UNIVERSITY OF CALIFORNIA
Los Angeles

Computational methods to elucidate post-transcriptional gene regulation using high-throughput sequencing data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioinformatics

by

Zijun Zhang

2019

ABSTRACT OF THE DISSERTATION

Computational methods to elucidate post-transcriptional gene regulation using high-throughput sequencing data

by

Zijun Zhang
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2019
Professor Yi Xing, Chair

Post-transcriptional regulation plays a central role in the flow of information from genotypes to phenotypes in the cellular machinery. Disruptions of post-transcriptional regulatory mechanisms underlie many human diseases. As high-throughput sequencing technology becomes the standard protocol in studying post-transcriptional regulation, large-scale data in public domain provides an unprecedented resource to understand the complex regulatory networks of gene regulation, while also presents challenges for the development of computational methods to analyze and interpret empirical data into biological knowledge. In this dissertation, novel statistical models and computational frameworks were developed to elucidate post-transcriptional gene regulation using high-throughput sequencing data. Utilizing these new tools, we demonstrated that we can robustly characterize the molecular signals and variations across diverse biological states, and more importantly, identify *bona fide* regulatory events that are inaccessible by conventional analyses.

The first part of the dissertation describes CLIP-seq Analysis of Multi-mapped reads (CLAM), a comprehensive computational pipeline for analyzing Crosslinking or RNA immunoprecipitation followed by sequencing (CLIP/RIP-seq) data. As CLIP-seq/RIP-seq reads are short, existing computational tools focus on uniquely mapped reads, while reads mapped to multiple loci are discarded. CLAM uses an expectation-maximization algorithm to assign multi-mapped reads and calls peaks combining uniquely and multi-mapped reads. CLAM recovered a large number of novel RNA regulatory sites inaccessible by uniquely mapped reads in datasets with different regulatory features, providing a useful tool to discover novel protein-RNA interactions and RNA modification sites from CLIP-seq and RIP-seq data.

The second part of the dissertation presents Deep-learning Augmented RNA-seq analysis of Transcript Splicing (DARTS), a novel computational framework that integrates deep learning-based predictions with empirical RNA-seq datasets to infer differential alternative splicing between biological conditions. A major limitation of RNA sequencing (RNA-seq) analysis of alternative splicing is its reliance on high sequencing coverage. DARTS employs a deep neural network (DNN) that predicts differential alternative splicing using *cis* RNA sequence features and *trans* RNA binding protein levels. DARTS DNN trained on public RNA-seq displays a high prediction accuracy and generalizability. Incorporating DARTS DNN prediction as an informative prior significantly improves the inference of differential alternative splicing. DARTS leverages public RNA-seq big data to provide a knowledge base of splicing regulation via deep learning, thereby helping researchers better characterize alternative splicing using RNA-seq datasets even with modest coverage.

The dissertation of Zijun Zhang is approved.

Douglas L. Black

Chuan He

Yingnian Wu

Yi Xing, Committee Chair

University of California, Los Angeles

2019

To my mother and father

and Liangke

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Yi Xing, who has been extremely supportive to my research and career development over the past five years. I came into his lab as an undergraduate student in 2013, and learned bioinformatics, statistics, RNA biology, writing, and everything I needed to do research from him throughout my PhD training. His critical thinking and scientific rigor have been invaluable in cultivating me to become a scientist.

I owe a special thanks to my wife, Liangke Gou, for her encouragement and love. She is always patient and supportive, and has devoted wholeheartedly to our family. I could not imagine how difficult it would be without her.

I would like to thank my committee members for their insightful advice and generous support: Douglas Black, Chuan He and Yingnian Wu. I was lucky to have them on my committee and owe them a debt of gratitude.

My life as a PhD student in the Xing lab would not have been so colorful without my friends and colleagues: Zhicheng Pan, Eddie Park, Yang Pan, Yida Zhang, Chengyang Wang, Samir Adhikari, Levon Demirdjian, Xinyuan Chen, Yuanyuan Wang, Harry Yang, Yongbo Wang, Yang Guo, Yan Gao, Amal Katrib, Yang Xu, Emad Bahrami-Samani, Ruijiao Xin, Jinkai Wang, Zhixiang Lu, Shaofang Li, Shayna Stein and Yu-ting Tseng.

Last but not least, I would like to thank my parents for their support and unconditioned love. I also want to apologize to them for not being able to keep them company during my PhD study in the US.

VITA

Education

2018-2019	Visiting Scholar, Center for Computational Genomics and Medicine, University of Pennsylvania
2014-2019	Graduate Student Researcher, Bioinformatics IDP, University of California, Los Angeles
2018	Teaching Assistant, MCDB 187AL: Bioinformatics Lab University of California, Los Angeles
2010-2014	B.S. in Biological Sciences, College of Life Sciences, Zhejiang University

Honors and Awards

2018-2019	Dissertation Year Fellowship, Grad Division, University of California, Los Angeles
Oct. 2016	Future of Science Scholarship Omics Strategies to Study the Proteome, Keystone Symposia
Oct. 2015	Travel Award for Best Poster, QCBio 1st Annual Retreat, University of California, Los Angeles
2010-2013	Outstanding Student's Award, Undergrad School Zhejiang University

Publications

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

Z Zhang, E Park, L Lin, Y Xing (2018). A panoramic view of RNA modifications: exploring new frontiers. *Genome biology*. 19 (1), 11

E Park, Z Pan, **Z Zhang**, L Lin, Y Xing (2018). The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics* 102 (1), 11-26

Z Zhang, Y Xing (2017). CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic acids research* 45 (16), 9260-9271

L Liu, **Z Zhang**, T Sheng, M Chen (2017). DEF: an automated dead-end filling approach based on quasi-endosymbiosis. *Bioinformatics* 33 (3), 405-413

L Liu, Q Mei, Z Yu, T Sun, **Z Zhang**, M Chen (2013). An integrative bioinformatics framework for genome-scale multiple level network reconstruction of rice. *Journal of integrative bioinformatics* 10 (2), 94-102

L Liu, **Z Zhang**, Q Mei, M Chen (2013). PSI: a comprehensive and integrative approach for accurate plant subcellular localization prediction. *PLOS One* 8 (10), e75826

Chapter 1 Introduction

Human transcriptomes consist of a versatile collection of RNA molecules. RNAs are transcribed from DNAs and subsequently translated into proteins, therefore play a central role in linking genotypes to molecular and phenotypical variations. Nascent RNAs undergo a variety of co-transcriptional and post-transcriptional processing steps to form the final mature RNA products. These steps involve 5'-capping, 3'-polyadenylation, RNA splicing, and RNA modifications including RNA editing and RNA base modifications. Understanding the molecular mechanisms in human transcriptomes is crucial for deciphering the regulatory basis of human cells, and ultimately for the therapeutics development for human diseases.

Among the RNA processing steps, RNA alternative splicing is an essential biological process where introns in nascent RNAs are excised, and exons are selectively included or excluded, to form mature mRNA products. The vast majority of multi-exon human genes are alternatively spliced (1). Through alternative splicing, one gene is capable to produce multiple RNA isoforms with distinct exon structures, hence greatly diversifies the human transcriptome. Different isoforms of the same gene can have vastly different functional characteristics (2, 3). Splicing is a key factor in translating genotypes to phenotypical variations (4), and disruption of splicing regulation underlies many human diseases (5).

Another emerging regulatory axis is RNA modifications. More than 140 types of chemical modifications have been found in RNA. N^6-methyladenosine (m^6A) is the most

prevalent type of reversible internal modification in eukaryotic mRNAs. m^6A-modified transcripts enter a separate track of fast decay and fast translation mediated through m^6A reader proteins (6). A set of reader proteins recognize the m6A-marked bases and affects the downstream metabolism of these RNAs. One reader protein, YTHDF2, promotes m^6A-mediated mRNA decay (7), while another reader protein, YTHDF1, promotes mRNA translation (8). The functional consequence of m^6A manifest particularly in fast-proliferating cells, including normal cell differentiation and development, as well as disease conditions like cancers (9). Mis-regulation of m^6A-associated processes leads to stalled cell differentiation and human diseases.

The complex RNA processing landscape is orchestrated by regulatory networks of RNA-binding proteins (RBP) and cross-talks between multiple processing pathways (10). The development of high-throughput sequencing technologies has provided powerful approaches to study the regulatory network as well as the downstream RNA molecular variations at large scale and at fine resolution. RNA-sequencing (RNA-seq) of Illumina short reads as well as different long reads technologies have enabled the characterization of the global transcriptomic landscape at large scale. Consortium efforts have profiled publicly-available RNA-seq across diverse tissue/cell-lines as well as under diverse perturbations. To study and characterize the RBPs, crosslinking immunoprecipitation and RNA immunoprecipitation followed by high-throughput sequencing (CLIP-seq and RIP-seq) has enabled transcriptome-wide discoveries of protein-RNA interactions as well as RNA modification sites.

As efforts to elucidate the human transcriptome regulatory patterns, Chapters 2 & 3 of this dissertation present two novel computational methods for analyzing high-

throughput transcriptome sequencing data. In detail, Chapter 2 describes a novel computational method for analyzing CLIP/RIP-seq data, called CLIP-seq Analysis of Multi-mapped reads (CLAM)(11). CLAM uses an expectation-maximization algorithm to assign multi-mapped reads and calls peaks combining uniquely and multi-mapped reads. To demonstrate the utility of CLAM, we applied it to a wide range of public CLIP-seq/RIP-seq datasets involving numerous splicing factors, microRNAs and m^6A RNA methylation. CLAM recovered a large number of novel RNA regulatory sites inaccessible by uniquely mapped reads. The functional significance of these sites was demonstrated by consensus motif patterns and association with alternative splicing (splicing factors), transcript abundance (AGO2) and mRNA half-life (m^6A). CLAM provides a useful tool to discover novel protein–RNA interactions and RNA modification sites from CLIP-seq and RIP-seq data, and reveals the significant contribution of repetitive elements to the RNA regulatory landscape of the human transcriptome.

Chapter 3 reports a novel computational and statistical framework for analyzing RNA alternative splicing events called Deep-learning Augmented RNA-seq analysis of Transcript Splicing (DARTS)(12). The rapid accumulation of RNA-seq data across diverse cell types and conditions provides an unprecedented resource for characterizing transcriptome complexity. However, the use of these large-scale data in routine RNA-seq studies to detect patterns of expression and thereby discover new regulatory events has been limited. DARTS integrates deep learning-based predictions with empirical evidence in specific RNA-seq datasets to infer differential alternative splicing between conditions. A core component of DARTS is a deep neural network (DNN) that predicts differential alternative splicing using cis RNA sequence features and trans RNA binding protein levels.

DARTS DNN trained on public RNA-seq datasets (ENCODE, Roadmap Epigenomics) displays a high prediction accuracy and generalizability. Incorporating DARTS DNN prediction as an informative prior significantly improves the inference of differential alternative splicing, especially from low-coverage RNA-seq datasets. In cellular models of the epithelial-mesenchymal transition, DARTS reliably predicted alternative splicing changes in lowly expressed genes, that were inaccessible by a conventional RNA-seq analysis even at a high sequencing depth. Thus, DARTS capitalizes on large-scale public RNA-seq resources to discover differential alternative splicing across diverse transcriptomes.

With these new tools at hand, we can more comprehensively analyze and understand the experimental data from human transcriptomes, and more importantly, look at the transcriptomic regions and events that are inaccessible by conventional methods previously. Because of the natural expression variations in the human transcriptome, studies are inherently biased towards ‘visible’ fractions of highly-expressed genes and/or high mappability regions, where the experimental information is ample, and the measurement is robust. Meanwhile, a non-trivial fraction of the transcriptome is not so highly-expressed and/or in less definitive regions. These are the ‘dark matters’ in the human transcriptome. To study the dark matters, we need more sophisticated computational methods -- CLAM was developed to uncover protein-RNA interaction sites in repetitive elements, and DARTS was developed to characterize splicing alterations in lowly-expression genes. In the final Chapter of this dissertation, we summarize these methodology and data analysis results. The implications of CLAM and DARTS are delineated, and ongoing and future directions of CLAM and DARTS are discussed.

References

1. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

2. Ellis,J.D., Barrios-Rodiles,M., Çolak,R., Irimia,M., Kim,T., Calarco,J.A., Wang,X., Pan,Q., O'Hanlon,D., Kim,P.M., *et al.* (2012) Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. *Mol. Cell*, **46**, 884–892.

3. Yang,X., Coulombe-Huntington,J., Kang,S., Sheynkman,G.M., Hao,T., Richardson,A., Sun,S., Yang,F., Shen,Y.A., Murray,R.R., *et al.* (2016) Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, **164**, 805–817.

4. Li,Y.I., van de Geijn,B., Raj,A., Knowles,D.A., Petti,A.A., Golan,D., Gilad,Y. and Pritchard,J.K. (2016) RNA splicing is a primary link between genetic variation and disease. *Science (80-.).*, **352**, 600–604.

5. Park,E., Pan,Z., Zhang,Z., Lin,L. and Xing,Y. (2018) The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.*, **102**, 11–26.

6. Zhang,Z., Park,E., Lin,L. and Xing,Y. (2018) A panoramic view of RNA modifications: exploring new frontiers. *Genome Biol.*, **19**, 11.

7. Wang,X., Lu,Z., Gomez,A., Hon,G.C., Yue,Y., Han,D., Fu,Y., Parisien,M., Dai,Q.,

Jia,G., *et al.* (2014) N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, **505**, 117–20.

8. Wang,X., Zhao,B.S., Roundtree,I.A., Lu,Z., Han,D., Ma,H., Weng,X., Chen,K., Shi,H. and He,C. (2015) N^6-methyladenosine modulates messenger RNA translation efficiency. *Cell*, **161**.

9. Frye,M., Harada,B.T., Behm,M. and He,C. (2018) RNA modifications modulate gene expression during development. *Science*, **361**, 1346–1349.

10. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.

11. Zhang,Z. and Xing,Y. (2017) CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res.*, **45**, 172–177.

12. Zhang,Z., Pan,Z., Ying,Y., Xie,Z., Adhikari,S., Phillips,J., Carstens,R.P., Black,D.L., Wu,Y. and Xing,Y. (2019) Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat. Methods 2019 164*, **16**, 307.

Chapter 2 CLIP-seq Analysis of Multi-mapped reads

2.1 Introduction

Mammalian genomes encode over a thousand RNA binding proteins (RBPs) that play important roles in RNA processing and metabolism (1,2). RBPs interact with their cognate sequences and/or structural elements within the RNA to impact diverse aspects of post-transcriptional regulation, including splicing, polyadenylation, transport, stability and translational control, as well as RNA base modifications (3). For example, many RBPs function as splicing factors through interactions with cis splicing regulatory elements within the pre-mRNA (4). In recent years, there have been intense efforts to identify and characterize RBPs using high-throughput methods. For example, technologies such as SELEX-seq (5), RNAcompete (1), and RNA Bind-n-Seq (6) have been developed to define the *in vitro* binding motifs of numerous RBPs.

A powerful strategy for transcriptome-wide mapping of RBP-RNA interactions and RNA regulatory elements is immunoprecipitation followed by high-throughput sequencing (7). Two popular approaches are CLIP-seq (crosslinking with immunoprecipitation followed by sequencing) (8-10) and RIP-seq (RNA immunoprecipitation followed by sequencing) (11). The standard protocol of CLIP-seq involves crosslinking protein-RNA interactions by UV, immunoprecipitating the RBP-RNA complexes by antibody, then sequencing cDNA library to generate short reads typically ranging between 35bp to 50bp. Three versions of CLIP-seq (HITS-CLIP, PAR-CLIP and iCLIP) deliver datasets with distinct features due to their technical differences and biases (12). RIP-seq experiments

are performed in similar procedures, but RIP-seq does not include the UV-crosslinking step, resulting in reduced resolution of the binding sites and lower signal-to-noise ratios (13). Besides detecting RBP-RNA interaction sites, RIP-seq and CLIP-seq have also been utilized for detecting RNA base modifications, in particular N^6-methyladenosine (m^6A) (14), revealing the prevalence and dynamic landscape of reversible RNA base modifications in the human transcriptome (14,15).

Despite the increasing popularity and widespread use of CLIP-seq and RIP-seq for mapping RBP-RNA interaction and RNA modification sites, existing computational approaches for analyzing these data still have important limitations. As CLIP-seq and RIP-seq reads are short (usually <50bp), in a conventional data analysis workflow, reads are mapped to the genome and transcriptome, uniquely mapped reads are retained, and RBP binding sites are identified by appropriate statistical models for peak calling (12) (**Figure 2.1A**). However, by restricting the analysis to uniquely mapped reads and removing reads mapped to multiple genomic loci, a potentially large catalog of regulatory sites residing in duplicated and repetitive regions of the transcriptome will be under-detected or inaccessible. Given that approximately half of the human genome is comprised of transposable elements (16), and a variety of RBPs such as hnRNPC, ADAR1, and STAU1 have binding sites derived from highly repetitive transposable elements (17-20), the restriction to uniquely mapped reads represents a significant source of false negatives in site identification from CLIP-seq and RIP-seq datasets.

Here, we present CLAM (CLIP-seq Analysis of Multi-mapped reads), a new computational method for CLIP/RIP-seq data analysis and peak calling utilizing multi-mapped reads. We applied CLAM to published CLIP-seq data of 18 RBPs, as well as

RIP-seq data of the m^{6}A RNA modifications. In all datasets, CLAM recovered a large number of novel RNA regulatory sites inaccessible by conventional analyses of uniquely mapped reads. We further demonstrated the physical and functional relevance of the identified CLAM sites based on consensus motif patterns as well as correlation with relevant RNA regulatory features. Altogether, CLAM provides a useful and widely applicable computational tool to discover novel functional protein-RNA interaction sites and RNA modification events from CLIP-seq and RIP-seq data, and reveals the significant contribution of repetitive elements to the RNA regulatory landscape of the human transcriptome.

2.2 Results

2.2.1 CLAM statistical model for multi-mapped reads in CLIP-seq and RIP-seq data

To utilize multi-mapped reads in CLIP-seq and RIP-seq data and improve peak calling in highly repetitive regions, we developed CLAM, which assigns multi-mapped reads using an Expectation-Maximization (EM) framework, followed by peak calling with a permutation-based procedure commonly used for CLIP-seq and RIP-seq data (Methods). The statistical model of CLAM was inspired by previous work on resolving multi-mapped reads in RNA-seq (24,41,42) and ChIP-seq data (43,44), while features specific for CLIP-seq and RIP-seq data were incorporated in the model. Below we briefly illustrate the CLAM algorithm, using one read mapped to two genomic regions as the example (**Figure 2.1B**). CLAM first collapses reads into genomic regions. The two genomic regions in **Figure 2.1B** have 6 and 2 uniquely mapped reads respectively, while sharing one multi-

mapped read which will be resolved by CLAM. As certain RBPs (e.g. AGO2) could have long footprints on mRNA transcripts due to multiple overlapping binding sites (45), we designed the EM algorithm in CLAM to assign a multi-mapped read based on the mapping status of other reads (uniquely + multi-mapped) within a defined local window surrounding the read of interest (Methods). The algorithm iterates between inferring the expected true origins of multi-mapped reads and deriving the Maximum Likelihood Estimates (MLE) for the probabilities of reads derived from specific regions, until it reaches convergence. In the hypothetical example in **Figure 2.1B**, for the multi-mapped read CLAM will assign 0.75 and 0.25 read to the left and right regions respectively to achieve the maximum likelihood. Once multi-mapped reads are re-assigned, a permutation test will be performed for peak calling combining uniquely-mapped reads and CLAM assignment of multi-mapped reads (Materials and Methods).

To systematically evaluate the behavior of the CLAM EM framework for re-assigning multi-mapped CLIP-seq reads, we generated a benchmark dataset by truncating the hnRNPC iCLIP (19) reads by 10bp from the 3' end then remapping the truncated reads to the genome. This strategy enabled us to assess the algorithm performance on "gold-standard" reads that were uniquely mapped in the full-length dataset but became multi-mapped in the truncated dataset. For comparison, we implemented and evaluated two alternative models: 1) assigning multi-mapped reads uniformly with equal weights for all mapped regions ("uniform" model); 2) assigning multi-mapped reads weighted by local counts of uniquely mapped reads, which corresponds to the first iteration of EM ("one-iter" model). Then we assessed the accuracy of re-assigning these reads to the known originating loci (positive loci) over the rest of multi-mapped loci

(negative loci) by AUROC (Area Under Receiver Operating Characteristic curve), AUPR (Area Under Precision-Recall curve), and the median/mean weight for positive vs negative loci. As illustrated in **Table 2.1**, uniform assignment of multi-mapped reads resulted in the poorest performance. Although the CLAM model and the one-iter model achieved comparable AUROC and AUPR values, detailed analyses indicated that the CLAM model as compared to the one-iter model assigned higher weights to positive loci (0.62 vs 0.50) and lower weights to the negative loci (0.02 vs 0.13), demonstrating its superior performance.

In sum, CLAM is a two-stage algorithm that first re-assigns the multi-mapped reads using a statistical model (i.e. EM), followed by peak calling using the information of both uniquely-mapped reads and multi-mapped reads. Compared to conventional CLIP-seq/RIP-seq peak calling procedures of using only uniquely-mapped reads, CLAM can discover a large number of novel sites inaccessible by conventional methods, as demonstrated by our systematic assessments using multiple datasets below. It should also be noted that while EM and permutation test could be slow, we used computational techniques to boost the speed of CLAM. For EM-based probabilistic read assignment, we implemented Binary Indexed Tree (BIT) for faster reading and updating of weights. For permutation test based peak calling, we implemented a multi-threading framework for parallel peak-calling on a gene-by-gene basis. As a result, CLAM has reasonable running time that scales well to the total library size.

2.2.2 CLAM rescues multi-mapped reads and discovers novel sites in CLIP-seq and RIP-seq data

To assess the utility of CLAM, we first applied it to three published datasets on hnRNPC, AGO2, and m^{6}A (**Table 2.2**). We chose these three datasets because they were associated with distinct RNA regulatory processes (alternative splicing, microRNA targeting, and m^{6}A methylation respectively), and included both CLIP-seq (hnRNPC, AGO2) and RIP-seq (m^{6}A) data. Each dataset had two biological replicates. After preprocessing and adapter trimming, the average read lengths were 40, 44, and 50 for the three datasets respectively. We then calculated the percentage of multi-mapped reads among all mapped reads. As shown in **Figure 2.2A**, approximately 10% to 18% of reads were multi-mapped across the six samples. Using CLAM, we were able to rescue the vast majority (83% to 92%) of multi-mapped reads, representing a significant gain in read coverage especially in repetitive regions of the transcriptome (see below). A small proportion (~10%) of multi-mapped reads were not analyzed by CLAM because they did not cluster to genomic regions (i.e. singleton reads with no other reads in vicinity), or were mapped to too many (≥100) regions and therefore discarded (see details in Methods).

The rescued multi-mapped reads were significantly enriched in repetitive regions. We obtained the annotation of repetitive elements in the human genome from the UCSC RepeatMasker track then calculated the percentage of uniquely-mapped and multi-mapped reads within different classes of repeats as well as non-repeat regions (**Supplementary Figure 2.7**). In all three datasets, the percentage of multi-mapped reads was much higher in repetitive regions as compared to non-repeat regions, thus creating a challenge for peak calling within repetitive regions. For example, in the hnRNPC dataset,

37% of reads mapped to antisense Alu elements were multi-mapped, as compared to only 3% for reads mapped to non-repeat regions. Overall, only 8% of multi-mapped reads in the hnRNPC dataset were mapped to non-repeat regions, while 60% of multi-mapped reads were mapped to antisense Alu elements (**Supplementary Figure 2.7**) , consistent with a previous report on widespread hnRNPC binding within antisense Alu elements (19). When we ranked the repeat family by their total number of multi-mapped reads, Alu, L1 and simple repeat were consistently among the top families with the highest number of multi-mapped reads across the three datasets (**Supplementary Figure 2.7**).

We next assessed CLIP-seq/RIP-seq peak calling by CLAM. We adopted a commonly used permutation procedure for CLIP-seq or RIP-seq peak calling (19,26,27), and defined genomic loci with gene-specific FDR < 0.001 as peaks. We performed peak calling using: a) uniquely mapped reads only; and b) uniquely mapped reads plus CLAM assignments of multi-mapped reads. We classified peaks called from the above procedures into three distinct categories: “common peaks” that were called in both procedures, “rescued peaks” that were called only with multi-mapped reads incorporated, and “lost peaks” that were called using uniquely-mapped reads but not with multi-mapped reads incorporated.

Compared to a naïve read mapping and peak calling procedure using only uniquely mapped reads, a substantial number of rescued peaks were identified from all three datasets by CLAM (**Table 2.3**). While a certain number of peaks called by the naïve peak calling procedure were lost in the CLAM results, these lost peaks were much smaller in number as compared to rescued peaks called with incorporating multi-mapped reads (**Table 2.3**). For example, in the hnRNPC dataset, we had 26,594 rescued peaks on

average in the two samples, as compared to 6,898 lost peaks on average. Moreover, the majority of these lost peaks can be recovered from the CLAM results of multi-mapped reads simply by using a relaxed (higher) FDR cutoff (**Figure 2.2B**), suggesting that these peaks were lost due to random statistical fluctuations. For example, by relaxing the gene-specific FDR cutoff from 0.001 to 0.005 in the hnRNPC dataset, we were able to recover 94% of lost peaks. The reverse was not true – only 25% of rescued peaks could be identified using only uniquely mapped reads at this higher FDR cutoff, demonstrating the importance of modeling multi-mapped reads in CLAM. We observed the similar trend in the AGO2 and m^6A datasets, in which we could recover a much higher percentage of lost peaks by relaxing the FDR cutoff, but much less so on rescued peaks if using only uniquely mapped reads (**Figure 2.2B**). We also noted that in the AGO2 and m^6A datasets, a number of "lost peaks" were the only visible peaks in their respective genes when only uniquely mapped reads were considered, but could not pass the gene-specific FDR cutoff when multi-mapped reads in these genes were rescued by CLAM.

As expected, the rescued peaks were strongly enriched in repetitive elements as compared to common peaks across all three datasets (**Figure 2.2C**). For example, rescued peaks for hnRNPC were strongly enriched in antisense Alu elements, consistent with previous findings about hnRNPC binding sites within antisense Alu (19). We noted that 76% of rescued peaks for hnRNPC were located in antisense Alu elements, as compared to only 36% for common peaks. Similarly, Alu elements also showed a significant enrichment in rescued peaks for AGO2 and m^6A.

2.2.3 Rescued peaks for hnRNPC were associated with alternative splicing

Next, we assessed the functional relevance of rescued CLAM peaks, by correlating these peaks with relevant RNA regulatory features. We first analyzed the rescued CLAM peaks for hnRNPC, a splicing factor known to bind to poly-U tracts within the pre-mRNA to regulate alternative splicing. Using the Zagros *de novo* motif finder (32) for CLIP-seq data, we found a significantly enriched poly-U motif within both common peaks and rescued peaks (**Figure 2.3A**), suggesting that the rescued peaks have the same binding properties with hnRNPC as the common peaks. We then evaluated the potential functions of these rescued peaks, by investigating whether they were in the vicinity of alternative exons regulated by hnRNPC. To identify hnRNPC-dependent exons, we re-analyzed the RNA-seq data of hnRNPC knockdown in the same cell type (19) using rMATS (34), and ranked all exon-skipping cassette exons with sufficient RNA-seq coverage for differential splicing by their rMATS $\Delta\psi$ values (i.e. the difference of exon inclusion level between hnRNPC control and knockdown; see Materials and Methods). We defined an alternative exon as being associated with a CLIP-seq peak, if the peak was located within the exon body or in intronic regions within 250bp of the exon. We hypothesized that if rescued CLAM peaks indeed represented functional protein-RNA interaction sites, we would observe an enrichment of exons associated with rescued peaks among hnRNPC-dependent alternative exons identified by RNA-seq. Specifically, as hnRNPC is known to repress exon inclusion (19), its direct target exons should have higher splicing levels upon hnRNPC knockdown. To test this hypothesis, we performed a Kolmogorov–Smirnov statistical test similar to the gene set enrichment analysis (GSEA) algorithm (35), by comparing the rankings of exons with or without hnRNPC peaks within the $\Delta\psi$ ranked list

of hnRNPC-dependent exons. Indeed, exons with rescued peaks were strongly enriched towards the left side ($\Delta\psi < 0$) of the list (p-value<2.2e-16, **Figure 2.3B**, top panel), with the enrichment score peaked around RNA-seq $\Delta\psi$ of 0 then decreased gradually. We observed an almost identical trend for exons associated with common peaks (**Figure 2.3B**, bottom panel). Two representative examples of hnRNPC-dependent exons associated with rescued peaks were shown in **Figure 2.3C-D**. In **Figure 2.3C** (*DDIAS*), RNA-seq data revealed an exon with significantly elevated splicing upon hnRNPC knockdown, but no peak was observed in the vicinity of this exon using uniquely mapped CLIP-seq reads. However, this exon had a number of multi-mapped reads. These reads mapped to distinct sets of other genomic loci, while all of them mapped to this *DDIAS* exon. Therefore, CLAM rescued and assigned these multi-mapped reads to this exon, resulting in the identification of a strong hnRNPC peak. Another example was provided for *SNHG17*, in which CLAM discovered a strong hnRNPC peak within an hnRNPC-dependent alternative exon, while the coverage by uniquely mapped CLIP-seq reads was low and no peak can be identified within the exon (**Figure 2.3D**). Of note, in both genes the rescued peaks were located within a primate-specific Alu retrotransposon, indicating the creation of species-specific splicing regulatory sequences from repetitive elements.

2.2.4 Rescued peaks for AGO2 were associated with microRNA-mediated mRNA repression

Next we used CLAM to analyze a CLIP-seq dataset of AGO2 in the GM12878 lymphoblastoid cell line (LCL) (30). AGO2 belongs to the Argonaute (AGO) protein family and plays a critical role in RNA silencing including microRNA-mediated mRNA repression

(46). CLIP-seq analysis of AGO2 allows transcriptome-wide identification of microRNA binding sites (47). CLAM rescued >2,000 peaks from the AGO2 CLIP-seq data (**Table 2.2**), with over half of these rescued peaks located within repetitive elements (**Figure 2.2C**).

To assess if these rescued peaks represented functional microRNA target sites, we ran TargetScan (36) to predict the microRNA target sites within each CLIP-seq peak. We then selected two microRNAs (miR-21 and miR-107) for detailed analyses of the predicted TargetScan microRNA target sites. These two microRNAs were selected because they both were abundantly expressed in the GM12878 LCL cell line according to small RNA sequencing data, and global microarray data of mRNA expression following ectopic expression or inhibition of the microRNA were available in the published literature (see Methods for details). For each microRNA, we obtained three categories of genes: genes with common peaks containing microRNA target sites; genes with rescued peaks containing microRNA target sites; and background genes with no AGO2 CLIP-seq peaks. We then calculated the fold change of gene expression level upon ectopic expression or inhibition of the microRNA, then plotted the cumulative density function of the log2 fold change values for the three categories of genes (**Figure 2.4**). For miR-21, genes with commons peaks and rescued peaks both had a significant increase in expression levels as compared to background genes following microRNA inhibition (p-value<2.2e-16 and p-value<2.2e-16 respectively, Kolmogorov–Smirnov test), consistent with de-repression of target mRNA levels (**Figure 2.4A**). By contrast, for miR-107, genes with common peaks and rescued peaks both had a significant decrease in expression levels as compared to background genes following microRNA overexpression (p-value=4.8e-7 and p-

value<2.2e-16 respectively, Kolmogorov–Smirnov test), consistent with repression of target mRNA levels (**Figure 2.4B**). These data are characteristic of microRNA's effects on target genes (48), suggesting that the rescued AGO2 peaks provide functional target sites for microRNA-mediated mRNA repression.

2.2.5 Rescued peaks for m^{6}A were associated with mRNA stability control

To test CLAM on RIP-seq data, we applied it to our published RIP-seq data of N6-methyladenosine (m^{6}A) in the H1 human embryonic stem cells (ESCs) (31). The m^{6}A modification involving the addition of a methyl group to the N6 position of adenosine is a widespread reversible RNA modification in mammalian cells. RNA immunoprecipitation by m^{6}A specific antibody followed by sequencing is a popular strategy to identify m^{6}A sites across the transcriptome (49). CLAM rescued >3,500 peaks from the m^{6}A RIP-seq data. Following an established procedure to identify the consensus m^{6}A motif from m^{6}A RIP-seq data (31), we ranked common or rescued m^{6}A peaks by the ratio of normalized read counts in the m^{6}A RIP-seq data versus the RNA-seq data of the input control, then performed *de novo* motif discovery using HOMER (33) in the top 1000 common or rescued peaks. We identified a significant GGACU motif that matched the known consensus m^{6}A motif (**Figure 2.5A**). Consistent with the observation that Alu elements were enriched in the rescued m^{6}A peaks (**Figure 2.2C**), we identified the consensus RRACU m^{6}A motif in the antisense and sense sequences of Alu subfamilies (**Figure 2.5B-C**). To test if these rescued CLAM peaks contained functional m^{6}A sites, we correlated the CLAM sites of human ESCs to published data of mRNA half-life in human induced pluripotent stem cells (iPSCs) (38). As m^{6}A has a well-established role in

regulating mRNA degradation and stability (50), we previously observed that genes with functional m^{6}A sites had reduced m^{6}A half-life (31). We classified genes into three categories: genes with common m^{6}A peaks; genes with rescued m^{6}A peaks; and background genes without m^{6}A peaks. Genes with common or rescued m^{6}A peaks both had significantly lower mRNA half-life as compared to background genes (p-value<2.2e-16 and p-value=1.3e-12 respectively, Kolmogorov–Smirnov test; see **Figure 2.5D** and **Supplementary Figure 2.8**), suggesting that the rescued peaks contained functional m^{6}A sites. Furthermore, we observed significant enrichment of both common and rescued m^{6}A sites near the stop codon (**Figure 2.5E-F**), demonstrating the similar topological feature of common and rescued m^{6}A sites that matched the previously reported pattern (49). An example of a rescued m^{6}A site was shown in the 3'-UTR of *NME6*, in which a strong RIP-seq peak derived from an Alu retrotransposon was identified by CLAM combining uniquely-mapped and multi-mapped reads (**Figure 2.5G**).

2.2.6 CLAM analysis of ENCODE CLIP-seq data of 17 splicing factors

To demonstrate the broad applicability of CLAM, we analyzed 17 splicing factors (**Supplementary Table 2.4**) with matching eCLIP (enhanced CLIP) data and shRNA knockdown followed by RNA-seq data on the HepG2 cell line from the ENCODE consortium (**Figure 2.6**). The ENCODE investigators have systematically performed eCLIP experiments on a large number of RBPs in the HepG2 cell line (39), along with RNA-seq analysis following shRNA knockdown of individual RBPs. For each of the 17 splicing factors, CLAM rescued thousands to tens of thousands of peaks (**Supplementary Table 2.4**). 12 of the 17 splicing factors had known consensus motifs

defined previously using the RNAcompete technology (1). For these splicing factors, we calculated the enrichment p-values of known consensus motifs within common or rescued peaks using a *t*-statistic procedure (Methods). The rescued and common peaks exhibited highly similar patterns of consensus motif enrichment in general for all 12 splicing factors (**Figure 2.6A**), despite that the p-value calculation could sometimes be skewed for rescued peaks due to their high content of repetitive elements and biased sequence compositions (**Figure 2.2C**).

To assess the functional relevance of rescued CLAM sites for these 17 splicing factors, we intersected the common and rescued eCLIP peaks with splicing factor dependent alternatively spliced cassette exons, identified from RNA-seq data of the HepG2 cell line following shRNA knockdown of the splicing factor. For each exon, we defined three non-overlapping regions as the 250bp upstream intronic region, the exon body, and the 250bp downstream intronic region. We then tested if exons containing eCLIP peaks (common or rescued) in these regions were significantly enriched towards the top or bottom of the $\Delta\psi$ ranked list of splicing factor dependent exons using the GSEA algorithm (see Materials and Methods). As the number of common peaks was generally substantially larger than the number of rescued peaks across all splicing factors (**Supplementary Table 2.4**), in order to control for the difference in statistical power in calculating the enrichment p-value, we used a down-sampling strategy to randomly sample a subset of common peaks for the enrichment analysis. Our data show that across the 17 splicing factors, splicing factor dependent alternative exons generally had similar patterns of enrichment for rescued and common peaks, and the –log10 p-value of rescued peaks in approximately half of the tested regions was within the mean ± standard

deviation of that of common peaks generated by 20 rounds of down-sampling (marked with an asterisk next to the bar, see **Figure 2.6B**), suggesting that rescued and common peaks had similar functional effects on regulating alternative splicing. Two detailed examples were provided for hnRNPC and U2AF2 (**Figure 2.6C-D**). For hnRNPC, we observed significant enrichment of common and rescued peaks around hnRNPC-repressed exons in the ENCODE HepG2 cells (**Figure 2.6C**), consistent with the pattern observed in the HeLa cells (**Figure 2.3B**). For U2AF2, we observed significant enrichment of common and rescued peaks around U2AF2-enhanced exons (**Figure 2.6D**), consistent with the well-established role of U2AF2 as a positive regulator of exon splicing (51).

2.3 Discussion

We report CLAM, a new computational method and software program for CLIP-seq/RIP-seq peak calling incorporating multi-mapped reads. Multi-mapped reads constitute an appreciable fraction of reads in CLIP-seq/RIP-seq experiments (**Figure 2.2A**), but conventional analytic tools for CLIP-seq/RIP-seq data do not properly handle multi-mapped reads. In contrast to naïve approaches of discarding multi-mapped reads or distributing fractional counts of multi-mapped reads equally to all mapped loci (20), CLAM utilizes an EM framework to assign reads based on the local information of all mapped reads in the vicinity of multi-mapped reads. Our evaluation using a synthetic benchmark dataset demonstrates that the CLAM EM model outperforms alternative models (**Table 2.1**). It should be noted that while the EM algorithm is widely used for resolving multi-mapped RNA-seq reads (24,41,42), existing RNA-seq-based tools are not suitable for CLIP/RIP-seq data. Specifically, RNA-seq-based tools only consider reads mapped to

annotated transcript regions and ignore reads in intronic regions, where a large number of CLIP/RIP-seq peaks reside. By contrast, CLAM is designed to account for unique features of CLIP/RIP-seq data. For example, CLAM assigns multi-mapped reads and calls peaks in local windows that match the size of RBP footprints. Collectively, CLAM provides a comprehensive and rigorous computational solution for CLIP/RIP-seq peak calling utilizing multi-mapped reads, and its performance is supported by comprehensive analyses of diverse datasets.

To demonstrate the utility of CLAM, we applied it to a wide range of public CLIP-seq/RIP-seq datasets involving splicing factors, microRNAs, and m^6A RNA methylation. Consistently across all datasets, CLAM rescued the vast majority of multi-mapped reads in CLIP-seq/RIP-seq libraries, and identified a large number of novel peaks that would otherwise be missed using only uniquely mapped reads. These rescued peaks show expected patterns of consensus motif enrichment. Moreover, analyses of RNA regulatory features suggest that these rescued CLAM peaks are functional, as evidenced by association with alternative splicing (hnRNPC and other splicing factors in ENCODE), steady-state transcript abundance (AGO2), and mRNA half-life (m^6A).

An important application of CLAM is to comprehensively discover novel RNA regulatory sites originating from transposable elements in the genome. Extensive research in the past few decades have demonstrated that transposable elements, initially considered as “genomic parasites” or “junk DNAs”, play important roles in essentially all aspects of gene regulation from transcription to RNA processing to protein synthesis (52). At the RNA level, transposable elements can contribute functional elements for post-transcriptional gene regulation (53). The CLIP-seq/RIP-seq technologies in principle

should enable large-scale discoveries of RNA regulatory sites derived from transposable elements, but the repetitive nature of these sequences combined with the short length of CLIP-seq/RIP-seq reads create computational challenges for peak identification. CLAM provides a statistically rigorous approach to identify CLIP-seq/RIP-seq peaks in repetitive regions of the transcriptome. Across multiple datasets, a significantly higher fraction of "rescued peaks" identified by CLAM are derived from transposable elements, as compared to "common peaks" that are readily identifiable using only uniquely mapped reads (**Figure 2.2C**). Of note, we identified numerous protein-RNA interaction events and RNA modification sites derived from Alu elements. As Alu elements are primate-specific retrotransposons (54), these Alu derived RNA regulatory sites have the potential to re-wire lineage-specific post-transcriptional regulatory networks, thus contributing to transcriptome diversification during primate and human evolution. For example, m^6A RNA methylation has recently emerged as a key player in RNA metabolism (49). While our previous m^6A RIP-seq analysis of human and mouse embryonic stem cells indicated significant conservation of m^6A patterns, we also discovered species-specific m^6A sites in over a thousand genes (31). However, the molecular mechanism and evolutionary source for these species-specific m^6A sites were unknown. In this work, using CLAM we identified 3,218 Alu-derived m^6A sites in human genes, revealing the significant contribution of Alu elements to human-specific m^6A sites and potentially m^6A-associated regulatory effects.

A potentially highly valuable feature of CLIP-seq data is the presence of diagnostic signals in CLIP-seq reads (e.g. read truncations and base substitutions) that may allow single-nucleotide-resolution mapping of protein-RNA interaction and RNA modification

sites (10,23,55). For example, iCLIP was designed to have single nucleotide resolution through read truncation at the cross-linking sites (10). However, recent literature (56-58) as well as our analysis of the ENCODE data suggest that the robustness of the truncation signals in iCLIP/eCLIP data varies among datasets as well as among individual sites in a single experiment, and could depend on various experimental, technical, and biological factors. One important future direction for CLAM is to model CLIP-seq diagnostic signals in a rigorous data-driven, probabilistic framework to further improve read re-assignment and site identification for CLIP-seq data.

In summary, by modeling and analyzing multi-mapped reads, CLAM provides a more comprehensive solution for CLIP-seq/RIP-seq peak identification beyond commonly used existing methods that focus on uniquely mapped reads. The CLAM software and user manual can be downloaded from https://github.com/Xinglab/CLAM. With the widespread application of CLIP-seq/RIP-seq technologies as well as the rapid accumulation of datasets in the public domain (7), we expect CLAM will be of broad interest to biomedical researchers studying post-transcriptional gene regulation in diverse biological and disease processes.

2.4 Methods

2.4.1 CLIP-seq/RIP-seq read preprocessing and mapping

A typical CLIP-seq library contains 3' adaptors due to the short length of RBP-protected fragments; and 5' random barcodes to discriminate PCR duplicates. The 3' adaptors were first removed by fastx_clipper from fastx toolbox (21), available at

http://hannonlab.cshl.edu/fastx_toolkit/. Low quality reads were discarded by requiring the minimum quality threshold of 30 and at least 50% of bases in a read above this quality threshold. Next, PCR duplicates were removed by collapsing the reads with the same random barcodes and identical sequences. After removal of PCR duplicates, barcodes were removed and the reads were aligned by Novoalign (available at http://www.novocraft.com/) to the human genome and transcriptome, using the hg19 version of the human genome as the genomic index and Gencode V19 (http://www.gencodegenes.org/releases/19.html) as the transcriptome annotations (22). The set of optimized Novoalign parameters for CLIP-seq data (23) was used. Specifically, the alignment cost score '–t 85' controls the mismatches as: two substitutions, two consecutive deletions, or one substitution plus one deletion. The option '-l 25' requires at least 25 high-quality matches. For multi-mapped reads, reads that map to <100 genomic loci were retained for downstream analyses.

All mapped reads (uniquely + multi-mapped) were then merged into genomic regions. Two reads were merged if the distance between them was smaller than a threshold d. By default, we set $d = 50$ for CLIP-seq and $d = 100$ for RIP-seq to match the size of RBP footprint or RNA fragment.

2.4.2 Expectation-Maximization analysis of multi-mapped reads

Distinct genomic regions were connected through multi-mapped reads as a graph. The connected subgraphs (i.e. regions sharing multi-mapped reads) were extracted and subsequently converted to a compatibility matrix *Y* representing the mapping relationships between reads and genomic regions. Each genomic region corresponded

to a column and each read corresponded to a row of the compatibility matrix *Y*. For read *i* uniquely mapped to genomic region *k*, $y_{i,\cdot} = 0$ except for $y_{i,k} = 1$. For read *i* mapped to multiple genomic regions $\{k_p, \ldots, k_q\}$, $y_{i,k} = 1$, for $k \in \{k_p, \ldots, k_q\}$ and 0 otherwise. Our goal was to resolve the rows with multiple 1's in the matrix *Y* using an Expectation-Maximization framework (24).

In other words, our goal is to infer another indicator matrix *Z* to represent the true origins of mapped reads. As certain RBPs (e.g. AGO2) could have long footprints on mRNA transcripts due to multiple overlapping binding sites, the statistical model of CLAM considers that for a potential binding site, the probability that a multi-mapped read originates from this region depends on the reads mapped to a defined local window surrounding the binding site. Hence given the vector $\widehat{\Theta}$ representing the relative abundance of multiple mapped genomic regions among RBP-bound RNAs and the compatibility matrix *Y*, the latent variable $\hat{z}_{i,k}$ that represents the true origin of read *i* from region *k*, is computed by taking the expectation at (*t*+1)-th iteration as the E-step:

$$\hat{z}_{i,k}^{(t+1)} = E\left[z_{i,k} \middle| Y, \widehat{\Theta}^{(t)}, c\right] = Pr\left(z_{i,k} = 1 \middle| Y, \widehat{\Theta}^{(t)}, c\right) = \frac{y_{i,k} \cdot \hat{\theta}_{k,c_{i,k}}^{(t)}}{\sum_k y_{i,k} \cdot \hat{\theta}_{k,c_{i,k}}^{(t)}}$$

where $c_{i,k}$ is the center position of read *i* on region *k*, $\hat{\theta}_{k,c_{i,k}}^{(t)}$ is the relative abundance of multiple mapped genomic regions estimated at the locus $c_{i,k}$ on region *k* in the previous iteration. For the starting condition *t*=0, the EM model converges to the optimal point regardless of its initial values, since the objective function to be maximized is concave (25). For simplicity, $\widehat{\Theta}$ was initialized uniformly for all regions.

Next in the (*t*+1*)*-th iteration of the M-step, for any particular column y_k in *Y* corresponding to a specific genomic region, we estimate its relative abundance $\hat{\theta}_{k,j}^{(t+1)}$ locally at each position *j* among multiple mapped regions using the true origin $\hat{Z}^{(t+1)}$ within the (2*w*+1) window:

$$\hat{\theta}_{k,j}^{(t+1)} = \frac{\sum_i \hat{z}_{i,k}^{(t+1)} \cdot \mathbf{1}(j - w \le c_{i,k} \le j + w)}{N}$$

where *N* is the total number of reads in these regions sharing multi-mapped reads, *w* is the window size defining the local window, $c_{i,k}$ is the center position for read *i* on region *k*, $\hat{z}_{i,k}^{(t+1)}$ is the estimated true origin of read *i* from region *k*, and $\mathbf{1}(\cdot)$ is the indicator function. By default, we set *w*=50 for CLIP-seq data and *w*=100 for RIP-seq data to match the size of RBP footprint or RNA fragment.

The E-step and M-step are iterated until convergence.

2.4.3 Peak calling

Peak calling was performed on a gene-by-gene basis, in order to control for the expression variability among genes as in previous work (19,26,27). Briefly, CLAM was applied to genes with multi-mapped reads. For a given gene, the mapped reads could be divided into two sets: uniquely mapped reads with probability of origin $p = 1$, and multi-mapped reads with $p \in [0,1)$. We used a random permutation procedure to obtain the background read count distribution. Specifically, uniquely mapped reads were randomly assigned a location along the gene for 1000 times. For multi-mapped reads, a uniform random variable $u \in [0,1]$ was first drawn; if $u \le p$, this multi-mapped read was randomly

assigned a location in the same manner as uniquely mapped reads; otherwise this read was discarded in the current permutation. For position *j* with height h_j>0, p-value = $\frac{\sum_k \mathbf{1}(k \geq h_j) \cdot n_k}{\sum_k n_k}$, where n_k is the number of positions with peak height $k \in (1,2,\dots)$ in permutation derived null distribution, and $\mathbf{1}(\cdot)$ is the indicator function. For each gene, multiple testing was corrected by the Benjamini-Hochberg FDR procedure (28). Positions with gene-specific FDR<0.001 were called as significant loci, and peaks were called as the most significant loci within 50bp windows. If a peak was less than 50bp, the peak was extended symmetrically to 50bp. For downstream analyses, the common or rescued peaks in individual replicates were then merged by taking the union respectively.

2.4.4 Analysis of m^{6}A RIP-seq data

We employed a slightly different processing pipeline as well as parameters for the m^{6}A data, given the differences between RIP-seq (for m^{6}A) and CLIP-seq. We first mapped the human m^{6}A RIP-seq reads using STAR (29) v2.4.2 to the hg19 genome with the Gencode v19 transcript annotations (22), retaining reads mapped to <100 loci. Then we ran CLAM with parameters: maximum distance for collapsing reads *d*=100; local window size *w*=100; p-value correction using the more stringent Bonferroni correction given the lower signal-to-noise ratio of RIP-seq; and peaks were called as the most significant loci within 500bp windows and extended to 100bp symmetrically.

2.4.5 Analysis of RNA motif and regulatory features

We applied CLAM to publicly available CLIP-seq/RIP-seq datasets listed in Table 1. We analyzed two iCLIP datasets, hnRNPC iCLIP on the HeLa cell line from Zarnack et al. (19), and AGO2 iCLIP on the GM12878 lymphoblastoid cell lines (LCL) from Wan et al. (30). We also analyzed one m^{6}A RIP-seq dataset on the H1 human embryonic stem cell line from Batista et al. (31). For each dataset, we analyzed RNA motif and regulatory features based on the known properties of the RBP or RNA modification. Annotations of repetitive elements for the hg19 human genome were downloaded from the UCSC RepeatMasker track, available at the UCSC table browser.

Motif finding for hnRNPC peaks was performed using Zagros (32), a specialized *de novo* motif finder for CLIP-seq data. For m^{6}A sites, motif finding was performed using HOMER (33) as in our previous m^{6}A work (31) on the top 1,000 peaks ranked by enrichment ratio over input control.

To assess the functional impact of hnRNPC CLAM sites on hnRNPC-dependent alternative splicing, hnRNPC shRNA knockdown followed by RNA-seq dataset in the same HeLa cell line (19) was analyzed by rMATS (34) (version 3.2.5) to detect differential alternative splicing events. The alternative exons were filtered by read counts (inclusion counts + skipping counts≥20) and then ranked by $\Delta\psi$ values (control – knockdown) from the most hnRNPC repressed exons ($\Delta\psi$=-1) to the most hnRNPC enhanced exons ($\Delta\psi = 1$). Each exon was extended symmetrically by 250bp on both sides to include the proximal intron regions. We applied a GSEA (Gene Set Enrichment Analysis)-like analysis (35) to test if exons with CLAM sites overlapping with the extended exon regions were enriched towards the top or the bottom of the $\Delta\psi$ ranked hnRNPC-dependent

differential alternative splicing events. Specifically, Enrichment Score (ES) was calculated as described previously (35) on the exons with CLIP-seq peaks as hits in this ranked exon list, and Kolmogorov–Smirnov test (K-S test) was performed to test for statistical significance.

To assess the effect of AGO2 CLAM sites on microRNA-mediated mRNA repression, microarray gene expression data of human cell lines upon ectopic expression or inhibition of two microRNAs were downloaded from GEO with accession number: GSE37213 (miR-21, T-lymphocytes) and GSE42823 (miR-107, H4 glioneuronal cells). We selected these two microRNAs because they were both abundantly expressed in the GM12878 cell line profiled by AGO2 iCLIP, based on small RNA-seq profiling data of microRNA abundance in the original study by Wan et al. (30). For each AGO2 peak, we predicted the targets of these two microRNAs using TargetScan (36) (http://www.targetscan.org/vert_71/). AGO2 target genes were then separated into two categories based on whether they had common or rescued peaks. Background genes were chosen as genes without any AGO2 peaks. Affymetrix microarray probesets were matched to corresponding transcripts using BiomaRt (37).

To assess the influence of m^6A modification on mRNA half-life, we used transcript half-life time measured in iPS cells as in our previous m^6A work (38). Genes were classified similarly as in the AGO2 analysis. We performed a meta-gene analysis to obtain the m^6A peak distributions in 5'-UTR, CDS, and 3'-UTR by binning the corresponding transcript region into 10 equal-sized bins then counting in each bin the frequency of top 1,000 common or rescued m^6A peaks respectively.

2.4.6 Analyses of ENCODE CLIP-seq and RNA-seq data on 17 splicing factors

We applied CLAM to 17 splicing factors with matching CLIP-seq (eCLIP) and shRNA knockdown followed by RNA-seq datasets in the HepG2 cell line from the ENCODE project. We followed the ENCODE SOP pipeline to remove adapters. We developed an in-house script for collapsing PCR duplicates based on the ENCODE SOP but preserved multi-mapped reads. Since eCLIP employed paired-end sequencing, only the second mate was extracted and fed into CLAM after mapping, following the same strategy adopted by the ENCODE consortium (39). CLAM was run using the same parameter set as in our analyses of the hnRNPC and AGO2 iCLIP data.

The CLAM sites for each splicing factor were validated in two aspects: enrichment of consensus motif (if available) and enrichment of splicing factor-dependent alternative exons upon shRNA knockdown of the splicing factor. Known consensus motifs of 12 splicing factors were retrieved from the RNAcompete database (1). Motif enrichment analysis was performed using a Z-score method as described previously (40), with minor modifications. Specifically, given a motif regular expression and a set of *n* CLIP-seq peaks, we first computed the number of peaks (sequences), denoted by X, containing the motif. Then we estimated the background frequency *p* of the given motif in a large collection of random genomic sequences of the same length as CLIP-seq peaks. The expected motif occurrence in the CLIP-seq peaks was hence $n \cdot p$, with the variance being $n \cdot p \cdot (1-p)$. We applied the Z-transformation as $Z = \frac{X-np}{\sqrt{np(1-p)}}$. To account for the over-dispersion in the above Z-score, we computed the Z-scores for an additional *m*=1000 randomers of the same length as the given motif, and estimated the sample standard deviation *s* of the

randomer Z-scores. Hence the final *t*-statistic is $t = \frac{Z}{s}$ with degree of freedom *m*-1, and p-value was given by Student's *t*-distribution.

RNA-seq data of splicing factor knockdown was publicly available in the ENCODE data portal and we used our rMATS pipeline (34) (version 3.2.5) to quantify the exon inclusion level (ψ) of cassette exon skipping events. We applied a read count filter to remove exon skipping events with < 20 combined (inclusion plus skipping) reads. As there were many more common peaks than rescued peaks, to account for the difference in statistical power in calculating the GSEA-like (35) K-S enrichment statistics, for each splicing factor we down-sampled the common peaks to the same number of rescued peaks and repeated the down-sampling procedure 20 times. For each exon, three non-overlapping regions were considered: upstream 250bp flanking intron, exon body, and downstream 250bp flanking intron. The rescued peaks and each set of down-sampled common peaks were tested for enrichment in each of these three regions separately, based on a ranked list of splicing factor dependent exons ranked by difference in exon inclusion levels ($\Delta\psi$) between control and knockdown, following the same procedure for calculating the K-S statistic as described above for the hnRNPC iCLIP data.

2.4.7 Code availability

The CLAM software and user manual can be downloaded from https://github.com/Xinglab/CLAM. All datasets used in this paper are publicly available in public repositories, i.e. SRA and ArrayExpress, with accession numbers listed in **Table 2.2**.

Acknowledgements

This study is supported by National Institutes of Health grant (R01GM088342 to Y.X.). Y.X. is supported by an Alfred Sloan Research Fellowship. We thank Emad Bahrami-Samani and Wang Xi for discussions and technical assistance. We thank the ENCODE Consortium and the ENCODE production laboratories especially Dr. Gene Yeo and Dr. Brenton Graveley for generating the eCLIP and RNA-seq data.

2.5 References

1. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172-177.

2. Gerstberger, S., Hafner, M. and Tuschl, T. (2014) A census of human RNA-binding proteins. *Nature Reviews Genetics*, **15**, 829-845.

3. Glisovic, T., Bachorik, J.L., Yong, J. and Dreyfuss, G. (2008) RNA-binding proteins and post-transcriptional gene regulation. *Febs Letters*, **582**, 1977-1986.

4. Fu, X.D. and Ares, M. (2014) Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, **15**, 689-701.

5. Dittmar, K.A., Jiang, P., Park, J.W., Amirikian, K., Wan, J., Shen, S., Xing, Y. and Carstens, R.P. (2012) Genome-wide determination of a broad ESRP-regulated

posttranscriptional network by high-throughput sequencing. *Mol Cell Biol*, **32**, 1468-1482.

6. Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A. and Burge, C.B. (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell*, **54**, 887-900.

7. Yang, Y.C.T., Di, C., Hu, B.Q., Zhou, M.F., Liu, Y.F., Song, N.X., Li, Y., Umetsu, J. and Lu, Z.J. (2015) CLIPdb: a CLIP-seq database for protein-RNA interactions. *Bmc Genomics*, **16**.

8. Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464-469.

9. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr., M., Jungkamp, A.C., Munschauer, M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129-141.

10. Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M. and Ule, J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, **17**, 909-915.

11. Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M. and Lee, J.T. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell*, **40**, 939-953.

12. Wang, T., Xiao, G., Chu, Y., Zhang, M.Q., Corey, D.R. and Xie, Y. (2015) Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res*, **43**, 5263-5274.

13. Bahrami-Samani, E., Vo, D.T., de Araujo, P.R., Vogel, C., Smith, A.D., Penalva, L.O. and Uren, P.J. (2015) Computational challenges, tools, and resources for analyzing co- and post-transcriptional events in high throughput. *Wiley Interdiscip Rev RNA*, **6**, 291-310.

14. Linder, B., Grozhik, A.V., Olarerin-George, A.O., Meydan, C., Mason, C.E. and Jaffrey, S.R. (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nature methods*, **12**, 767-772.

15. Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N. and Rechavi, G. (2013) Transcriptome-wide mapping of N6-methyladenosine by m6A-seq based on immunocapturing and massively parallel sequencing. *Nature protocols*, **8**, 176-189.

16. de Koning, A.J., Gu, W., Castoe, T.A., Batzer, M.A. and Pollock, D.D. (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet*, **7**, e1002384.

17. Gong, C. and Maquat, L.E. (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, **470**, 284-288.

18. Nishikura, K. (2016) A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol*, **17**, 83-96.

19. Zarnack, K., Konig, J., Tajnik, M., Martincorena, I., Eustermann, S., Stevant, I., Reyes, A., Anders, S., Luscombe, N.M. and Ule, J. (2013) Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, **152**, 453-466.

20. Kelley, D.R., Hendrickson, D.G., Tenen, D. and Rinn, J.L. (2014) Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome biology*, **15**, 1-16.

21. Gordon, A. and Hannon, G.J. (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools. *Unpublished*.

22. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, **22**, 1760-1774.

23. Zhang, C. and Darnell, R.B. (2011) Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol*, **29**, 607-614.

24. Xing, Y., Yu, T., Wu, Y.N., Roy, M., Kim, J. and Lee, C. (2006) An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res*, **34**, 3150-3160.

25. Pachter, L. (2011) Models for transcript quantification from RNA-Seq. *arXiv preprint arXiv:1104.3889*.

26. Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.S., Zhang, C., Yeo, G., Black, D.L., Sun, H. *et al.* (2009) Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*, **36**, 996-1006.

27. Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D. and Gage, F.H. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*, **16**, 130-137.

28. Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289-300.

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

30. Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706-709.

31. Batista, P.J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., Bouley, D.M., Lujan, E., Haddad, B., Daneshvar, K. *et al.* (2014) m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell*, **15**, 707-719.

32. Bahrami-Samani, E., Penalva, L.O., Smith, A.D. and Uren, P.J. (2015) Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Res*, **43**, 95-103.

33. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, **38**, 576-589.

34. Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q. and Xing, Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*, **111**, E5593-5601.

35. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.

36. Agarwal, V., Bell, G.W., Nam, J.-W. and Bartel, D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.

37. Durinck, S., Spellman, P.T., Birney, E. and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, **4**, 1184-1191.

38. Neff, A.T., Lee, J.Y., Wilusz, J., Tian, B. and Wilusz, C.J. (2012) Global analysis reveals multiple pathways for unique regulation of mRNA decay in induced pluripotent stem cells. *Genome research*, **22**, 1457-1467.

39. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K. *et al.* (2016)

Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*, **13**, 508-514.

40. Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W., Ares, M. and Fu, X.-D. (2013) Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol Cell*, **50**, 223-235.

41. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

42. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511-515.

43. Chung, D., Kuan, P.F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E.H., Dewey, C. and Keles, S. (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput Biol*, **7**, e1002111.

44. Wang, J., Huda, A., Lunyak, V.V. and Jordan, I.K. (2010) A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics*, **26**, 2501-2508.

45. Boudreau, R.L., Jiang, P., Gilmore, B.L., Spengler, R.M., Tirabassi, R., Nelson, J.A., Ross, C.A., Xing, Y. and Davidson, B.L. (2014) Transcriptome-wide discovery of microRNA binding sites in human brain. *Neuron*, **81**, 294-305.

46. Hutvagner, G. and Simard, M.J. (2008) Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol*, **9**, 22-32.

47. Chi, S.W., Zang, J.B., Mele, A. and Darnell, R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479-486.

48. Guo, H., Ingolia, N.T., Weissman, J.S. and Bartel, D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835-840.

49. Fu, Y., Dominissini, D., Rechavi, G. and He, C. (2014) Gene expression regulation mediated through reversible m6A RNA methylation. *Nature Reviews Genetics*, **15**, 293-306.

50. Wang, X. and He, C. (2014) Reading RNA methylation codes through methyl-specific binding proteins. *RNA biology*, **11**, 669-672.

51. Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., Li, H. *et al.* (2014) Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nat Struct Mol Biol*, **21**, 997-1005.

52. Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*, **9**, 397-405.

53. Elbarbary, R.A., Lucas, B.A. and Maquat, L.E. (2016) Retrotransposons as regulators of gene expression. *Science*, **351**, aac7247.

54. Hasler, J. and Strub, K. (2006) Alu elements as regulators of gene expression. *Nucleic Acids Research*, **34**, 5491-5497.

55. Konig, J., Zarnack, K., Luscombe, N.M. and Ule, J. (2012) Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet*, **13**, 77-83.

56. Haberman, N., Huppertz, I., Attig, J., Konig, J., Wang, Z., Hauer, C., Hentze, M.W., Kulozik, A.E., Le Hir, H., Curk, T. *et al.* (2017) Insights into the design and interpretation of iCLIP experiments. *Genome Biology*, **18**.

57. Hauer, C., Curk, T., Anders, S., Schwarzl, T., Alleaume, A.M., Sieber, J., Hollerer, I., Bhuvanagiri, M., Huber, W., Hentze, M.W. *et al.* (2015) Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. *Nature Communications*, **6**.

58. Sugimoto, Y., Konig, J., Hussain, S., Zupan, B., Curk, T., Frye, M. and Ule, J. (2012) Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol*, **13**, R67.

2.6 Figures

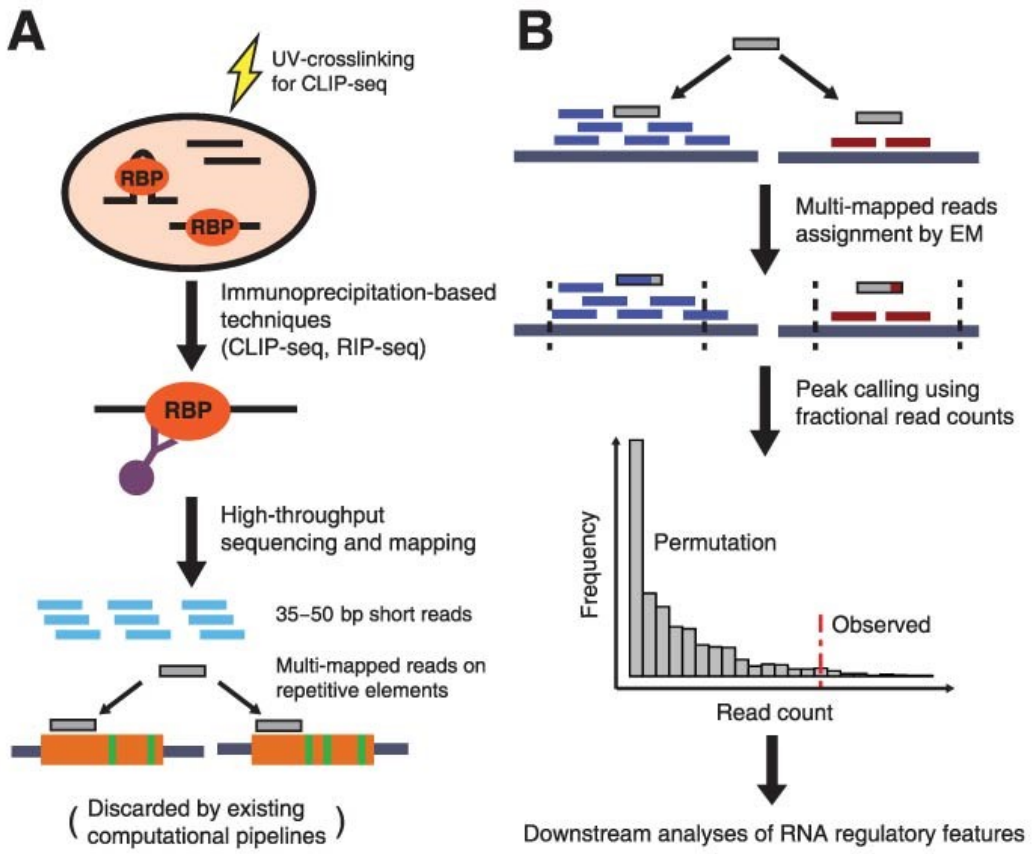


Figure 2.1 Motivation and schematic overview of CLAM.

(a) In immunoprecipitation based techniques for analyzing RBP-RNA interactions (CLIP-seq, RIP-seq), RNA associated with the target RBP is subject to RNase lysis and fragmentation after the RBP-RNA complex is immunoprecipitated by specific antibody, followed by high-throughput sequencing to generate short reads typically ranging between 35bp to 50bp. An appreciable fraction of reads, such as those originated from repetitive element derived RBP-RNA interaction sites, are mapped to multiple genomic regions and subsequently discarded by conventional data analysis pipelines. Shown here is a read mapped to two genomic copies of a repetitive element (orange boxes), which have identical sequences where the read is aligned but have mutations elsewhere

between these two copies (green vertical lines). **(b)** CLAM identifies a set of genomic regions sharing multi-mapped reads. It then uses an Expectation-Maximization algorithm to rescue multi-mapped reads and assign them to specific genomic regions, followed by a permutation-based procedure for peak calling with gene-specific FDR control. The rescued peaks are then assessed via downstream analyses of RNA regulatory features, including enrichment of consensus motifs and evaluations of RBP-specific regulatory features.

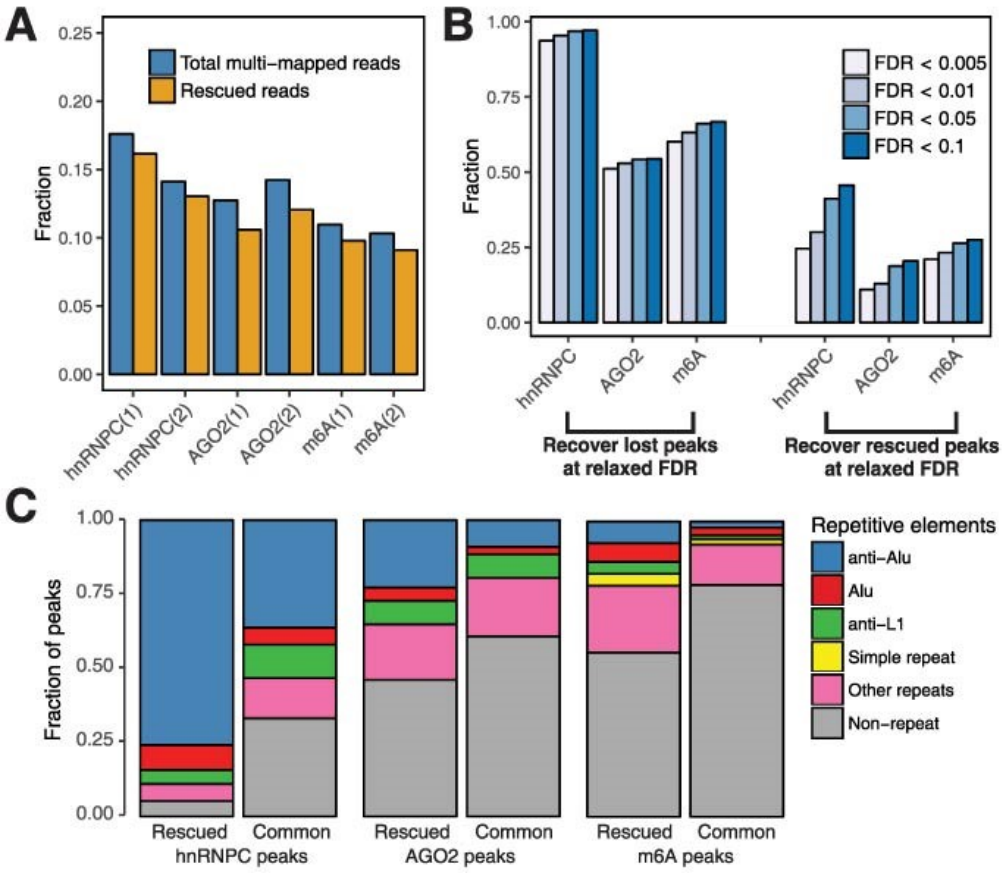


Figure 2.2 Summary statistics of CLAM results on three CLIP-seq/RIP-seq datasets.

(a) Percentage of multi-mapped reads (blue) and percentage of multi-mapped reads rescued by CLAM (orange) among all mapped reads in analyzed datasets. **(b)** Sensitivity analysis at various FDR thresholds. The majority of lost peaks can be recovered using the combination of uniquely and multi-mapped reads at higher (more relaxed) FDR thresholds (bar graphs on the left), while a significantly smaller fraction of rescued peaks can be identified using only uniquely mapped reads at higher FDR thresholds (bar graphs on the right). **(c)** Fraction of rescued and common peaks derived from various types of repetitive elements. A significantly higher fraction of rescued peaks are derived from repetitive elements across all three datasets.

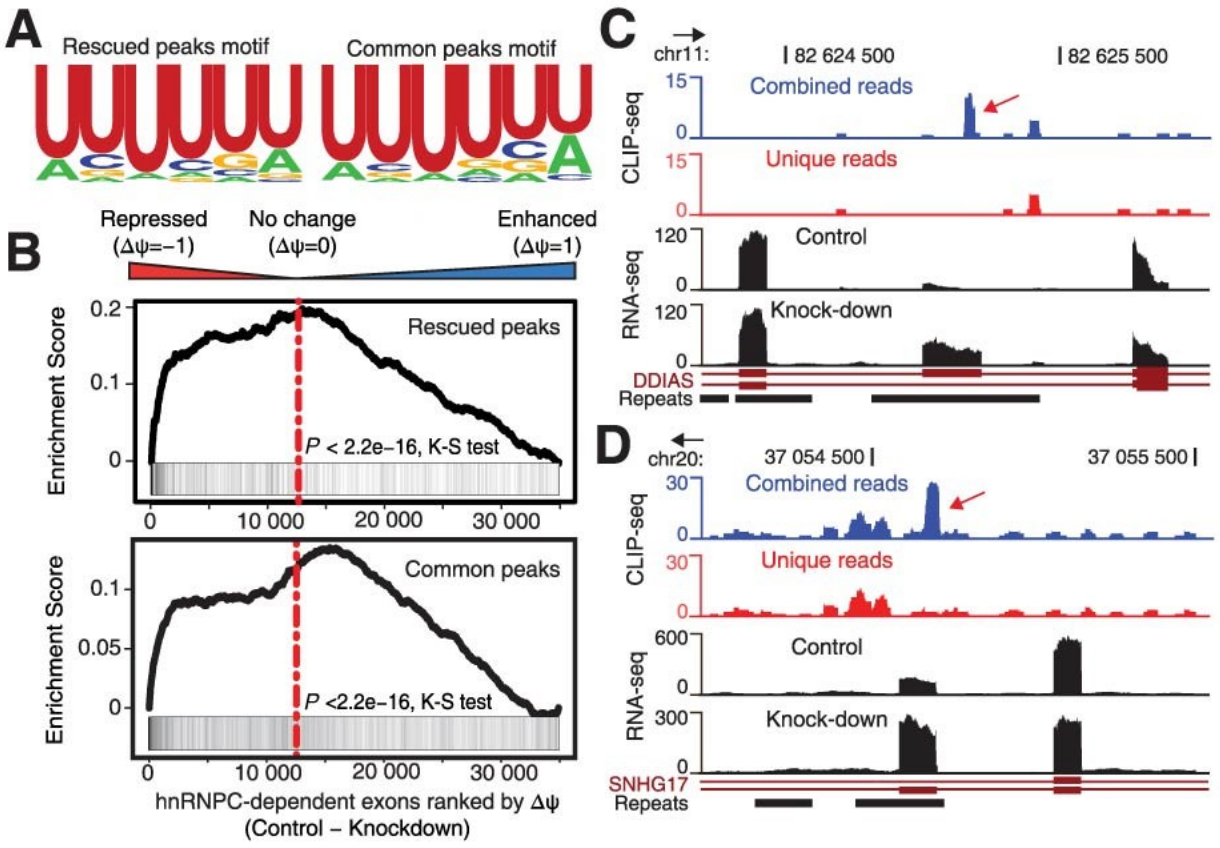


Figure 2.3 Functional evaluation of CLAM on the hnRNPC CLIP-seq data.

(a) Identification of the known consensus hnRNPC motif by de novo motif discovery in rescued and common hnRNPC peaks. **(b)** Enrichment analysis of hnRNPC dependent alternative exons for rescued and common hnRNPC peaks. X-axis represents alternative exons ranked by rMATS Δψ values (the difference in exon inclusion levels between control and knockdown). Y-axis is the Enrichment Score (ES) calculated via the Kolmogorov-Smirnov statistic. Both rescued and common hnRNPC peaks are strongly enriched for hnRNPC-repressed alternative exons. **(c)** Example of a rescued hnRNPC peak in DDIAS. **(d)** Example of a rescued hnRNPC peak in SNHG17.

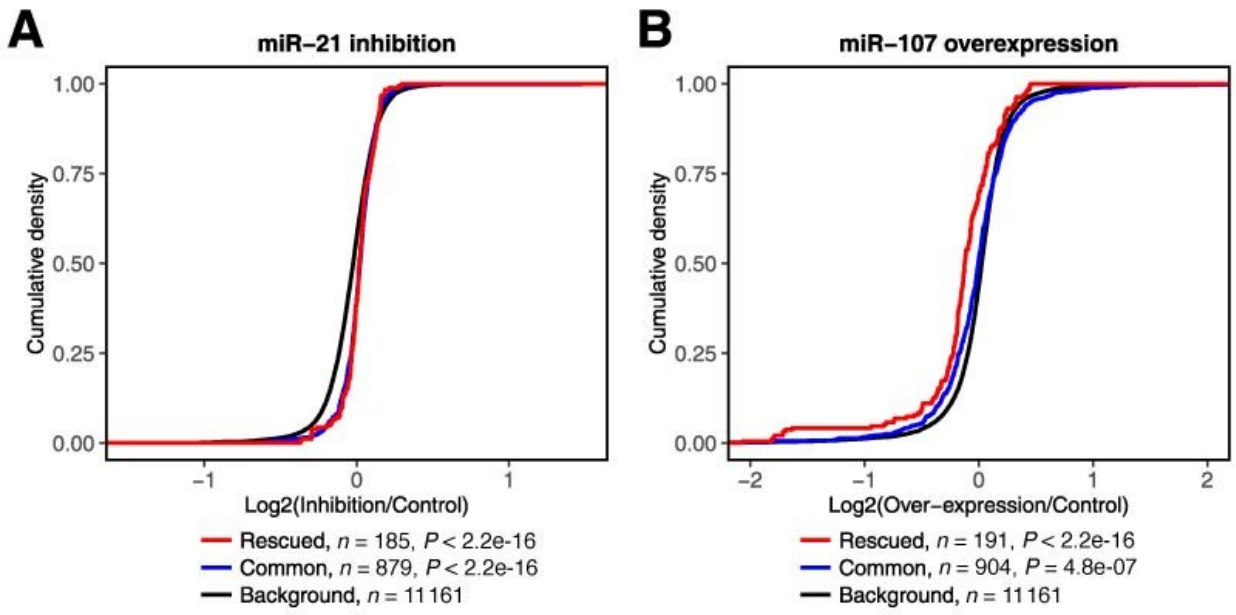


Figure 2.4 Functional evaluation of CLAM on the AGO2 CLIP-seq data.

For each microRNA, three classes of genes are compiled: genes with common peaks containing microRNA target sites (Common, blue); genes with rescued peaks containing microRNA target sites (Rescued, red); and background genes without AGO2 CLIP-seq peaks (Background, black). Cumulative density function is plotted for the log2 gene expression fold change upon **(a)** inhibition of miR-21 or **(b)** ectopic expression of miR-107. For both microRNAs, rescued and common target genes show the same significant shift in cumulative density function as compared to background genes.

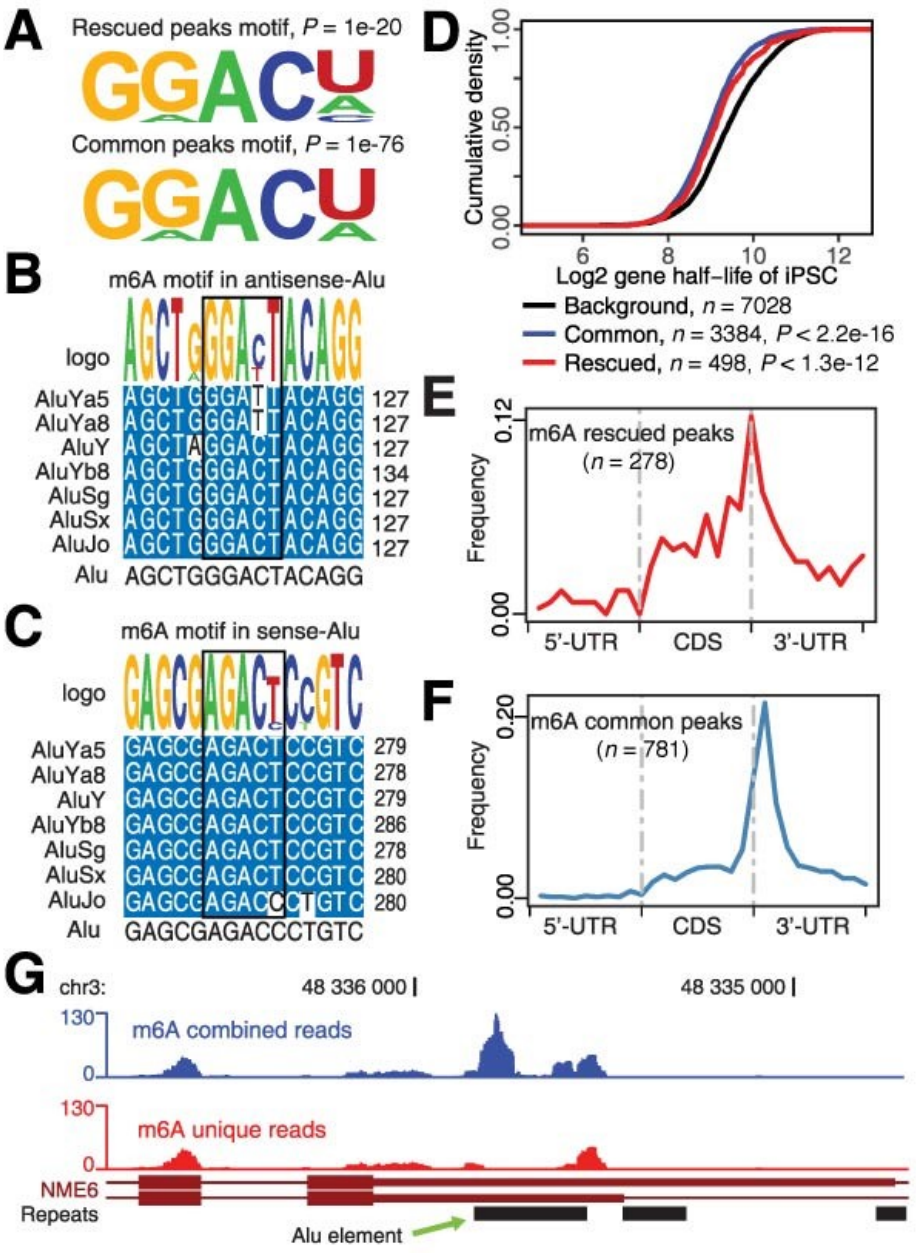


Figure 2.5 Functional evaluation of CLAM on the m^{6}A RIP-seq data.

(a) Identification of the known consensus m^{6}A motif by de novo motif discovery in rescued and common m^{6}A peaks. The conserved m6A RRACU motif in **(b)** anti-sense and **(c)** sense sequences of major Alu subfamilies. **(d)** Cumulative density function of mRNA half-life in iPSCs. Both genes with common and rescued m^{6}A peaks have significantly lower mRNA half-life as compared to background genes without m^{6}A peaks. Topological

distribution of **(e)** rescued and **(f)** common m^6A peaks across the 5'-UTR, CDS, and 3'-UTR of protein-coding genes. **(g)** Example of a rescued Alu-derived m^6A peak in the 3'-UTR of NME6.

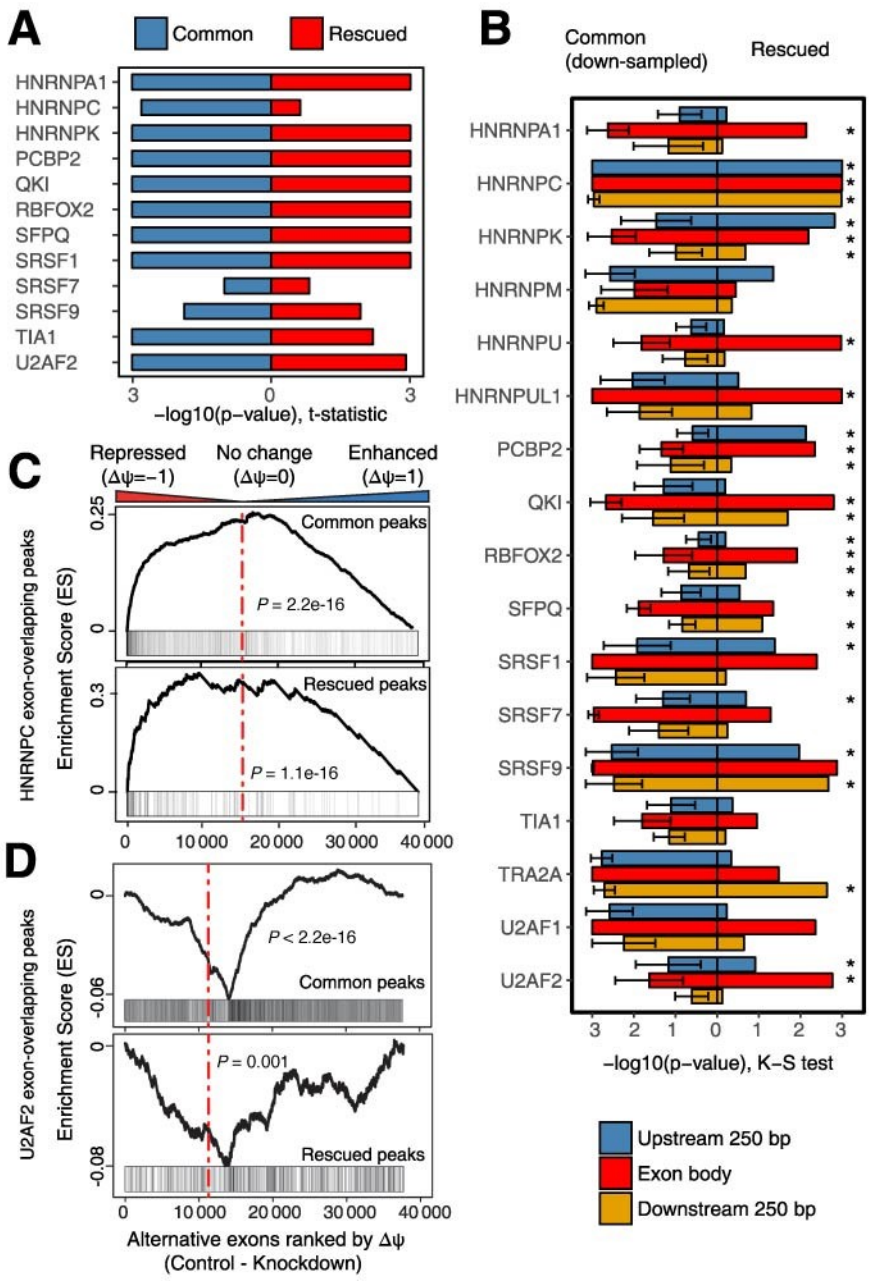
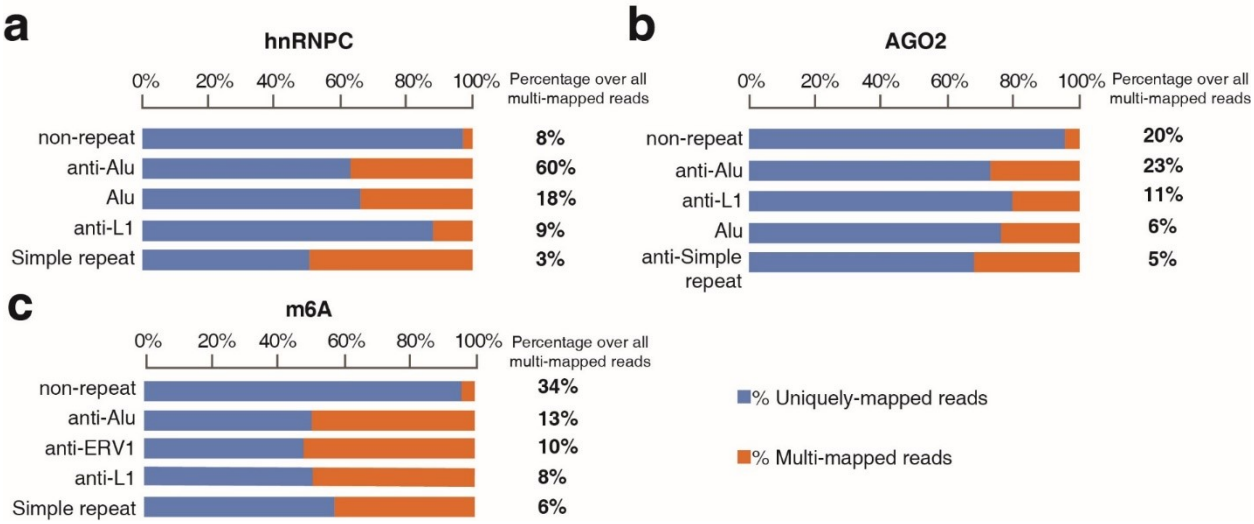


Figure 2.6 CLAM analysis of 17 splicing factors with ENCODE eCLIP data and matching RNA-seq data following splicing factor knockdown in the HepG2 cell line.

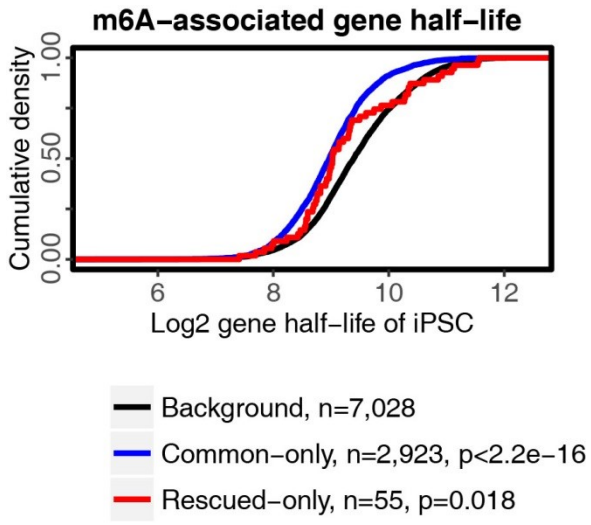
In visualizing the -log10(p-value), we added a pseudo-count of 1e-3 to all p-values to truncate the –log10(p-value) at an upper limit of 3, while the same pattern was observed

for pseudo-count of 1e-4 and 1e-5. **(a)** Negative log10 enrichment p-values of known splicing factor motifs within common (blue) and rescued (red) peaks. The frequency of motif occurrences were compared to randomly sampled genomic sequences and Student's t-distribution was fitted to measure the statistical significance of enrichment. **(b)** Barplots of negative log10 p-values of GSEA test on the enrichment of splicing factor dependent alternative exons for common or rescued peaks within the upstream 250bp intronic region (blue), the exon body (red), and the downstream 250bp intronic region (orange). For common peaks, the -log10 p-value of enrichment was calculated as the average from 5 random iterations of down-sampling to the same number of rescued peaks. **(c)** Enrichment analysis of hnRNPC dependent exons for common and rescued hnRNPC exon-overlapping peaks in the ENCODE HepG2 data. Both common and rescued hnRNPC peaks are strongly enriched for hnRNPC-repressed exons. **(d)** Enrichment analysis of U2AF2 dependent exons for common and rescued U2AF2 exon-overlapping peaks. Both common and rescued peaks are strongly enriched for U2AF2-enhanced exons in the ENCODE HepG2 data.



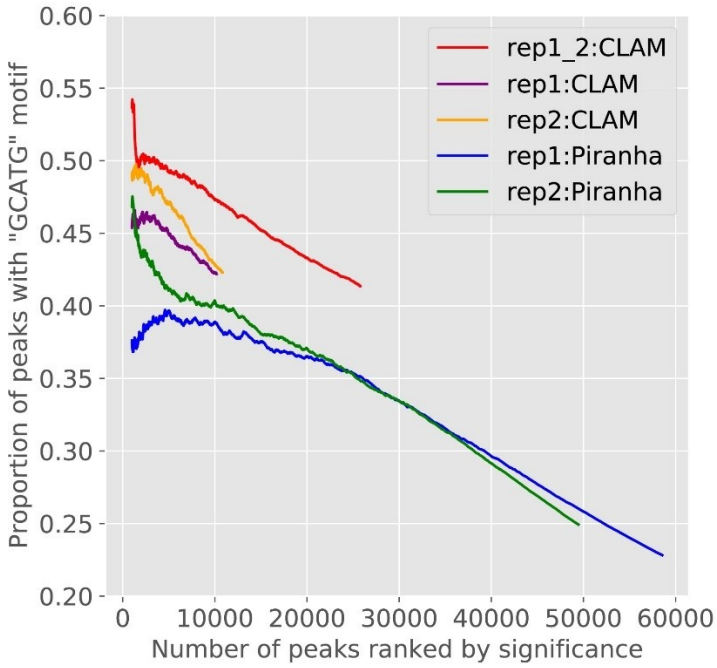
Supplementary Figure 2.7 Multi-mapped reads are enriched in repetitive elements.

Displayed are the most abundant repetitive element families with reads in: **(a)** hnRNPC; **(b)** AGO2; **(c)** m^{6}A dataset. Compared to non-repeat background regions, all of these repetitive element families are enriched for multi-mapped reads. “Percentage over all multi-mapped reads” indicates the percentage of multi-mapped reads derived from specific repetitive element families or non-repeat regions among all multi-mapped reads in the library.



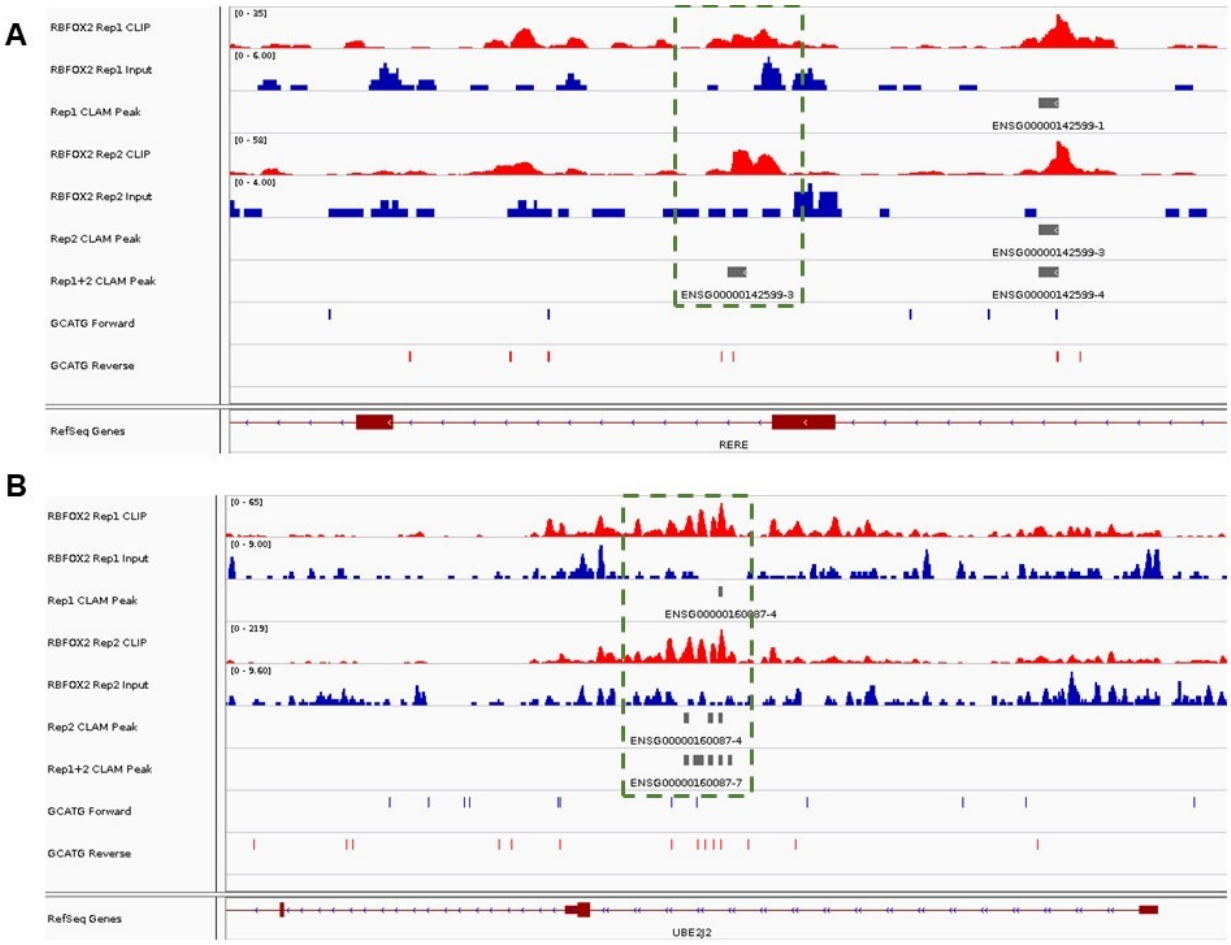
Supplementary Figure 2.8 Cumulative density function of mRNA half-life in iPSCs for different groups of genes.

Both genes with common but no rescued m^{6}A peaks (“common-only”, blue), and genes with rescued but no common m^{6}A peaks (“rescued-only”, red) have significantly lower mRNA half-life as compared to background genes without m^{6}A peaks.



Supplementary Figure 2.9 Benchmarking the performance of different CLAM run modes and a baseline method.

CLAM and Piranha were run on the eCLIP data of RBFOX2 in HEK293T cell line (accession ID: GSE77629). The peaks were ranked by significance reported by peak-callers, then the proportions of peaks with GCATG motif were plotted.



Supplementary Figure 2.10 Examples of CLAM peaks called by modelling multiple replicates.

Two examples of RBFOX2 eCLIP peaks called by modelling multiple replicates were shown **(a)** in RERE and **(b)** UBE2J2 genes. Dashed boxes highlighted the peaks called by aggregating weak signals from two replicates.

2.7 Tables

Table 2.1 Performance of CLAM and two alternative models on a synthetic benchmark dataset.

Model	AUROC	AUPR	Positive loci weight (median, mean)	Negative loci weight (median, mean)
CLAM	0.88	0.79	0.62, 0.65	0.02, 0.15
One-iter	0.88	0.78	0.50, 0.54	0.13, 0.20
Uniform	0.75	0.48	0.50, 0.42	0.20, 0.25

Table 2.2 Three representative datasets analyzed by CLAM.

Dataset	Predominant binding region	Motif	Technology	Cell line	Accession ID
hnRNPC	intronic	poly-U	iCLIP	HeLa	E-MTAB-1371
AGO2	3'-UTR	microRNA seeds	iCLIP	LCL	GSE50676
m^6A	3'-UTR	RRACU	RIP	H1-ESC	GSE52600

Table 2.3 Summary of CLAM peak calling on the hnRNPC, AGO2 and m^{6}A datasets.

Dataset	Replicate	Rescued	Common	Lost
hnRNPC	1	24,976	99,890	6,027
	2	28,211	133,708	7,769
AGO2	1	2,169	32,494	546
	2	2,243	29,774	536
m^{6}A	1	3,598	36,000	1,790
	2	3,702	39,153	2,151

Supplementary Table 2.4 Summary of CLAM peak calling and motif scores on ENCODE eCLIP data of 17 splicing factors in the HepG2 cell line.

RBP	Motif	Rescued peaks			Common peaks		
		No. of peaks	Motif t-statistic	Motif p-value	No. of peaks	Motif t-statistic	Motif p-value
HNRNPA1	[AGT]TAGGG[AT]	14,106	5.09	0.000	135,877	10.10	0.000
HNRNPC	[ACT]TTTTT[GT]	55,924	0.72	0.236	291,065	3.25	0.001
HNRNPK	CCA[AT][AC]CC	8,457	9.78	0.000	77,416	10.64	0.000
HNRNPM	NA	23,599	NA	NA	142,135	NA	NA
HNRNPU	NA	9,836	NA	NA	72,094	NA	NA
HNRNPUL1	NA	6,866	NA	NA	38,257	NA	NA
PCBP2	CC[CT][CT]CC[ACT]	7,769	15.66	0.000	79,038	20.96	0.000
QKI	ACTAAC[ACG]	5,715	18.49	0.000	66,022	48.17	0.000
RBFOX2	TGCATG	7,139	10.27	0.000	79,327	12.77	0.000
SFPQ	[GT]T[AG][AG]T[GT][GT]	4,794	5.57	0.000	42,723	9.83	0.000
SRSF1	GG[AG]GGA[ACG]	10,575	7.38	0.000	131,502	8.45	0.000
SRSF7	ACGACG	4,889	1.04	0.150	29,124	1.30	0.097
SRSF9	A[GT]GA[ACG][AC][AG]	8,112	2.30	0.011	64,304	2.25	0.012
TIA1	[AT]TTTTT[CGT]	9,504	2.55	0.005	83,158	5.80	0.000
TRA2A	NA	2,618	NA	NA	14,951	NA	NA
U2AF1	NA	10,455	NA	NA	107,983	NA	NA
U2AF2	TTTTT[CT]C	8,602	3.49	0.000	107,890	12.79	0.000

2.8 Appendix

2.8.1 Methodology updates since the publication

There have been a few major updates to the model and processing modules in the CLAM software since its initial release of v1.0. These changes are detailed below.

In release v1.1, we first introduced a preprocessing module to CLAM. The preprocessing module is designed to prepare the raw inputs (e.g. BAM or fastq reads) for the downstream CLAM realigning, peak-calling analyses and visualizations. A core function in the preprocessing steps is a read tagging function to account for the different technology biases from CLIP experiment variants. The read tagging function returns a single genomic locus to represent the crosslinking sites for a given CLIP/RIP-seq read. For iCLIP/eCLIP reads, we implemented the read starting site to represent the truncation events. For MeRIP-seq reads, users could choose between the median site of the read or extend certain base pairs from the read start site to account for fragmentation lengths. Similarly, for HITS-CLIP and PAR-CLIP where the crosslinking signatures are mutations embedded in the reads, we will implement a read tagging function by searching specific mutational types as compared to the reference genome. In sum, the preprocessing module provides an easy and extensible way for parsing raw inputs to CLAM analyses.

Moreover, as demonstrated in the eCLIP technology, it is now evident that the incorporation of a properly matched control experiment is crucial for peak detection. In the case of eCLIP, the control experiment is Size-Matched Input; in MeRIP-seq experiments, the control experiment can be a regular RNA-seq without any immunoprecipitation. To consider these control experiments in a generic statistical

framework for peak calling, we implemented a negative binomial model-based peak caller to supplement the permutation-based peak caller in the original publication.

Specifically, to test for significantly enriched IP signal over Input control, we first divided each gene into *n*-bp non-overlapping bins. We then used a negative binomial model to model the observed counts in IP and Input sample and test the effect of IP in a given bin while controlling for the gene-specific expression levels. Specifically, for the i-th bin, we modeled the observed read counts in each bin as:

$$K_{ijl} \sim NB(\lambda_{ijl}, \alpha_{ij})$$

$$log(\lambda_{i0l}) = \mu_{i0} + l * \beta$$

$$log(\lambda_{i1l}) = \mu_{i1} + l * \beta + l * \delta$$

Where K_{ijl} is the observed read counts for the test bin (l=1) or other bins (l=0) in IP sample (j=1) or in Input sample (j=0), which effectively constructs a 2x2 contingency table. β denotes the bin-specific effect, and δ is the effect-size of immunoprecipitation. We first estimated gene-level over-dispersion parameters α_j from each sample by steepest damping, then used likelihood ratio test to compare the likelihoods between the constrained model $H_0: \delta = 0$ versus the unconstrained model $H_1: \delta \neq 0$ to determine the significance level for the enrichment of observed read counts. The bins with gene-level corrected FDR <0.05 were called as peaks. By default, we set *n*=100 bp for MeRIP-seq and *n*=50 bp for CLIP-seq experiments.

While the above model formulation is for one IP vs one Input comparison, it is straightforward to extend the model to consider multiple replicates. In fact, it is increasingly prevalent that multiple replicated IP and/or control experiments are performed for more robust detection of protein-DNA/RNA interaction signals. When

modelling multiple replicates, we let each replicate have replicate-specific baseline parameters $\mu_{\cdot}$, while the bin-specific parameter β and effect-size δ parameters are shared among all replicates for any given bin *i*.

In the release of CLAM v1.2, we further added an alternative approach for the background bin count K_{i0l}. Instead of the summing over the read counts across the same gene as the background bin count, we fed the total library read count to K_{i0l}. This approach is complementary to the gene-based background bin count, and proves useful when the input signal is very sparse and the gene-based count is unstable. As Python 2 is near the end of its maintenance, the CLAM package is also migrated for Python 3 compatibility as of release v1.2.

We systematically benchmarked the performance of CLAM using different run modes, and subsequently compared it to other popular peak-calling tools, as detailed in the next section.

2.8.2 Benchmarking CLAM and other peak callers

To evaluate the peak calling performance, the RBFOX2 dataset published in the original eCLIP paper (accession ID: GSE77629; Van Nostrand et al., 2016) was re-analyzed with CLAM and another popular CLIP-seq peak-caller software Piranha (version 1.2.1; Uren et al., 2012). The preprocessing steps were performed following the ENCODE eCLIP SOP. CLAM was run to call peaks using uniquely-mapped reads only on two individual replicates, as well as both replicates in a replicate statistical model, with bin size=50bp and FDR<0.05. Piranha was run on two individual replicates, with covariate logarithm transformed, bin size=50bp and FDR<0.05. The detailed pipeline implementation is

deposited as a Snakemake workflow in https://github.com/zj-zhang/CLAM_ENCODE_Snakemake.

We leveraged the conserved RBFOX2 motif GCAUG to benchmark the peak calling efficiency. Specifically, we ranked the peaks by the significance reported by each peak caller, then computed the proportion of peaks with the RBFOX2 motif at varying significance thresholds (**Supplementary Fig. 2.9**). As expected, the enrichment of motif was highest for the most stringent thresholds and decreased gradually for all peak callers. CLAM consistently performs better than Piranha as demonstrated by higher motif enrichment. Moreover, by comparing the motif enrichment for replicate 1 versus replicate 2, we observed that the performances of both CLAM and Piranha in replicate 2 were better than those of replicate 1. The data/replicate quality variability plays an important role in peak calling regardless of computational methods used.

To reduce such experimental variabilities and model the between-replicate variances in a principled statistical framework, we ran CLAM using multi-replicates mode on the two replicates of RBFOX2 eCLIP and two replicates of RBFOX2 SM-Input data. Indeed, multi-replicates mode CLAM achieved best specificity and sensitivity in peak calling (**Supplementary Fig. 2.9**). The multi-replicates CLAM called ~1.5 fold more peaks at the same FDR<0.05 threshold with a significantly better motif enrichment, compared to the single-replicate CLAM. Overall, the motif enrichment for all called peaks from multi-replicates CLAM was 41.4%, comparable to single-replicate CLAM on replicate 1 (42.2%) and replicate 2 (42.3%), demonstrating validity of FDR control in CLAM.

Two examples to demonstrate the benefit of considering multi-replicates were shown in **Supplementary Fig. 2.10**. In both cases, the CLAM multi-replicate statistical

model aggregated the relatively weak and noisy signals from individual replicates to call statistically enriched genomic bins as peaks. These peaks were putative RBFOX2 binding sites as they were embedded with RBFOX2 motifs, and would not be called by using simple intersections with the peaks called from individual replicates. Strong peaks called in both replicates were preserved in the multi-replicates results (rightmost peak in RERE gene in **Supplementary Fig. 2.10A**), while bogus peaks caused by random statistical and experimental fluctuations called in one replicate were down-weighed by considering additional replicates in the multi-replicates output (data not shown).

In sum, the statistical model to consider multiple replicates in CLAM provides a more robust and sensitive peak caller for replicated CLIP-seq data. The multi-replicates CLAM implementation can be found in v1.2 release.

Chapter 3 Deep-learning Augmented RNA-seq analysis of Transcript Splicing

3.1 Introduction

The rapid accumulation of RNA-seq data across diverse cell types and conditions provides an unprecedented resource for characterizing transcriptome complexity. However, the use of these large-scale data in routine RNA-seq studies to detect patterns of expression and thereby discover new regulatory events has been limited. Here we report DARTS, Deep-learning Augmented RNA-seq analysis of Transcript Splicing, a computational framework that integrates deep learning-based predictions with empirical evidence in specific RNA-seq datasets to infer differential alternative splicing between conditions. A core component of DARTS is a deep neural network (DNN) that predicts differential alternative splicing using *cis* RNA sequence features and *trans* RNA binding protein levels. DARTS DNN trained on public RNA-seq datasets (ENCODE, Roadmap Epigenomics) displays a high prediction accuracy and generalizability. Incorporating DARTS DNN prediction as an informative prior significantly improves the inference of differential alternative splicing, especially from low-coverage RNA-seq datasets. In cellular models of the epithelial-mesenchymal transition, DARTS reliably predicted alternative splicing changes in lowly expressed genes, that were inaccessible by a conventional RNA-seq analysis even at a high sequencing depth. Thus, DARTS capitalizes on large-scale public RNA-seq resources to discover differential alternative splicing across diverse transcriptomes.

Alternative splicing is a major mechanism for generating regulatory complexity (1) as well as linking genotypes to phenotypes in eukaryotes (2). RNA sequencing (RNA-seq) is a widely used technology for transcriptome-wide profiling of alternative splicing. The general workflow of RNA-seq alternative splicing analysis involves counting RNA-seq reads mapped to exons and splice junctions; estimating relative abundances of splice isoforms; and detecting differential alternative splicing events between biological conditions using appropriate statistical models (3,4). An inherent limitation of this approach is that it solely relies on empirical evidence in RNA-seq data, and thus is restricted by sequencing depth and cost. Moreover, even at a high sequencing depth the detection of splicing changes is biased against lowly or moderately expressed genes (5).

The availability of large-scale RNA-seq data in public repositories has enabled quantitative measurement of alternative splicing across diverse biological states. For example, the Roadmap Epigenomics consortium has generated deep RNA-seq data across over 100 human tissues and cell types (6), while the ENCODE consortium has systematically performed RNA-seq of two human cell lines upon knockdown of over 250 RNA binding proteins (RBPs) (7). Motivated by the recent success in using machine learning techniques to predict exon inclusion/skipping levels in bulk tissues or single cells (8-11), we hypothesized that large-scale public RNA-seq resources can be utilized to construct a deep learning model of differential alternative splicing. This deep learning model could in turn generate a predictive prior to augment alternative splicing analysis of a specific RNA-seq dataset.

To test this hypothesis, we developed DARTS (Deep-learning Augmented RNA-seq analysis of Transcript Splicing), a Bayesian computational framework for statistical

inference of differential alternative splicing. The DARTS framework consists of two core components (**Figure 3.1**): a Deep Neural Network (DNN) model that predicts differential alternative splicing between two biological conditions based on exon-specific sequence features and sample-specific regulatory features; and a Bayesian Hypothesis Testing (BHT) statistical model that infers differential alternative splicing by integrating empirical evidence in a specific RNA-seq dataset (as likelihood) with prior probability of differential alternative splicing, which can be either uninformative or informative (see below). During the training process (green), large-scale RNA-seq data (e.g. ENCODE, Roadmap Epigenomics) are analyzed by the DARTS BHT with an uninformative prior (i.e. DARTS BHT(flat), with only RNA-seq data used for the inference) to generate training labels of high-confidence differential or unchanged splicing events between biological conditions, which are then used to train the DARTS DNN. In the application process (pink), our goal is to leverage the deep learning prediction to analyze user-specific RNA-seq data. Therefore the trained DARTS DNN is used to predict differential alternative splicing in a user-specific dataset based on exon-specific sequence features and sample-specific regulatory features. This prediction is incorporated as an informative prior with the observed RNA-seq read counts by the DARTS BHT (i.e. DARTS BHT(info)) to perform deep learning augmented splicing analysis.

3.2 Results

3.2.1 DARTS deep neural network accurately predicts differential splicing

A core component of DARTS is a deep neural network (DARTS DNN) that uses predictive features to generate a probability of differential alternative splicing between two biological conditions. Unlike existing methods that use only *cis* sequence features to predict exon splicing patterns in specific samples (8-11), the DARTS DNN incorporates both *cis* sequence features and the mRNA levels of *trans* RNA binding proteins (RBPs) in two biological conditions (**Figure 3.2a**). This design allows the DARTS DNN to consider how altered expression of RBPs, including well-annotated splicing factors, affects alternative splicing in response to perturbations or stimuli. As a starting point, we initially focused on exon skipping, the most frequent type of alternative splicing events in mammalian cells (5). Specifically, we compiled a list of 2,926 *cis* sequence features including evolutionary conservation, splice site strength, regulatory motif composition, and RNA secondary structure, and a list of 1,498 annotated RBPs (12), whose mRNA levels were treated as RBP features. We designed the DARTS DNN to have 4 hidden layers and 7,923,402 parameters. This DARTS DNN can be trained using high-confidence differentially spliced and unchanged exons in a large compendium of pairwise RNA-seq comparisons between distinct biological states. A schematic diagram of the DARTS DNN is shown in **Supplementary Figure 3.5**.

To train the DARTS DNN, we utilized RNA-seq data from a large collection of RBP-depletion experiments in two human cell lines (K562 and HepG2) generated by the ENCODE consortium (13) (**Figure 3.2b**). Specifically, we used RNA-seq data corresponding to 196 RBPs that were depleted by at least one shRNA in both the K562 and HepG2 cell lines, corresponding to a total of 408 shRNA knockdown vs. control pairwise comparisons (**Figure 3.2b**). The remaining ENCODE data, corresponding to 58

RBPs that were depleted in only one cell line, were excluded from training and used as leave-out data to independently evaluate the DARTS DNN (**Figure 3.2b**; see below). Note that throughout this work, we used such independent leave-out datasets that had never been seen during training, to avoid issues of overfitting in benchmarking the performance of the DARTS DNN. To generate training labels for differentially spliced vs. unchanged exons in each pairwise comparison, we applied DARTS BHT(flat) to calculate the probability of an exon being differentially spliced or unchanged between two conditions. The performance of DARTS BHT(flat) was benchmarked using simulation datasets, and compared favorably to two of the state-of-the-art statistical models for differential splicing inference MISO and rMATS (**Supplementary Figure 3.6 and 3.7**). We should also note that a unique feature of DARTS BHT(flat) is that it quantifies the statistical evidence for both differentially spliced (positive) and unchanged (negative) exons. By contrast, conventional methods only test the statistical significance of differential splicing (positive) but the “insignificant” events contain a mixture of events that are truly unchanged (negative), and events that are called insignificant due to lack of RNA-seq coverage and power (inconclusive), hence are not suitable for generating training labels. From the high-confidence differentially spliced vs. unchanged exons called by DARTS BHT(flat) on the training RNA-seq data, we used 90% of the labeled events for training and 5-fold cross validation of the DARTS DNN, and the remaining 10% of events for testing the trained DARTS DNN (Methods). The performance of the DARTS DNN increased as the training progressed, reaching a maximum Area Under the Receiver Operating Characteristic curve (AUROC) of 0.97 during cross-validation and 0.86 during testing (**Figure 3.2c**).

To test the performance and general applicability of the DARTS DNN using independent datasets, we used the trained DARTS DNN to predict differentially spliced vs. unchanged exons from the leave-out data, which included 58 RBPs that were knocked-down in only one of the two cell lines and had not been used for training the DARTS DNN (**Figure 3.2b**). These leave-out data were derived from cell lines depleted for a different set of RBPs, therefore the performance of the DARTS DNN on the leave-out data would indicate model generalizability. The trained DARTS DNN model showed a high accuracy (average AUROC=0.87) on the leave-out data. We also used the leave-out data to compare the DARTS DNN to three alternative baseline methods: the identical DNN structure trained on individual leave-out datasets (DNN), logistic regression with L2 penalty (Logistic), and Random Forest (**Figure 3.2d**). We trained the baseline methods using 5-fold cross-validation in each leave-out dataset and plotted the average AUROC for each method (**Figure 3.2d**). Additionally, we implemented another alternative baseline method, by predicting sample-specific exon inclusion levels (11) then taking the absolute difference of the predicted exon inclusion levels (PSI values) between the two conditions as the metric for differential splicing ($|\hat{\psi}_{KD} - \hat{\psi}_{CTRL}|$; **Figure 3.2d**). We found that the DARTS DNN trained on the large-scale ENCODE data outperformed baseline methods by a large margin in 57/58 experiments, with the sole exception being AQR knockdown in K562. The best performance of AUROC=0.95 by the DARTS DNN was achieved for RPL23A knockdown in HepG2. We note that the DARTS DNN model trained on individual leave-out datasets was the worst performer, illustrating the importance of training the DARTS DNN on large-scale data comprising diverse perturbation experiments.

Collectively, these results indicate that the DARTS DNN can predict differential splicing upon RBP depletion in the two ENCODE cell lines.

3.2.2 DARTS Bayesian hypothesis testing model incorporates informative prior with empirical RNA-seq data to improve inference efficiency

Having demonstrated the performance of the DARTS DNN model, we set out to evaluate the ability of the DARTS framework to infer differential splicing from a specific RNA-seq dataset, by incorporating the DARTS DNN prediction score as the informative prior and observed RNA-seq read counts as the likelihood (see Methods). Specifically, the posterior ratio of differential splicing consists of two components: the prior ratio, generated by the DARTS DNN model based on *cis* sequence features and *trans* RBP expression levels; and the likelihood ratio, determined by modelling the biological variation and estimation uncertainty of splice isoform ratio based on observed RNA-seq read counts (**Figure 3.3a**). We performed simulation studies with varying strengths of informative prior and observed RNA-seq read counts (see Methods). These studies demonstrated that the informative prior improves the inference when the observed data is limited, for instance due to low gene expression levels or limited RNA-seq depth, but does not overwhelm the evidence in the observed RNA-seq read counts (**Supplementary Figure 3.8**). Specifically, for true differential splicing events in the simulation, a considerable number of true positives in the low RNA-seq coverage regions can be rescued via a strong informative prior, whereas the effect of the prior was diminished when the observed RNA-seq read counts were large (**Supplementary Figure 3.8**). We refer to this method as DARTS BHT(info), and compared it to DARTS BHT(flat) that uses the same BHT statistical model but only

considers the RNA-seq data without incorporating the DARTS DNN prediction as the informative prior (i.e. flat prior).

To investigate the utility of incorporating the DARTS DNN prediction as the informative prior on a real dataset, we used DARTS BHT(info) and DARTS BHT(flat) to infer cell-type-specific differential splicing events between two ENCODE cell lines (HepG2 and K562). ENCODE generated paired-end RNA-seq data on 24 and 28 biological replicates of HepG2 and K562 respectively, with on average 66 million read pairs per replicate. We confirmed by cluster analysis of gene expression levels that these 24 and 28 biological replicates clustered into two distinct groups that matched their cell type labels (**Supplementary Figure 3.9**). To obtain high-confidence differential and unchanged splicing events between the two cell types, we aggregated all replicates of HepG2 or K562 and applied DARTS BHT(flat) to this ultra-deep RNA-seq dataset. Next, we applied DARTS BHT(info) and DARTS BHT(flat) to all possible (24x28) pairwise comparisons between individual replicates of HepG2 and K562. We computed the Area Under Precision Recall Curve (AUPR) for the two methods to evaluate their performance in detecting cell-type specific alternative splicing. DARTS BHT(info) outperformed DARTS BHT(flat) in all pairwise comparisons, and the gain in inference accuracy had a significant negative correlation with the RNA-seq depth of individual replicates (Spearman's rho=-0.69, p-value<2.2e-16), with the largest gain coming from pairwise comparisons involving low-coverage RNA-seq samples (**Figure 3.3b**). We should also note that by using DARTS BHT(flat) to obtain the lists of high-confidence differential and unchanged exons between the two cell types from the ultra-deep RNA-seq data, the comparison at the low sequencing depth was inherently biased towards DARTS BHT(flat)

and against DARTS BHT(info). Therefore, the consistently superior performance of DARTS BHT(info) demonstrates the advantage of incorporating the DNN prediction as prior information when analyzing low-coverage RNA-seq data.

3.2.3 Expanding DARTS to diverse types of splicing events and cellular conditions

Next, we determined whether the DARTS DNN model can be extended to additional tissues and cell types, and whether and how the choice of training datasets influences the performance of the DARTS DNN on other datasets. For this analysis we utilized deep RNA-seq data from diverse cell types and tissues generated by the Roadmap Epigenomics consortium (6). We performed 253 pairwise comparisons of Roadmap samples by DARTS BHT(flat) to generate training data for the DARTS DNN, following the same procedure as for the ENCODE data. We held-out all pairwise comparisons involving the thymus tissue in Roadmap (22 comparisons) so we could use these later for model testing. We used three DARTS DNN models, trained on ENCODE data only, Roadmap data only, or both, to evaluate model performance with the held-out data from ENCODE or Roadmap (**Figure 3.3c,d**). We found that DARTS DNN trained on ENCODE data exhibited high predictive power for differential splicing in held-out ENCODE data and modest predictive power for differential splicing in held-out Roadmap data (**Figure 3.3c**). Conversely, DARTS DNN trained on Roadmap data had high predictive power for held-out Roadmap data and modest predictive power for held-out ENCODE data (**Figure 3d**). Finally, the best model performance was achieved with DARTS DNN trained on the combination of ENCODE and Roadmap data (**Figure 3c,d**).

Having demonstrated the performance of the DARTS DNN on predicting differential exon skipping events, we extended the DNN model to different classes of alternative splicing patterns. Specifically, we compiled 2,973, 2,971, and 1,748 *cis* sequence features and trained three DNN models for predicting differential alternative 5' splice sites (A5SS), alternative 3' splice sites (A3SS), and retained intron (RI) events, respectively. The training behavior of these DNN models was similar to the DNN model trained for exon skipping events (**Supplementary Figure 3.10**). Trained by ENCODE and Roadmap data, these DNN models also exhibited a high prediction accuracy and generalizability in the independent leave-out datasets and outperformed baseline methods by a large margin (**Supplementary Figure 3.10**). These data extend the utility of the DARTS DNN beyond exon skipping to diverse types of alternative splicing patterns.

3.2.4 DARTS analysis of Epithelial-Mesenchymal Transition

As a further proof-of-principle study, we applied DARTS DNN trained on ENCODE and Roadmap datasets to uncover alternative splicing during the epithelial-mesenchymal transition (EMT), a key cellular process in embryonic development and cancer metastasis (14). We previously published RNA-seq data of an H358 lung cancer cell line that underwent EMT through a 7-day time course via inducible expression of the mesenchymal EMT driver ZEB1 (15). We re-analyzed this RNA-seq dataset and used DARTS BHT(flat) to compare the splicing profiles of each day to Day 0. We then assessed the ability of the DARTS DNN model to predict high-confidence differential vs. unchanged splicing events during the EMT time course. As EMT progressed, the number of differential splicing events called by DARTS BHT(flat) increased (**Figure 3.4a**). We found

that the DARTS DNN trained on ENCODE+Roadmap data displayed the best model performance, followed closely by the DARTS DNN trained on Roadmap data, whereas the DARTS DNN trained on ENCODE data performed least well (**Figure 3.4a**). These findings were not unexpected, given that the Roadmap RNA-seq data cover diverse tissues and cell types including various epithelial and mesenchymal cell types, whereas the ENCODE data are restricted to HepG2 and K562 cell lines. The best prediction accuracy of AUC=0.82 was achieved by the DARTS DNN trained on ENCODE+Roadmap for the Day 6 versus Day 0 comparison.

To further investigate the validity of the DARTS predictions, we compiled a set of 449 “DARTS DNN rescued” events from the Day 6 vs. Day 0 comparison. These are splicing events that displayed a high DARTS DNN score of differential splicing (FPR<5%) and a non-trivial splicing change (over 10% difference in exon inclusion level), but did not pass the significance threshold by DARTS BHT(flat) using observed RNA-seq read counts alone. We uncovered a subset of “DARTS DNN rescued” events with significantly reduced exon inclusion during EMT, and found that the intronic regions downstream of these exons were enriched for the previously defined consensus motif of the splicing factors ESRP1 and ESRP2 (16) (**Figure 3.4b**). A similar pattern of ESRP motif enrichment was observed for differential splicing events called by DARTS BHT(flat) using RNA-seq data alone (**Figure 3.4b**). By contrast, we also found 123 events that were called significant by DARTS BHT(flat) but fell below the significance threshold (posterior probability<0.9) after incorporating the informative prior; these events were not enriched for the ESRP motif (**Supplementary Figure 3.11**). ESRPs are epithelial-specific splicing factors whose downregulation is a major driver of alternative splicing during EMT (15).

The pattern of ESRP motif enrichment in the subset of “DARTS DNN rescued” events with reduced exon inclusion during EMT is consistent with previous findings that ESRP binding downstream of alternative exons enhances exon inclusion (14). These motif data provide transcriptome-wide regulatory evidence for the validity of the DARTS DNN prediction. As a specific example, DARTS DNN predicted the EMT-associated alternative splicing change in the PLEKHA1 gene (**Figure 3.4c**). The DARTS DNN score for this exon is 0.94 in Day 5 versus Day 0, increasing the posterior probability of differential splicing to 0.73 over 0.42 when using RNA-seq data alone. The differential splicing pattern of this exon was apparent throughout the time course and was validated by RT-PCR (15).

To extend our observations on DARTS analysis of EMT-associated alternative splicing in the H358 lung cancer cell line, we compared other epithelial and mesenchymal cell lines. We performed paired-end RNA-seq of the PC3E and GS689 prostate cancer cell lines, which we previously showed have contrasting epithelial vs. mesenchymal characteristics respectively (4,17). We generated a deep RNA-seq dataset with on average 125 million read pairs per replicate for three biological replicates per cell type. We found that several EMT-relevant splicing factors (ESRP1, ESRP2, RBM47) were differentially expressed in both the GS689-PC3E “EMT system” and during EMT of H358; a few other RBPs were differentially expressed in only one of the two comparisons (**Figure 3.4d**). The DARTS DNN scores of these two EMT systems were highly correlated (Spearman’s rho=0.87, p-value<2.2e-16; **Figure 3.4e**). This correlation was higher than the correlation of the GS689-PC3E scores with the DARTS DNN scores for any ENCODE RBP-depletion experiment (median=0.69, interquartile range IQR=[0.67, 0.73]). These

data suggest that there is a core differential alternative splicing signature between epithelial and mesenchymal cells, and that the DARTS DNN model can capture this signature.

Finally, to assess if DARTS can uncover *bona fide* differential splicing events from moderately or lowly expressed genes, we performed RASL-seq on the same PC3E and GS689 RNA samples to generate high-confidence measurements of exon inclusion levels (PSI values) for these two cell types. RASL-seq is a sequencing method for targeted amplification and quantitative profiling of alternative splicing events (18) (Methods). The absolute difference of PSI values between the two cell types (denoted as RASL-ΔPSI) was computed for each alternative splicing event (n=1,058) passing the RASL-seq read coverage filter (see Methods). From the RNA-seq data and DARTS DNN prediction, we compiled four groups of alternative splicing events and compared their distributions of RASL-ΔPSI values (**Figure 3.4f**). For this analysis, we restricted to events with RASL-ΔPSI value <0.3. As expected, alternative splicing events called as differential or unchanged using RNA-seq data alone (by DARTS BHT(flat)) displayed the highest or lowest RASL-ΔPSI values, respectively. For the remaining alternative splicing events called as inconclusive by DARTS BHT(flat) using RNA-seq data alone, we compiled two additional groups: DARTS DNN-predicted differential events, with high DARTS DNN scores (FPR<5%); and DARTS DNN-predicted unchanged events, with low DARTS DNN scores (FPR>80%). The RASL-ΔPSI values of the DARTS DNN-predicted differential events were significantly larger than those of the DARTS DNN-predicted unchanged events (p-value=0.035, one-sided Wilcoxon test), with the DARTS DNN-predicted differential events similar to the RNA-seq differential events and the DARTS DNN-

predicted unchanged events similar to the RNA-seq unchanged events (**Figure 3.4f**). These events were in genes with significantly lower gene expression levels (p-value=0.001, Wilcoxon test) and had significantly lower RNA-seq coverage (p-value=2.1e-7, Wilcoxon test) compared to differential events called by DARTS BHT(flat) (**Supplementary Figure 3.12a,b**), and were similarly called as insignificant by rMATS (4) when using RNA-seq data alone. Collectively, among the events analyzed by RASL-seq, DARTS DNN predicted 52 additional differential splicing events, beyond the 77 events called using RNA-seq data alone (**Figure 3.4f**). Importantly, alternative splicing events in moderately or lowly expressed genes with high DARTS DNN prediction scores had a comparable shift in RASL-seq PSI values as the differential splicing events called from RNA-seq data alone in highly expressed genes. Additionally, on this set of RNA-seq inconclusive events with high or low DARTS DNN scores, we used the RASL-seq data to define the ground truth with RASL-$|\Delta PSI|>5\%$ as differential and RASL-$|\Delta PSI|<1\%$ as unchanged events respectively. We then benchmarked the performance of DARTS BHT(info), DARTS BHT(flat), DARTS DNN, as well as two existing methods rMATS (4) and SUPPA2 (19). We chose rMATS and SUPPA2 because they represented two distinct strategies (alignment-based vs alignment-free) for quantifying alternative splicing using RNA-seq data. As shown in **Supplementary Figure 3.12c**, DARTS BHT(info) consistently outperformed baseline methods that use RNA-seq data alone to call differential splicing: AUC of 0.76 for DARTS BHT(info), versus 0.68, 0.63, and 0.61 for DARTS BHT(flat), rMATS, and SUPPA2 respectively. We also observed a consistent gain by DARTS BHT(info) over baseline methods at different FPR thresholds for DARTS DNN-predicted differential events, with the maximum gain observed for the most confidently

predicted events of FPR=1% (**Supplementary Figure 3.12d**). Together, these data suggest that DARTS can reliably predict and uncover differential splicing events from moderately or lowly expressed genes and expand the findings beyond a conventional RNA-seq splicing analysis, even on a deep RNA-seq dataset.

3.3 Discussion

In summary, we present DARTS, a deep-learning augmented statistical framework for RNA-seq analysis of differential alternative splicing. DARTS leverages massive RNA-seq datasets across diverse cell types and perturbation conditions to predict differential alternative splicing using exon-specific *cis* sequence features and sample-specific *trans* RBP expression levels. We extend the DARTS deep learning model beyond exon skipping and show that it can reliably predict differential alternative splicing involving diverse types of alternative splicing patterns. We demonstrate that DARTS can improve differential splicing analysis from user-specific RNA-seq data and predict alternative splicing changes in lowly expressed genes that are inaccessible by a conventional RNA-seq analysis. This addresses a fundamental limitation in RNA-seq studies of alternative splicing in their overly reliance on high sequencing coverage (20). Conceptually, the DARTS framework transforms existing RNA-seq big data into a knowledge base of splicing regulation via deep learning, which can in turn help individual investigators better characterize alternative splicing profiles in their specific RNA-seq studies (**Figure 3.1**). The DARTS software as well as the associated training data and predictive features are available at https://github.com/Xinglab/DARTS.

3.4 Methods

3.4.1 DARTS Bayesian hypothesis testing (BHT) framework

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

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

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

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

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

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

Where I_{ij}, S_{ij} and ψ_{ij} are the exon inclusion read count, the exon skipping read count, and the exon inclusion level for exon i in sample group $j \in (1,2)$, respectively; f_i is the length normalization function for exon i that accounts for the effective lengths of the exon inclusion and skipping isoforms (4); μ_i is the baseline inclusion level for exon i, and δ_i is the expected difference of the exon inclusion levels between the two conditions. The goal of the differential splicing analysis is to test whether the difference in exon inclusion levels between the two conditions δ_i exceeds a user-defined threshold c (e.g. 5%) with a high probability, i.e.

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

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

In the DARTS BHT statistical framework, we impose a spike prior H_0 with a small variance $\tau = \tau_0$, such that the probability of δ concentrates around 0, to account for random biological or technical fluctuations in PSI values between two biological conditions for unchanged splicing events. We impose a slab prior H_1 with a much larger variance $\tau = \tau_1$ to model the difference in PSI values between two conditions for differential splicing events. We set $\tau_0 = 0.03$, corresponding to 90% density constrained within $\delta \in [-0.05, 0.05]$, and $\tau_1 = 0.3$; we note that the final inference is robust to choice of τ values (for more details, see Appendix and **Supplementary Figure 3.13**). The posterior probability of a splicing event being generated by the two models can be written as:

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

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

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

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

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

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

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

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

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

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

Hence, the final inference is performed on the probability $P(|\delta_i| > c|I_{ij}, S_{ij})$. In our analysis, we set c=0.05 (i.e. a 5% change in exon inclusion level) and call events with $P(|\delta_i| > 0.05\ |I_{ij}, S_{ij}) > 0.9$ as significant differential splicing events and $P(|\delta_i| > 0.05\ |I_{ij}, S_{ij}) < 0.1$ as significant unchanged splicing events. Events with $0.1 \leq P(|\delta_i| >$

$0.05|I_{ij}, S_{ij}) \leq 0.9$ are deemed as inconclusive. In the following text, we omit the subscripts and use $P(|\delta_i| > c\ |I_{ij}, S_{ij})$ and $P(|\Delta\psi| > c)$ interchangeably.

3.4.2 DARTS deep neural network (DNN) model for predicting differential alternative splicing

A core component of the DARTS BHT framework is a deep neural network (DNN) model that generates a probability of differential splicing between two biological conditions using exon- and sample-specific predictive features. We designed the DARTS DNN to predict differential splicing of a given exon based on exon-specific *cis* sequence features and sample-specific *trans* RBP expression levels in two biological conditions.

As noted above, a useful feature of the DARTS BHT framework is its capability to determine the statistical significance of both differential splicing events and unchanged splicing events. Specifically, for a splicing event i in the comparison k between RNA-seq datasets from two distinct biological conditions $(j \in (1,2))$, let $Y_{ik} = 1$ if this event is differentially spliced (i.e. H_1 is true); and $Y_{ik} = 0$ if H_0 is true as labels for differential or unchanged splicing events respectively. The task of predicting differential splicing can be formulated as:

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

Where Y_{ik} is the label for event i in the comparison k; E_i is a vector of 2,926 *cis* sequence features for exon i, including evolutionary conservation, splice site strength, regulatory motif composition, and RNA secondary structure. G_k is a vector of 2,996 (=1,498x2) normalized gene expression levels of 1,498 RBPs in the two conditions. The prediction

of $P(Y_{ik} = 1)$ based on the features from any specific RNA-seq dataset can then be incorporated as an informative prior for $P(H_1)$ in the DARTS BHT framework.

We implemented a deep learning model (DARTS DNN) to learn the unknown function F that maps the predictive features to splicing profiles (differential vs. unchanged). We designed the DARTS DNN with 4 hidden layers and 7,923,402 parameters. The configuration of the DNN was: an input layer with 5922(=2926+1498*2) variables; 4 fully-connected hidden layers with 1200, 500, 300, 200 variables and the ReLU activation function; and an output layer with 2 variables and the Softmax activation function. We implemented the DARTS DNN using Keras (https://github.com/keras-team/keras) with the Theano backend.

To mitigate potential overfitting of the DARTS DNN, we added a drop-out probability (21) for connections between hidden layers. Specifically, the variables in the four hidden layers were randomly turned off during the training process with probability 0.6, 0.5, 0.3, and 0.1, respectively. We also added batch-normalization layers (22) for all hidden layers to help the model converge and generalize. Finally, we used the RMSprop optimizer to adaptively adjust for the magnitudes of the components of the gradient in this deep architecture and chose 1000 labeled alternative splicing events as one mini-batch to obtain a more stable gradient. In each mini-batch we balanced the composition of positive and negative labels by adding more positive events in the mini-batch such that positive : negative = 1:3 in the mini-batch. Since there were significantly more negative (unchanged) events compared to positive (differential) events, such a balanced composition will provide a gradient for learning the positive events in different mini-batches.

To monitor the training loss and validation loss, we computed the loss every 10 mini-batches and saved the current model parameters if the validation loss was lower than the previous best model. We trained the DARTS DNN on Tesla K40m.

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

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

We processed the RNA-seq alignments (bam files) using rMATS (4) (v4.0.1). Given RNA-seq alignment files, rMATS constructs splicing graphs, detects annotated and novel alternative splicing events, and counts the number of RNA-seq reads for each exon and splice junction. Given the modest depth of the ENCODE RNA-seq data (32 million read pairs per replicate on average), the read counts from the two replicates were pooled together for downstream analyses.

We processed the raw RNA-seq reads with Kallisto (23) (v0.43.0) to quantify gene expression levels using Gencode (24) (v19) protein coding transcripts as the index. For each of the two biological conditions in a given comparison (i.e. shRNA knockdown vs. control), we extracted the Kallisto derived gene-level TPM values of 1,498 known RBPs (12). The TPM value of each RBP was normalized by dividing by its maximum observed

TPM value of all comparisons, then used as RBP expression features by the DARTS DNN.

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

The vast majority of the RBPs (n=196) in the ENCODE data were knocked-down by at least one shRNA in both HepG2 and K562 cell lines, corresponding to a total of 408 comparisons between knockdown and control. We set aside 10% of the labelled positive events and the same number of labeled negative events in each comparison as the testing data for estimating the generalization error of the trained DNN model. For the remaining 90% of the labelled events, we further split them into 5-fold cross-validation subsets for the purposes of training, monitoring overfitting, and early-stopping. We also collected ENCODE RBP knockdown experiments performed in only one cell line (either HepG2 or K562, n=58) as leave-out datasets. All labelled events in these leave-out datasets were only utilized for evaluating the trained DARTS DNN and were never used during training.

We randomly drew 4 RBPs without replacement for a training batch, and iterated through all 196 RBPs as an epoch. The performance of the DARTS DNN was measured by Area Under the Receiver Operating Characteristics curve (AUROC). The model with the best performance during training and cross-validation was selected, and subsequently benchmarked using the testing data and leave-out data.

3.4.4 Rank-transformation of the DARTS informative prior

In a typical RNA-seq study, the number of unchanged splicing events can be orders of magnitude larger than differential splicing events, and machine learning algorithms may be biased to the majority class. To mitigate this potential bias, we used an unsupervised rank-transformation to rescale DARTS DNN scores to derive the informative prior for the DARTS BHT framework. Specifically, we first fit a two-component Gaussian mixture model for all the DARTS DNN scores to derive the mean and variance of the two mixed Gaussian components as well as the posterior probability λ of each DARTS DNN score belonging to a specific component. Setting the new mean and variance of the two Gaussian components to μ_0 and μ_1, σ_0 and σ_1, respectively, each DARTS DNN score was rank-transformed to the new Gaussian components and then averaged by the weight parameter λ. Finally, to maintain a valid prior probability, the transformed DARTS DNN scores were rescaled to $[\alpha, 1-\alpha]$, where $\alpha \in [0,0.5)$ sets the desired prior strength for the DARTS BHT framework and a smaller α value corresponds to a stronger strength of the informative prior. Using this rescaling scheme, the entire ranks of the DARTS DNN scores are preserved while the potential bias for negative over positive events is reduced. In practice, we set $\mu_0 = 0.05$, $\mu_1 = 0.95$, $\sigma_0 = \sigma_1 = 0.1$, and $\alpha = 0.05$.

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

We generalized the DARTS framework to incorporate diverse tissues and cell types by utilizing RNA-seq resources from the Roadmap Epigenomics project (6). The Roadmap data was processed following the same protocol as for the ENCODE data. We took all

Roadmap data with 101bp x 2 or 100bp x 2 paired-end RNA-seq, and truncated reads from the 101bp x 2 datasets to 100bp for rMATS. In total, this represented 23 distinct tissues or cell types. All possible pairwise comparisons (n=253) between these 23 RNA-seq samples were performed. Comparisons involving thymus were held out as Roadmap leave-out data, and all remaining comparisons were used as training datasets.

We trained three DARTS DNN models using different training datasets: i) ENCODE data only, ii) Roadmap data only, and iii) the combination of ENCODE+Roadmap data. The performances of the three models were subsequently benchmarked by using ENCODE or Roadmap leave-out datasets.

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

We applied the trained DARTS model to study EMT-associated alternative splicing events in two distinct human cell culture systems: H358 lung cancer cell line induced to undergo EMT through a 7-day time course (15), and PC3E/GS689 prostate cell lines that had contrasting epithelial versus mesenchymal characteristics (4,17).

For the H358 time-course RNA-seq data (GSE75492), we used DARTS BHT(flat) to compare RNA-seq data from Day 1 to Day 7 against Day 0. Motif analysis was performed by calculating the average percentage of nucleotides covered by any of the top 12 ESRP SELEX-seq hexamer motifs (16) in a 45bp sliding window. Background sequences were significant unchanged events by DARTS BHT(flat). For the PC3E and GS689 cell lines, we conducted RASL-seq (18) and RNA-seq experiments on the same batch of RNA samples, each with 3 replicates and on average 125 million read pairs per replicate (raw data deposited as GSE112037). RASL-seq reads were aligned to the pool

of target splice junctions in the RASL-seq library using Blat (25). RASL-PSI values were calculated as $\frac{I}{I+S}$, where I is the number of exon inclusion splice junction reads and S is the number of exon skipping splice junction reads. Alternative splicing events with total RASL-seq read counts larger than 5 in every replicate were used for downstream analyses. Gene expression levels of RBPs in the two datasets were quantified using Kallisto v0.43.0.

3.4.7 RASL-seq library preparation and sequencing

RASL-seq was performed as described (26) with some modifications. Total RNA from PC3E and GS689 cell lines were extracted with Trizol (Thermo Fisher Scientific). RASL-seq oligonucleotides (a gift from Xiang-Dong Fu, UCSD) were annealed to 1 μg of total RNA, followed by selection by oligo-dT beads. Paired probes templated by polyA$^+$ RNA were ligated and then eluted. 5 μl of the eluted ligated oligos were used for 8 cycles of PCR amplification using primers F1: 5'-CCGAGATCTACACTCTTTCCCTACACGACGGCGACCACCGAGAT-3' and R1: 5'-GTGACTGGAGTTCAGACGTGTGCGCTGATGCTACGACCACAGG-3'. One third of the resulting PCR products were used in the second round of PCR amplification (9 cycles) using primers F2: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG-3' and R2: 5'-CAAGCAGAAGACGGCATACGAGAT[index]GTGACTGGAGTTCAGACGTGTGC-3'; indexes used in this study were Illumina indexes D701-D706. The indexed PCR products were pooled and sequenced on a Miseq with a custom sequencing primer 5'-

ACACTCTTTCCCTACACGACGGCGACCACCGAGAT-3' and a custom index sequencing primer 5'-TAGCATCAGCGCACACGTCTGAACTCCAGTCAC-3'.

3.4.8 Data Availability

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

3.4.9 Code Availability

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

Acknowledgements

I would like to thank my collaborators and co-authors on the publication of this work, they are: Zhicheng Pan, Yi Ying, Zhijie Xie, Samir Adhikari, John Phillips, Russ P. Carstens, Douglas L. Black, Yingnian Wu, and Yi Xing.

This study is supported by National Institutes of Health grants (R01GM088342 and R01GM117624 to Y.X.). Z.Z. is partially supported by a UCLA Dissertation Year Fellowship.

3.5 References

1. Nilsen, T.W. & Graveley, B.R. Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457-463 (2010).

2. Manning, K.S. & Cooper, T.A. The roles of RNA processing in translating genotype to phenotype. Nature reviews. Molecular cell biology 18, 102-114 (2017).

3. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods 7, 1009-1015 (2010).

4. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences of the United States of America 111, E5593-5601 (2014).

5. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. American journal of human genetics 102, 11-26 (2018).

6. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317-330 (2015).

7. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).

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

9. Barash, Y. et al. Deciphering the splicing code. Nature 465, 53-59 (2010).

10. Leung, M.K., Xiong, H.Y., Lee, L.J. & Frey, B.J. Deep learning of the tissue-regulated splicing code. Bioinformatics 30, i121-129 (2014).

11. Huang, Y. & Sanguinetti, G. BRIE: transcriptome-wide splicing quantification in single cells. Genome biology 18, 123 (2017).

12. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. Nature reviews. Genetics 15, 829-845 (2014).

13. Van Nostrand, E.L. et al. A large-scale binding and functional map of human RNA binding proteins. bioRxiv, 179648 (2017).

14. Warzecha, C.C. et al. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. The EMBO journal 29, 3286-3300 (2010).

15. Yang, Y. et al. Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition. Molecular and cellular biology 36, 1704-1719 (2016).

16. Dittmar, K.A. et al. Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. Molecular and cellular biology 32, 1468-1482 (2012).

17. Lu, Z.X. et al. Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization. Molecular cancer research : MCR 13, 305-318 (2015).

18. Li, H., Qiu, J. & Fu, X.D. RASL-seq for massively parallel and quantitative analysis of gene expression. Current protocols in molecular biology Chapter 4, Unit 4 13 11-19 (2012).

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

20. Cieslik, M. & Chinnaiyan, A.M. Cancer transcriptome profiling at the juncture of clinical translation. Nature reviews. Genetics 19, 93-109 (2018).

21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929-1958 (2014).

22. Ioffe, S. & Szegedy, C. in International conference on machine learning 448-456 (2015).

23. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology 34, 525-527 (2016).

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

25. Kent, W.J. BLAT--the BLAST-like alignment tool. Genome research 12, 656-664 (2002).

26. Ying, Y. et al. Splicing Activation by Rbfox Requires Self-Aggregation through Its Tyrosine-Rich Domain. Cell 170, 312-323 e310 (2017).

3.6 Figures

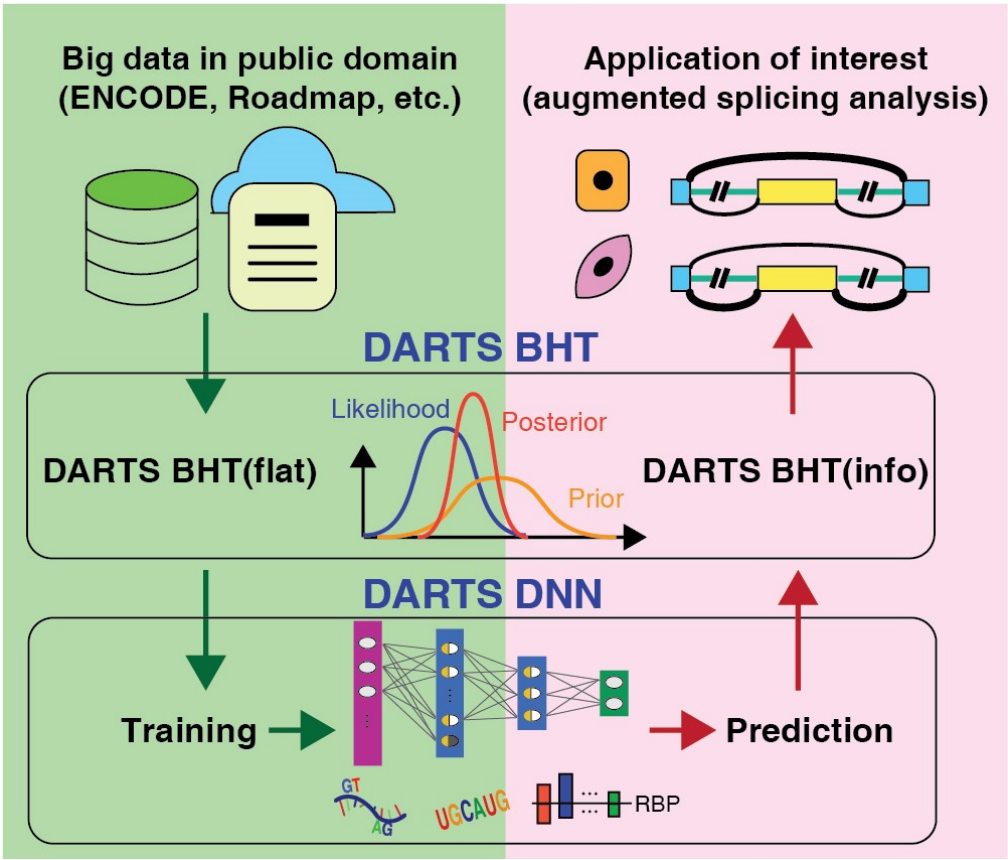


Figure 3.1 Overall workflow of the DARTS computational framework.

During the training process (green), large-scale RNA-seq data (e.g. ENCODE, Roadmap) are analyzed by DARTS BHT(flat), then used to train the DARTS DNN using exon-specific sequence features and sample-specific regulatory features (i.e. gene expression levels of RNA binding proteins) in the training datasets. In the application process (pink) to a user-specific dataset, the trained DARTS DNN is used to generate an informative prior of differential alternative splicing based on exon-specific sequence features and sample-specific regulatory features in the user-specific dataset. The informative prior is subsequently incorporated with the observed RNA-seq read counts by DARTS BHT (info) to perform deep learning-augmented splicing analysis.

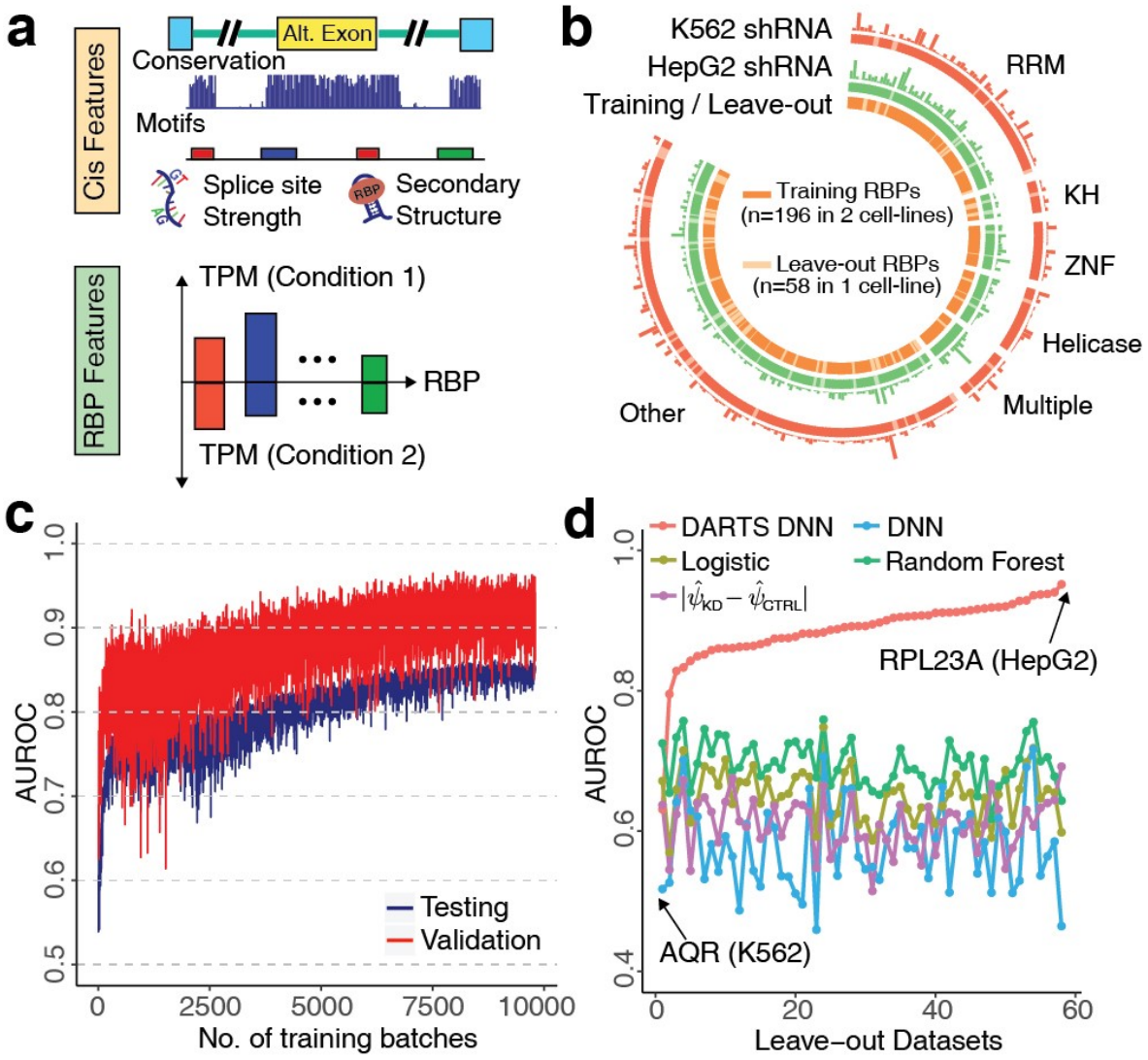


Figure 3.2 The DARTS Deep Neural Network (DNN) model of differential alternative splicing.

(a) Schematic illustration of the DARTS features, including cis sequence features and trans RBP features. **(b)** Overview of training/testing and leave-out RBPs, and the number of significant differential splicing events called by DARTS BHT(flat) on the ENCODE data (illustrated by bar charts above the outer and middle circles). 196 RBPs knocked-down in both the K562 and HepG2 cell lines are used for training and testing (orange), while the

remaining 58 RBPs are leave-out data (light orange) (illustrated in the inner circle). **(c)** AUROC values of validation and testing data increase as the training progresses. **(d)** Comparison of DARTS DNN with baseline methods in leave-out datasets. DARTS DNN outperforms baseline methods trained on individual leave-out datasets by a large margin. The identical deep neural network (DNN) trained on individual leave-out datasets performs poorly due to severe overfitting without being trained on big data.

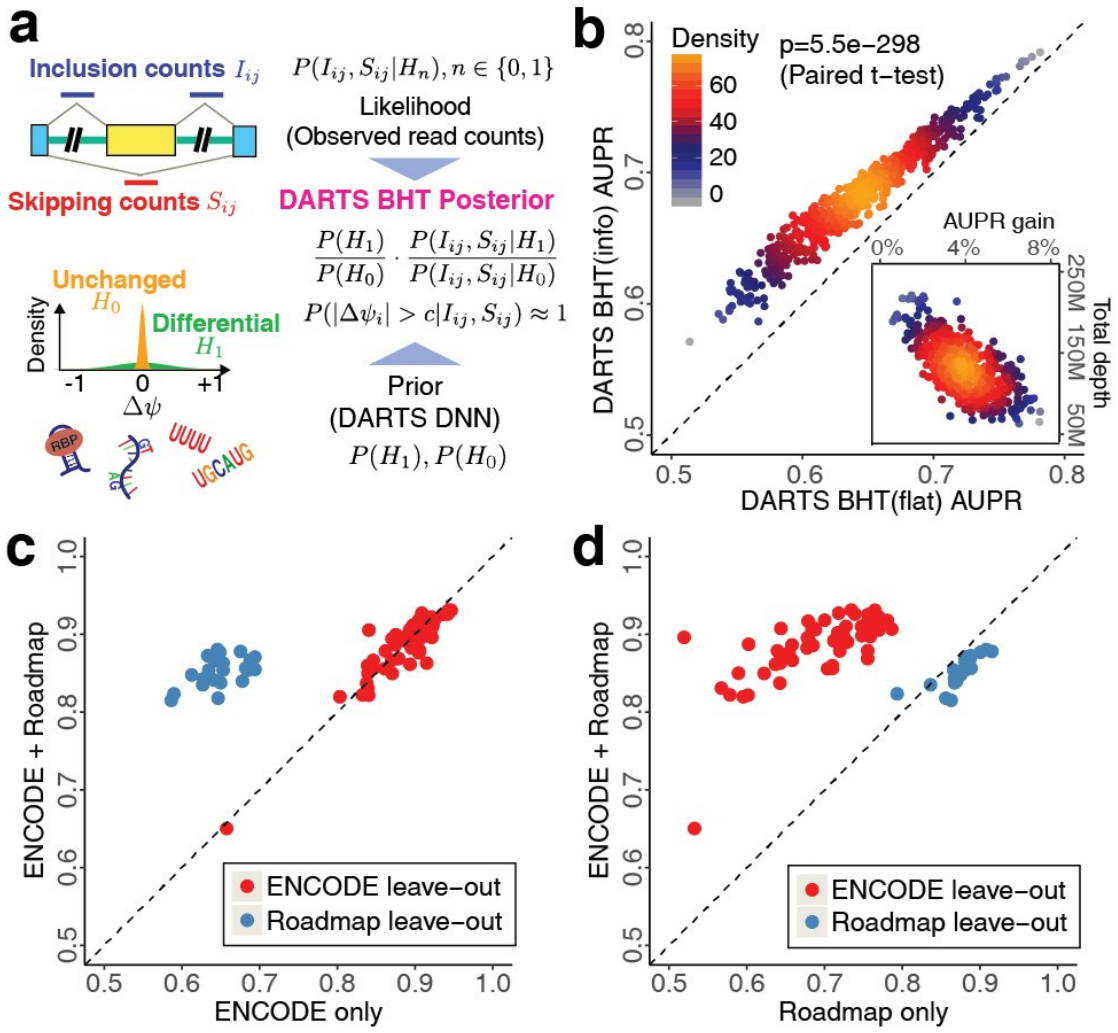


Figure 3.3 The schematic illustration and performance evaluation of the DARTS Bayesian Hypothesis Testing (BHT) framework.

(a) DARTS BHT infers differential alternative splicing by combining the likelihood of the observed RNA-seq read counts and the prior probability determined by the DARTS DNN using exon-specific cis sequence features and sample-specific trans RBP features. **(b)** DARTS BHT(info) consistently outperforms DARTS BHT(flat) in the cell-type-specific differential splicing analysis of HepG2 and K562. The gain in accuracy by DARTS BHT(info) was more prominent in pairwise comparisons of low-coverage replicates (inset).

(c) DARTS DNN trained on ENCODE+Roadmap data outperforms DARTS DNN trained on ENCODE data only when applied to Roadmap leave-out data. Plotted are the AUROC values. **(d)** DARTS DNN trained on ENCODE+Roadmap data outperforms DARTS DNN trained on Roadmap data only when applied to ENCODE leave-out data. Plotted are the AUROC values.

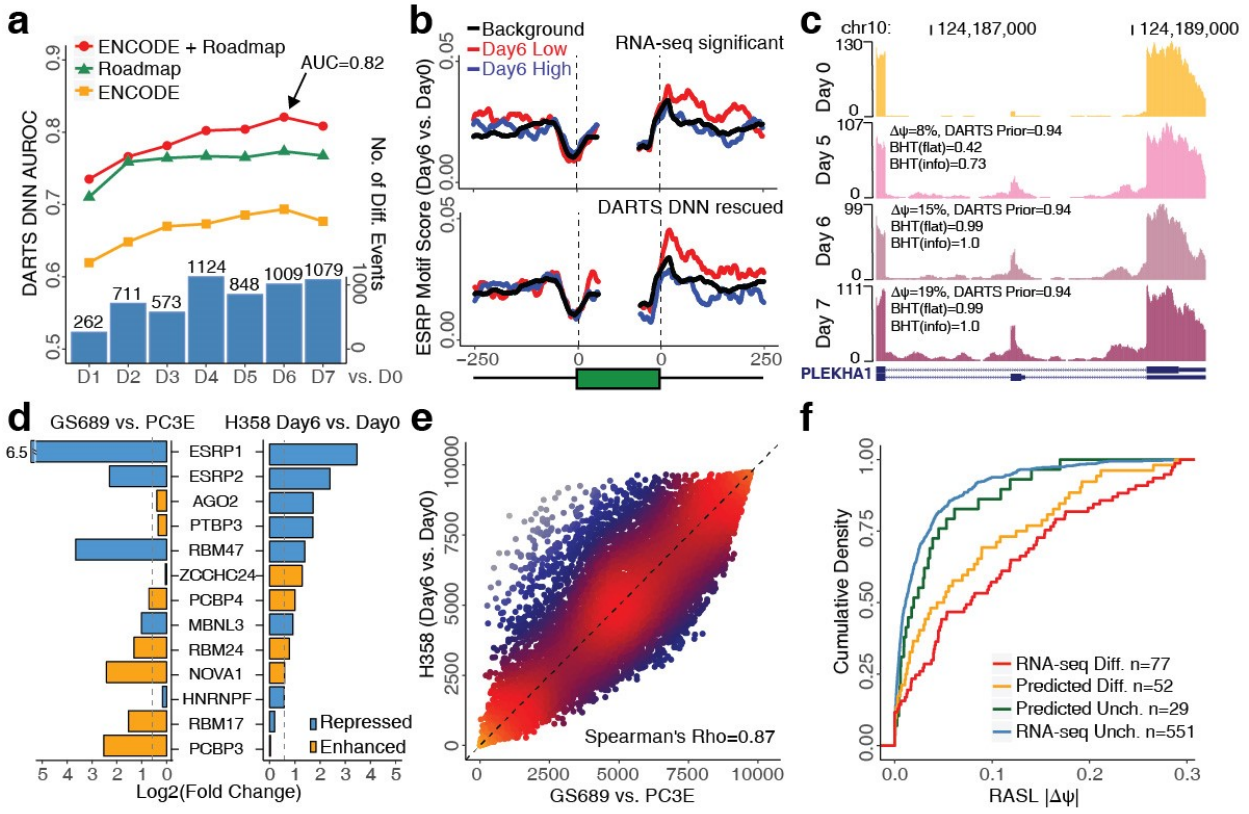
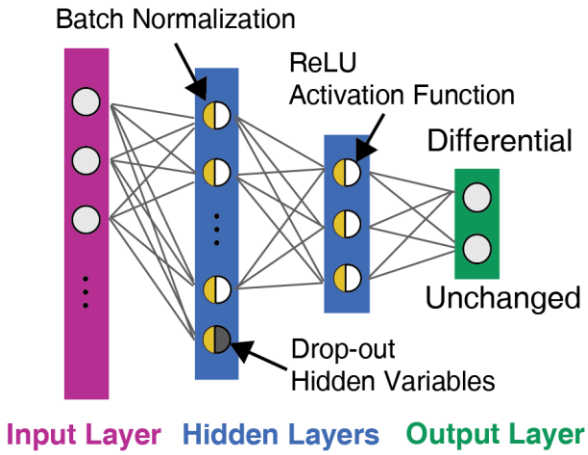


Figure 3.4 DARTS splicing analysis of EMT-associated RNA-seq data.

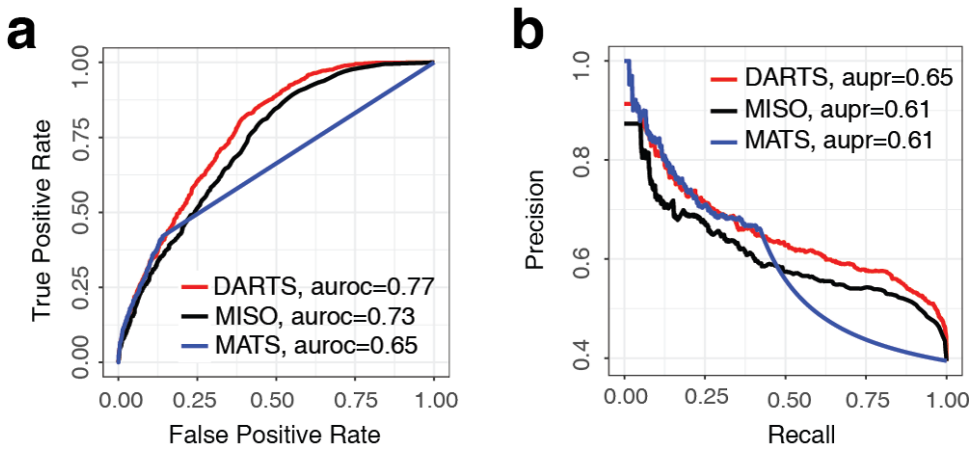
(a) The performance of the DARTS DNN on the H358 lung cancer cell line time-series EMT data. The numbers of differential splicing events called by DARTS BHT(flat) were shown as bar plots at the bottom. **(b)** Meta-exon motif analysis of the known ESRP motif for RNA-seq differential events and DARTS DNN rescued events in the Day 6 vs. Day 0 comparison. In both exon sets, an enrichment of the ESRP motif was observed in the downstream intronic region for exons with low splicing levels in Day 6 as compared to Day 0. **(c)** An example of the DARTS prediction for the PLEKHA1 gene. **(d)** Differential RBP expression signatures in GS689 vs. PC3E prostate cancer cell lines and in H358 Day 6 vs. Day 0. Repressed or enhanced genes have lower or higher gene expression levels in the mesenchymal state, respectively. **(e)** DARTS predictions in the H358 EMT

time-series (Day 6 vs. Day 0) and in GS689 vs. PC3E are highly correlated. Plotted are the ranks of predicted DARTS DNN scores. **(f)** RASL-seq validation of RNA-seq called events and DARTS DNN predicted events. The RNA-seq inconclusive events with high DARTS DNN scores (FPR<5%) had a large change in RASL-PSI values (orange line) compared to RNA-seq inconclusive events with low DARTS DNN scores (FPR>80%) (green line).



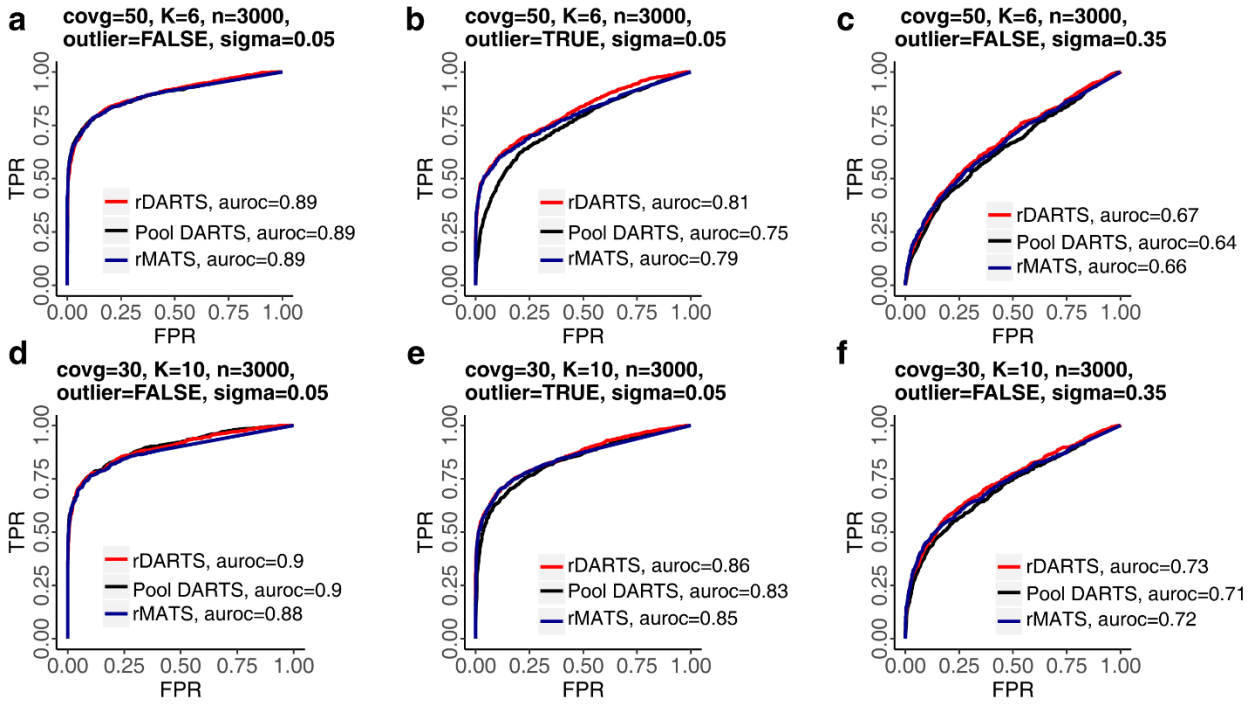
Supplementary Figure 3.5 Schematic overview of the DARTS DNN model.

The DARTS DNN model consists of 4 hidden layers and 7,923,402 parameters. Batch normalization and drop-out of hidden variables are implemented during training to mitigate overfitting.



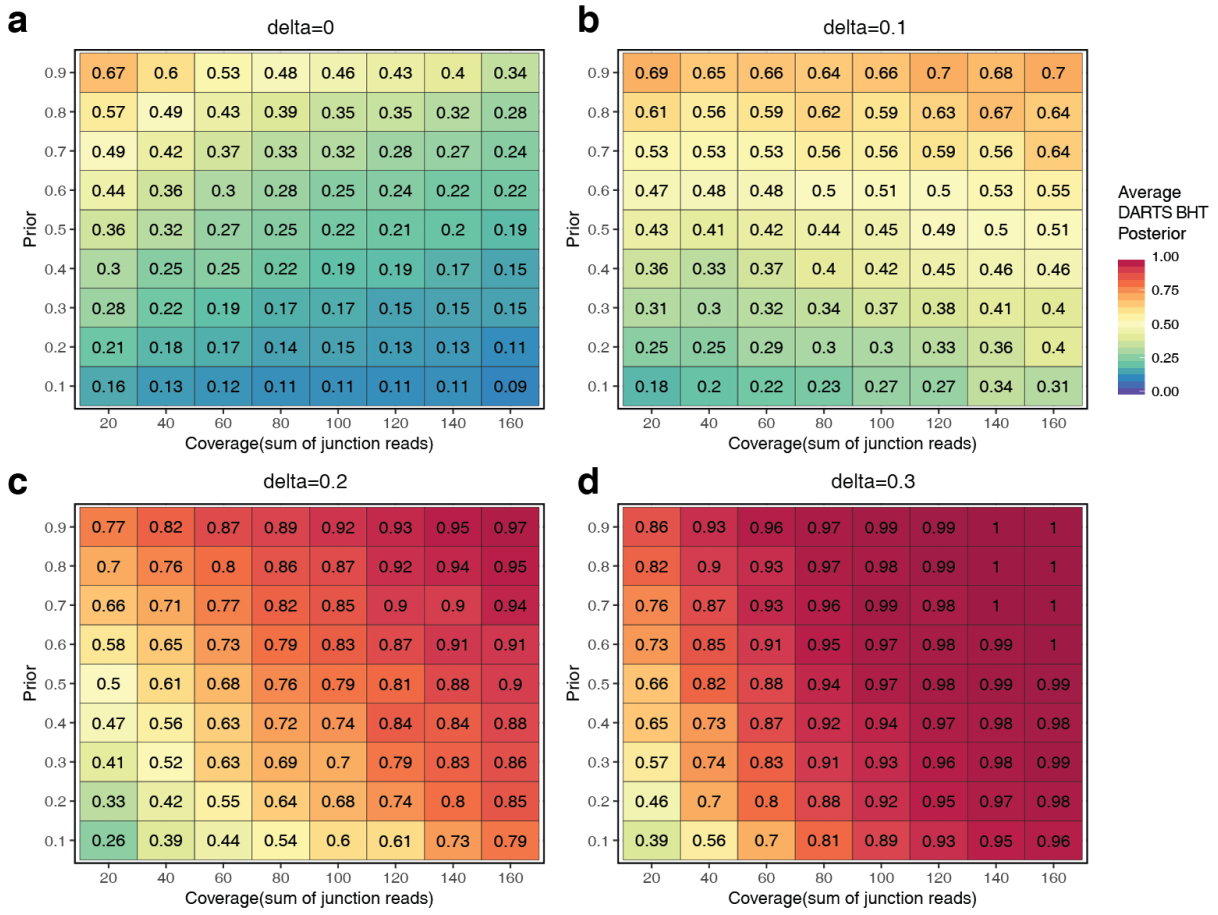
Supplementary Figure 3.6 Performance comparison of DARTS BHT(flat), MISO, and MATS using simulated RNA-seq data generated by Flux simulator.

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



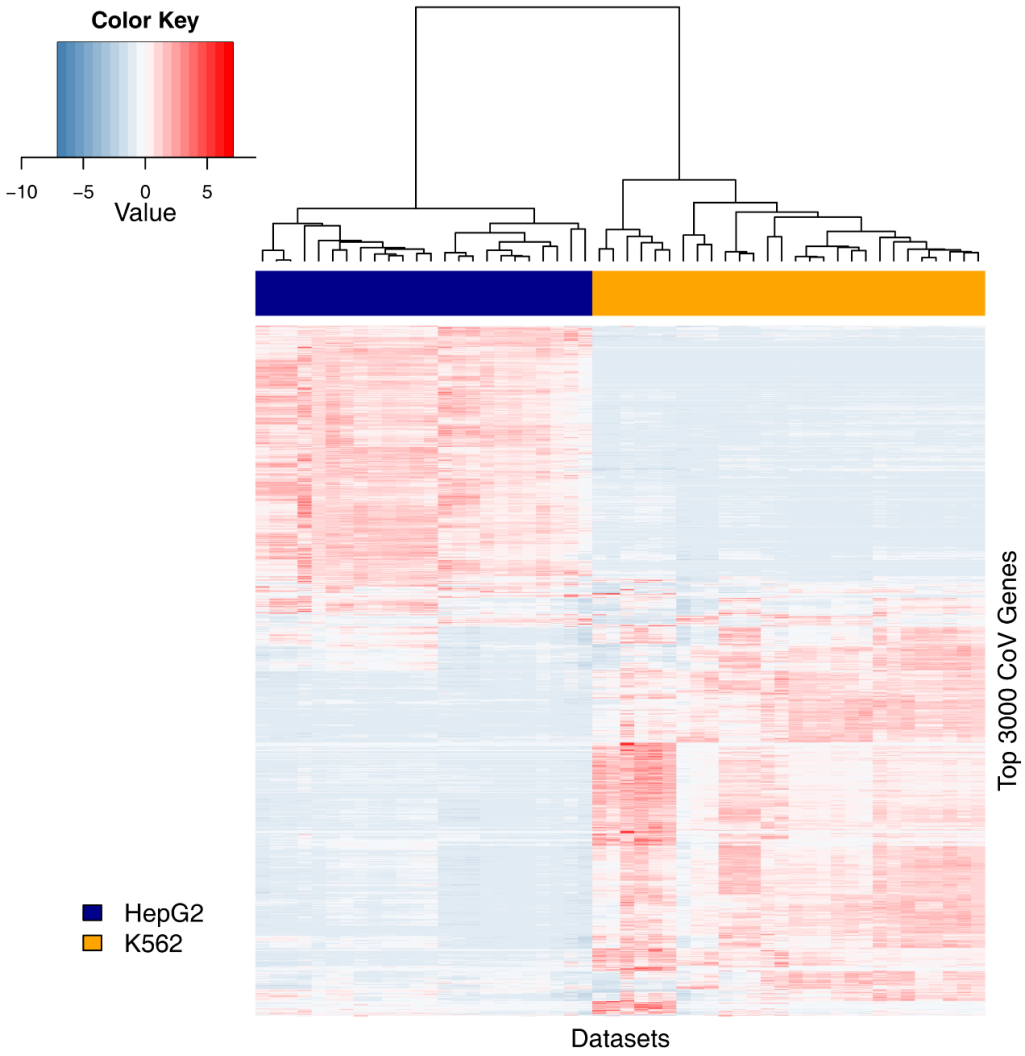
Supplementary Figure 3.7 Performance comparison of DARTS BHT(flat) with replicates versus DARTS BHT(flat) on pooled data and rMATS with replicates.

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



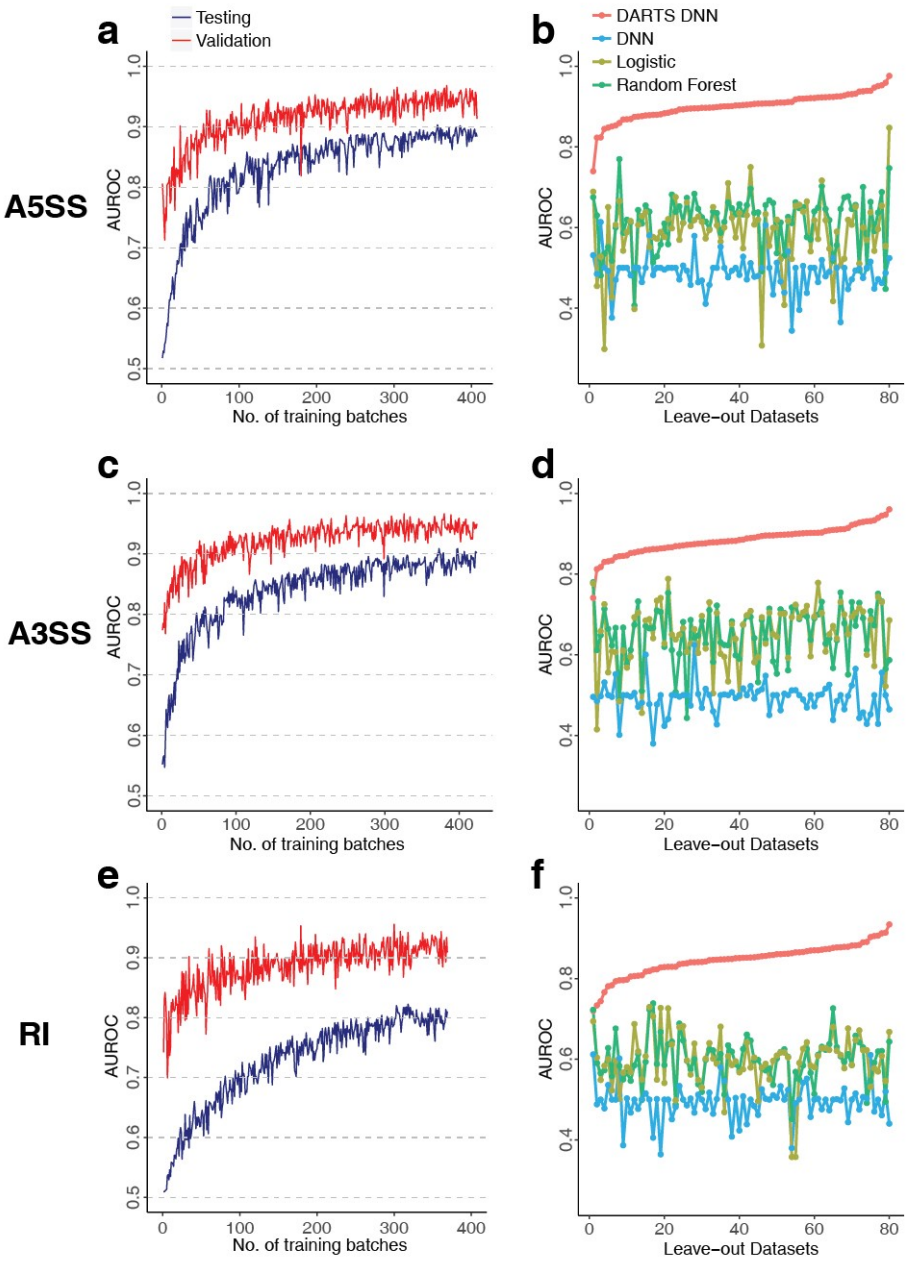
Supplementary Figure 3.8 Relationship of DARTS posterior, prior, and the amount of observed RNA-seq read counts.

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



Supplementary Figure 3.9 Cluster analysis of top 3,000 genes with the highest coefficient of variation (CoV) in gene expression in the ENCODE HepG2 and K562 cell lines.

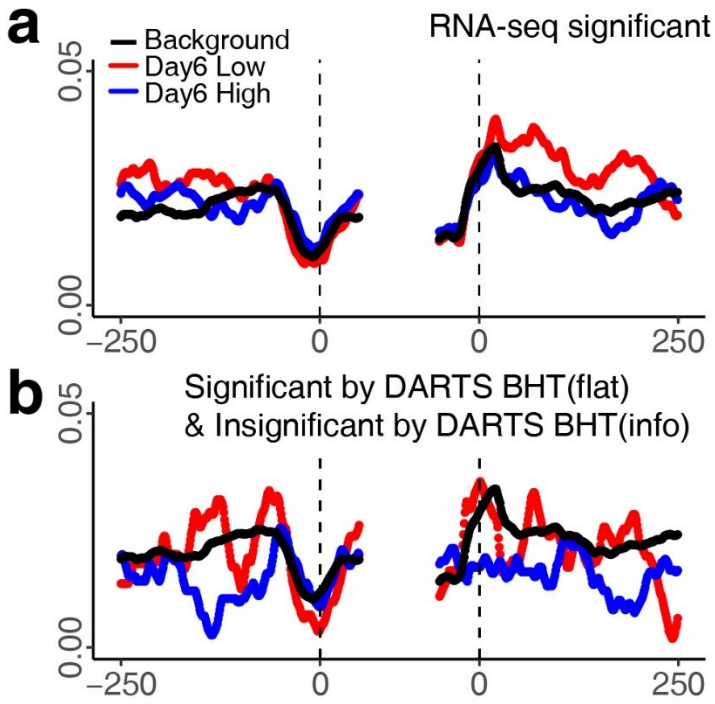
The 24 and 28 biological replicates of HepG2 and K562 clustered into two distinct groups that matched their cell type labels. Plotted for each gene are the normalized Z-scores.



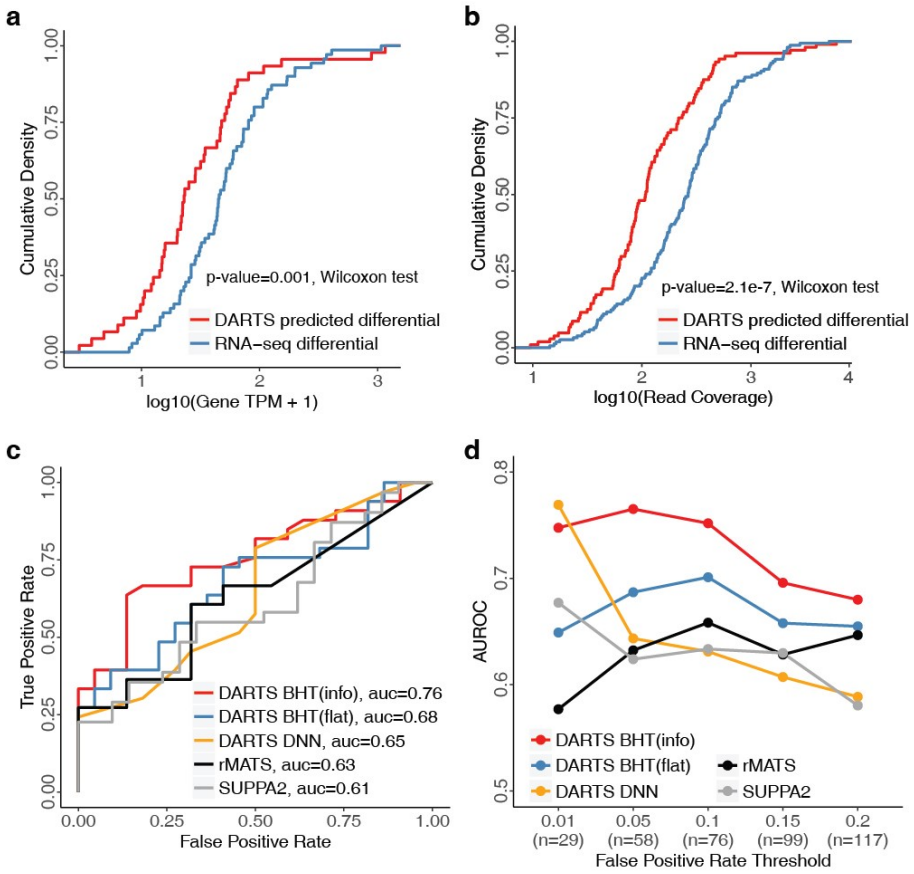
Supplementary Figure 3.10 Application of the DARTS DNN to different classes of alternative splicing patterns.

(a, c, e) The performance of the DARTS DNN on validation and testing data as training progresses for alternative 5' splice sites (A5SS), alternative 3' splice sites (A3SS), and retained introns (RI) as measured by AUROC. **(b, d, f)** Comparison of the DARTS DNN

with baseline methods in independent leave-out datasets. DARTS DNN outperforms baseline methods trained on individual leave-out datasets by a large margin. Note that in these analyses the DARTS DNN is trained using combined ENCODE + Roadmap RNA-seq datasets, with certain pairwise comparisons held-out for benchmarking as independent leave-out datasets.

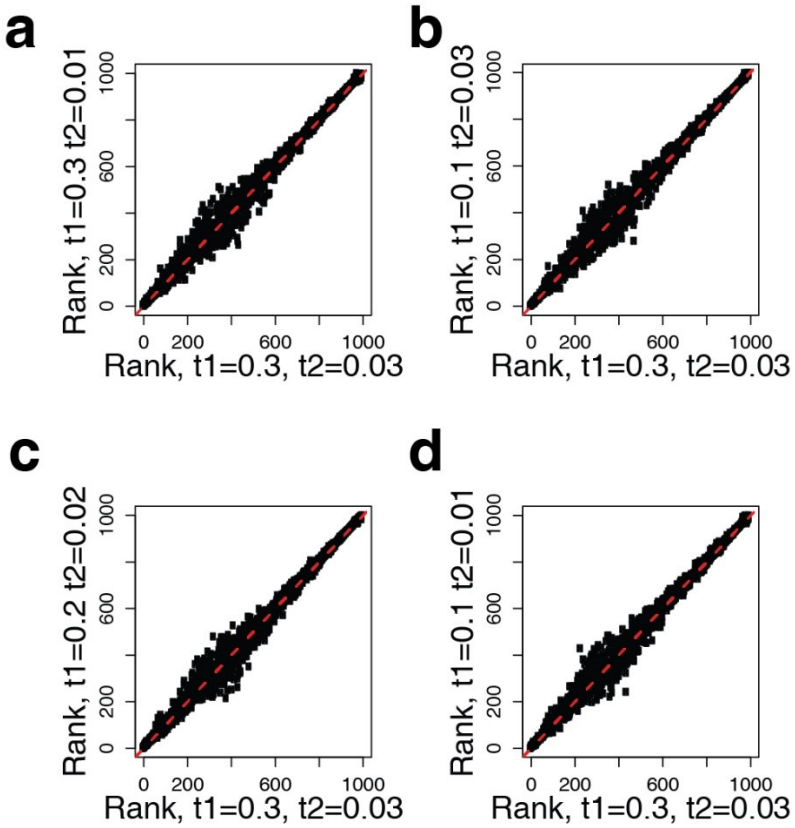


Supplementary Figure 3.11 Additional ESRP motif analysis of DARTS BHT events. Motif scores are shown for **a)** all DARTS BHT(flat) significant events and **b)** DARTS BHT(flat) significant events that become insignificant in DARTS BHT(info). The latter set of events does not have enrichment of the ESRP motif.



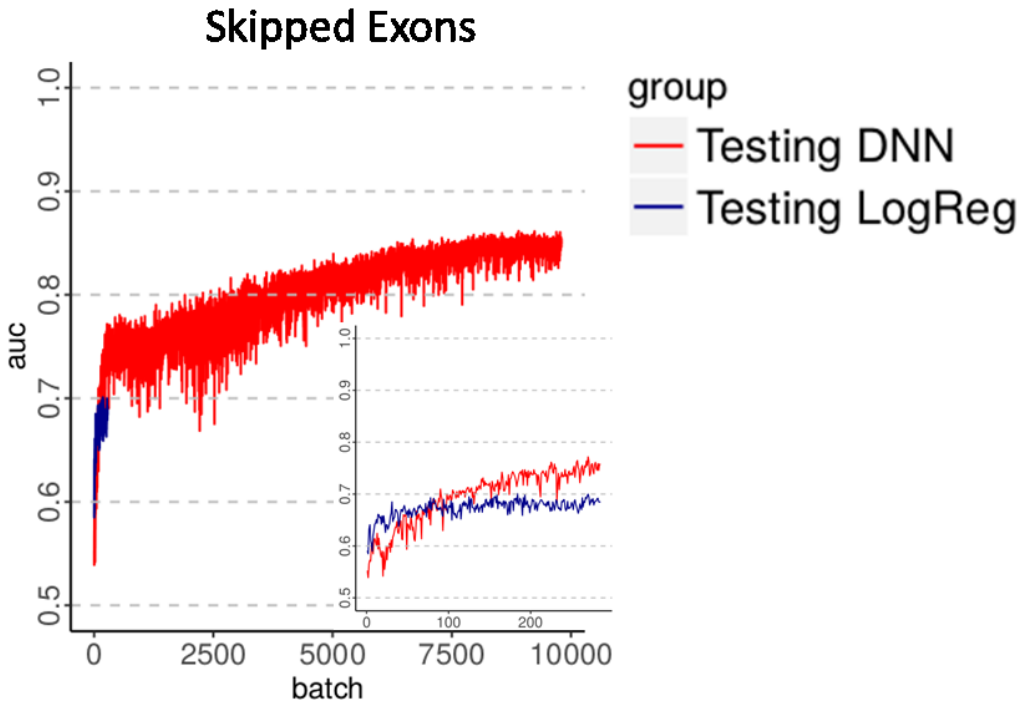
Supplementary Figure 3.12 Characteristics of the DARTS DNN predicted events.

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



Supplementary Figure 3.13 Ranking by DARTS BHT on simulated data when using different t_1, t_2 values.

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



Supplementary Figure 3.14 Testing data performance comparison of DARTS DNN and logistic regression using all ENCODE data.

Inner set is a zoom-in of the performance of logistic regression before it early-stopped.

3.7 Appendix

3.7.1 DARTS BHT statistical modeling

Benchmarking DARTS BHT on simulated data

Generation of benchmark dataset

In order to better represent the variability inherent in real experimental datasets, while knowing the ground-truth, we employed the flux-simulator (1) software (v1.2.1) to simulate RNA-seq reads. Flux-simulator is a specialized simulation software program that models RNA-seq experiments using a set of modules for different experimental procedures, including RNA fragmentation, library preparation and high-throughput sequencing. The major advantage of simulating data using flux-simulator as opposed to directly drawing reads from a statistical distribution is that the former approach takes the variances/noises at different stages into consideration, whereas the latter assumes all reads are generated by a simple stochastic process and are counted correctly; hence, our approach better captures the real-world variances compared to a naïve simulator.

Flux-simulator simulates RNA-seq reads based on a given molecular profile that contains “number of molecules” for each transcript. We derived the molecular profile from a previously published dataset, E-MTAB-1147 from ArrayExpress, which is an RNA-seq experiment of HeLa cell line upon hnRNPC knockdown (2). We chose this dataset because our previous analysis had demonstrated that it contained abundant splicing changes (3), and that its sequencing depth was sufficiently deep to ensure robust estimation of the transcript expression. We used Kallisto (4) (v 0.43.0) to estimate the transcript TPM from the raw reads using Gencode (5) V19 as reference GTF. The

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

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

Evaluation of DARTS BHT, MISO and MATS

We processed the alignment files with rMATS (6) (v4.0.1) to count the junction-spanning reads with Gencode v19 as reference annotation. The inclusion junction counts and skipping junction counts for all detected events were then fed into the DARTS BHT model with a flat prior as input. We ran DARTS BHT with $\tau_1 = 0.3, \tau_0 = 0.03$ and testing for C=0.05. The output of DARTS BHT was subsequently benchmarked using the true delta-PSI values between two conditions.

Note that we only considered simple skipping events in the simulation study, because the complex events are often combinations of multiple alternative splicing events and the true PSI values are often ambiguous to define and hence hard to compute. We define the simple events as events with a unique one-to-one mapping for the 5'- and 3'-splice sites of the middle skipping exon, upstream exon 5'- to middle exon 3'- splice site,

and middle exon 5'- to downstream 3'- splice site. After filtering for simple events, we had 7,678 simple exon-skipping events out of 16,676 exons detected by rMATS.

As a comparison, we also ran MISO (7) (v0.5.3) and MATS (v4.0.1) on the simulated datasets. To run MISO exon-centric analysis, we downloaded the Human genome (hg19) annotation file v1.0 from the MISO website (https://miso.readthedocs.io/en/fastmiso/annotation.html) and built the index for MISO using the Skipping Events (SE) in the annotated folder by "index_gff". Next, we ran MISO to quantify the splicing level under each condition then used "compare_miso" to compute the Bayes Factor for the skipping events in MISO annotation files. We ran the MATS statistical model with setting C=0.05 on the read counts generated by rMATS.

Since MISO analyzes its own internal skipping events annotation which is different from the simple events definition in DARTS and MATS, we took the intersection of the events from these software programs. There were 3,407 common events between the two software programs, with 1,344 events' absolute delta-PSI larger than 5% which we labeled as positive. We measured the accuracy of DARTS BHT, MISO and MATS by AUROC and AUPR. As shown in **Supplementary Figure 3.6**, DARTS compares favorably to MISO and MATS, demonstrating its superior inference power to the state-of-the-art splicing inference tools when using only empirical RNA-seq data.

DARTS BHT statistical model for unpaired or paired replicates

Illustration of replicate DARTS BHT statistical model

Thanks to the rapid development of sequencing technology, it has become practical and common for transcriptomic studies to carry out RNA-seq experiments with multiple

replicates to quantify biological variances and improve reproducibility. Previously we had demonstrated that pooling the reads from different replicates is not recommended (6). Motivated by the replicate analysis, we sought to develop the replicate DARTS model that considers replicates in its likelihood function while still being capable of taking the informative prior into account.

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

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

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

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

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

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

[S1]

I_{ijk}, S_{ijk} and ψ_{ijk} are the inclusion read counts, the skipping read counts and the exon inclusion level for exon i, sample group j=1,2, in replicate k; f_i is the length normalization function for exon i; μ_i is the baseline inclusion level for exon i, and δ_i is the difference of the exon inclusion levels between the two conditions. Without loss of generality, we let $\psi_{i1\text{k}} = \mu_i + \epsilon_{ik}$, $\psi_{i2\text{k}} = \mu_i + \delta_i + \epsilon_{ik}$; that is, we assume that the effect

size δ_i is the same across different replicates; and that ψ_{ijk} values in each replicate k have a random replicate-specific deviation from the group mean μ_i by ϵ_{ik}. The term ϵ_{ik} captures the within group variance of PSI values in different replicates and has an expectation of 0.

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

Simulated read counts and evaluation

Next, we simulated read counts by drawing reads from binomial distributions as in Eq. **S1**. We did not use flux-simulator for this analysis because it is non-trivial to define the within-group variances at the "number of molecules" level; instead, we imposed a normal distribution to the simulated group mean PSI value, then drew read counts from this hierarchical generating process.

We performed extensive simulation studies using different combinations of parameters. Specifically, we set the model parameters equal to the following values: $\sigma \in \{0.05, 0.35\}$, the within group variance, smaller values of σ indicated more consistent patterns across replicates; $K \in \{6,10\}$, the number of replicates, more replicates would

help better capture the within group variance; $n \in \{30, 50\}$, the coverage of each replicate, deeper coverage would help estimation of sample-wise PSI; presence of outlier, outlier PSI value was draw randomly from [0,1] to represent one unrelated sample out of the all replicates. We benchmarked the performances of pooled DARTS, replicate DARTS (rDARTS), and rMATS, using the AUROC and AUPR. To obtain a reliable performance estimate, we randomly sampled n=3,000 events under each simulation configuration, with the expected differentially spliced events (positive cases) at 50%.

As shown in **Supplementary Figure 3.7**, replicate DARTS showed a consistent gain in power under two specific situations, regardless of the number of replicates: i) when the within-group variance σ is large, and ii) when there is an outlier sample. This is consistent with our previous observation in the rMATS paper (6). Notably, in all simulations, we fixed the total coverage at 300, i.e. when K=6, each sample has 50 read counts per event; when K=10, each sample has 30 read counts per event. Such configurations emulate a fixed sample-size budget, where researchers hope to get the best scientific outcomes using the optimal experimental design. It is not surprising that increasing the number of 6 replicates by 4 would significantly reduce the loss of power caused by introducing 1 outlier sample. The same effect was true for larger within-group variances, demonstrating the better group variance estimation captured by more replicates with less coverage per replicate. In all comparisons, the replicates DARTS model outperforms the pooled DARTS model under certain conditions, while inflicting no loss of power under regular conditions. Hence, we recommend using the replicate DARTS model whenever possible, and advise against pooling reads from replicates.

Technical notes on statistical model optimization

Laplacian approximation

The optimization of the DARTS model involves two major steps: i) calculating the Bayes Factor of two competing models/hypotheses, ii) sampling the posterior distribution given the non-conjugate priors. In this part we will first deal with the calculation of the Bayes Factor, where we utilized Laplace's method to approximate the intractable integrals.

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

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

$$\propto \int_{-\infty}^{+\infty} \int_{0}^{1} f_i(\psi_{i1})^{I_{i1}} \cdot \left(1 - f_i(\psi_{i1})\right)^{S_{i1}} \cdot f_i(\psi_{i2})^{I_{i2}} \cdot \left(1 - f_i(\psi_{i2})\right)^{S_{i2}}$$

$$\cdot \mathbf{1}(|2 \cdot (\mu_i + \delta_i - 0.5)| < 1) \cdot e^{-\delta^2/\tau_n^2} d\mu_i\, d\delta_i$$

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

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

[S2]

The above integral cannot be solved in closed form. Instead, we employ Laplace's method to approximate the integral. Let $g_1 = \log g$ be the log posterior density function,

the Laplacian approximation can be viewed as the Gaussian approximation to any (posterior) distribution that is smooth and well-peaked around its maximal point. To implement Laplacian approximation for DARTS BHT, we compute both the maximal point of the posterior probability as well as the local curvature/Hessian matrix around the maximal point using the "optim" function in R by feeding its objective function and the gradient function. Then the approximation for the integral, denoted by Z_n, is

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

[S3]

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

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

[S4]

MCMC sampling

Next we seek to sample from the posterior distribution of the parameters given the data/observations under a specific hypothesis. Since we do not have the conjugate prior for the likelihood, we employ an MCMC random walk to draw samples from the posterior distribution. Specifically, we designed the transition probability q as a normal distribution

with mean equal to the current state and a small variance corresponding to a small step size. For each proposed state, we accept the proposal by a Metropolis-Hasting acceptance probability:

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

[S5]

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

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

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

Justification on different values of τ parameter

In DARTS BHT, the choice of the parameter τ specifies the two competing hypotheses: differential splicing and unchanged splicing between two biological conditions. Here we show that since the final inference is performed on the probability of $P(|\Delta\psi| > c)$ marginalizing over the hypotheses, DARTS BHT is robust to different choices of τ_k. We started with an example by comparing the inference results on a set of simulated splicing events (n=1000) when setting $\tau_1 = 0.3, \tau_2 = 0.03$ (default setting in our paper) with $\tau_1 = 0.4, \tau_2 = 0.02$ (alternative setting here). We observed the ranks of the final inference $P(|\Delta\psi| > c)$ under these two settings were highly consistent (Spearman's rho=0.99), demonstrating the robustness of DARTS BHT to difference choices of τ. Additionally, comparing the actually posterior probability of these two settings, we observed the values were highly similar for $P(|\Delta\psi| > c) \approx 1$, where is the major region of interest for inference of differential splicing. The alternative setting has a negative bias (more conservative) around $P(|\Delta\psi| > c) \approx 0$ due to stronger regularization effect from a smaller $\tau_2 = 0.02$. This will allow users to reflect their beliefs on data quality through the choices of τ as regularization strength. For example, when data is noisy, users would preferably specify the alternative setting over our default setting. To further understand the impact of the parameter τ, we examined another four alternative settings of τ using various combinations of different τ values. Indeed, the inference results are robust in different scenarios, especially for the ranking/inference of differential alternative splicing events (upper right corner of each panel in **Supplementary Figure 3.13**). The model of DARTS BHT is designed to be robust to different specifications as well as flexible enough to account for different dataset-specific requirements.

Running time analysis

The computation of the DARTS BHT model is demanding because of the random sampling of the non-conjugate posterior. Compared to conventional inference methods that only estimate point estimates for the parameters of interest, the DARTS BHT model needs to derive the whole posterior probability distribution using an MCMC sampling. Hence, we re-wrote the MCMC sampler in Rcpp (8). The source code was compiled during the installation of the DARTS R package and the resulting speed gain was around 10 fold. We also tuned the MCMC sampling (see subsection above) to shorten the burn-in period.

In general, the optimized optimization procedure runs in a reasonable amount of time. For the DARTS BHT without replicate mode, an individual event takes 0.23s wall-clock time on average to finish the optimization on an Intel i7-4790 3.60GHz CPU. For the DARTS BHT with replicate, the running time scales linearly with the number of replicates for an individual event. In our benchmarking, an event with 6 replicates takes around 1.38s and an event with 10 replicates takes 2.07s on average.

3.7.2 DARTS DNN machine learning

Sequence feature extraction and normalization

The DARTS DNN cis sequence features are built upon a previous report (9) that curated 1,393 RNA features. Furthermore, we expanded the feature set by including 1,533 additional features on RBP binding motifs and conservation scores. We compiled cis sequence features for four different types of alternative splicing events, i.e. exon skipping, alternative 5’ splice sites, alternative 3’ splice sites, and retained introns. Below we briefly describe all the feature annotations of exon skipping events as an example; the full lists

of all cis sequence features for the four types of alternative splicing events can be found in the GitHub repository.

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

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

2) Nucleosome occupancy scores are computed using NuPoP (10) for the skipping exon and flanking introns. The features are defined as predicting the nucleosome positioning in the first 100 nucleotides of each intron and in the first and last 50 nucleotides of skipping exons.

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

4) We include 111 curated RBP-binding motifs and count motifs in each of the 7 intronic and exonic regions. In addition to the counting procedure, we also download the RBP binding PSSM matrix from RBPmap (11) and calculate the PSSM scores of each RBP-binding profile.

5) We run two different tools, one from Itoh et al. (12), and maxent (13), to estimate the splicing strength between the three exon-exon junctions: C1-C2, C1-A and A-C2.

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

7) The secondary structure score is predicted by the maximum availability of intron regions using RNAfold (14).

8) Short motifs are integrated from Xiong et al. (9).

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

10) ESE (exon splicing enhancer), and ESS (exon splicing silencer) are from Burge's and Chasin's work (15,16). ISS (intron splicing silencer) and ISE (intron splicing enhancer) are from Wainberg's work (17).

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

Although certain classifiers (e.g. tree-based models) are robust to the feature scaling, it is important to scale the features for neural networks. We followed the feature scaling method described previously (9) and divided each feature by its maximum absolute value across all training sets. This rescaled the features to [-1,1] while preserving the zero values, which has specific biological indications.

ENCODE data processing

Extraction of junction counts and detection of novel events

Following the descriptions in the Method section, we had downloaded all the alignment files from the ENCODE data portal (18) and processed the bam files with rMATS. Aside from the annotated events in the reference GTF Gencode v19, rMATS detected novel splicing events where edges not annotated in the GTF splicing graph connect two annotated exons. These novel events consist of a large proportion of our training dataset and are crucial for learning the regulatory code between RBP perturbations and alternative splicing. Note that our definition of novel events are novel edges or junction reads that are not present in GTF; we do not detect novel splice sites or novel exons.

RBP expression estimation

The robust performance of DARTS DNN is dependent on the robust estimation of RBP expression levels, given that all sequence features are static. A previous report has demonstrated that 10 million reads per sample was a good depth for differential gene expression analysis (19), hence we reasoned that the gene expression estimates are fairly robust to reduction in sequencing depth, unlike the exon inclusion level estimates that depends on junction-spanning read counts. In practice we re-analyzed gene expression using Kallisto (v.0.43.0) from raw fastq reads downloaded from the ENCODE data portal. We extracted the TPM of all RBPs from the annotated list. The estimated TPM was subsequently divided by the maximum value across all datasets to rescale it range to [0,1].

Implementation of other machine learning strategies and comparison to DARTS DNN

Logistic regression and Random Forest using individual leave-out datasets

To benchmark the performance of our trained DARTS DNN model to other machine learning strategies, we implemented two baseline methods, Logistic regression with L2 penalty and Random Forest. Because these baseline methods were unable to scale up to big data (see below), they were trained and benchmarked on individual ENCODE leave-out datasets by cross-validation. The identical events with their corresponding labels and features were fed into the baseline classifiers through 5-fold cross-validation and we recorded the performance measured by AUROC in each of the validation sets.

We implemented the two methods using scikit-learn in python. For the logistic regression, we need to tune one parameter, i.e. the penalty strength, or the inverse of the penalty strength C. This parameter controls the complexity of the classifier and hence the severity of overfitting. We chose C=0.1 for our implementation of logistic regression because in practice such a penalty achieves good reasonable generalization over different datasets. Although logistic regression is easy to interpret and a good baseline method for most classification tasks, it cannot effectively detect high-order interaction terms, diminishing its predictive power for such complex tasks.

Another more powerful and robust machine learning strategy we employed as a baseline method was Random Forest. Random Forest is an ensemble learning method where each base classifier is a decision tree that over-fits a set of bootstrapped training samples with a subset of features. The Random Forest classifier has several desirable properties, including being robust to feature scaling and irrelevant features, and being capable of dividing the feature space more flexibly than more conventional partitioning based classification methods. We tuned the hyper-parameter of Random Forest, i.e. the number of trees in the forest. Typically, the more trees in a random forest, the better

predictive power it renders to the ensemble classifier. We noticed that for our datasets, 500 trees achieved the best testing accuracy while increasing the number of trees further did not grant much more gain.

As shown in **Figure 3.1d**, Random Forest almost always outperformed Logistic regression given the same training datasets. We can also observe a positive correlation between the performance of Random Forest and Logistic regression, indicating the internal structure of the training data plays an important role in the learning efficiency, despite the fact that the two learning algorithms are based on dramatically different underlying structures. Nevertheless, DARTS DNN showed superior performance compared to the baseline methods, even though these knock-down datasets have never been trained in DARTS DNN. Furthermore, the performance of DARTS DNN does not show strong correlations with the base learners, indicating its generalization over the single datasets to a more generic regulatory code.

Logistic regression using all perturbation data

In the original publication of the DARTS work, we compared DARTS DNN to baseline methods only trained on individual leave-out datasets using cross-validation. The comparison was inherently biased against DARTS DNN because the leave-out datasets were never trained by DARTS but were cross-validated in the baseline methods. In the meanwhile, there has been questions about how a simpler model (as compared to DARTS DNN) will perform when trained on the entire data, and the contribution of non-additive interactions between the features representing cis-elements and trans-acting factors. As a primer to evaluate the non-linear interactions in the task of differential

splicing prediction, a logistic regression model was fit using the entire ENCODE shRNA knockdown data, with the identical training-validation-testing split as in DARTS DNN. Note that the logistic regression model can be viewed as a special form of neural network with no hidden layers, hence the model is so simple that it is unable to capture the feature-feature interactions. The testing performance of DARTS DNN and logistic regression is shown in **Supplementary Figure 3.14**. The simple logistic regression model's learning power was saturated in less than 500 training batches, and its predictive power was significantly worse than the DARTS DNN model, which has complex non-linear interactions and also requires significant amount of training. Pinpointing the interaction is non-trivial, because the feature interactions are easily entangled and disentangling requires sophisticated modeling and neural architecture searching. Therefore, the interaction detection in differential splicing prediction remains an important future direction.

Technical notes on DNN training

Below we briefly describe some technical details in training the DARTS DNN model using the ENCODE data. DARTS DNN was implemented in Keras with Theano backend. The DNN model was a 4-hidden layer fully-connected neural network with drop-out (with probability 0.6, 0.5, 0.3 and 0.1, respectively) and batch-normalization layers, and each neuron had ReLU (rectifier linear unit) activation function that maps the input vector x to a non-linear output:

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

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

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

We optimized the model parameters using the RMSprop optimizer. RMSprop is a variant of the stochastic gradient descent algorithm, which accounts for the recent momentums of the gradient and adaptively adjusts the learning rate. In our experiments, RMSprop works better than other optimizers for most network architectures.

Because the training dataset was huge and took too much memory (>100G) to be loaded at once, we divided the training samples into different data batches by the knock-down experiments. In each data batch, we randomly picked two different RBP knockdown experiments; due to the way the training datasets were constructed, every RBP selected must have been knocked down in both HepG2 and K562 cell lines. Hence in each data batch, we had at least 4 different datasets, sometimes more if this RBP was knocked-down by more than one shRNA in a certain cell line. The pairing of the same RBP in two different cell lines ensured that there was sufficient variance in the RBP expression features, hence facilitating the classifier to learn from the trans-acting factors.

Next, we mixed the training skipping-exon events from the data batch and held-out 20% of these events as validation set, and the remaining 80% as training set. The training set was then split into positive and negative stacks of cases, and we aimed to construct mini-batches of size 400 to feed into training the model sequentially. Because

the training set was very imbalanced, and the number of negative cases outweighed the number of positive cases, we balanced the composition of each mini-batch by first extracting 100 (25%) positive cases from the positive stack, then compensating 300 (75%) negative cases from the negative stack. Such biased composition of mini-batches will generate the back-propagation of errors from positive cases and reduce strong negative bias caused by the imbalanced data.

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

3.7.3 References for Appendix

1. Griebel, T. et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res 40, 10073-10083 (2012).

2. Zarnack, K. et al. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell 152, 453-466 (2013).

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

4. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525-527 (2016).

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

6. 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 (2014).

7. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7, 1009-1015 (2010).

8. Eddelbuettel, D. Seamless R and C++ integration with Rcpp. (Springer, New York; 2013).

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

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

11. Paz, I., Kosti, I., Ares, M., Jr., Cline, M. & Mandel-Gutfreund, Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res 42, W361-367 (2014).

12. Itoh, H., Washio, T. & Tomita, M. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. RNA 10, 1005-1018 (2004).

13. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11, 377-394 (2004).

14. Bindewald, E. & Shapiro, B.A. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA 12, 342-352 (2006).

15. Fairbrother, W.G. et al. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res 32, W187-190 (2004).

16. Zhang, X.H. & Chasin, L.A. Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18, 1241-1250 (2004).

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

18. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).

19. Liu, Y., Zhou, J. & White, K.P. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 30, 301-304 (2014).

Chapter 4 Concluding Remarks

The past decade has witnessed the fast developments of genomics and biotechnology. High-throughput sequencing methods of DNA and RNA provide global profiling of the molecular landscape and facilitate the discoveries of regulatory basis of molecular and phenotypical variations. Functional genomics assays have characterized and annotated numerous gene and protein functions. Consortia efforts have systematically performed the cellular profiles across diverse perturbations and conditions. Altogether, the end-products of these experimental efforts are massive data of diverse forms.

The ever-growing mountains of data provide unprecedented opportunities for researchers to study the mechanisms of cellular and transcriptional regulation, while also poses significant challenges for making sense of the big data and extract useful knowledge and principles from it. This dissertation has been focused on uncovering novel biological insights from the rapid-accumulating biological data, with a focus particularly on the post-transcriptional regulation and RNA splicing. RNA splicing is an essential biological process that greatly expands the transcriptome diversity. It is regulated by an extensive protein-RNA interaction network involving cis elements within the pre-mRNA and trans-acting factors that bind to these cis elements. Towards understanding the functional and evolutionary impacts of RNA post-transcriptional regulation, computational tools were developed to utilize the big human transcriptomic data as well as answer essential biological questions, with the following two major components.

In Chapter 2, repetitive elements-derived RNA-protein interaction sites and RNA modification sites were studied, shedding new lights in the evolutionary re-wiring of the transcriptional regulatory networks. We developed the first repeat-aware peak caller for CLIP/RIP-seq data called CLIP-seq Analysis of Multimapped reads (CLAM). CLAM uses an expectation-maximization algorithm to assign multi-mapped CLIP-seq reads and calls peaks combining uniquely and multi-mapped reads. To demonstrate the utility of CLAM, we applied it on hnRNPC iCLIP and m6A RIP-seq datasets and showed that CLAM is capable of uncovering novel RNA regulatory sites that are inaccessible by existing methods.

Since the initial publication and release of the CLAM software, we have closely kept the maintenance and update of CLAM. The probabilistic re-aligning module was updated with a read-tagging function to better capture the cross-linking sites in different CLIP experiments (e.g. the base transition events in HITS-CLIP vs truncation events in iCLIP). The peak calling module in the initial development is based on a permutation test, which is unable to account for uncertainties from the paired control experiments. As reported in the eCLIP benchmarks, the inclusion of the Size-Matched Input control experiment can significantly improve the false positives in peak detection. Hence, we developed a new negative binomial model-based peak calling module in CLAM for better analyzing eCLIP and RIP-seq data. We showed that the CLAM peak caller is both sensitive and robust, especially when running the multi-replicate mode to aggregate replicate information. The benchmarking results of the new modules can be found in the appendix of Chapter 2.

In Chapter 3, we aimed to improve the quantification and analysis accuracy for RNA splicing and relieve its over-reliance on ample RNA-seq coverage, a major limitation of

existing tools encountered during the development of CLAM. Hence, we developed a Deep-learning Augmented RNA-seq analysis of Transcript Splicing (DARTS) framework by leveraging big data in the public domain, and demonstrated its superior performance, especially when the observed data is limited or biologically the gene of interest is lowly expressed. DARTS uses big-data powered deep learning to augment any user-specific RNA splicing analysis, representing a major advance over existing computational tools.

The conceptual innovation in DARTS is arguably more profound than the analyses. Prior to DARTS, the majority of deep learning applications in genomics are based on genomic sequences. For example, hand-made features based on exonic and proximal intronic sequence were used to predict exon inclusion levels (Xiong et al. 2015). More recently, raw genomic sequences were fed as inputs in convolutional neural networks to predict transcription factor binding, chromatin states (Zhou and Troyanskaya 2015) as well as cryptic splice sites (Jaganathan et al. 2019). In addition to genomic sequence features, DARTS used the molecular experimental measurements, i.e. gene expression estimates of the RBP genes, from the same RNA-seq data as features. This design scales up with massive data, and is readily extendable from transcriptomic to multi-omics analyses.

Furthermore, a natural extension for DARTS is to study the interactions between the genomic sequence motifs and RBP gene expression levels. In a broad sense, DARTS modelling of the cis- genomic sequences and trans- gene expression is similar to the genetic-environment interaction modelling. By definition these interactions are highly non-linear and cannot be fit by simple linear models. To understand the contribution of these non-linearity in DARTS model, a simple logistic regression without any hidden layers was

fit using all exon-skipping events and we observed it could not perform nearly as well as DARTS DNN. These results are in the appendix of Chapter 3. Hence, a somewhat missed opportunity in DARTS development is that we were not able to pinpoint the interactions between cis- and trans- features. This of course is non-trivial and will be a future direction.

Moving forward, a unified and interpretable cellular model will connect multi-omics functional data in bulk tissues and single cells, and ultimately provide guidance in clinical applications and therapeutics developments. With the development of technologies in the next decade, pure predictive modeling will be less urgent in demand, as most measurements can be easily performed at affordable costs; instead, the knowledge we learned from the mountains of data, and the data-driven principles provided by that knowledge to help individual studies, will only be more prevailing.

Deep learning provides an ideal workhorse in this realm, given its scalability to massive data, as well as its flexibility to diverse data types. Harnessing the power of massive functional data with quantitative modeling, we can explain and predict complex biological traits better than ever before. Computational biologists should rethink how to design models stemmed from biological regulatory mechanisms, fully utilized the functional assays and the power of technologies we already process, and shed new lights on biological knowledge discoveries as well as new experimental designs.