53895843 | 53895968 | 53895089 | 53895245 | 53896810 | 53896996 |
| ENSG00000064655 | EYA2 | - | YES | - | 0.191078217 | 4.14E-05 | chr20 | + | 45702796 | 45702974 | 45644819 | 45644936 | 45717877 | 45718020 |
| ENSG00000139546 | TARBP2 | dsrm | YES | NMD* | -0.06649497 | 6.33E-05 | chr12 | + | 53895798 | 53895968 | 53894704 | 53895245 | 53896810 | 53896913 |
| ENSG00000101849 | TBL1X | LisH* | YES | - | -0.273448155 | 6.54E-05 | chrX | + | 9621584 | 9621729 | 9608312 | 9608400 | 9622254 | 9622328 |
| ENSG00000198176 | TFDP1 | DP | YES | - | 0.075586604 | 8.95E-05 | chr13 | + | 114292132 | 114292211 | 114290848 | 114291015 | 114294434 | 114294999 |
| ENSG00000005801 | ZNF195 | - | NO | - | 0.194226144 | 0.000115099 | chr11 | - | 3394806 | 3394905 | 3392806 | 3392933 | 3400267 | 3400369 |
| ENSG00000005889 | ZFX | Zfx_Zfy_act* | YES | - | 0.149878128 | 0.000137914 | chrX | + | 24193505 | 24193560 | 24190872 | 24190917 | 24197299 | 24197667 |
| ENSG00000102103 | PQBP1 | - | YES | - | -0.08218082 | 0.000148055 | chrX | + | 48759509 | 48759661 | 48759206 | 48759319 | 48760008 | 48760070 |
| ENSG00000085274 | MYNN | BTB* | YES | - | 0.183656738 | 0.000159271 | chr3 | + | 169491818 | 169491885 | 169491214 | 169491250 | 169492052 | 169492349 |
| ENSG00000198176 | TFDP1 | DP | YES | - | 0.090569578 | 0.000308031 | chr13 | + | 114292132 | 114292199 | 114290848 | 114291015 | 114294434 | 114294713 |
| ENSG00000115207 | GTF3C2 | - | YES | - | 0.129756297 | 0.000350349 | chr2 | - | 27573180 | 27573499 | 27566197 | 27566445 | 27579605 | 27579866 |
| ENSG00000111653 | ING4 | ING | NO | NMD | -0.060823976 | 0.000376274 | chr12 | - | 6764803 | 6765079 | 6762395 | 6762562 | 6765892 | 6765964 |
| ENSG00000100084 | HIRA | - | NO | - | -0.06751115 | 0.000488248 | chr22 | - | 19365391 | 19365589 | 19363153 | 19363315 | 19371142 | 19371228 |
| ENSG00000134852 | CLOCK | - | YES | - | -0.135910675 | 0.000584389 | chr4 | - | 56376078 | 56376232 | 56355540 | 56355632 | 56412648 | 56412813 |
| ENSG00000177981 | ASB8 | Ank_3 | NO | - | 0.067965541 | 0.00060343 | chr12 | - | 48543869 | 48543923 | 48543311 | 48543781 | 48544983 | 48545088 |
| ENSG00000116731 | PRDM2 | - | NO | - | 0.203212682 | 0.000658775 | chr1 | + | 14099572 | 14099683 | 14075855 | 14075982 | 14142921 | 14143065 |
| ENSG00000090060 | PAPOLA | - | NO | - | -0.054834451 | 0.000925408 | chr14 | + | 97026985 | 97027048 | 97022511 | 97022750 | 97029155 | 97029230 |
| ENSG00000054267 | ARID4B | Tudor-knot | NO | - | 0.173855012 | 0.000945713 | chr1 | - | 235377083 | 235377341 | 235359345 | 235359430 | 235383107 | 235383283 |
| ENSG00000112983 | BRD8 | - | YES | - | -0.057979735 | 0.001008484 | chr5 | - | 137499775 | 137499822 | 137498818 | 137499033 | 137500008 | 137500102 |
| ENSG00000187079 | TEAD1 | TEA | NO | - | -0.205878924 | 0.001083303 | chr11 | + | 12900435 | 12900447 | 12886384 | 12886447 | 12901254 | 12901389 |
| ENSG00000179094 | PER1 | - | NO | - | -0.073428664 | 0.001235771 | chr17 | - | 8049275 | 8049455 | 8047372 | 8048311 | 8049689 | 8049806 |
| ENSG00000106459 | NRF1 | Nrf1_DNA-bind | YES | - | -0.105166536 | 0.001602051 | chr7 | + | 129311268 | 129311383 | 129297182 | 129297414 | 129317471 | 129317598 |
| ENSG00000185024 | BRF1 | TFIIB | NO | - | -0.126536098 | 0.001684363 | chr14 | - | 105707601 | 105707751 | 105695156 | 105695250 | 105718843 | 105718916 |
| ENSG00000198176 | TFDP1 | DP | YES | - | 0.139016818 | 0.0018896 | chr13 | + | 114291934 | 114292211 | 114290848 | 114291015 | 114294434 | 114294625 |
| ENSG00000102081 | FMR1 | - | NO | - | 0.162169694 | 0.002627816 | chrX | + | 147019617 | 147019680 | 147018984 | 147019119 | 147022094 | 147022181 |
| ENSG00000171456 | ASXL1 | ASXH | NO | - | -0.145845325 | 0.003703502 | chr20 | + | 31017703 | 31017856 | 31017140 | 31017234 | 31019123 | 31019287 |
| ENSG00000186448 | ZNF197 | KRAB* | YES | - | -0.086552457 | 0.003889162 | chr3 | + | 44673596 | 44673688 | 44672553 | 44672713 | 44673964 | 44674091 |
| ENSG00000137309 | HMGA1 | - | NO | - | -0.055388456 | 0.004489747 | chr6 | + | 34204980 | 34205094 | 34204649 | 34204740 | 34208513 | 34208659 |
| ENSG00000100429 | HDAC10 | Hist_deacetyl | NO | - | -0.235606061 | 0.005589482 | chr22 | - | 50687756 | 50687945 | 50687531 | 50687597 | 50688063 | 50688132 |
| ENSG00000111206 | FOXM1 | - | NO | - | 0.072458342 | 0.010560464 | chr12 | - | 2970464 | 2970578 | 2966848 | 2968829 | 2973485 | 2973661 |
| ENSG00000067066 | SP100 | SAND | NO | - | 0.168979088 | 0.023504136 | chr2 | + | 231372707 | 231372746 | 231371017 | 231371160 | 231375839 | 231375881 |
| ENSG00000152784 | PRDM8 | - | YES | - | 0.203431373 | 0.039352881 | chr4 | + | 81115266 | 81115393 | 81112625 | 81112723 | 81117166 | 81117278 |
| ENSG00000138785 | INTS12 | - | NO | - | 0.113845432 | 0.085586846 | chr4 | - | 106624042 | 106624180 | 106621006 | 106621171 | 106624804 | 106624966 |
| ENSG00000134982 | APC | Arm | YES | - | -0.017115841 | 0.825678682 | chr5 | + | 112170647 | 112170862 | 112164531 | 112164669 | 112173249 | 112173419 |

* affected PFAM/NMD is related to frame-shifted downstream exon (downstream exon which is frame-shifted by inclusion/exclusion of target exon).

**Supplementary Table 3.1 Transcription factors with exon skipping events are often affected by frame shift and/or disruption of functional domains.**

Shown are transcription factors with significant exon skipping changes during osteogenic differentiation. Column 3 reports the protein family domain (Pfam) affected by inclusion/exclusion of target exon, due to the presence/absence of the domain, encoded by the target exon or the downstream exon (*) which is frame-shifted by incorporation/exclusion of target exon. Column 4 indicates whether the target exon

incorporation results in a frameshift. Column 5 indicates whether the target exon or the frameshifted downstream exon (*) is involved in nonsense mediated decay.

# 3.7 References


2 Fu, X. D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet **15**, 689-701, doi:10.1038/nrg3778 (2014).

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, doi:10.1016/j.ajhg.2017.11.002 (2018).

7 Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol **18**, 437-451, doi:10.1038/nrm.2017.27 (2017).

35 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**, E5593-5601, doi:10.1073/pnas.1419161111 (2014).

70 Duque, G. Bone and fat connection in aging bone. Curr Opin Rheumatol **20**, 429-434, doi:10.1097/BOR.0b013e3283025e9c (2008).

71 Meunier, P., Aaron, J., Edouard, C. & Vignon, G. Osteoporosis and the replacement of cell populations of the marrow by adipose tissue. A quantitative study of 84 iliac bone biopsies. Clin Orthop Relat Res **80**, 147-154, doi:10.1097/00003086-197110000-00021 (1971).

72 Blume, S. W. & Curtis, J. R. Medical costs of osteoporosis in the elderly Medicare population. Osteoporos Int **22**, 1835-1844, doi:10.1007/s00198-010-1419-7 (2011).

73 Kapinos, K. A., Fischer, S. H., Mulcahy, A., Hayden, O. & Barron, R. Medical Costs for Osteoporosis-Related Fractures in High-Risk Medicare Beneficiaries. J Am Geriatr Soc **66**, 2298-2304, doi:10.1111/jgs.15585 (2018).

74 Augello, A., Kurth, T. B. & De Bari, C. Mesenchymal stem cells: a perspective from in vitro cultures to in vivo migration and niches. Eur Cell Mater **20**, 121-133, doi:10.22203/ecm.v020a11 (2010).

75 Pittenger, M. F. et al. Multilineage potential of adult human mesenchymal stem cells. Science **284**, 143-147, doi:10.1126/science.284.5411.143 (1999).

76 Otto, F. et al. Cbfa1, a candidate gene for cleidocranial dysplasia syndrome, is essential for osteoblast differentiation and bone development. Cell **89**, 765-771, doi:10.1016/s0092-8674(00)80259-7 (1997).

77 Komori, T. et al. Targeted disruption of Cbfa1 results in a complete lack of bone formation owing to maturational arrest of osteoblasts. Cell **89**, 755-764, doi:10.1016/s0092-8674(00)80258-5 (1997).

78 Barak, Y. et al. PPAR gamma is required for placental, cardiac, and adipose tissue development. Mol Cell **4**, 585-595, doi:10.1016/s1097-2765(00)80209-9 (1999).

79 Rosen, E. D. et al. PPAR gamma is required for the differentiation of adipose tissue in vivo and in vitro. Mol Cell **4**, 611-617, doi:10.1016/s1097-2765(00)80211-7 (1999).

80 Chan, C. K. F. et al. Identification of the Human Skeletal Stem Cell. Cell **175**, 43-56 e21, doi:10.1016/j.cell.2018.07.029 (2018).

81 Dominici, M. et al. Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. Cytotherapy **8**, 315-317, doi:10.1080/14653240600855905 (2006).

83 Soleimani, M. & Nadri, S. A protocol for isolation and culture of mesenchymal stem cells from mouse bone marrow. Nat Protoc **4**, 102-106, doi:10.1038/nprot.2008.221 (2009).

84 Wagey, R. & Short, B. Isolation, enumeration, and expansion of human mesenchymal stem cells in culture. Methods Mol Biol **946**, 315-334, doi:10.1007/978-1-62703-128-8_20 (2013).

85 Ciuffreda, M. C., Malpasso, G., Musaro, P., Turco, V. & Gnecchi, M. Protocols for in vitro Differentiation of Human Mesenchymal Stem Cells into Osteogenic, Chondrogenic and Adipogenic Lineages. Methods Mol Biol **1416**, 149-158, doi:10.1007/978-1-4939-3584-0_8 (2016).

86 Jaiswal, N., Haynesworth, S. E., Caplan, A. I. & Bruder, S. P. Osteogenic differentiation of purified, culture-expanded human mesenchymal stem cells in vitro. J Cell Biochem **64**, 295-312 (1997).

87 Chen, Q. et al. Fate decision of mesenchymal stem cells: adipocytes or osteoblasts? Cell Death Differ **23**, 1128-1139, doi:10.1038/cdd.2015.168 (2016).

88 Han, L. et al. The shift in the balance between osteoblastogenesis and adipogenesis of mesenchymal stem cells mediated by glucocorticoid receptor. Stem Cell Res Ther **10**, 377, doi:10.1186/s13287-019-1498-0 (2019).

89 Rauch, A. et al. Osteogenesis depends on commissioning of a network of stem cell transcription factors that act as repressors of adipogenesis. Nat Genet **51**, 716-727, doi:10.1038/s41588-019-0359-1 (2019).

90 Wolock, S. L. et al. Mapping Distinct Bone Marrow Niche Populations and Their Differentiation Paths. Cell Rep **28**, 302-311 e305, doi:10.1016/j.celrep.2019.06.031 (2019).

91 Zhong, L. et al. Single cell transcriptomics identifies a unique adipose lineage cell population that regulates bone marrow environment. Elife **9**, doi:10.7554/eLife.54695 (2020).

92 Boonrungsiman, S. et al. The role of intracellular calcium phosphate in osteoblast-mediated bone apatite formation. Proc Natl Acad Sci U S A **109**, 14170-14175, doi:10.1073/pnas.1208916109 (2012).

93 Feng, Y. & Walsh, C. A. The many faces of filamin: a versatile molecular scaffold for cell motility and signalling. Nat Cell Biol **6**, 1034-1038, doi:10.1038/ncb1104-1034 (2004).

94 Norris, R. A. & Kern, M. J. The identification of Prx1 transcription regulatory domains provides a mechanism for unequal compensation by the Prx1 and Prx2 loci. J Biol Chem **276**, 26829-26837, doi:10.1074/jbc.M100239200 (2001).

95 Reichert, M. et al. The Prrx1 homeodomain transcription factor plays a central role in pancreatic regeneration and carcinogenesis. Genes Dev **27**, 288-300, doi:10.1101/gad.204453.112 (2013).

96 Takano, S. et al. Prrx1 isoform switching regulates pancreatic cancer invasion and metastatic colonization. Genes Dev **30**, 233-247, doi:10.1101/gad.263327.115 (2016).

97 Wang, J. et al. Paired Related Homeobox Protein 1 Regulates Quiescence in Human Oligodendrocyte Progenitors. Cell Rep **25**, 3435-3450 e3436, doi:10.1016/j.celrep.2018.11.068 (2018).

98 Lu, X. et al. Identification of the homeobox protein Prx1 (MHox, Prrx-1) as a regulator of osterix expression and mediator of tumor necrosis factor alpha action in osteoblast differentiation. J Bone Miner Res **26**, 209-219, doi:10.1002/jbmr.203 (2011).

99 Erhard, F. et al. scSLAM-seq reveals core features of transcription dynamics in single cells. Nature **571**, 419-423, doi:10.1038/s41586-019-1369-y (2019).

100 Dominguez, D. et al. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol Cell **70**, 854-867 e859, doi:10.1016/j.molcel.2018.05.001 (2018).

101 Ray, D. et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature **499**, 172-177, doi:10.1038/nature12311 (2013).

102 Gabut, M., Chaudhry, S. & Blencowe, B. J. SnapShot: The splicing regulatory machinery. Cell **133**, 192 e191, doi:10.1016/j.cell.2008.03.010 (2008).

103 Danilenko, M. et al. Binding site density enables paralog-specific activity of SLM2 and Sam68 proteins in Neurexin2 AS4 splicing control. Nucleic Acids Res **45**, 4120-4130, doi:10.1093/nar/gkw1277 (2017).

104 Farini, D. et al. A Dynamic Splicing Program Ensures Proper Synaptic Connections in the Developing Cerebellum. Cell Rep **31**, 107703, doi:10.1016/j.celrep.2020.107703 (2020).

105 Feracci, M. et al. Structural basis of RNA recognition and dimerization by the STAR proteins T-STAR and Sam68. Nat Commun **7**, 10355, doi:10.1038/ncomms10355 (2016).

106 Traunmuller, L., Gomez, A. M., Nguyen, T. M. & Scheiffele, P. Control of neuronal synapse specification by a highly dedicated alternative splicing program. Science **352**, 982-986, doi:10.1126/science.aaf2397 (2016).

107 Zhang, D. et al. Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. Nat Commun **11**, 2089, doi:10.1038/s41467-020-15815-7 (2020).

108 Wilsker, D. et al. Nomenclature of the ARID family of DNA-binding proteins. Genomics **86**, 242-251, doi:10.1016/j.ygeno.2005.03.013 (2005).

109 Clark, M. D. et al. Structural insights into the assembly of the histone deacetylase-associated Sin3L/Rpd3L corepressor complex. Proc Natl Acad Sci U S A **112**, E3669-3678, doi:10.1073/pnas.1504021112 (2015).

110 Fleischer, T. C., Yun, U. J. & Ayer, D. E. Identification and characterization of three new components of the mSin3A corepressor complex. Mol Cell Biol **23**, 3456-3467, doi:10.1128/mcb.23.10.3456-3467.2003 (2003).

111 Lasko, P. Tudor domain. Curr Biol **20**, R666-667, doi:10.1016/j.cub.2010.05.056 (2010).

112 Shimojo, H. et al. Novel structural and functional mode of a knot essential for RNA binding activity of the Esa1 presumed chromodomain. J Mol Biol **378**, 987-1001, doi:10.1016/j.jmb.2008.03.021 (2008).

113 Monroe, D. G., Hawse, J. R., Subramaniam, M. & Spelsberg, T. C. Retinoblastoma binding protein-1 (RBP1) is a Runx2 coactivator and promotes osteoblastic differentiation. BMC Musculoskelet Disord **11**, 104, doi:10.1186/1471-2474-11-104 (2010).

114 Wu, R. C., Jiang, M., Beaudet, A. L. & Wu, M. Y. ARID4A and ARID4B regulate male fertility, a functional link to the AR and RB pathways. Proc Natl Acad Sci U S A **110**, 4616-4621, doi:10.1073/pnas.1218318110 (2013).

115 Bonnet, N., Conway, S. J. & Ferrari, S. L. Regulation of beta catenin signaling and parathyroid hormone anabolic effects in bone by the matricellular protein periostin. Proc Natl Acad Sci U S A **109**, 15048-15053, doi:10.1073/pnas.1203085109 (2012).

116 Duchamp de Lageneste, O. et al. Periosteum contains skeletal stem cells with high bone regenerative potential controlled by Periostin. Nat Commun **9**, 773, doi:10.1038/s41467-018-03124-z (2018).

117 Litvin, J. et al. Expression and function of periostin-isoforms in bone. J Cell Biochem **92**, 1044-1061, doi:10.1002/jcb.20115 (2004).

118 Aprile, M. et al. PPARgammaDelta5, a Naturally Occurring Dominant-Negative Splice Isoform, Impairs PPARgamma Function and Adipocyte Differentiation. Cell Rep **25**, 1577-1592 e1576, doi:10.1016/j.celrep.2018.10.035 (2018).

119 Huot, M. E. et al. The Sam68 STAR RNA-binding protein regulates mTOR alternative splicing during adipogenesis. Mol Cell **46**, 187-199, doi:10.1016/j.molcel.2012.02.007 (2012).

120 Li, H. et al. SRSF10 regulates alternative splicing and is required for adipocyte differentiation. Mol Cell Biol **34**, 2198-2207, doi:10.1128/MCB.01674-13 (2014).

121 Vernia, S. et al. An alternative splicing program promotes adipose tissue thermogenesis. Elife **5**, doi:10.7554/eLife.17672 (2016).

122 Makita, N. et al. Two of four alternatively spliced isoforms of RUNX2 control osteocalcin gene expression in human osteoblast cells. Gene **413**, 8-17, doi:10.1016/j.gene.2007.12.025 (2008).

123 Johnson, C. et al. Tracking COL1A1 RNA in osteogenesis imperfecta. splice-defective transcripts initiate transport from the gene but are retained within the SC35 domain. J Cell Biol **150**, 417-432, doi:10.1083/jcb.150.3.417 (2000).

124 Wang, Q., Forlino, A. & Marini, J. C. Alternative splicing in COL1A1 mRNA leads to a partial null allele and two In-frame forms with structural defects in non-lethal osteogenesis imperfecta. J Biol Chem **271**, 28617-28623, doi:10.1074/jbc.271.45.28617 (1996).

125 Xia, X. Y. et al. A novel RNA-splicing mutation in COL1A1 gene causing osteogenesis imperfecta type I in a Chinese family. Clin Chim Acta **398**, 148-151, doi:10.1016/j.cca.2008.07.030 (2008).

126 Yee, B. A., Pratt, G. A., Graveley, B. R., Van Nostrand, E. L. & Yeo, G. W. RBP-Maps enables robust generation of splicing regulatory maps. RNA **25**, 193-204, doi:10.1261/rna.069237.118 (2019).

127 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).

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

129 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res **4**, 1521, doi:10.12688/f1000research.7563.2 (2015).

130 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

131 Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

132 Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. Nat Rev Genet **15**, 829-845, doi:10.1038/nrg3813 (2014).

133 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

## Uncategorized References

82 Robey, P. G. & Riminucci, M. in Principles of Bone Biology (Fourth Edition) (eds John P. Bilezikian, T. John Martin, Thomas L. Clemens, & Clifford J. Rosen) 45-71 (Academic Press, 2020).

# 4 PRMT9 AFFECTS NEURON DEVELOPMENT BY REGULATING SPLICING THOUGH SF3B2 METHYLATION

## 4.1 Introduction

Protein arginine methylation is a widespread post-translational modification that plays a pivotal role in many biological processes. Mammalian genomes encode a family of 9 protein arginine methyltransferases (PRMTs), which can be categorized into three types based on their catalytical products. Type I, type II and type III PRMTs can deposit mono-methylarginine (MMA) mark; while type I and type II (including PRMT5 and PRMT9) enzymes can further establish asymmetric dimethylarginine (ADMA) and symmetric dimethylarginine (SDMA), respectively [134,135]. PRMTs are involved in many fundamental

cellular processes and exhibit physiological roles liking them to various diseases [134], such as cancer, metabolic diseases and neurodegenerative [29,136,137] disorders.

One of the representative processes regulated by PRMTs is pre-mRNA splicing, which is also a highly regulated mechanism in brain development [11,12,64,138]. For example, RNA binding proteins (RBPs), most of which are trans-acting splicing factors [2,100], are the most enriched cellular substrates of PRMTs [30]. Both type II enzymes, PRMT5 and PRMT9, contribute to the methylation of key components of the splicing machinery [28,139]. Moreover, they have been shown to methylate distinct substrates and do not display redundancy [28,139,140]. PRMT9 can catalyze the symmetric dimethylation at R508 (R508me2s) of spliceosome-associated protein 145 (SAP145, also known as SF3B2), which is a core component of U2 snRNP, linking PRMT9 to U2 snRNP maturation and alternative splicing regulation [28]. The R508me2s of SF3B2 is required for the interaction between SF3B2 and survival motor neuron (SMN) [28]. The SMN protein is regarded as the direct cause of neurodegenerative disorder spinal muscular atrophy [141], thus indicating a physiological role of PRMT9 in maintaining normal brain function.

Loss-of-function mutations in PRMTs are rare, but could lead to strong pathophysiological contributions [135,142]. A rare missense mutation (dbSNP rs769164317, variant frequency = 0.000008) in PRMT9 causing G to R amino acid substitution at site 189 (G189R) is identified as a causative variant for autosomal recessive intellectual disability (ARID) [143]. In addition, the aberrant overexpression of PRMT9 has also been demonstrated to promote hepatocellular carcinoma invasion and metastasis [144]. These observations further highlight the importance of PRMT9 in the pathogenesis of brain-related diseases.

In this study, we sought to dissect the catalytical consequences of PRMT9 G189R mutant and decode the regulatory mechanism of PRMT9 on brain development to shed light on the pathogenesis of ARID and other neurological diseases. We found that PRMT9 G189R mutant has eliminated methyltransferase activity and diminished stability. In addition, we directly investigated the consequences of Prmt9 depletion in the hippocampus tissue and found that mice with tissue-specific Prmt9 cKO in excitatory neurons exhibit impaired learning, memory and synapse maturation. From a transcriptome-wide analysis using RNA-seq data generated from hippocampus tissue of wild-type (WT) and Prmt9 whole body KO mice, we revealed a PRMT9-SF3B2-splicing-synapse regulatory cascade that links PRMT9 to brain development. We also inspected the sequence features of alternative splicing events altered by Prmt9 KO, and proposed a working model that PRMT9-mediated SF3B2 R508me2s regulates splicing through 3' splice site competition by altering SF3B2/pre-mRNA interaction.

# 4.2 Results

## 4.2.1 PRMT9 G189R mutant is catalytically inactive and unstable

To deepen our insights into the ARID-causative PRMT9 G189R mutant [143], we performed amino-acid sequence alignment of PRMTs, which revealed that the G189 is located at a highly conserved motif I region (**Supplementary figure 4.6**). This underlines the importance to evaluate the enzymatic activity of PRMT9 G189R mutant protein, especially the methyltransferase activity for the formation of SDMA on SF3B2. We performed in vitro methylation assays by incubating SF3B2 fragments with either wild-type, G189R mutant PRMT9, or a previously reported catalytically inactive mutant PRMT9 with quadruple

mutations (L182A, D183A, I184A, and G185A, denoted as 4A) [28]. It shows that the methyltransferase activity is completely abolished in G189R mutant and 4A mutant PRMT9 in vitro (**Figure 4.1A**). In vivo methylation assays were conducted by rescuing the expression of either wild-type, G189R, or 4A mutant PRMT9 to PRMT9 KO Hela cells followed by western blotting by SF3B2 R508me2s methyl-specific antibody. Wild-type, but not the G189R or 4A mutant, PRMT9 can restore the methylation of SF3B2 in vivo (**Figure 4.1B**).

To determine the effects of G189R mutant on protein stability of PRMT9 in vivo, the half-life of wild type or the G189R mutant PRMT9 were calculated in cycloheximide (CHX)-treated Hela cells. As shown in **Figure 4.1C**, the protein level of G189R mutant PRMT9 decreased to 50% in ~1-2 hours and is totally degraded within 4 hours, while the protein level of wild-type PRMT9 remains largely unchanged. These data demonstrated that the G189R mutant significantly shortens the half-life of PRMT9 protein.

### 4.2.2 Prmt9 cKO in excitatory neurons resulted in impaired learning, memory and synapse maturation in mice

PRMT9 G189R mutant was identified as disease-causing variant in ARID [143]. It has also been shown previously that aberrant overexpression of PRMT9 promote hepatocellular carcinoma invasion and metastasis [144]. However, studies directly linking PRMT9 and brain development were scarce and the function of PRMT9 on ARID remains enigmatic. To address this question, we generated Prmt9 tissue-specific conditional KO (cKO) mice using Emx1-Cre recombinase, which depleted the expression of Prmt9 in excitatory neurons of the developing and adult cerebral cortex and hippocampus [145,146].

We then evaluated the behaviour and learning ability of Prmt9 cKO mice. The Morris water maze is a gold standard procedure for assessment of spatial learning and memory [147] (**Figure 4.2A**). Prmt9 cKO mice displayed a significantly delayed acquisition compared to wild-type mice (**Figure 4.2B**). Pavlovian fear conditioning is an associative learning task in which mice learn to pair a neutral conditional stimulus (tone) with an aversive unconditional stimulus (electrical foot shock) (**Figure 4.2C**) [148]. Prmt9 cKO mice froze significantly less than the wild-type mice (**Figure 4.2D**), demonstrating the importance of Prmt9 in associative learning.

To determine the cellular changes underlining impaired learning and memory, we cultured neurons from hippocampus tissue. Functional excitatory synapses were quantified by double staining of glutamate receptors (GluN1, NMDA receptor; GluA1, AMPA receptor) and pre-/post-synaptic markers (Synapsin I, pre-synaptic marker; PSD95, post-synaptic marker) (**Figure 4.2E**). Prmt9 cKO neurons exhibit decreased puncta density for GluN1, GluA1 and PSD95 and reduced colocalization of NMDA/AMPA receptors and pre-/post-synaptic markers (**Figure 4.2E**). These results demonstrated a reduction of functional excitatory synapses in Prmt9 cKO mice.

### 4.2.3 Alternative splicing acts independently to define brain-specific transcriptome in Prmt9 KO mice

The effect of PRMT9 on SF3B2 methylation, and the methylation-dependent SF3B2-SMN interaction reported by Yang et al [28] strongly suggest that the impaired brain function might result from dysregulation of pre-mRNA splicing by unmethylated SF3B2 in Prmt9 KO mice.

To test this hypothesis, we extracted RNA samples from hippocampus tissue of two-week-old wild-type control mice and whole body Prmt9 KO mice, and generated high-depth RNA-seq data (**Supplementary figure 4.7A**). The expression of the floxed exon 5 is completely vanished and expression of other regions of Prmt9 is also dramatically diminished in Prmt9 KO samples (**Supplementary figure 4.7B**). The RNA-seq data were subject to comprehensive gene expression and alternative splicing analysis. Five types of alternative splicing events were analysed, including exon skipping (SE), alternative 5' or 3' splice sites (A5SS, A3SS), mutual exclusive exons (MXE) and intron retention (RI).

Extensive alternative splicing alterations were observed from the Prmt9 KO samples (**Figure 3A**, **Supplementary figure 4.7C**), with enrichment of A3SS events (**Supplementary figure 4.7C**), indicating 3'-splice-site-related regulation of splicing by Prmt9. In contrast, no differentially expressed genes (except for Prmt9 gene itself) were identified in wild-type versus Prmt9 KO hippocampus (**Figure 4.3B**). Similar phenomenon has been observed in a previous study where no significant steady-state gene expression changes were detected in wild-type versus neuronal splicing factor Nova2 knockout neocortex [149]. These results suggest that alternative splicing acts independently to define brain-specific transcriptome in hippocampus tissue of Prmt9 KO mice.

In addition to the unchanged gene expression profile in both Prmt9 KO and Nova2 KO, we found that lots of the validated Nova-regulated synaptic genes [149] are differentially spliced in Prmt9 KO samples (**Figure 4.3C**). For example, changes of exon 5 and exon 21 skipping is observed in Glutamate Ionotropic Receptor NMDA Type Subunit 1 (Grin1, also known as GluN1) gene (**Figure 4.3C**), a core component of NMDA receptor complex at the glutamatergic synapse [150]. The alternative splicing of exon 5 and exon 21 in Grin1 has

already been demonstrated to regulate maturation of excitatory synapse [151] and control long-lasting synaptic potentiation, learning and memory in mice [152]. A3SS events is detected in Rap1 GTPase-activating protein 1 (Rap1gap) gene whose expression is enriched in striatal medium spiny neurons [153]. The proximal 3' splice site is used more frequently on Prmt9 KO samples compared to wild-type, resulting in increased ratio of the longer transcript (**Figure 4.3C**). Other Nova-regulated synaptic genes with alternative splicing changes includes Potassium Inwardly Rectifying Channel Subfamily J Member 6 (Kcnj6, also known as Girk2) [154], Erythrocyte Membrane Protein Band 4.1 Like 3 (EPb41l3, also known as 4.1B) [155] and Calsyntenin 1 (Clstn1) [156].

To confirm the differential splicing detected from RNA-seq, we exploit RT-PCR to validate the isoform switch of selected genes with changes in exon skipping, alternative 3' splice sites or intron retention. Changes in the abundance of RT-PCR products corresponding to each splicing isoform is consistent with detected splicing changes from RNA-seq (**Figure 4.3C**, **Supplementary figure 4.7E**).

### 4.2.4 Splicing alterations are highly associated with excitatory synapse-related pathways

To further examine the relationship between dysregulation of splicing and impaired brain function in Prmt9 KO mice, we conducted Gene Ontology (GO) and pathway enrichment analysis of genes with splicing changes in the form of SE or A3SS. Most of the top enriched pathways are related to brain function (**Figure 4.4A**). For example, pathways involved in the activation or unblocking of NMDA receptor from BioPlanet database [157] represent an active area of brain research [158,159]. Considering that we have already matched splicing

alternations in several Nova-regulated synaptic genes, it is expected that genes with splicing changes are also significantly enriched in the splicing factor Nova regulated synaptic proteins pathway [149] from WikiPathway database [160]. We also performed enrichment analysis on GO terms. Similarly, genes with SE or A3SS changes are also enriched in neurotransmission-related biological processes, cellular component and molecular functions (**Supplementary figure 4.7D**).

Several other mutations identified in the same study as PRMT9 G189R variant involves genes with neuron- or brain-specific functions, for example, genes in the intellectual-disability-associated Ras/Rho/PSD95 (Postsynaptic density protein-95) network [143]. Interestingly, many of the genes with causative variants for ARID detected in that study, or their paralogs, are responsive to Prmt9 KO in mice (**Figure 4.4B**). PSD95, also known as Discs Large MAGUK Scaffold Protein 4 (DLG4), is an abundant scaffold protein of excitatory synapses, which organize synaptic signal transduction in Ras signalling and Rho signalling pathways [161]. Variants in DLG4 leads to DLG4-related synaptopathy, a new rare brain disorder [162]. Post-transcriptional repression of PSD95 cellular expression though splicing has also been proved critical during early neural development [163,164]. We revealed that most genes in the Ras/Rho/PSD95 network, or their paralogous genes, displays splicing changes upon Prmt9 KO in mouse hippocampus tissue (**Figure 4.4C**). These findings indicate that Prmt9 KO-induced splicing changes might contribute to the pathogenesis of ARID and related disorders.

### 4.2.5 PRMT9-mediated SF3B2 R508me2s regulates splicing through 3' splice site competition by altering SF3B2/pre-mRNA interaction

The cis-acting elements strictly required by the spliceosome consist of 5' and 3' splice site, branch site and polypyrimidine track upstream of the 3' splice site [2,6]. Also, previous studies indicate that sequence-dependent binding of SF3B complex to the anchoring site 6-25 nt upstream of the branch point site is essential for anchoring U2 snRNP to pre-mRNA [165]. SF3B2 is a core component of U2 snRNP complex that recognizes and binds to sequences near 3' splice site [166]. Thus, we hypothesized that splicing changes induced by PRMT9 KO might result from altered 3' splice site usage mediated by unmethylated SF3B2 R508. Indeed, the transcriptome-wide splicing analysis discovered that differential splicing events are enriched in the A3SS category, implying a regulatory mechanism related to 3' splice site usage (**Supplementary figure 4.7C**).

To further inspect the selective features of splicing changes induced by Prmt9 KO, we compared sequence features near the differentially spliced cassette exons (SE events) (**Figure 4.5A**) against a transcriptome-wide background cassette exon set. We first examined the splice strength for the two 5' splice sites and two 3' splice sites involved in the definition of SE event. No significant difference was observed for 5' splice sites in either upstream or downstream intron, whereas cassette exons more included and excluded upon Prmt9 KO are associated with weaker 3' splice sites in upstream and downstream introns, respectively (**Figure 4.5B**). We also extracted sequences of introns upstream and downstream of the cassette exons for prediction of branch point sites using the BPP software [167]. The branch point A predicted in introns upstream of more included cassette exons are more distal to their corresponding 3' splice sites (**Figure 4.5C**), and have lower

predicted branch point sequence scores (**Figure 4.5D**). Also, as shown in the nucleotide frequency logos near the predicted branch point adenosine, more degenerate sequences were observed within the anchoring site in upstream and downstream introns for cassette exons more included and excluded upon Prmt9 KO, respectively (**Figure 4.5E**). Interestingly, in the human spliceosomal $B^{act}$ complex (PDB: 6FF4 [168]), SF3B2 R508 is spatially closest to the uracil (U 381) nucleotide of pre-mRNA, which is 13 nt upstream of branch point adenosine (A 194) (**Figure 4.5F**). The spatial location of R508 is consistent with the location of SF3B anchoring sites, which further indicates that SF3B2 R508 methylation might impact Sf3B2/pre-mRNA interaction. The length of cassette exons and their flanking introns are less distinguishable between included and excluded SE events (**Supplementary figure 4.8A, 4.8B**).

Overall, we revealed from the sequence feature comparisons that more included cassette exons were always associated with weaker features near 3' splice sites in upstream intron (3' splice site strength, branch point sequences, more degenerate anchoring sites); while more excluded cassette exons were always associated with weaker features near 3' splice sites in downstream intron (3' splice site strength, more degenerate anchoring sites). This suggested that the regulation of splicing by Prmt9 is related to 3' splice site competition [2], which could be affected by the methylation of SF3B2 R508me2s. Moreover, the spatial location of SF3B2 R508 and the sequence changes in SF3B anchoring site on pre-mRNA indicated that the methylation of SF3B2 R508me2s might affect its interaction with pre-mRNA. Inspired by these observations, we proposed a working model that PRMT9-mediated SF3B2 R508me2s regulates splicing through 3' splice site competition by altering SF3B2/pre-mRNA interaction (**Figure 4.6**).

To test this hypothesis, we examined the SF3B2/pre-mRNA interaction near 3' splice sites in upstream and downstream of cassette exons in Stxbp5l and Grin1, both of which showed more exclusion in Prmt9 KO mice (**Figure 4.3C**). Indeed, the CLIP-qPCR demonstrated that interactions of SF3B2 with sequences near 3' splice sites in downstream introns were relatively more enhanced in Prmt9 KO mice compared to its interactions with sequences near 3' splice sites in upstream introns (**Figure 4.5G**). Although these results validated the 3' splice site competition model, additional experiments need to be further conducted to investigate whether the differential interaction of SF3B2 with pre-mRNA is related to sequence feature changes, especially changes of sequences in the SF3B complex anchoring site.

# 4.3 Discussion

PRMT9 G189R mutant was identified as disease-causing variant in ARID [143], which is located in a conserved motif I region. However, little is known on its pathogenic contributions in ARID. In this study, both in vitro and in vivo methylation assay proved that the PRMT9 G189R mutant is catalytically inactive and cannot methylate SF3B2, the well-known substrate of PRMT9 methyltransferase [28,140].

The methylation of SF3B2 at R508 by PRMT9 is required for SF3B2-SMN interaction, indicating a brain-related function of PRMT9 [28]. Abnormal expression of PRMT9 has also been proved to promote hepatocellular carcinoma invasion and metastasis [144]. However, in vivo functional outcomes and the cellular changes associated with PRMT9 has never been examined. We generated a tissue-specific Prmt9 KO mouse model, and demonstrated impairment of learning, memory and synapse maturation when Prmt9 is

knocked out from excitatory neurons. These results further shed light on the pathogenesis of ARID and other related disorders, and also highlighted the clinical relevance of targeting arginine methylation in brain tumors and neurodegenerative diseases [30,135,169].

Considering the fact that SF3B2 is a core component of U2 snRNP complex [166], and splicing is misregulated in PRMT9 KO Hela cells [28,140], it is reasonable to further investigate the splicing alterations and their relationship with brain-related genes. We generated a whole-body Prmt9 KO mouse model and performed RNA-seq followed by transcriptome-wide gene expression and splicing analysis, and revealed a PRMT9-SF3B2-splicing-synapse regulatory cascade that links PRMT9 to brain development. Surprisingly, no steady-state change in gene expression was observed from our data. In contrast, many genes involved in synapse development and brain function were subject to splicing changes. Moreover, genes with alternative splicing changes were significantly enriched in neuron-/synapse-related pathways, Gene Ontology terms, and signalling networks. It also provided a valuable resource for splicing variations that can be further experimentally investigated.

Furthermore, inspired by the results from sequence feature comparisons and the spatial structure of SF3B2 in spliceosomal $B^{act}$complex, we proposed a working model that PRMT9-mediated SF3B2 R508me2s regulates splicing through 3' splice site competition by altering SF3B2/pre-mRNA interaction (**Figure 4.6**). Splice site selection not only depends on the intrinsic properties of single splice site, but is often involved with the competition of two or more splice sites for the assembly of splicing complexes [2,170]. For example, it has been shown that 5' splice sites competition could facilitate proximal splicing [171], and is involved in the commitment of splice site pairing [172]. Splice site competition can also explain the non-monotonicity of the mutation-splicing map in the mathematical prediction

of mutational effects on splicing [170]. Our model captures this fundamental component of splice site choice. Our CLIP-qPCR assay identified differential SF3B2/pre-mRNA interaction between 3' splice sites in the upstream and downstream intron, supporting the 3' splice site competition model. However, further experiments are still needed to figure out whether the differential SF3B2/pre-mRNA interaction was associated with sequence features, especially sequence changes in SF3B complex anchoring site.

# 4.4 Methods

## 4.4.1 Knockout mice

#### 4.4.1.1 cKO mice

Prmt9 were conditionally knocked out in mice using Emx1-Cre recombinase, whose expression is restricted to excitatory neurons in the developing and adult cerebral cortex and hippocampus [145,146]. Prmt9 cKO mice are subject to behaviour test. Hippocampal neurons from Prmt9 cKO mice were also isolated and cultured for immunocytochemistry.

#### 4.4.1.2 Whole body KO mice

Prmt9 whole body KO mice were generated by CMV-Cre recombinase. RNA samples were extracted from hippocampus tissue of two-week-old wild-type or Prmt9 whole body KO mice followed by RNA sequencing.

## 4.4.2 Amino-acid sequence alignment using ClustalW

The parameters for the alignment using ClustalW [173] were set as follows: Gap Penalty: 10, Gap Length Penalty: 0.2, Delay Divergent Seqs (%) 30, Protein Weight Matrix: Gonnet

Series for multiple alignment parameters, and for pairwise alignment, Gap Penalty: 10, Gap Length 0.1, Protein Weight Matrix: Gonnet 250.

### 4.4.3 RNA-seq

RNA samples were extracted from hippocampus tissue of two-week-old wild-type or Prmt9 whole body KO mice (n = 3 for each group). Quality of RNA samples was ensured by calculation of RNA integrity number as well as degradation measurement using 2200 TapeStation system. Poly(A)$^+$ cDNA libraries were subsequently generated using TruSeq stranded mRNA Library Prep Kit. Libraries were sequenced on NovaSeq 6000 System using S4 flow cell with a PE 2x100 kit at the Translational Genomics Research Institute (TGen).

### 4.4.4 Gene expression and alternative splicing analysis from RNA-seq data

The quality of raw RNA-seq datasets were inspected using FastQC. Reads were aligned to the mouse genome (mm10/GRCm38) by STAR (v2.7.1a) [69] using two-pass mode with Ensembl release 97 annotations.

Gene expression values were quantified in TPM (Transcripts Per Million) using kallisto (v0.43.1) [128] and subsequently summarized to gene expression matrix using tximport (v1.6.0, R package) [129]. Differential expression analysis was performed with the count-based tool DeSeq2 (v1.18.1, R package) [130]. Genes with fold change > 1.5 and FDR < 0.01 were identified as differentially expressed genes between wild-type and Prmt9 KO mice. UCSC genome browser track was utilized to visualize and confirm the change of Prmt9 expression in wild-type and Prmt9 KO mice.

Alternative splicing events detection, quantification and differential splicing analysis were conducted using rMATS-turbo (v4.1.0). Five types of alternative splicing events were detected, including exon skipping (SE), alternative 5' or 3' splice sites (A5SS, A3SS), mutual exclusive exons (MXE) and intron retention (RI). Novel splice site detection feature of rMATS-turbo was turned on to identify alterations in both annotated and cryptic splicing events. Exon inclusion levels were calculated as PSI (Percent Spliced In) value between 0 and 1, which is the ratio of reads supporting the inclusion isoform to total reads. To enhance the robustness and reliability of the analysis, events with low read support (75 percentile of read count < 10 in either group) or constitutively spliced (average PSI value < 0.05 or > 0.95 in both groups) were excluded from downstream analysis. Differentially spliced events were further filtered by the cut-offs of FDR ($\leq 0.01$) and PSI value difference ($\geq 0.05$). Virtualization of selected differential splicing events was achieved by rmats2sashimiplot software.

### 4.4.5 Validation of differential splicing events using RT-PCR and agarose gel electrophoresis

Selected differential splicing events were validated by semi-quantitative reverse transcription–polymerase chain reaction (RT–PCR) assay followed by agarose gel electrophoresis. PCR primers were designed to amplify the region around target exon for each splicing event. PCR products were separated by gel-electrophoresis with different bands representing different isoforms. Expression abundance shift between different isoforms were visualized by the change of intensity for different bands.

### 4.4.6 Gene set enrichment analysis

Genes with differential splicing were tested for enrichment in both Gene Ontology (GO) terms and biological pathways (BioPlanet [157], Elsevier and WikiPathway [160]), which were retrieved from Enrichr [174] libraries (https://maayanlab.cloud/Enrichr/#libraries). To eliminate the bias resulted from gene expression on differential splicing analysis, a customized background excluding lowly expressed genes (DeSeq2 baseMean value ≥ 5) were used instead of leveraging all genes in the mouse genome as background gene list. Genes with splicing alterations in the top 2 categories (SE and A3SS) were selected as foreground gene list. The significance of enrichment was then evaluated by hypergeometric test, and adjusted p values were calculated from Benjamini-Hochberg procedure.

### 4.4.7 Ras/Rho/PSD95 network analysis

Protein nodes of the Ras/Rho/PSD95 network were curated from two resources [143,150]. Protein-protein interaction edges were collected from STRING (v11.5) [175] database, with active interaction sources extracted from experimental data (BIND, DIP, GRID, HPRD, IntAct, MINT, and PID) or databases (eg. Biocarta, BioCyc, GO, KEGG, and Reactome).

### 4.4.8 Sequence feature analysis for differential exon skipping events

Comparisons were performed between differentially spliced cassette exons (exon skipping events) and a transcriptome-wide background. The background cassette exon set, or native-exons, were defined as exons that are alternatively spliced under normal conditions (0.05 < mean PSI < 0.95 in wild-type samples).

#### 4.4.8.1 Splice site strength

Splice site sequences were extracted from 5' splice sites and 3' splice sites in both upstream introns and downstream introns. Splice site strengths were calculated using MaxEntScan [176]. The statistical significance of splice site strength differences between differentially spliced cassette exons and native background cassette exons were assessed using Wilcoxon's rank-sum test.

#### 4.4.8.2 Branch point prediction and comparison

Branch point prediction is performed using BPP [167] software in both upstream intron and downstream intron. It reports both the specific position of predicted branch point relative to the corresponding 3' splice site as well as the score of predicted branch point. The statistical significance of branch point score differences between differentially spliced cassette exons and native background cassette exons were assessed using Wilcoxon's rank-sum test.

#### 4.4.8.3 Anchoring sites ahead of branch point

Sequences ranging from -25 nt to 4 nt relative to the predicted branch points were extracted, which includes the anchoring sites for SF3B complex [165]. Observed frequencies of nucleotides at each specific position were visualized by WebLogo [177].

### 4.4.9 RBP motif enrichment analysis for differential exon skipping events

189 RBP binding motifs with position weight matrix information for 129 RBPs (including many well-characterized splicing factors) were curated from two different sources and

screened in this analysis. This includes 78 6-mer motifs for 78 RBPs from RNA Bind-n-Seq (RBNS) [100] and 111 7-mer motifs for 82 RBPs from RNAcompete [101].

Significant and background alternative splicing events were defined as described before (Gene expression and alternative splicing analysis from RNA-seq data section). To identify region-specific RBP regulatory patterns for exon skipping events, we evaluated three regions around the alternatively spliced exons: 1) 300 nt of intronic sequence upstream of the target exon; 2) the exon body sequences; and 3) 300 nt of intronic sequence downstream the target exon. Scores for each motif were calculated by sliding window scanning of the position weight matrix at each possible binding position. Region-specific motif occurrence was then determined by comparing the calculated motif scores with a threshold score (80% of the maximum PWM score). If there was any position with a calculated motif score ≥ the threshold score for a particular exon skipping event, then the motif occurrence was marked as "True" for this event in the corresponding region; otherwise it was marked "False".

To determine whether a motif occurred in a specific region more often in foreground event sets than in the background event set, a one-tailed Fisher's exact test was used to test the null hypothesis that the number of events with motif occurrence at a specific region was not different between the foreground and the background event set. P values were adjusted by Benjamini-Hochberg Procedure.

# 4.5 Figures

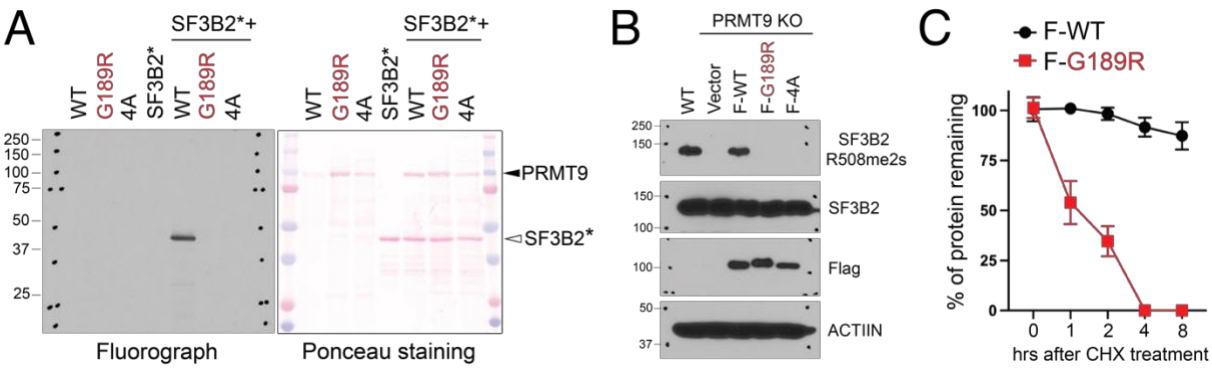



**Figure 4.1 PRMT9 G189R mutant is catalytically dead and unstable**

**(A)** In vitro methylation assay. Wild type, but not the G189R/4A mutant, PRMT9 can methylate SF3B2 fragment in vitro. The in vitro methylation was performed by incubating either wild-type or mutant PRMT9 with SF3B2* fragment (a.a. 400-500) as substrate. Loading of protein was checked by ponceau staining. 4A represents a previously reported enzymatic mutant PRMT9 with 4 amino acids (LDIG) within the conserved motif I mutated to AAAA.

**(B)** In vivo methylation assay. Wild type, but not the G189R/4A mutant, PRMT9 restores SF3B2 R508me2s in PRMT9 KO Hela cells. PRMT9 KO HeLa cells were transfected with Flag-PRMT9 (WT), Flag-G189R and Flag-4A mutant PRMT9. Total cell lysates were subject to western blotting detection with R508 methylation-specific antibody (αSF3B2 R508me2s), αSF3B2, αFlag and αActin antibodies.

**(C)** Stability measurement of wild-type PRMT9 and G189R mutant PRMT9 in Hela cells. Hela cells were treated with CHX to prevent new protein synthesis. Cell lysates were collected at indicated time after CHX treatment and analyzed by western blotting. The

percentage of remaining proteins were quantified by the intensity of bands from western blotting.

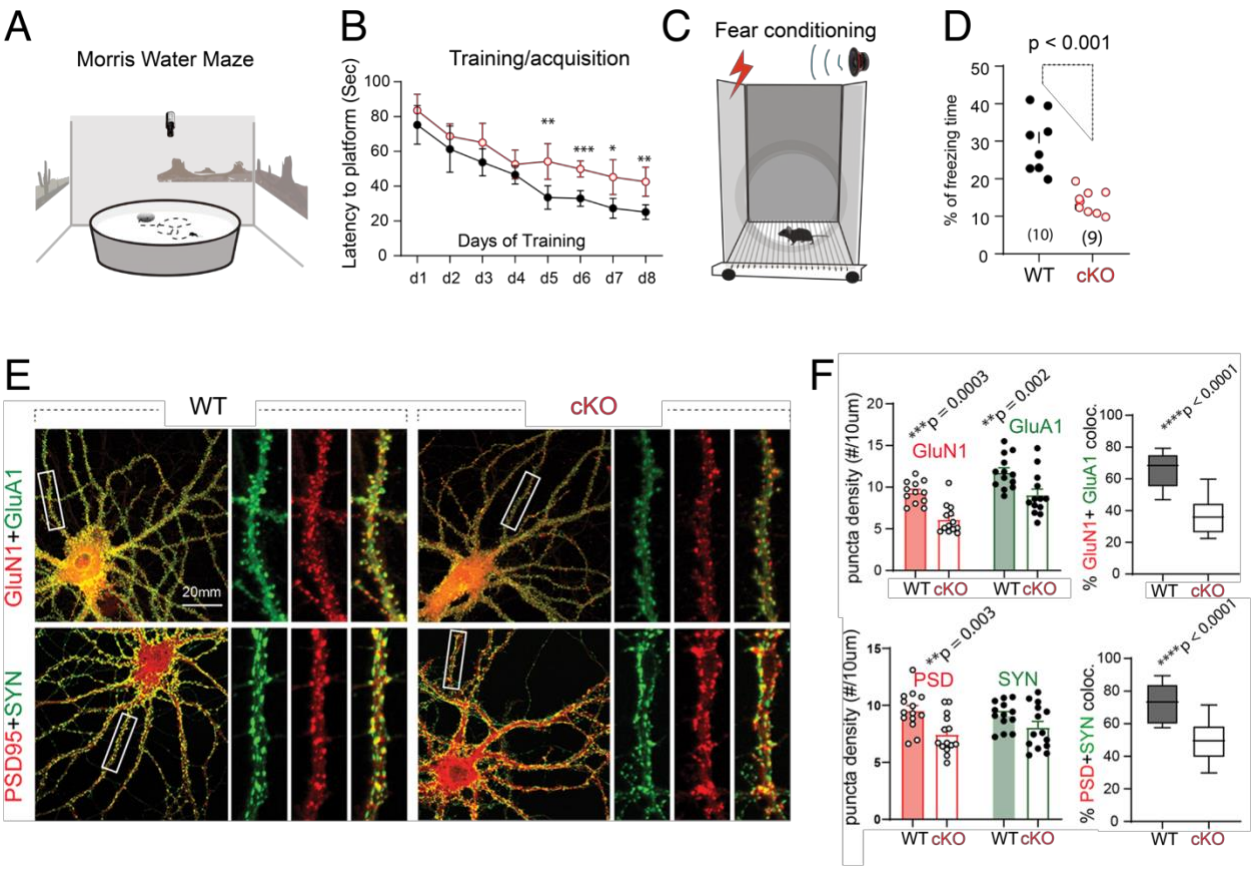



**Figure 4.2 Prmt9 cKO in excitatory resulted in impaired learning, memory and synapse maturation in mice**

**(A)** Schematic diagram of the Morris water maze test [147].

**(B)** Acquisition of the spatial memory shown by mean (±SEM) escape latency (time used to reach a hidden escape platform) over 8 consecutive days.

**(C)** Schematic diagram of the fear conditioning procedure [148]. Auditory tone is given as neutral conditional stimulus and electrical foot shock is given as aversive unconditional stimulus.

**(D)** Graphic representation of percent time freezing during fear conditioning procedure.

**(E)** Representative photomicrographs of functional-synapse markers in cultured hippocampal neurons of wild-type and Prmt9 cKO mice. On the top shows the double labeling of two different glutamate receptors, GluN1 (red, NMDA receptor) and GluA1 (green, AMPA receptor). On the bottom shows the double labeling of post-synaptic scaffold protein PSD95 (red) and pre-synaptic vesicle marker SYN (synapsin I, green).

**(F)** Quantification of protein staining detected in **(E)**. Puncta density (mean ± SEM) were quantified to indicate the protein abundance in mouse hippocampal neurons.

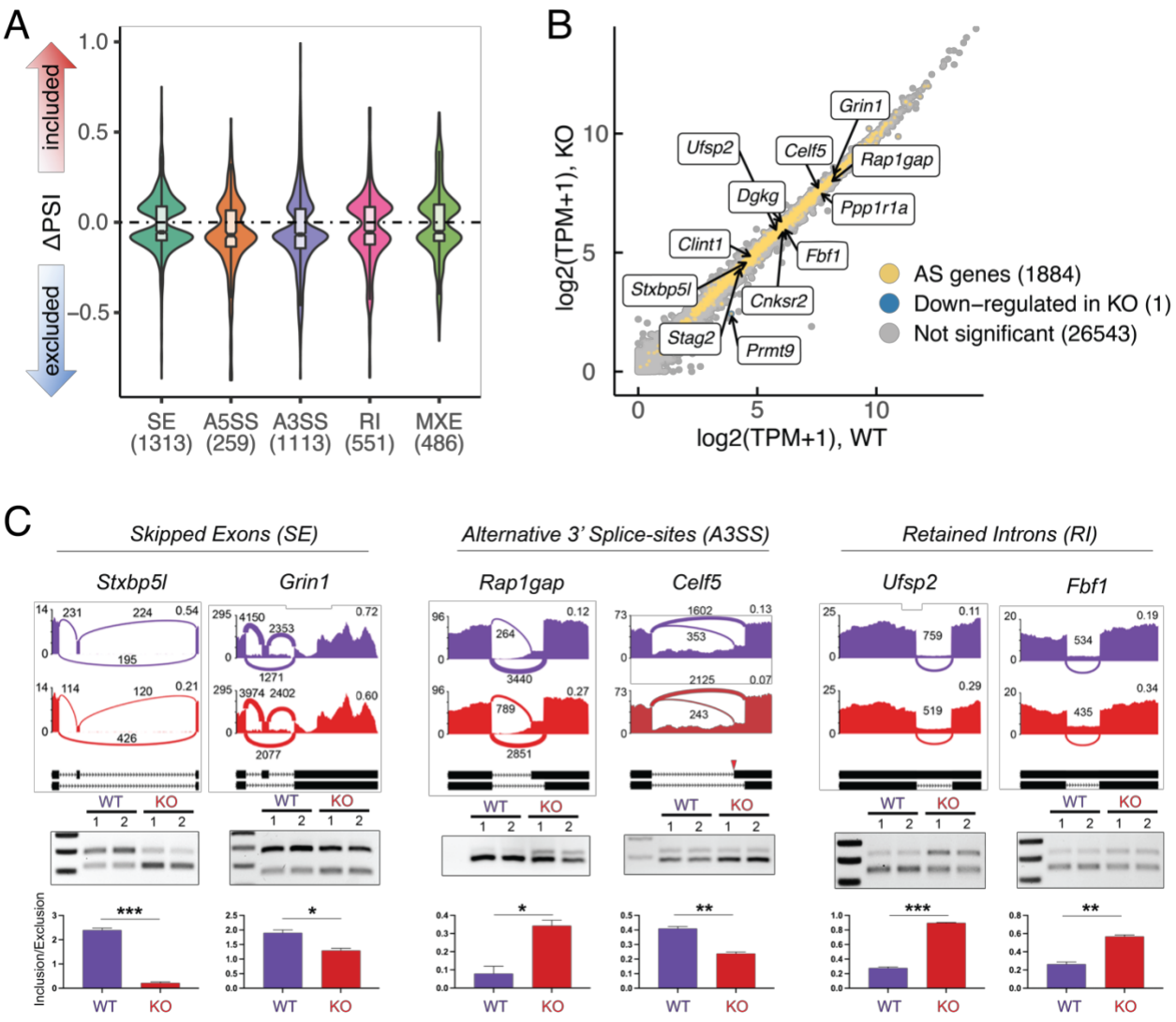



**Figure 4.3 Alternative splicing acts independently to define brain-specific transcriptome**

**(A)** Violin plot of alternative splicing events significantly changed upon Prmt9 KO in mice. Positive ΔPSI represents increased inclusion of events upon Prmt9 KO, whereas negative ΔPSI represents more exclusion of events upon Prmt9 KO. Number of significant events within each category were indicated within parentheses along x-axis labels. SE, exon

skipping / skipped exon; A5SS, alternative 5’ splice sites; A3SS, alternative 3’ splice sites; RI, intron retention / retained intron; MXE, mutually exclusive exon.

**(B)** Scatter plot of gene expression levels in wild-type and Prmt9 KO mice. Genes with significant alternative splicing changes are depicted in yellow.

**(C)** Sashimi plot visualization and experimental validation of selected differential splicing events. The black bars and dashed lines in the middle represent exons and introns, respectively. The splice site highlighted by red triangle is a cryptic splice site not annotated in the reference genome. Purple and red sashimi plots illustrate the splicing patterns in wild-type and Prmt9 KO mice, with solid peaks representing RNA-seq read coverages in RPKM (reads per kilobase per million mapped), arches representing splice junctions, and the numbers representing number of reads mapped to each splice junction. PSI values are also indicated on the right side of the sashimi plot. In the bottom, RT-PCR validations were shown. Higher band intensity of PCR products indicates higher expression of corresponding splicing isoform. The ratios of exon inclusion isoform to exon exclusion isoform are quantified and displayed in bar plots. The cassette exon in Grin1 is exon 21.

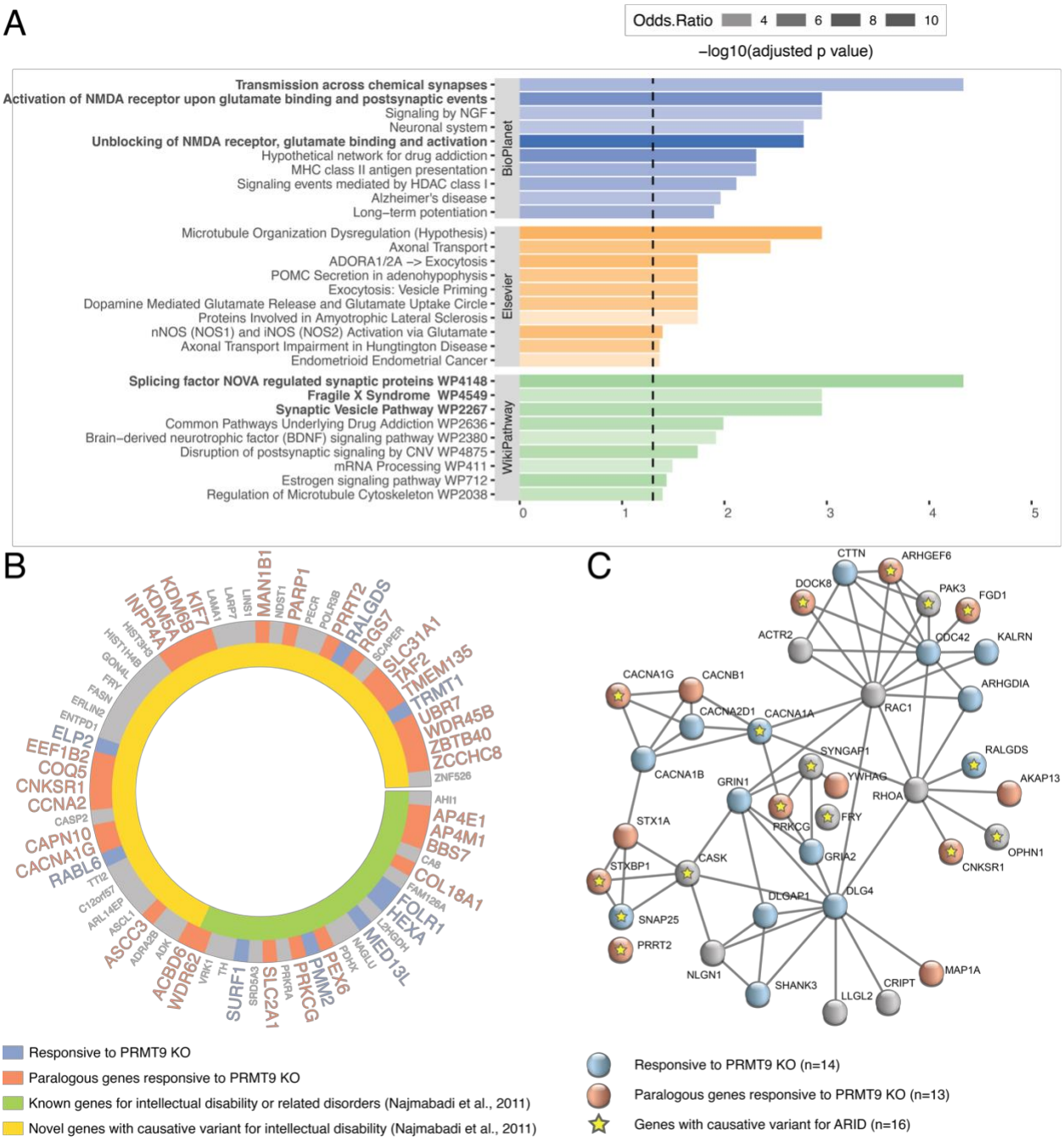



**Figure 4.4 Genes with alternative splicing changes are significantly enriched in brain-related pathways**

**(A)** Pathway enrichment analysis of alternatively spliced genes (SE or A3SS). Top 10 enriched pathways from 3 database resource origins (BioPlanet pathway, Elsevier pathway

collection, or the WikiPathway) are shown in the bar charts. The length of bars depicts the Benjamini-Hochberg adjusted p values calculated from hypergeometric test. Odds ratio of the enrichment were indicated by opacity of bars.

**(B)** Circular plot of genes with causative variants for autosomal recessive intellectual disabilities (ARID). Genes exhibit splicing alterations upon Prmt9 KO are annotated in blue, while genes which are not differentially spliced themselves but whose paralogous genes were differentially spliced are annotated in red.

**(C)** Ras/Rho/PSD95 network. Connecting edges collected from STRING database stands for protein-protein interactions. ARID-related genes were highlighted by the star shape in the center of the nodes. Genes exhibit splicing alterations upon Prmt9 KO are annotated in blue, while genes which are not differentially spliced themselves but whose paralogous genes were differentially spliced are annotated in red.

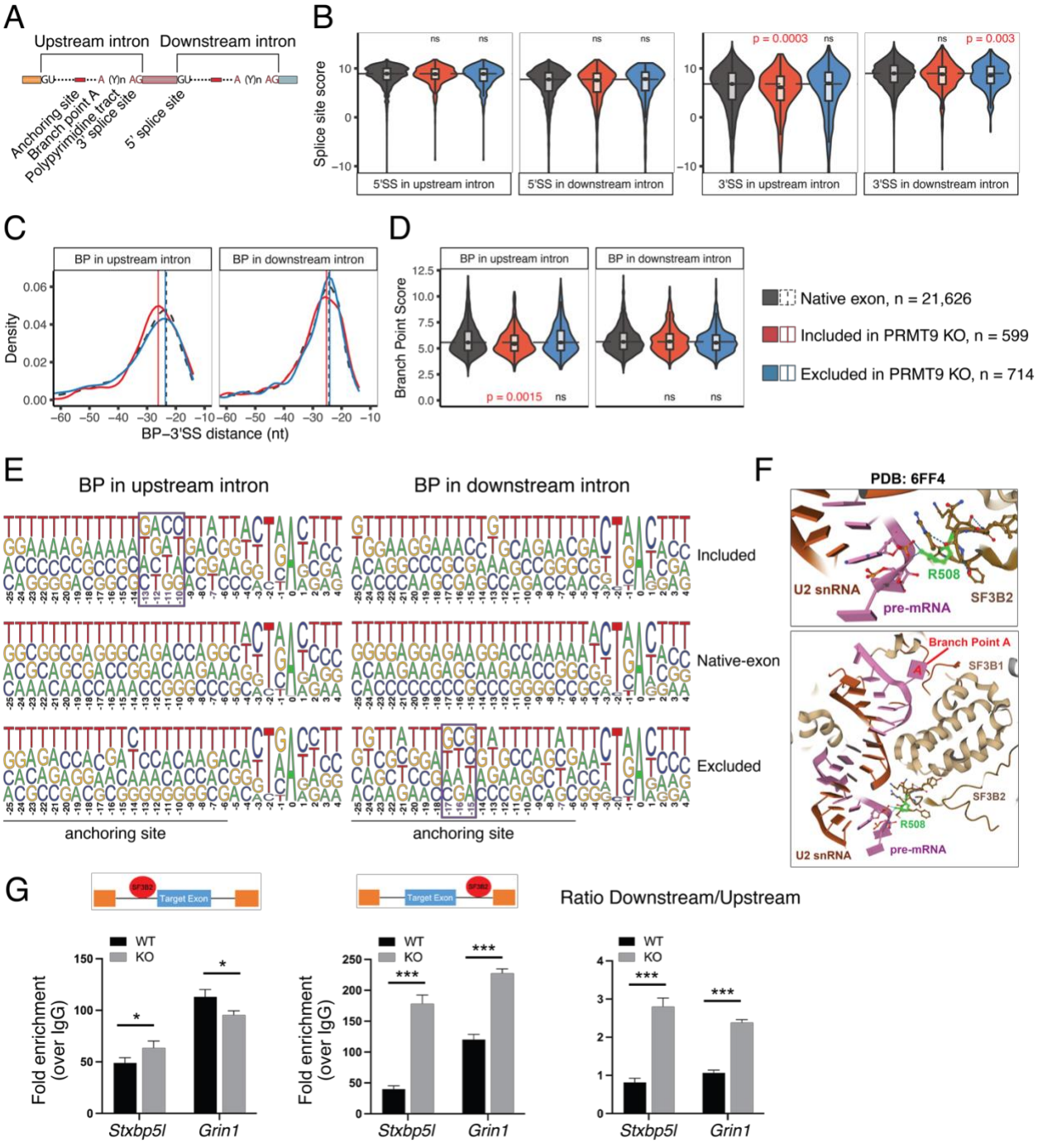



**Figure 4.5 Sequence features of exon skipping events affected by Prmt9 KO**

**(A)** Schematic illustration of cis-acting elements on pre-mRNA.

**(B**) Violin plot showing the maximum entropy score of 5' and 3' splice sites in upstream or downstream introns of differentially spliced cassette exons. The statistical significance against background native cassette exons were assessed using Wilcoxon's rank-sum test. ns, not significant.

**(C)** Density plot of the relative positions of predicted branch point sites to their corresponding 3' splice sites.

**(D)** Violin plot of BPP scores from predicted branch point sites. The statistical significance against background native cassette exons were assessed using Wilcoxon's rank-sum test. ns, not significant

**(E)** Sequence logo showing the frequency of nucleotide near the predicted branch point sites. The height of the symbols within the stack indicates the observed frequency of the corresponding nucleotide at that position. The 0 point demarks the position of the branch point adenosine (BPA). Sequences were shown from -25 nt to 4 nt relative to the BPA, which include the reported anchoring site of SF3B complex 6 to 25 nt ahead of BPA [165].

**(F)** Protein structure of human spliceosomal $B^{act}$ complex (PDB: 6FF4 [168]) with focus on the PRMT9-methylated R508 of SF3B2 protein. BPA on the pre-mRNA (A 394) is shown in red. R508 of SF3B2 is spatially close to the uracil (U 381) on pre-mRNA, which is 13 nt upstream of BPA.

**(G)** CLIP-qPCR assay of cells from hippocampus tissue of wild-type and Prmt9 KO mice. The SF3B2-bound RNA segments near 3' splice sites in the upstream and downstream introns of cassette exons were quantified using qPCR and normalized against the IgG-bound RNAs. Asterisks represent a significant difference in SF3B2/pre-mRNA interaction

at indicated regions between wild-type and KO samples. The cassette exons for Stxbp5l and Grin1 are from differential exon skipping events shown in **Figure 4.3C**.

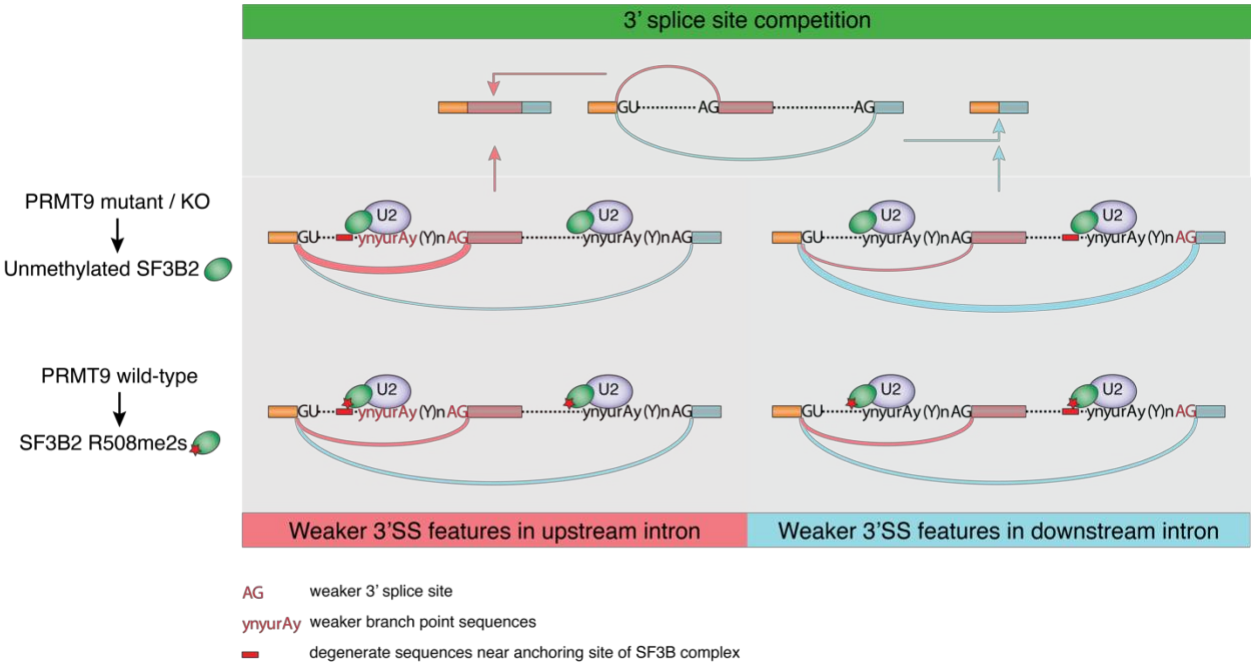
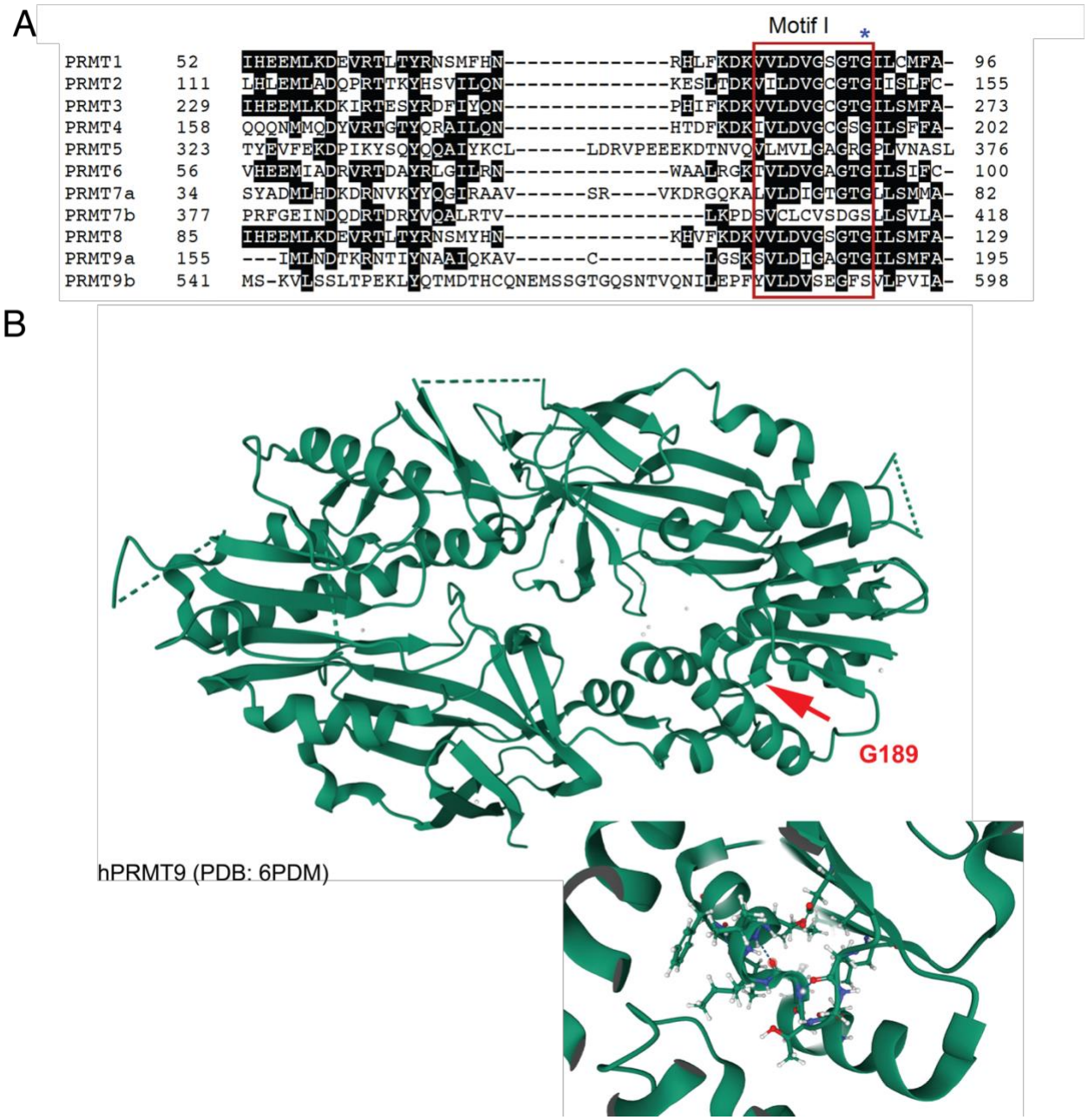



**Figure 4.6 Proposed working model: PRMT9-mediated SF3B2 R508me2s regulates splicing through 3' splice site competition by altering SF3B2/pre-mRNA interaction**

The top panel shows the competition between the two 3' splice sites involved in exon skipping events. The two 3' splice sites are engaged in competition with each other for the assembly of splicing complexes, which causes the alternative exon in the middle to be either included or excluded in the final transcript [2]. For exons whose upstream intron is associated with weaker 3' splice site features (e.g. weaker 3' splice site, weaker branch point sequences, more degenerate sequences in anchoring site of SF3B complex), compared to the downstream 3' splice site, the selection of upstream 3' splice site would be relatively enhanced by unmethylated SF3B2 in the PRMT9 loss-of-function mutant or KO, leading to more exon inclusion isoform. Vice versa, for exons associated with weaker 3' splice site features in downstream introns, the selection of downstream 3' splice site would be relatively enhanced while the selection of upstream 3' splice site would be relatively

diminished by unmethylated SF3B2, resulting in more exon skipping in PRMT9 loss-of-function mutant or KO.

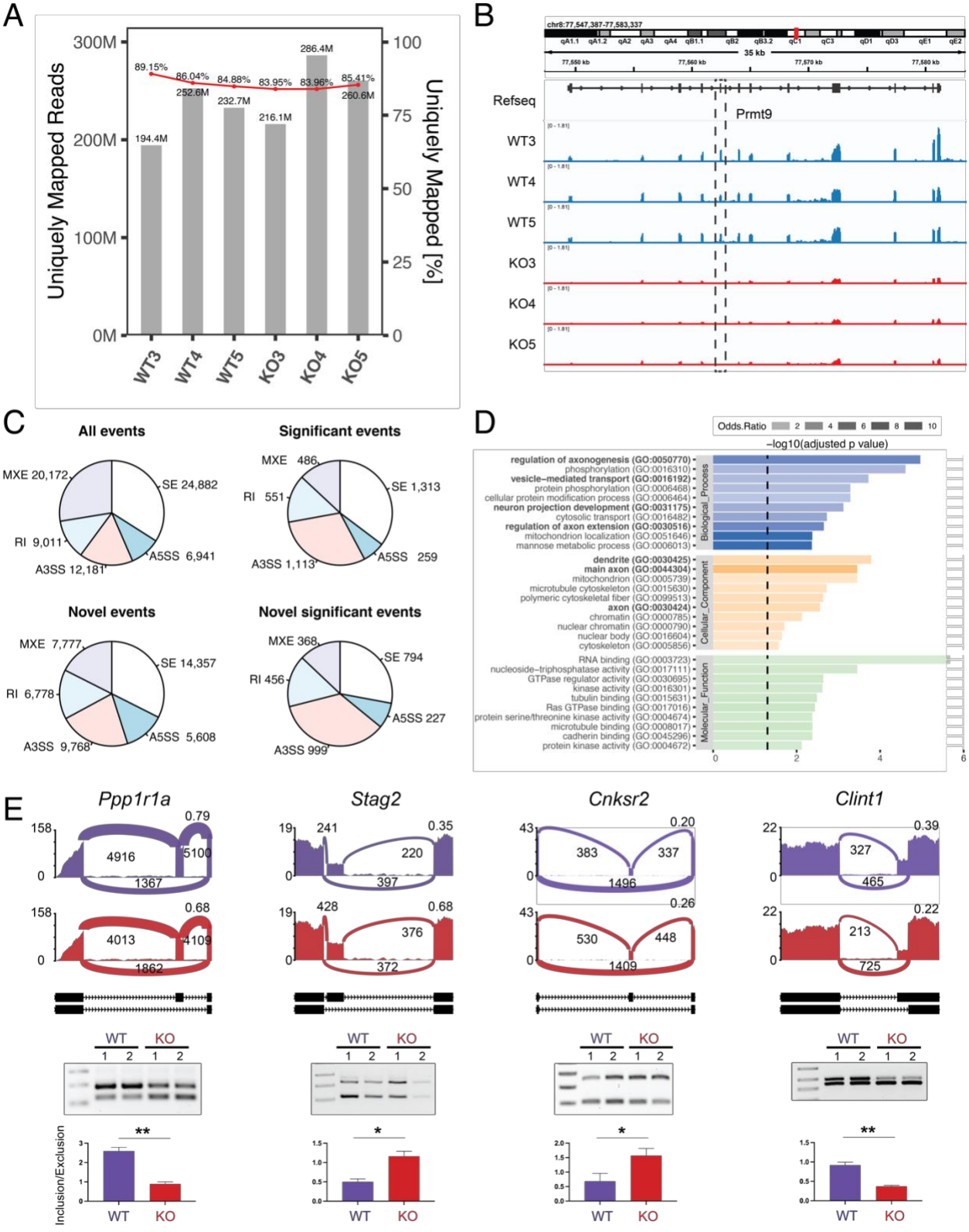



**Supplementary Figure 4.7 Amino-acid sequences and protein structure of human PRMT9**

**(A)** Amino-acid sequences of human PRMTs are aligned using ClustalW. The conserved Motif I is boxed with black squares. The blue asterisk indicates the G189 amino acid in PRMT9a. The number on the left indicates the positions of amino acid of individual PRMT,

starting at the initiator methionine. Human PRMT sequences used included PRMT1: NP_001527.3; PRMT2: NP_996845.1; PRMT3: NP_005779.1; PRMT4: NP_954592.1; PRMT5: NP_006100.2; PRMT6: NP_060607.2; PRMT7: NP_061896.1; PRMT8: NP_062828.3; and PRMT9: NP_612373.2. (b)

**(B)** Protein structure of human PRMT9 (PDB: 6PDM [178]) with focus on the G189 amino acid.


A
Uniquely Mapped Reads
Uniquely Mapped [%]
300M
200M
100M
0M
100
75
50
25
0
89.15%
86.04%
84.88%
83.95%
83.96%
85.41%
194.4M
252.6M
232.7M
216.1M
286.4M
260.6M
WT3
WT4
WT5
KO3
KO4
KO5
B
chr8:77,547,387-77,583,337
35 kb
Refseq
Prmt9
WT3
WT4
WT5
KO3
KO4
KO5
C
All events
MXE 20,172
SE 24,882
RI 9,011
A5SS 6,941
A3SS 12,181
Significant events
MXE 486
SE 1,313
RI 551
A5SS 259
A3SS 1,113
Novel events
MXE 7,777
SE 14,357
RI 6,778
A5SS 5,608
A3SS 9,768
Novel significant events
MXE 368
SE 794
RI 456
A5SS 227
A3SS 999
D
Odds.Ratio 2 4 6 8 10
−log10(adjusted p value)
regulation of axonogenesis (GO:0050770)
phosphorylation (GO:0016310)
vesicle−mediated transport (GO:0016192)
protein phosphorylation (GO:0006468)
cellular protein modification process (GO:0006464)
neuron projection development (GO:0031175)
cytosolic transport (GO:0016482)
regulation of axon extension (GO:0030516)
mitochondrion localization (GO:0051646)
mannose metabolic process (GO:0006013)
Biological_Process
dendrite (GO:0030425)
main axon (GO:0044304)
mitochondrion (GO:0005739)
microtubule cytoskeleton (GO:0015630)
polymeric cytoskeletal fiber (GO:0099513)
axon (GO:0030424)
chromatin (GO:0000785)
nuclear chromatin (GO:0000790)
nuclear body (GO:0016604)
cytoskeleton (GO:0005856)
Cellular_Component
RNA binding (GO:0003723)
nucleoside−triphosphatase activity (GO:0017111)
GTPase regulator activity (GO:0030695)
kinase activity (GO:0016301)
tubulin binding (GO:0015631)
Ras GTPase binding (GO:0017016)
protein serine/threonine kinase activity (GO:0004674)
microtubule binding (GO:0008017)
cadherin binding (GO:0045296)
protein kinase activity (GO:0004672)
Molecular_Function
0
2
4
6
E
Ppp1r1a
Stag2
Cnksr2
Clint1
WT
KO
Inclusion/Exclusion
**
*
*
**

**Supplementary Figure 4.8 Alternative splicing analysis of RNA-seq data from wild-type and Prmt9 KO mice.**

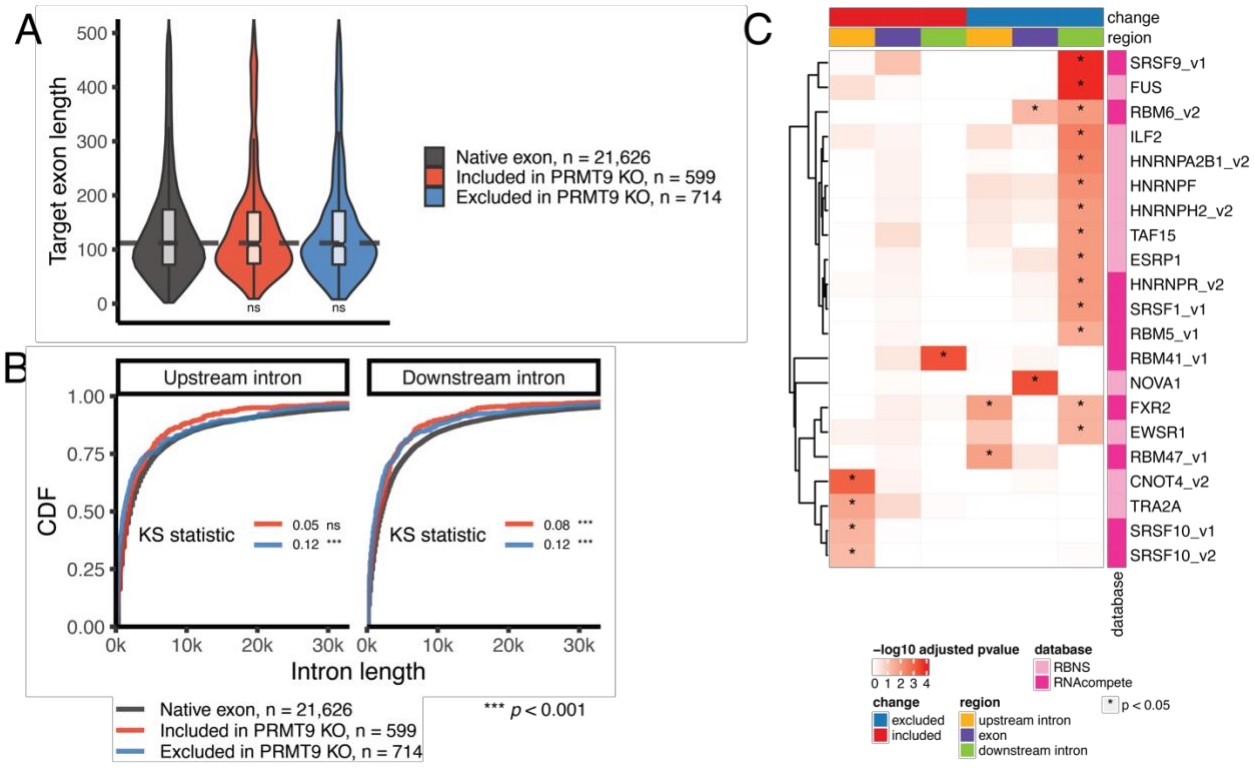
**(A)** Summary of sequencing depth and mapping statistics of RNA-Seq dataset.

**(B)** Genome browser tracks of Prmt9 gene in wild-type and Prmt9 KO mice. Refseq annotation of Prmt9 gene is diagrammed at the top. The exon 5 highlighted by the black box is the floxed exon. Knockout of exon 5 creates a pre-mature stop codon in exon 6, resulting in global reduction of Prmt9 mRNA.

**(C)** Pie charts depict the number of alternative splicing events in each category. There is an enrichment of A3SS events in significantly changed splicing events compared to all splicing events detected from the RNA-seq. All events, all alternative splicing events detected from the RNA-seq after filtering by number of supporting reads; Novel events, alternative splicing events harboring at least one cryptic splice site which is not annotated from the reference genome; (Novel) significant events, (novel) alternative splicing events significantly changed upon Prmt9 KO. SE, exon skipping / skipped exon; A5SS, alternative 5' splice sites; A3SS, alternative 3' splice sites; RI, intron retention / retained intron; MXE, mutually exclusive exon.

**(D)** Gene Ontology (GO) enrichment analysis of alternatively spliced genes (SE or A3SS). Top 10 enriched GO terms for Biological Process, Cellular Component and Molecular Function are shown in the bar charts. The length of bars depicts the Benjamini-Hochberg adjusted p values calculated from hypergeometric test. Odds ratio of the enrichment were indicated by opacity of bars.

**(E)** Sashimi plot visualization and experimental validation of additional differential splicing events. The black bars and dashed lines in the middle represent exons and introns, respectively. The splice site highlighted by red triangle is a cryptic splice site not annotated in the reference genome. Purple and red sashimi plots illustrate the splicing patterns in wild-type and Prmt9 KO mice, with solid peaks representing RNA-seq read coverages in RPKM (reads per kilobase per million mapped), arches representing splice junctions, and the numbers representing number of reads mapped to each splice junction. PSI values are also indicated on the right side of the sashimi plot. In the bottom, RT-PCR validations were shown. Higher band intensity of PCR products indicates higher expression of corresponding splicing isoform. The ratios of exon inclusion isoform to exon exclusion isoform are quantified and displayed in bar plots.




**Supplementary Figure 4.9 Additional sequence features of exon skipping events affected by Prmt9 KO**

**(A)** Violin plot of cassette exon length. The statistical significance against background native cassette exons were assessed using Wilcoxon's rank-sum test. ns, not significant.

**(B)** Cumulative density function plot of intron length upstream or downstream of the cassette exon. The statistical significance against background native cassette exon set was assessed by two-sample Kolmogorov–Smirnov test (2KS) test. The KS statistic and p values were shown in the middle of the plot. ns, not significant.

**(C)** Heatmap depicting the log-transformed Benjamini-Hochberg adjusted p values of RNA binding protein (RBP) motifs enriched in the target cassette exon or its flanking introns. Adjusted p values < 0.05 are marked by asterisk. Origins of motif information are indicated

by the database column. Included, cassette exons more included upon Prmt9 KO; excluded, cassette exons more excluded upon Prmt9 KO. Upstream intron, intronic region 300 nt upstream of the cassette exon; exon, the exon body region; downstream intron, intronic region 300 nt downstream the cassette exon.

# 4.6 References


2 Fu, X. D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet **15**, 689-701, doi:10.1038/nrg3778 (2014).

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, doi:10.1016/j.ajhg.2017.11.002 (2018).

11 Raj, B. & Blencowe, B. J. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron **87**, 14-27, doi:10.1016/j.neuron.2015.05.004 (2015).

12 Vuong, C. K., Black, D. L. & Zheng, S. The neurogenetics of alternative splicing. Nat Rev Neurosci **17**, 265-281, doi:10.1038/nrn.2016.27 (2016).

28 Yang, Y. et al. PRMT9 is a type II methyltransferase that methylates the splicing factor SAP145. Nat Commun **6**, 6428, doi:10.1038/ncomms7428 (2015).

29 Sachamitr, P. et al. PRMT5 inhibition disrupts splicing and stemness in glioblastoma. Nat Commun **12**, 979, doi:10.1038/s41467-021-21204-5 (2021).

30 Fong, J. Y. et al. Therapeutic Targeting of RNA Splicing Catalysis through Inhibition of Protein Arginine Methylation. Cancer Cell **36**, 194-209 e199, doi:10.1016/j.ccell.2019.07.003 (2019).

64 Irimia, M. et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell **159**, 1511-1523, doi:10.1016/j.cell.2014.11.035 (2014).

69 Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

100 Dominguez, D. et al. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol Cell **70**, 854-867 e859, doi:10.1016/j.molcel.2018.05.001 (2018).

101 Ray, D. et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature **499**, 172-177, doi:10.1038/nature12311 (2013).

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

129 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res **4**, 1521, doi:10.12688/f1000research.7563.2 (2015).

130 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

134 Blanc, R. S. & Richard, S. Arginine Methylation: The Coming of Age. Mol Cell **65**, 8-24, doi:10.1016/j.molcel.2016.11.003 (2017).

135 Guccione, E. & Richard, S. The regulation, functions and clinical relevance of arginine methylation. Nat Rev Mol Cell Biol **20**, 642-657, doi:10.1038/s41580-019-0155-x (2019).

136 Bryant, J. P., Heiss, J. & Banasavadi-Siddegowda, Y. K. Arginine Methylation in Brain Tumors: Tumor Biology and Therapeutic Strategies. Cells **10**, doi:10.3390/cells10010124 (2021).

137 Scaglione, A. et al. PRMT5-mediated regulation of developmental myelination. Nat Commun **9**, 2840, doi:10.1038/s41467-018-04863-9 (2018).

138 Zhang, X. et al. Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. Cell **166**, 1147-1162 e1115, doi:10.1016/j.cell.2016.07.025 (2016).

139 Bezzi, M. et al. Regulation of constitutive and alternative splicing by PRMT5 reveals a role for Mdm4 pre-mRNA in sensing defects in the spliceosomal machinery. Genes Dev **27**, 1903-1916, doi:10.1101/gad.219899.113 (2013).

140 Hadjikyriacou, A., Yang, Y., Espejo, A., Bedford, M. T. & Clarke, S. G. Unique Features of Human Protein Arginine Methyltransferase 9 (PRMT9) and Its Substrate RNA Splicing Factor SF3B2. J Biol Chem **290**, 16723-16743, doi:10.1074/jbc.M115.659433 (2015).

141 Lefebvre, S. et al. Correlation between severity and SMN protein level in spinal muscular atrophy. Nat Genet **16**, 265-269, doi:10.1038/ng0797-265 (1997).

142 Akawi, N. et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nat Genet **47**, 1363-1369, doi:10.1038/ng.3410 (2015).

143 Najmabadi, H. et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature **478**, 57-63, doi:10.1038/nature10423 (2011).

144 Jiang, H. et al. PRMT9 promotes hepatocellular carcinoma invasion and metastasis via activating PI3K/Akt/GSK-3beta/Snail signaling. Cancer Sci **109**, 1414-1427, doi:10.1111/cas.13598 (2018).

145 Guo, H. et al. Specificity and efficiency of Cre-mediated recombination in Emx1-Cre knock-in mice. Biochem Biophys Res Commun **273**, 661-665, doi:10.1006/bbrc.2000.2870 (2000).

146 Gorski, J. A. et al. Cortical excitatory neurons and glia, but not GABAergic neurons, are produced in the Emx1-expressing lineage. J Neurosci **22**, 6309-6314, doi:20026564 (2002).

147 Vorhees, C. V. & Williams, M. T. Morris water maze: procedures for assessing spatial and related forms of learning and memory. Nat Protoc **1**, 848-858, doi:10.1038/nprot.2006.116 (2006).

148 Saxe, M. D. et al. Ablation of hippocampal neurogenesis impairs contextual fear conditioning and synaptic plasticity in the dentate gyrus. Proc Natl Acad Sci U S A **103**, 17501-17506, doi:10.1073/pnas.0607207103 (2006).

149 Ule, J. et al. Nova regulates brain-specific splicing to shape the synapse. Nat Genet **37**, 844-852, doi:10.1038/ng1610 (2005).

150 Zhu, J., Shang, Y. & Zhang, M. Mechanistic basis of MAGUK-organized complexes in synaptic development and signalling. Nat Rev Neurosci **17**, 209-223, doi:10.1038/nrn.2016.18 (2016).

151 Liu, H. et al. N-terminal alternative splicing of GluN1 regulates the maturation of excitatory synapses and seizure susceptibility. Proc Natl Acad Sci U S A **116**, 21207-21212, doi:10.1073/pnas.1905721116 (2019).

152 Sengar, A. S. et al. Control of Long-Term Synaptic Potentiation and Learning by Alternative Splicing of the NMDA Receptor Subunit GluN1. Cell Rep **29**, 4285-4294 e4285, doi:10.1016/j.celrep.2019.11.087 (2019).

153 McAvoy, T., Zhou, M. M., Greengard, P. & Nairn, A. C. Phosphorylation of Rap1GAP, a striatally enriched protein, by protein kinase A controls Rap1 activity and dendritic spine morphology. Proc Natl Acad Sci U S A **106**, 3531-3536, doi:10.1073/pnas.0813263106 (2009).

154 Clarke, T. K. et al. KCNJ6 is associated with adult alcohol dependence and involved in gene x early life stress interactions in adolescent alcohol drinking. Neuropsychopharmacology **36**, 1142-1148, doi:10.1038/npp.2010.247 (2011).

155 Rangel, L., Lospitao, E., Ruiz-Saenz, A., Alonso, M. A. & Correas, I. Alternative polyadenylation in a family of paralogous EPB41 genes generates protein 4.1 diversity. RNA Biol **14**, 236-244, doi:10.1080/15476286.2016.1270003 (2017).

156 Ou, M. Y. et al. The CTNNBIP1-CLSTN1 fusion transcript regulates human neocortical development. Cell Rep **35**, 109290, doi:10.1016/j.celrep.2021.109290 (2021).

157 Huang, R. et al. The NCATS BioPlanet - An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and

Chemical Genomics. Front Pharmacol **10**, 445, doi:10.3389/fphar.2019.00445 (2019).

158 Zhu, S. et al. Mechanism of NMDA Receptor Inhibition and Activation. Cell **165**, 704-714, doi:10.1016/j.cell.2016.03.028 (2016).

159 Zhang, Y. et al. Structural basis of ketamine action on human NMDA receptors. Nature **596**, 301-305, doi:10.1038/s41586-021-03769-9 (2021).

160 Slenter, D. N. et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res **46**, D661-D667, doi:10.1093/nar/gkx1064 (2018).

161 Feng, W. & Zhang, M. Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. Nat Rev Neurosci **10**, 87-99, doi:10.1038/nrn2540 (2009).

162 Rodriguez-Palmero, A. et al. DLG4-related synaptopathy: a new rare brain disorder. Genet Med **23**, 888-899, doi:10.1038/s41436-020-01075-9 (2021).

163 Zheng, S. et al. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. Nat Neurosci **15**, 381-388, S381, doi:10.1038/nn.3026 (2012).

164 Zheng, S., Damoiseaux, R., Chen, L. & Black, D. L. A broadly applicable high-throughput screening strategy identifies new regulators of Dlg4 (Psd-95) alternative splicing. Genome Res **23**, 998-1007, doi:10.1101/gr.147546.112 (2013).

165 Gozani, O., Feld, R. & Reed, R. Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. Genes Dev **10**, 233-243, doi:10.1101/gad.10.2.233 (1996).

166 Wahl, M. C. & Luhrmann, R. SnapShot: Spliceosome Dynamics I. Cell **161**, 1474-e1471, doi:10.1016/j.cell.2015.05.050 (2015).

167 Zhang, Q. et al. BPP: a sequence-based algorithm for branch point prediction. Bioinformatics **33**, 3166-3172, doi:10.1093/bioinformatics/btx401 (2017).

168 Haselbach, D. et al. Structure and Conformational Dynamics of the Human Spliceosomal B(act) Complex. Cell **172**, 454-464 e411, doi:10.1016/j.cell.2018.01.010 (2018).

169 Wu, Q., Schapira, M., Arrowsmith, C. H. & Barsyte-Lovejoy, D. Protein arginine methylation: from enigmatic functions to therapeutic targeting. Nat Rev Drug Discov **20**, 509-530, doi:10.1038/s41573-021-00159-8 (2021).

170 Baeza-Centurion, P., Minana, B., Schmiedel, J. M., Valcarcel, J. & Lehner, B. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. Cell **176**, 549-563 e523, doi:10.1016/j.cell.2018.12.010 (2019).

171 Hicks, M. J., Mueller, W. F., Shepard, P. J. & Hertel, K. J. Competing upstream 5' splice sites enhance the rate of proximal splicing. Mol Cell Biol **30**, 1878-1886, doi:10.1128/MCB.01071-09 (2010).

172 Yu, Y. et al. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. Cell **135**, 1224-1236, doi:10.1016/j.cell.2008.10.046 (2008).

173 Larkin, M. A. et al. Clustal W and Clustal X version 2.0. Bioinformatics **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

174 Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res **44**, W90-97, doi:10.1093/nar/gkw377 (2016).

175 Szklarczyk, D. et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res **49**, D605-D612, doi:10.1093/nar/gkaa1074 (2021).

176 Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol **11**, 377-394, doi:10.1089/1066527041410418 (2004).

177 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res **14**, 1188-1190, doi:10.1101/gr.849004 (2004).

178 Halabelian, L., Tempel, W., Zeng, H., Li, Y., Seitova, A., Hutchinson, A., Bountra, C., Edwards, A.M., Arrowsmith, C.H., Structural Genomics Consortium (SGC). Crystal structure of Human Protein Arginine Methyltransferase 9 (PRMT9), <https://www.rcsb.org/structure/6PDM> (2019).

# 5 CONCLUDING REMARKS

A wealth of studies, comprising high-throughput sequencing, protein structure visualization, functional assays, cell culture, and transgenic mouse models, have helped to define the physiological roles of splicing in cell fate commitment and diseases such as cancer, neurodegenerative and other related disorders. Also, our knowledge of pre-mRNA splicing has remarkably expanded with advances in the sequencing technologies and development of computational tools to detect and quantify splicing variations.

In the meantime, the accumulation of RNA-seq data, especially those generated from consortium studies, also presented new challenges for the global profiling of splicing landscapes using massive datasets. It urges the development of new computational tools that are time- and memory-efficient for splicing analysis of large-scale datasets. Novel discoveries of associations between genetic variants and disease, or between splicing and other biological processes are continuously emerging under various settings. Examination of the splicing-related regulatory mechanisms underlying those associations would be

extremely helpful to narrow down the inventory for further experimental validation and enlighten the pathogenesis of diseases.

To resolve the computational challenge on big data analysis, in Chapter 2, we developed the rMATS-turbo, a computational tool designed for splicing analysis that exhibits dramatic improvement in processing speed and memory efficiency. A simple single-command differential analysis on relative smaller dataset produced robust identification of splicing alterations between cell lines, including those derived from cryptic splice site usage. The decoupling of splicing graph generation steps and splicing event detection step enables parallel processing of input samples and can strikingly reduce the time and memory consumption. Multi-command analysis on 1,019 RNA-seq datasets (18.58 T base) from the CCLE database can be finished in ~ 3 days, when parallelly processed. These results demonstrated that rMATS-turbo can facilitate robust, straightforward, and ultra-fast analysis of alternative splicing, which is suitable for splicing profiling in large-scale dataset.

In Chapter 2 and 3, we try to utilize rMATS-turbo, as well as other computational platforms and approaches to computationally elucidate the splicing regulatory mechanism underlying tissue development and disease. In Chapter 3, we used the osteogenic differentiation of MSPCs, to study the regulation of splicing by temporal expression of trans-acting RBPs using time-course RNA-seq data. Extensive splicing changes has been identified, which shows temporal correlation with expression of specific RBPs. We developed a new computational framework, combining correlation analysis and RBP motif enrichment analysis, to determine key splicing regulators of osteogenic differentiation. Perturbation of two out of nine RBP candidates identified by our computational framework

lead to reduced osteogenesis differentiation in vitro. Overall, this work highlights a high degree of splicing regulation network during osteogenic differentiation. And the computational framework can be generalized to other time-course data to elucidate the splicing regulations in other biological processes and disease trajectories.

In Chapter 4, we attempt to figure out whether PRMT9 contributes to brain development though SF3B2 methylation mediated splicing regulation. We showed that the PRMT9 G189R mutant, which is previously reported as causative variant for autosomal recessive intellectual disability, has eliminated methyltransferase activity and diminished stability. Also, the tissue-specific knockout of Prmt9 in excitatory neurons exhibit impaired learning, memory, and functional synapse maturation in hippocampus tissue. We also generated a deep RNA-seq dataset from wild-type and Prmt9 whole body knockout mice for global gene expression and alternative splicing analysis. We revealed a PRMT9-SF3B2-splicing-synapse regulatory cascade that associates Prmt9 to brain development in mice, as evidenced by: 1) Except for Prmt9 itself, no steady-state gene expression change was observed in Prmt9 KO mice; 2) Many of the synaptic genes were subject to splicing changes, which has been validated by RT-PCR; 3) Genes with splicing changes were significantly enriched in synapse-related pathways and GO terms; 4) ARID-causative genes and genes in the Ras/Rho/PSD95 network, or their paralogous genes, displayed splicing changes upon Prmt9 KO. Moreover, computational comparison of sequencing features associated with SE event suggested that Prmt9 affects splicing by regulating 3' splice site competition. This is confirmed by the spatial proximity of the SF3B2 R508, which can be methylated by PRMT9, to the anchoring site on the pre-mRNA. A working model is proposed that PRMT9-mediated SF3B2 R508me2s regulates splicing through 3' splice site

competition by altering SF3B2/pre-mRNA interaction. This model is supported by the differential interaction of SF3B2 with 3' splice site sequences between upstream and downstream introns. It can be further verified by additional experimental assays inspecting whether the differential interaction results from changes in sequence features, especially changes in SF3B complex anchoring site.

To summarize, the development of new computational approaches and data analysis can be inspired by advances in technologies and emerge of new biological discoveries. The design of computational analysis should take pre-existing biological knowledges into account. Reciprocally, it would facilitate the understanding of molecular changes underlying biological processes, and guide the design of functional assays to validate computational findings. In terms of alternative splicing, with the reinforced understanding of its regulation in development and disease, it is expected that the clinical relevance of splicing to disease diagnosis, prognosis, and therapy will be emphasized. For example, cancer cells with specific genetic backgrounds may respond differently to the perturbation of upstream regulators in the splicing regulatory cascade; Neoepitopes arising from individual splicing alterations can be targeted by immunotherapy in personalized medicine. Overall, we anticipate that computational approaches studying splicing regulatory mechanisms could deepen our insights into the pathogenesis of diseases, and guide new biological and clinical discoveries.