

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

A probabilistic approach to studying alternative splicing in single cell data

Permalink

<https://escholarship.org/uc/item/4b72343b>

Author

Buen Abad Najar, Carlos Fernando

Publication Date

2021

Peer reviewed|Thesis/dissertation

A probabilistic approach to studying alternative splicing in single cell data

By

Carlos F Buen Abad Najjar

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Liana Lareau, Co-Chair

Professor Nir Yosef, Co-Chair

Professor Sandrine Dudoit

Professor Donald Rio

Spring 2021

Abstract

A probabilistic approach to studying alternative splicing in single cell data

by

Carlos F Buen Abad Najar

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Liana Lareau, Co-Chair

Professor Nir Yosef, Co-Chair

Alternative splicing is a post-transcriptional gene regulatory process that increases transcriptomic diversity through the differential selection of exons. Over the past decade, single cell RNA sequencing (scRNA-seq) experiments have allowed the study of the complex heterogeneity of tissues, and the gene expression regulation mechanisms that determine the identity of individual cells. However the study of alternative splicing has largely lagged behind in single-cell approaches. This is in part due to a number of technical limitations that make this study particularly difficult. In this dissertation, I first demonstrate that the technical distortion of splicing observations in single cells produces the appearance of bimodality in single cell splicing. I discuss how this limitation has affected our understanding of single cell splicing, and what computational approaches can be used to extract meaningful biological information from this exceptionally noisy data. I then present Psix: a computational tool that implements an updated model for single cell splicing observations that takes into account this distortion. Using Psix, I identify alternative splicing that changes across a phenotypic landscape, as well as groups of potentially co-spliced exons. Finally, I explore potential regulatory roles for these poison exons in the production of Serine/Arginine-rich (SR) splicing factor proteins.

Para mi esposa Ariadna

Por Rosario

Y así, del poco dormir y del mucho leer, se le secó el cerebro, de manera que vino a perder el juicio.

El Ingenioso Hidalgo Don Quijote de la Mancha
- Miguel de Cervantes Saavedra

Acknowledgments

Thank you first and foremost to my beautiful and brilliant wife Ariadna. You are my greatest inspiration and my reason for doing everything that I do. Thank you, from the bottom of my heart, for all the love, patience, devotion, and support through the years. You have given me the happiest moments of my life, and you have supported me through the toughest. I dedicate my PhD to you.

Thank you to my advisors Liana Lareau and Nir Yosef, for your guidance, mentorship, unwavering support and counsel both in the lab and outside of it. You believed in me and in our work, even in the times when I did not, and for that I will always be grateful. You are the best mentors I have ever had. Special thanks to Nick Ingolia, for sharing lab space and group meetings, and for being an unofficial mentor during my PhD. Thanks to Sandrine Dudoit and Don Rio for their insight and commentary at the committee meetings.

Thanks to the person that I admire the most in the world: my father. You inspired my intellectual curiosity, and who taught me how to be a man of virtue. Thank you to my mom, for her infinite and unconditional love, tenderness and support. I would not be the man that I am without you. Thanks to my little sister Mony, for all the sweetness and sisterly love, and for the legendary endurance of all the teasing from her big brothers. Thanks to my sister-in-law Steph, for loving and supporting my brother, and for taking such good care of my sobrino and ahijado Manuelito, the joy of the family. And of course, thanks to the Patron Saint of older brothers: Manuel, my absolute best friend. Because your faith, guidance, support, and fraternal love in the toughest times have helped me get through life. Thanks as well to all the Buen Abad and the Najar family: to el abuelo in Heaven, and to Ticha. To my abuelo Don Ciri3n in Heaven, and to abuelita Elena. Thanks to my aunts, uncles and cousins, especially to la t3a Rosalba, for all her support, and for welcoming when I came to California. Thanks also to my in-law family: to Don Alfredo and Do3a Ariadna in Heaven, to Alf and Mar3a, and to the many aunts, uncles and cousins of the Murgu3a-Berthier family, for welcoming me with open arms.

Thanks to the community of the St Joseph the Worker Church in Berkeley, and to the community of the Five Wounds Parish in San Jose. Especially thanks to Father Antonio, for your spiritual guidance and friendship for me and my wife. Thanks also to Gabriel, and all the volunteers of St Isabel's Kitchen, for giving me the opportunity to serve our community.

Thank you to the current and former members of the Lareau, Yosef and Ingolia labs, for all the support and fun through the years. Thanks to my great friend Marek Kudla and to his family, for the invaluable friendship, support, and insight into both machine learning and family life. Thank you for the help assembling the desktop computer,

and also for teaching me how to pronounce the names of the characters in *With Fire and Sword*. Thanks to Jim Kaminski, Rob Tunney, and David DeTomaso, for the help with all my questions about science and about life. Thanks to Alejandro Padron, for the friendship, the vivid discussions, and for his advocacy for intellectual freedom. Thanks to Chenling Xu for the friendship through class and research. Thanks to Amanda Mok, for the exciting scientific discussions and incredible insights in the lab. To Karinna Vivanco, for taking the torch on the poison exons project. Thanks to Nick McGlincy for sharing his knowledge of RNA, and to Wren Kim for teaching me about neurons. Thanks to the single cell crew: Xiuwei Zhang, Michael Cole, Matt Jones, Romain Lopez, Adam Gayoso, Nick Everetts and Allon Wagner, for all the helpful discussions at lab meetings, and for enduring my incessant questions on Slack. Thanks to Alex Khodaverdian for his help in classes, with algorithms, and for co-teaching CS176. Thanks to Helen, Vikram, Maria, Anna, Monica, Nicki, and to the Ingolia lab members, for making the lab so much fun. Especially thanks to Prakruthi Burra, for the contribution to my research.

I would be remiss if I did not extend my thanks to the friends and companions outside of Academia that have made the last six years of my life so enjoyable. Thanks to my amazing friend, and brother from another mother, Utsav, for all the great times we spent together, for the support, for the guidance, for bringing kilos of Indian sweets, and for manning the wheel when sailing the seas for treasure. Thanks to my good friend Lalo, who walked with me through all these years (and football seasons) all the way from Mexico City. Thanks to the Judo community in the Bay Area, including Cal Judo, Do Martial Arts, and SJSU Judo. Especially thanks to Dr Bertsch and Chishiro, for helping me to finally get my black belt. Thanks to my good amigo Jorge, to the fierce Chris Moulton, to his lordship the Count Maxime de Kouchkovsky, to “cool Maxime” Langevin, to Daví, to Ivan, and to Father Vitaly, for the great rivalries on the mat, and for the friendships outside of it. Thanks also to Jackie and to Victoria for administering the Dojo. Thanks also to Devo and to everyone in Zpaz Town, even when we never met in person, because the virtual gaming sessions were an important part of my PhD life.

Finally, thank you to Conacyt and UC Mexus for the financial support, and to Susana Hidalgo and Veronica Sandoval for all their help and support.

Contents

Introduction	1
1 Coverage-dependent bias creates the appearance of binary splicing in single cells	6
1.1 The modality of alternative splicing in single cells	7
1.2 Single cell binary splicing observations are inversely correlated with read coverage	10
1.3 Simulations of RNA sequencing reveal technical sources of distortion of splicing estimates	11
1.4 Accounting for mRNA recovery improves analysis of alternative splicing	14
1.5 Discussion	19
1.6 Methods and Materials	20
1.6.1 Analysis of single cell RNA-seq datasets	20
1.6.2 Linear regression on transcriptional burst kinetics	26
1.6.3 Theoretical analysis of the observed $\hat{\Psi}$ with limited capture rate	26
1.6.4 Probabilistic simulator of splicing in single cell data	29
1.7 Chapter appendix: supplementary tables and figures	35
2 Identifying cell-state associated alternative splicing events and their co-regulation	45
2.1 Cell-type associated alternative splicing	46
2.2 Psix: a probabilistic approach to cell-state associated alternative splicing	47
2.3 Validation of Psix	48
2.4 Finding modules of co-regulated splicing in single cells	50
2.5 Chapter appendix: supplementary tables and figures	67
3 Regulation of SR poison exons in neurogenesis	72
3.1 Background: Nonsense mediated decay and SR poison exons in neurogenesis	72
3.2 Increased poison exon usage in SR genes during neurogenesis	74
3.3 A network of SR splicing factor auto and cross-regulation through poison exon splicing	76
3.3.1 A CLIP-seq data metaclassifier	76

3.3.2	Poison exon binding maps	77
3.4	Supplementary figures and tables	80

Introduction

Alternative splicing is a process in which differential selection of exons results in the production of different RNA molecules from the same gene [1]. This results in an increased functional diversity of the transcriptome. E.g., some alternative splicing events might add or remove protein domains, leading to differences in protein function, or in protein transport and location [2]. Other events might result in unproductive mRNA molecules, resulting in a post-transcriptional layer of gene regulation [3–6]. A common approach to measuring alternative splicing is to estimate usage of a particular alternative exon in mRNAs of a gene. Exon usage is quantified as percent spliced-in (Ψ), which is the proportion of RNA molecules that include the exon, out of all RNA molecules expressed from the gene. The majority of mammalian genes are alternatively spliced, and the molecular mechanisms through which this process is regulated have been the subject of extensive research [7]. Such mechanisms extend from cis-regulatory sequences that regulate the recruitment of the splicing machinery, to trans-acting protein factors that bind to those sequences to promote or inhibit the inclusion of exons, and changes in chromatin state [8]. Alternative splicing plays a crucial role in diverse biological processes, such as neurogenesis, immune response and the cell cycle [9, 10]. As a result, disruption of this process or its regulation is known to be associated with complex diseases such as cancer, autoimmune disorders, amyotrophic lateral sclerosis, and more [11–17]. For this reason, studying the dynamics of alternative splicing regulation is paramount for our understanding of the biology of multicellular organisms.

Mechanisms of splicing regulation

Splicing of pre-mRNA is conducted through a complex of small nuclear ribonucleoproteins (snRNP) known as the spliceosome. This complex is formed by five snRNPs (U1, U2, U4/6 and U5), which recognize the splice sites and catalyze splicing. The assembly, catalysis and disassembly of the spliceosome is required for the removal of each intron, and it is regulated by a plethora of cis and trans acting elements [7].

The cyclic process of assembly and disassembly of the spliceosome is regulated by a variety of mechanisms that coordinate the alternative splicing patterns. These

mechanisms can be positive (enhancers) or negative (inhibitors), and they range from cis-acting elements to trans-acting splicing factor proteins and RNAs. Cis-acting elements are sequences that can act as enhancers or silencers of splicing, and they can be located either on exons or introns. These sequences are recognized by trans-acting elements such as non-spliceosomal RNA-binding splicing factor proteins, which subsequently recruit and coordinate the assembly of the spliceosome. These protein splicing factors can be loosely classified as Serine/Arginine Repeat (SR) proteins, heterogeneous nuclear RNP (hnRNP), and tissue specific splicing factors. Tissue specific splicing factors can act as either enhancers or repressors, while SR splicing factors usually bind to enhancers and promote splicing, and hnRNPs usually bind to silencers and repress it. However, protein splicing factors can play different roles as activators or repressors of splicing, often depending on other factors such as the cis-acting elements and the pre-mRNA position on which they bind. It has been suggested that as a general rule, the binding of splicing factors in the upstream intron promotes exon skipping, while binding in the downstream exon promotes exon inclusion, but this is not always necessary the case, as upstream binding of certain splicing factors reportedly promotes exon inclusion [18]. Histone marks and chromatin accessibility has also been observed to be linked to alternative splicing patterns. Histone modifications can alter the speed of pre-mRNA transcription by the RNA polymerase II, which influences co-transcriptional splicing. Adapters linked to specific histone marks, as well as other elements such as chromatin binding/remodeling proteins, can also influence the recruitment of splicing factors. Transcription associated histone marks such as H3K4me3 interact with proteins that recruit spliceosomal snRNPs, while repressors such as H3K9me3 and H3K27me3 can slow the transcription rate, modifying alternative splicing patterns.

Alternative splicing regulation in neurogenesis

Alternative splicing is highly prevalent and more evolutionary conserved in the nervous system of vertebrates than in other tissues [19, 20]. Transcriptome profiling and the detection of protein-RNA interactions with high-throughput sequencing have advanced our understanding of neural-specific splicing networks, and their role in the proper differentiation, maintenance and function of neurons [9, 18, 21]. During neurogenesis, a number of neural-enriched splicing factors are activated at specific stages during neural differentiation, and often regulate, enhance or antagonize the effects of each other [9, 18]. Extensive studies have revealed four families of splicing factors as key regulators of alternative splicing in neurogenesis: Neuro-Oncological Ventral Antigen (NOVA) [22]; Polypyrimidine Tract Binding Protein (PTBP) [23, 24]; RNA-Binding Protein, Fox Homolog (RBFOX) [25]; and Muscleblind Like Splicing Regulator (MBNL) [26]. Proteins from these families regulate alternative splicing at specific times during brain development, and their activity is characteristic of neuronal type and developmental stage [27, 28]. The misregulation and mutations in

splicing factors have been linked to defects in the proper development and function of the nervous system [29, 30].

Next generation sequencing methods for alternative splicing

Next generation sequencing of RNA (RNA-seq), is a technology that profiles entire transcriptomes, allowing us to observe and quantify the transcript diversity of biological samples [31, 32]. The rapid popularization of this technology has provided an unprecedented framework for the study of alternative mRNA isoform usage in eukaryotes. This has aided to the discovery of new mRNA isoforms [33–35], and to the quantification of isoform and exon usage across different tissues with unprecedented accuracy [36–39]. Through a combination of computational and experimental developments, RNA-seq has also allowed us to uncover the mechanism of alternative splicing regulation at high detail [22, 25, 27]. Further development of these technologies over the last decade permitted the profiling of transcriptomes at the level of individual cells through single cell RNA sequencing (scRNA-seq). Such an approach provides a powerful insight into the factors that determine each cell’s unique identity [40–42]. The wealth of information of scRNA-seq also harbors several challenges. The small amount of RNA present in each individual cell, for example, results in extremely low coverage compared to bulk RNA-seq experiments, creating high technical noise and making it difficult to study differential gene expression and isoform variation [43, 44]. To meet these challenges, intense computational efforts have been made to extract meaningful biological information about gene expression and its regulation from this data. These efforts have improved the resolution of studies of continuous biological processes such as cell differentiation [45, 46], gene regulation [47], and each cell’s biological identity [48–50].

Despite the enormous potential of scRNA-seq approaches, the study of alternative splicing in single cells remains relatively underdeveloped due to technical limitations, such as limited sequencing coverage and low capture efficiency [51–53]. Moreover, many single cell methods that rely on unique molecular identifiers (UMI) can only cover splicing events occurring in the last 100-800 nucleotides of a transcript [54]. This precludes the study of many alternative splicing events of interest that occur upstream from the read coverage. Short-read single cell methods such as smart-seq2 provide full-transcript coverage, as well as better capture efficiency. As a result, these methods have been used in many single-cell alternative splicing studies [55–57]. However smart-seq2 data is still noisy, and the lack of UMIs makes mRNA quantification challenging [53]. An unexpected yet recurrent observation in scRNA-seq studies is that the majority of alternatively spliced exons show a bimodal distribution of Ψ in the data: most cells seem to produce only isoforms including the exon, or excluding the exon, but they rarely produce both at the same time [40, 56, 58]. These ob-

servations led to the conclusion that alternative splicing is truly bimodal in single cells, and thus to the development of computational tools that characterize changes in alternative splicing as mere qualitative changes in the empirical distribution of Ψ [56, 57].

Bias analysis and computational methods for single-cell alternative splicing

In this dissertation, we first demonstrate that the appearance of bimodality in single cell alternative splicing is almost entirely the result of technical limitations [59]. Chapter 1 addresses how the aforementioned coverage-dependent biases distort the observed distribution of Ψ . This distortion results in the inflation of the proportion of observed extreme values ($\Psi \in \{0, 1\}$), leading to the appearance of bimodality in the observed Ψ distribution of exons. We show evidence that single-cell splicing observations derived from exons in highly expressed genes tend to show unimodal distributions, and constitute more reliable splicing observations. We also show that datasets with higher capture efficiency present a higher proportion of unimodally distributed exons. We propose using the number of captured mRNA molecules as an indirect measure of the quality of splicing observations in single cells.

A major goal in alternative splicing studies is to find exons that present different alternative splicing programs across different biological conditions. This approach has been especially challenging at the single-cell level due to the technical distortion of the data. Our study of the coverage-dependent biases showed that alternative splicing can be reliably observed in high-quality single-cell observations. However, the quality restriction only allowed us to study a handful of exons per dataset. This limitation leads to another question: How can we extract meaningful biological information from highly distorted splicing observations? In Chapter 2, we present Psix: a computational tool inspired by autocorrelation that selects exons with significant biological variance and correlation across a transcriptomic space of single cells, without being misled by the coverage-dependent biases. Using Psix allows us to study the regulation of alternative splicing at the single cell level.

The work presented in this dissertation bridges the gap between powerful experimental approaches to study transcriptomes, and the study of an important biological source of transcriptome diversity. Our analysis of the technical sources of variance in single-cell splicing data will give biologists a better perspective on the challenges and possibilities for the study of splicing at the individual cell level. With the development of Psix, we provide a framework for the study of autocorrelated exons in single cells. We also lay the ground for a systems biology approach to the study of alternative splicing in single cells.

Alternative splicing systems biology

Understanding the regulatory interactions between genes is a key step in understanding a biological process. Regulatory network models are built to capture the dependence relationships between molecular entities to represent the dynamics of biological processes. Probably the most widely studied regulatory networks are systems of transcriptional regulation. These studies have benefited greatly from high-throughput sequencing approaches such as the profiling of protein-DNA interactions with ChIP-seq, and transcriptome profiling with RNA-seq, which have allowed to study the dynamics and temporal changes of continuous biological processes such as cell differentiation [60]. The subsequent development of single cell approaches improved the resolution needed to describe complex multicellular systems and the mechanistic details that underlie cell identity and differentiation, but these advances have also come with new challenges that demand the integration of experimental and computational frameworks [41, 42, 61]. The study of splicing regulatory mechanisms has also benefited from the integration of high-throughput data has also been possible. An important high-throughput experimental technique that has particular relevance for the study of splicing regulatory mechanisms is cross-linking immunoprecipitation followed by sequencing, or CLIP-seq. This approach combines UV cross-linking of proteins and RNA with immunoprecipitation to generate precise footprints of the interactions between RNA and RNA-binding proteins [62].

The final Chapter of this dissertation follows from the observation in Chapter 2 that the highly conserved poison exons in SR genes [4] are regulated in neurogenesis. We develop a machine learning method for integrating the results of multiple classifiers on CLIP-seq data and predict splicing factor binding at a nucleotide resolution. We identify potential cis-regulatory regions in SR poison exons. We propose that the highly conserved SR poison exons are a post-transcriptional mechanism of self-regulation. Through the inclusion of their own poison exons, splicing factors of this family regulate their own protein production.

Chapter 1

Coverage-dependent bias creates the appearance of binary splicing in single cells

Publication and authorship permissions

This work was published in eLife in June 29, 2020 as:

Buen Abad Najar CF, Yosef N, Lareau L. *Coverage-dependent bias creates the appearance of binary splicing in single cells*. eLife 2020, 9:e54603.

The co-authors of this paper granted me their authorization for its reproduction in this dissertation.

Single cell RNA sequencing provides powerful insight into the factors that determine each cell's unique identity. Previous studies led to the surprising observation that alternative splicing among single cells is highly variable and follows a bimodal pattern: a given cell consistently produces either one or the other isoform for a particular splicing choice, with few cells producing both isoforms. Here we show that this pattern arises almost entirely from technical limitations. We analyze alternative splicing in human and mouse single cell RNA-seq datasets, and model them with a probabilistic simulator. Our simulations show that low gene expression and low capture efficiency distort the observed distribution of isoforms. This gives the appearance of binary splicing outcomes, even when the underlying reality is consistent with more than one isoform per cell. We show that accounting for the true amount of information recovered can produce biologically meaningful measurements of splicing in single cells.

1.1 The modality of alternative splicing in single cells

Single-cell RNA sequencing (scRNA-seq) has provided impressive temporal resolution to our understanding of continuous biological processes such as cell differentiation [41, 42]. It has uncovered hidden heterogeneity among cells and exposed the factors that determine each cell’s unique identity. One broad source of transcriptomic diversity is alternative splicing, and several studies have uncovered compelling evidence of changes in alternative splicing among single cells during differentiation [55, 56, 63–65].

A particularly surprising conclusion of several scRNA-seq studies was the observation that splicing was often bimodal among supposedly homogeneous cells [40, 52, 56, 58]. Individual cells had binary outcomes in splicing: some cells always spliced in a particular cassette exon, and some cells never spliced in the exon, but few cells showed truly intermediate inclusion within one cell. This unexpected result contrasted with previous single molecule imaging studies of several alternative exons that showed that cell-to-cell variability is minimized and tightly regulated by the splicing machinery in single cells [66, 67]. Curiosity about this result led to investigations of mechanisms that might be responsible for stochastic splicing variability among apparently homogeneous cells, such as variation in DNA methylation [57].

We propose that the observed bimodality does not generally reflect a binary nature of splicing biology, but rather, that it exposes a technical limitation of the scRNA-seq data that have been collected so far. Because alternative isoforms of a gene share much of the same sequence, only the few RNA-seq reads mapping to the exact alternative splice junctions, or to the alternative exon itself, reveal its alternative splicing. When combined with the low mRNA capture efficiency of scRNA-seq and the PCR amplification of small amounts of starting material into a full-length sequencing library, these circumstances create the risk of bottlenecks that lose all but a few individual mRNAs of most genes in each cell.

The limitations of scRNA-seq are a known obstacle to studying splicing in single cells [51]. Similar concerns have arisen with the use of scRNA-seq to infer allelic expression; a careful analysis showed that stochastic patterns resulted from technical noise [68]. A recent study observed and modeled the high dropout rate of individual isoforms in scRNA-seq and advised that scRNA-seq is fundamentally unsuitable for measuring changes in alternative splicing [53]. Others have implemented workarounds, e.g., using sequence features to predict splicing outcomes in lieu of sufficient sequencing coverage [64, 65], or attempting to identify excess variance beyond technical noise [55, 57]. These studies have identified true examples of differential splicing in single cells, but they fundamentally do not explain how scRNA-seq limitations have caused qualitative, not just quantitative, distortions in our understanding of alternative splic-

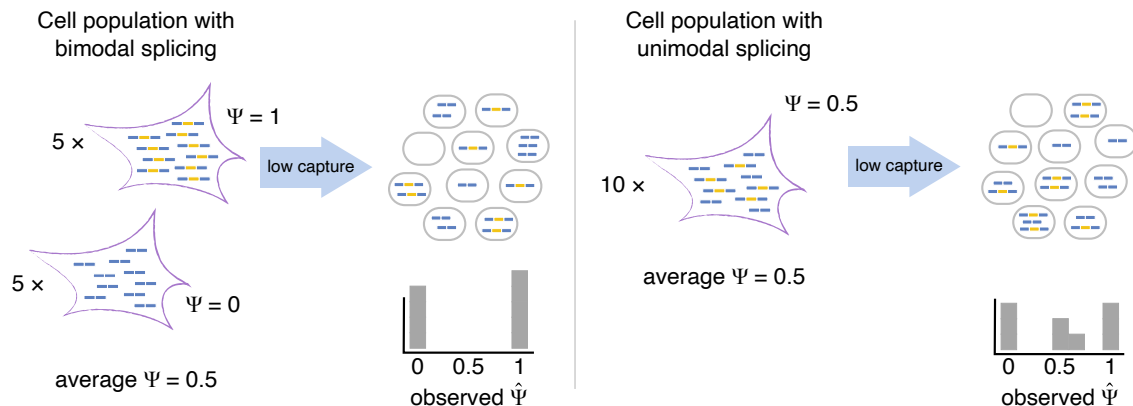


Figure 1.1: Bimodal vs unimodal models of cassette exon splicing. In the bimodal model, some cells consistently splice in the exon, while others consistently skip it. After mRNA capture and sequencing, observations of Ψ are almost exclusively binary. In the unimodal model, individual cells express some mRNAs that splice in the cassette exon and some that skip it. Low mRNA capture dramatically reduces the number of cells in which both isoforms are observed, artificially inflating binary Ψ values.

ing.

Here, we show that scRNA-seq splicing data are consistent with a simple model (Figure 1.1). Consider a particular cassette exon whose true pattern of exclusion follows a unimodal distribution of isoform ratios across cells (i.e., most cells express both isoforms, with a ratio revolving around the same mean). This distribution can be distorted by extreme information loss during library preparation and sequencing, creating the illusion that individual cells only produce one isoform or the other. Our simulations make it clear that the reliability of splicing measurements is a function of the initial amount of mRNA, the efficiency of its recovery, the underlying splicing rate, and further distortions from PCR amplification of cDNA. These effects should be considered when interpreting previous studies that used qualitative changes in the observed distribution of the splicing rates [56] or changes in their variance [57] as evidence for regulation of alternative splicing. Considering the true amount of information available for a cassette exon can allow for accurate observations of alternative splicing. Using a data normalization and filtering method to identify cassette exons with sufficient information, we are able to draw biologically relevant conclusions about alternative splicing in single cells.

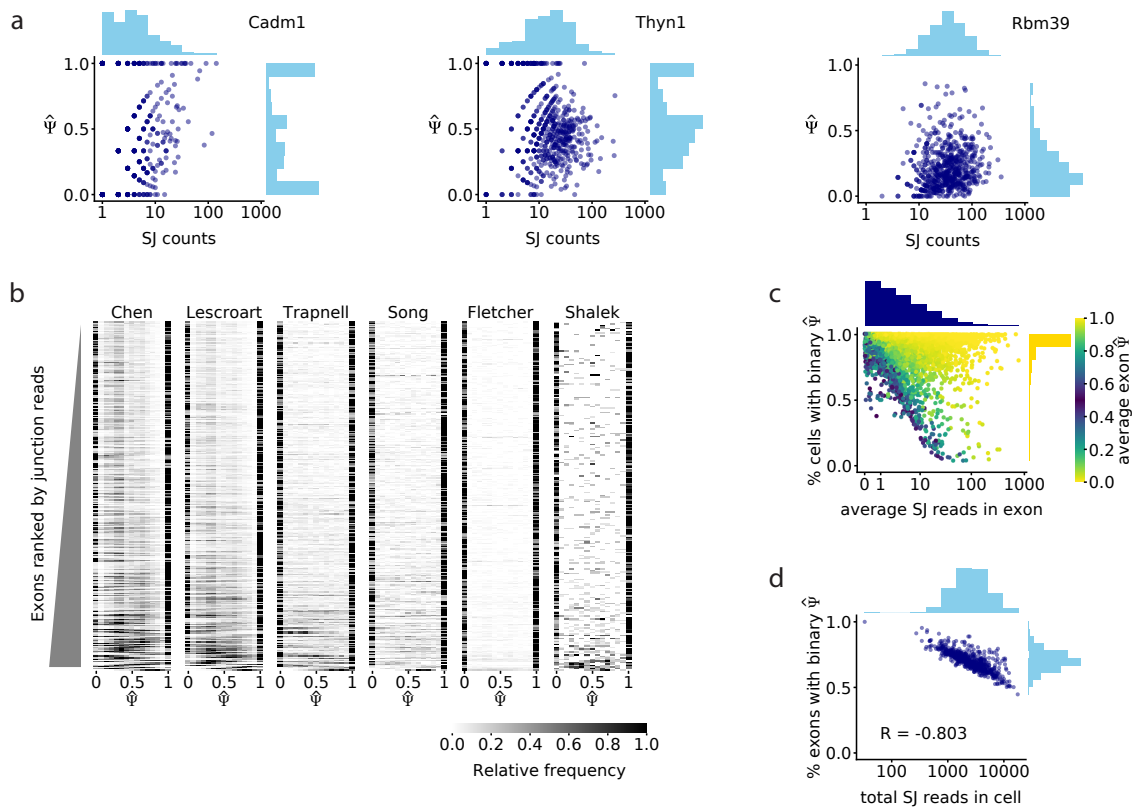


Figure 1.2: a) Comparison of splice junction read coverage and observed $\hat{\Psi}$ for three cassette exons in the Chen dataset, with low (Cadm1 exon 8; chr9: 47829377–47829409), medium (Thyn1 exon 6, chr9: 27006801–27006951), and high coverage (Rbm39 exon 3; chr2: 156178880–156178952). Each dot represents the $\hat{\Psi}$ of that exon in one cell. b) $\hat{\Psi}$ distribution of the 300 highest coverage cassette exons with intermediate splicing (average $\hat{\Psi}$ between 0.2 and 0.8) in each of the six analyzed datasets. Each row in the heatmap shows the distribution of $\hat{\Psi}$ for one exon across all cells. c) Relationship between the average read coverage and proportion of binary observations for each cassette exon in the Chen dataset. d) Correlation between the total number of splice junction reads captured in each cell, and proportion of cassette exons with intermediate splicing that show binary $\hat{\Psi}$ in that cell.

1.2 Single cell binary splicing observations are inversely correlated with read coverage

Our interest in splicing regulation led us to examine alternative splicing in several single cell differentiation datasets from mice and humans that were generated with methods that recover sequence from along the full length of mRNAs. To investigate the reported high variability of splicing between cells more closely, we began by examining the splicing of cassette exons in a high-coverage mouse scRNA-seq dataset [69], estimating their percent spliced-in as the fraction of splice junction reads that show exon inclusion (out of all reads that cover the junction). We use $\hat{\Psi}$ to denote these estimated rates, while Ψ denotes the actual rate as it is in the cell. For clarity we define a single $\hat{\Psi}$ observation (which pertains to a specific cassette exon in an individual cell) as *binary* if it is close to 0 or 1 (i.e., the respective cell tends to express transcripts that either include the exon or exclude it, but not both). We then describe the distribution of an exon’s $\hat{\Psi}$ across cells as *bimodal* when its individual values are predominantly binary, where some cells have a $\hat{\Psi}$ close to 1 (most observed transcripts include the exon) and others have $\hat{\Psi}$ close to 0 (most observed transcripts do not include the exon). Strikingly, when we inspected several exons, we saw that they had more binary outcomes in cells with fewer reads covering their splice junctions, while cells with more reads were more likely to show non-binary $\hat{\Psi}$ values (Figure 1.2a). We realized that this effect of coverage may reflect a non-binary reality, since even if both isoforms are expressed in a certain cell, the likelihood of observing both isoforms is reduced as the number of captured mRNAs decreases. In contrast, if the underlying distribution were indeed bimodal with binary modes, as previously proposed [40, 56, 58], then the read coverage would have little effect on the proportion of binary $\hat{\Psi}$ observations across cells.

To further explore this phenomenon, we extended our analysis to the full scRNA-seq datasets. In all cases, we found a strong effect of coverage on the observed binary $\hat{\Psi}$ in exons with intermediate splicing (average $\hat{\Psi}$ between 0.2 and 0.8). We consistently found that exons with low junction read coverage had more binary $\hat{\Psi}$ values and bimodal $\hat{\Psi}$ distributions (Figure 1.2b,c; Supplementary Figure 1.5a,b). We found that the association between binary values and read coverage was not observed in exons that are binary but not bimodal (i.e., nearly constitutively excluded or included exons, with average $\hat{\Psi}$ close to 0 or 1; 1.2c). Taken together, these observations suggest that the abundance of binary observations in exon inclusion patterns may reflect a distortion of an underlying unimodal splicing distribution (i.e., when cells in fact express both isoforms), rather than a truly bimodal splicing pattern in the analyzed cells.

We then asked if the association of binary splicing outcomes with low read coverage could be due to some biological consequence of low transcription, or if it was a technical consequence of low sequencing coverage. We found that, among the cells in a

single experiment, the cells with an overall higher number of splice junction reads also tended to have a smaller fraction of exons with binary values (Figure 1.2d, Supplementary Figure 1.5b), suggesting an influence of technical coverage rather than transcription level. We further considered the effect of biological variations in transcription. Transcription at a single locus occurs in intermittent bursts of RNA synthesis [70], suggesting a possible effect of size and frequency of transcriptional bursts on binary splicing. Using transcriptional bursting data from [71], we found that genes with either high burst frequency or large bursts did exhibit fewer binary splicing observations (Supplementary Figure 1.5c), but a linear regression showed that burst frequency and size did not contribute to binary observations beyond the effect of read coverage (Supplementary Figure 1.5d). This suggests that transcriptional bursting contributes to splicing variability only by determining the overall abundance of an mRNA, consistent with single molecule fluorescence observations [66].

1.3 Simulations of RNA sequencing reveal technical sources of distortion of splicing estimates

A simple probabilistic exercise shows the potential loss of splicing information during sequencing. Single cell RNA-seq experiments that capture full-length transcripts have an estimated capture efficiency of only $\sim 10\%$, due to RNA degradation and inefficient reverse transcription [44, 58, 63, 72]. For instance, a gene that expresses 20 mRNA molecules in a cell might only have two mRNAs recovered, and if that gene is alternatively spliced with a true splicing rate Ψ of 0.5, there is approximately a 50% chance that those few recovered mRNAs will only represent one of the two isoforms that were originally present in the cell (Supplementary Figure 1.6a,b). As many genes are expressed at just a few RNA molecules per cell, low recovery might affect many alternative splicing events [73, 74]. Furthermore, while the empirically observed $\hat{\Psi}$ provides a maximum likelihood estimate for the true splicing rate, the uncertainty of this estimate (i.e., the range of alternative values with a nearly similar likelihood) decreases substantially with the number of observed molecules (Supplementary Figure 1.6c and Methods). As a result, the probability that the observed $\hat{\Psi}$ is close to the true underlying Ψ increases when more mRNA molecules are captured (Supplementary Figure 1.6d).

Our theoretical reasoning above relied on a simple model where the number of observed mRNA molecules (rather than number of reads) is known and the only distorting factor is a limited capture efficiency. In practice, both of these assumptions are challenged due to additional factors, such as PCR amplification and variability in the capture efficiency across cells. To investigate the pertinent effects on $\hat{\Psi}$ distributions in this more complex setting, we designed a probabilistic simulator of alternative splicing in single cells (Supplementary Figure 1.7). The model has two main components: we begin by simulating the underlying molecular content of each cell, by drawing

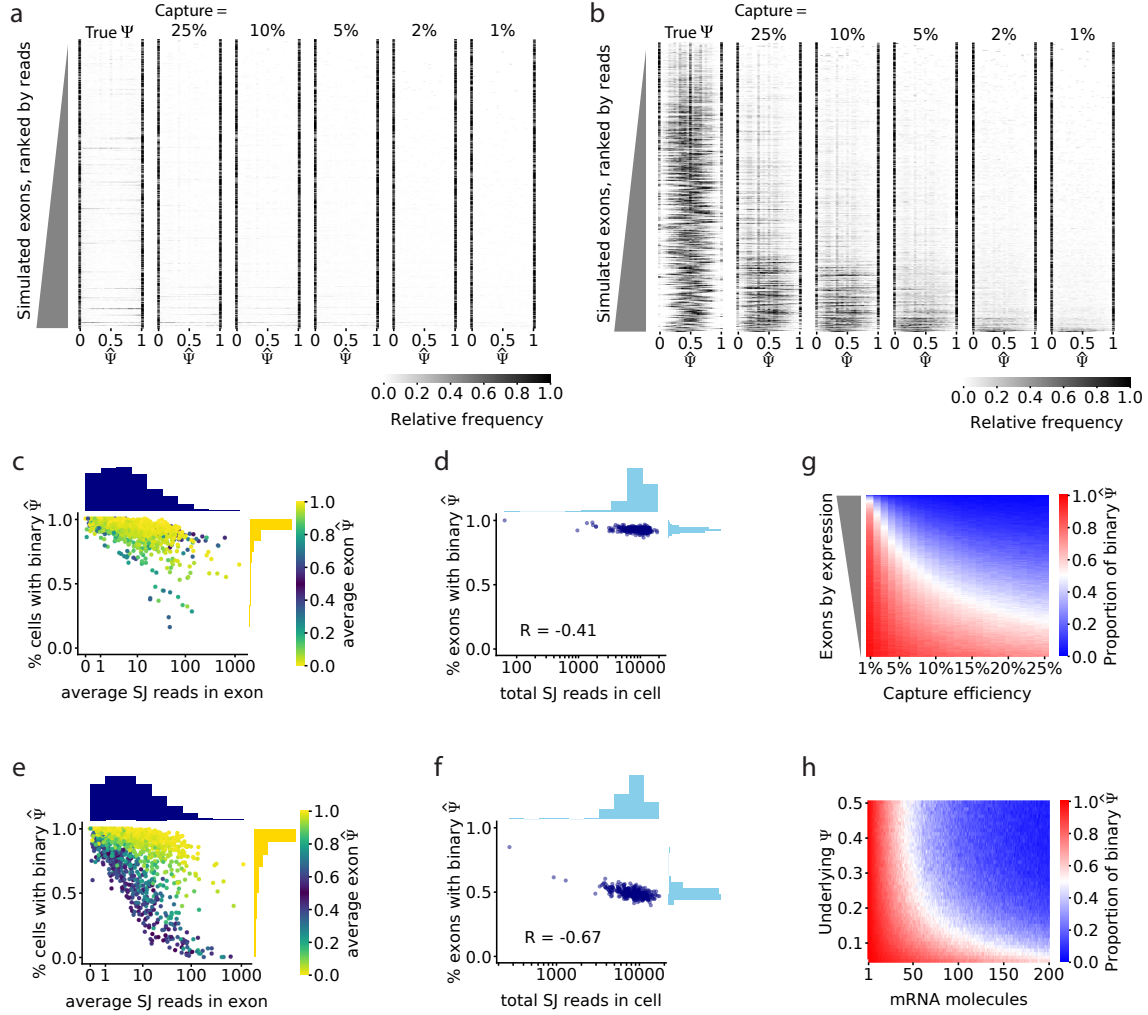


Figure 1.3: a) Simulations of alternative splicing and scRNA-seq under the binary-bimodal model, in which each cell produces one isoform or the other, but rarely both. As in Figure 1.2b, each row of the histogram shows $\hat{\Psi}$ for one intermediate exon across all cells. The observed $\hat{\Psi}$ distribution is similar to the true Ψ distribution, and its shape is largely unaffected by capture efficiency. b) Simulations with the non-binary, unimodal model, in which most cells present a mixture of the two alternative isoforms. Exons with high expression have a unimodal distribution of true Ψ . Low capture efficiency results in an increase in binary observations (only one isoform observed), leading to a distortion of the observed distribution of $\hat{\Psi}$ to look bimodal. Only a handful of the highest expressed exons maintain a unimodal distribution of $\hat{\Psi}$. Fewer exons show unimodal splicing as the capture efficiency is reduced. c) Under the binary-bimodal model, exons with high coverage have slightly fewer binary $\hat{\Psi}$ observations, and d) simulated cells with a high number of total splice junction reads have slightly fewer exons with binary $\hat{\Psi}$. e) Under the unimodal model, exons with intermediate splicing show a strong decrease in binary observations as coverage increases, as seen in real data (Figure 1.2c). f) Similarly, simulated cells with high read coverage have a decrease of the proportion of binary $\hat{\Psi}$. g) Effect of capture efficiency on the proportion of binary observations of cassette exons with underlying $\Psi = 0.5$. h) Effect of the initial number of mRNA molecules and underlying Ψ on the proportion of binary $\hat{\Psi}$ observations.

gene expression levels and cassette exon splicing rates from a probabilistic model of cell state. We then simulate the technical process of extracting data from each cell using single cell RNA sequencing with full transcript coverage. This part accounts for variability in capture rates, and the effects of PCR amplification, fragmentation and sequencing. It relies on SymSim, a simulation tool for single cell RNA sequencing data [75]. The final product of our simulation is the number of splice junction reads that either span or skip each exon in each cell. These numbers are distorted in a way that reflects real nuisance factors. For instance, two reads could have originated from the same molecule due to amplification effects.

We used our simulator to investigate how the observed inclusion ($\hat{\Psi}$) of cassette exons differs from the underlying Ψ , under different average capture rates, and setting the other technical parameters to values that are characteristic of Smart-seq2 datasets (see Methods). We considered either a binary-bimodal regime of Ψ (i.e., both isoforms are expressed in the population, but rarely by the same cell; Figure 1.1a), or a non-binary regime (cells tend to express both isoforms; Figure 1.1b). We simulated splicing of cassette exons in 1500 genes, in a population of 300 single cells. In the bimodal simulation, 500 exons were modeled to have alternative splicing with a bimodal distribution across cells, and in the unimodal simulation, these 500 exons were modeled to have a unimodal distribution. To reflect the patterns seen in real data, we also simulated splicing of 500 exons that were nearly constitutively included and 500 that were nearly constitutively skipped.

As expected, in the binary-bimodal simulation, the observed $\hat{\Psi}$ reflected the underlying process well, independent of the average capture efficiency (Figure 1.3a). In contrast, when we modeled a non-binary splicing regime, the observed $\hat{\Psi}$ distributions were strikingly similar to the splicing distributions of cassette exons in real single cell RNA-seq datasets (Figure 1.3b). Specifically, the loss of information due to mRNA recovery and library generation led many of the observed $\hat{\Psi}$ to become binary, and their observed distribution across cells to become bimodal. This tendency again correlated with coverage, whereby lowly covered exons showed the strongest effect, while exons with high coverage maintained a non-binary, unimodal distribution. Consistently, in this non-binary regime, the average of $\hat{\Psi}$ was similar to the true average of Ψ , but the variance of $\hat{\Psi}$ increased (Supplementary Figure 1.6e,f). Furthermore, as in the real data sets (Figure 1.2c), we also found that the dependency between read coverage and the chance of observing a binary $\hat{\Psi}$ is more pronounced in exons with an underlying Ψ that is far from binary (Figure 1.3c-f), highlighting again that such an association likely indicates an artifact.

To address the extent of the distortion from mRNA recovery, we ran the simulator with varying capture efficiency and a fixed underlying $\Psi = 0.5$. We found that decreasing the average capture efficiency dramatically increased the number of binary $\hat{\Psi}$ observations, particularly for exons with low expression, although even highly expressed genes suffered great distortion in the observed $\hat{\Psi}$ when the average capture

efficiency was very low (Figure 1.3g). These results reinforced the hypothesis that capture efficiency is a main technical factor that creates the appearance of bimodality in single cell splicing.

Finally, we estimated the chance of observing only one type of isoform in any given cell (i.e., a binary $\hat{\Psi}$) as a function of the underlying Ψ and the number of transcripts that are present in the cell (Figure 1.3h). For this analysis, we set an average mRNA capture rate of 10%. Our results delineate the range of values in which an artifact is less expected. For instance, for an exon with 50% inclusion rate, a non binary estimate is more likely if the respective gene has at least 50 transcripts in the cell. Notably, these estimates are more conservative than the theoretical analysis (Supplementary Figure 1.6a,b), due to the effect of the technical nuisance factors we modeled in this simulation.

1.4 Accounting for mRNA recovery improves analysis of alternative splicing

We sought criteria that would identify reliable measurements of splicing in single cells and avoid distortions from low mRNA recovery. While ideally we should rely on the actual number of mRNA molecules recovered, full-length RNA-seq experiments generally do not report an absolute mRNA count. Previous studies assessed the quality of splicing observations based on the number of reads covering alternative splice junctions [56, 57], but the number of splice junction reads is influenced by the extent of PCR amplification and sequencing depth, and so it does not directly reflect the number of recovered mRNAs.

We realized that we could estimate the number of mRNA molecules that were captured into cDNA by using the Census normalization approach proposed by [63]. This method infers a per-cell scaling factor between the relative abundance of each transcript, inferred from RNA-seq, and the actual number of mRNAs recovered. We found that some datasets with many reads per cell, such as the Song dataset [56], nonetheless had few mRNAs recovered per cell (Figure 1.4a, Supplementary Figure 1.8c), which may explain the extensive splicing bimodality in this dataset. The dataset with the highest recovery of mRNAs [69] indeed showed less binary splicing (Figure 1.2b).

Next, we sought a threshold for mRNA recovery that would suppress spurious observations of binary splicing. Our simulations and our analytical calculations suggested that, for a capture efficiency of 10%, recovering 7–10 mRNA molecules would be more likely than not to capture both isoforms for exons with intermediate splicing (Figure 1.3h), and that the observed $\hat{\Psi}$ would most likely be within 0.1 of the real value (Supplementary Figure 1.6d). In keeping with this, we saw in the real data that exons with an average of at least 10 mRNAs recovered per cell generally had substantially fewer

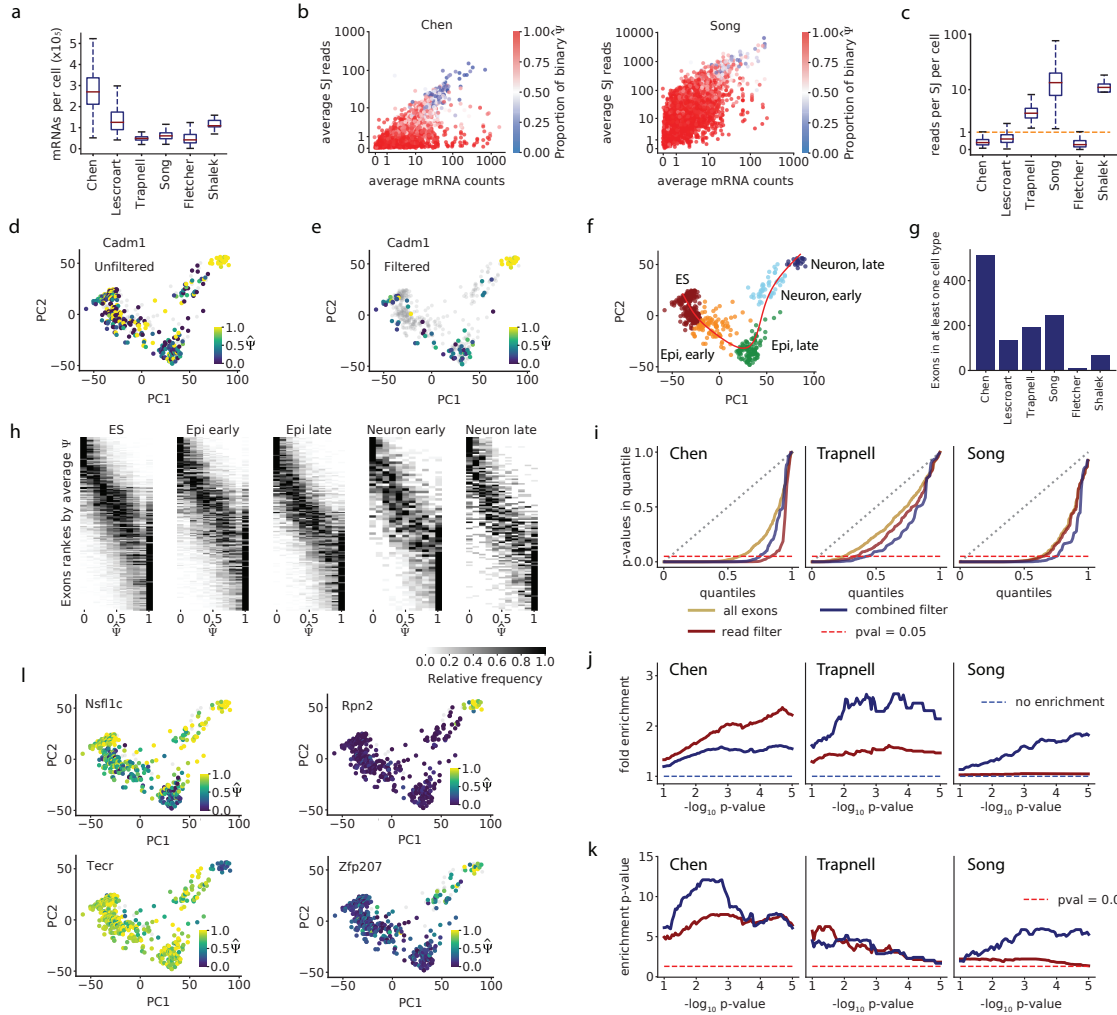


Figure 1.4: a) Estimated total mRNA molecules captured per cell. b) Estimated number of recovered mRNAs vs splice junction reads for cassette exons, averaged across cells. Each dot corresponds to an exon, and its color indicates the proportion of cells in which it has a binary observation (only one isoform observed). We analyzed exons with average $\hat{\psi}$ between 0.05 and 0.95. c) Per-cell splice junction coverage rate in each dataset. d) *Cadm1* exon 8 alternative splicing appears binary in many cells in the Chen dataset. Correlation with lineage pseudotime: Spearman's $r_s = 0.1$. e) *Cadm1* exon after removing cells with fewer than 10 recovered *Cadm1* mRNA molecules and fewer splice junction reads than expected from 10 mRNAs (grey). Spearman's $r_s = 0.52$. f) PCA projection and clustering of single cells in the Chen dataset, showing differentiation of mouse ES cells into neurons. Red line, lineage inferred with Slingshot. g) Number of cassette exons with observations from at least 10 mRNA reads in at least 50% of cells in any cluster. h) Stacked histograms showing the distribution of observed $\hat{\psi}$ of exons as in (g), in each cell cluster of the Chen dataset. Observations with fewer than 10 mRNA molecules were removed. We show exons with average $\hat{\psi}$ ranging from 0.1 to 0.9 per cluster.

Figure 1.4: i) QQ-plot comparing the quantiles of a uniform distribution (x-axis) with the quantiles of the distributions of p-values from the Kruskal-Wallis test (y-axis). A diagonal line (gray dotted line) would mean the p-values are uniformly distributed. A lower area under the curve indicates enrichment for low p-values. The point on the x-axis at which each line crosses the dotted red line indicates the proportion of p-values that are below 0.05 in the distribution. j) Fold enrichment of exons with a Kruskal-Wallis $p < x$ in the set of exons selected with the mRNA-based filter (blue), and exons selected with a flat read minimum filter (red). k) Significance p-value of the enrichment, estimated with the hypergeometric test and adjusted for FDR. l) Example exons that pass the overall filter criteria in the Chen dataset and have $p < 0.05$ in the Kruskal-Wallis test.

binary observations (Figure 1.4b, Supplementary Figure 1.8d; limited to exons with average $\hat{\Psi}$ between 0.05 and 0.95). Nonetheless, a subset of these exons still showed binary splicing in most cells. These exons had few splice junction reads relative to the estimated mRNA count. We expect that these represent anomalously low recovery of reads from the specific splice junctions of interest, perhaps due to annotation errors or poor recovery of fragments with particular sequence composition. We reasoned that a data-driven splice junction read criterion would prevent distortions arising from this low read recovery. To find an appropriate read minimum, we determined the number of splice junction reads expected to arise from a cassette exon in each cell in different datasets, based on that cell’s overall mRNA recovery (Figure 1.4c; see Methods). This provided a second filtering criterion: we excluded exons with fewer splice junction reads than would be expected from 10 mRNAs, given that exon’s $\hat{\Psi}$ and the coverage rate in that particular cell. This metric is calculated for each exon separately, driven by the actual information in each cell, and it varies substantially between datasets.

We then selected the exons from each cell that met the combined requirement of 10 mRNAs and the splice junction reads that would arise from 10 mRNAs, and asked what effect this filter had on binary observations of splicing. As an example, our filter removed seemingly spurious binary observations for alternatively spliced exon 8 of *Cadm1* (Figure 1.4d). The observations that remained after filtering showed a clear pattern of increased inclusion along pseudotime, with exon skipping in stem cells and exon inclusion in fully differentiated neurons (Spearman’s $r_s = 0.52$, $p = 1.2 \times 10^{-7}$; Figure 1.4e). The correlation between splicing observations and differentiation pseudotime highlighted that cell type differences could contribute to observations of bimodality that are not the result of an artificial inflation of binary observations, and this was an important factor to consider in our analysis.

Bimodality in filtered exons is explained by differential splicing between cell types

Earlier papers had reported a surprising amount of bimodal splicing within a cell type, with extensive binary outcomes in individual cells, i.e., similar cells with different

dominant isoforms. This suggested that some aspects of gene expression stochastically locked individual cells into producing primarily molecules of one or the other isoform. However, our results so far suggest that much of this observed variability may be an artifact of low mRNA recovery. In contrast, differences in splicing between cells of different types are consistent with conventional mechanisms such as regulation by cell-type-specific splicing factors. Therefore, when considering heterogeneous cell samples, large shifts in splicing between cell types may result in truly bimodal Ψ . In contrast to the first scenario, we would expect the corresponding observations $\hat{\Psi}$ to remain bimodal even after low-quality data points were removed.

We therefore sought to distinguish between these two scenarios and estimate the extent of bimodal splicing within a cell type that is strongly supported by the data. Most of the datasets in our analysis came from differentiation time courses, and so to examine homogeneous cell subsets, we stratified the cells into groups indicative of their developmental stage. The groups were identified by clustering the cells using the normalized sum of reads from every gene (i.e., not directly representing their variation in splicing; Figure 1.4f) and labeled based on expression of known marker genes (Supplementary Figure 1.8e).

We found that, prior to filtering, hundreds of exons with intermediate splicing (average $\hat{\Psi}$ between 0.2 and 0.8) seemed to have at least weakly bimodal splicing distributions (using a heuristic definition of at least 25% of cells with $\hat{\Psi} \leq 0.25$ and 25% with $\hat{\Psi} \geq 0.75$), a rate that roughly matched previous descriptions [56] (Table 1.2). We then discarded the observations that did not meet our filtering criteria, and retained for every cluster only those exons with observations remaining in at least 50% of cells (Figure 1.4g). With that filtering we found that most remaining exons had fewer binary observations than the discarded exons (Supplementary Figure 1.8f). Furthermore, cell clusters had no or very few remaining exons with bimodal splicing distributions (Figure 1.4h, Supplementary Figure 1.8g,h, Table 1.2). For instance, in the Chen dataset, there were two remaining cases of bimodality, and both seem to be the result of sex-specific splicing (Supplementary Figure 1.8i). This trend persisted when reanalyzing all datasets and clusters using a range of cutoffs for qualifying a splicing pattern as binary (Supplementary Figure 1.8e).

Taken together, these results challenge the idea that widespread bimodality among homogeneous cells arises from a binary nature of splicing in single cells, and instead suggest that splicing bimodality reflects biological differences between cell types or subtypes. Indeed, experimental observations of splicing bimodality support the importance of differences between cell types, rather than stochastic differences between homogeneous cells. The two cases of splicing bimodality confirmed with smFISH by Song et al. both showed bimodality between two characterized cell types [56]. Similarly, Shalek et al. confirmed one case of splicing bimodality, in a gene whose expression they showed to be a marker of a hidden difference between mature and maturing cells, and so its splicing may reflect cell state difference at a wider context

as well [40].

Filtering improves the identification of differentially spliced exons

Finally, we asked if our filtering criteria could help identify biologically relevant changes in splicing between cell types, by selecting exons whose biological variance is not overwhelmed by high technical variance. Focusing on three datasets with multiple cell type clusters (Chen, Trapnell, and Song), we examined the effect of retaining only exons that passed our filter in at least 50% of cells in each cluster. To avoid biases depending on the number of observations available for each remaining exon, we retained all data points for these exons, rather than removing individual cell data points that did not pass the filter. As a benchmark for comparison, we used a simpler criterion, akin to previous studies, of at least 10 splice junction reads in at least 50% of cells (Figure 1.4i). To focus on exons with substantial alternative splicing, we omitted exons from this analysis that were consistently excluded (average $\hat{\Psi} \leq 0.05$) or included (average $\hat{\Psi} \geq 0.95$) across all cells.

For each remaining exon in the two filtering schemes, we performed a Kruskal–Wallis one-way analysis of variance, which assigned it a p-value (Methods). This test determines if the exon’s median Ψ changed significantly between any clusters. Both filtering schemes enriched substantially for exons that show significant changes across different clusters, compared with the unfiltered set of all exons with average Ψ between 0.05 and 0.95. Furthermore, the extent of enrichment increased with the strictness of the differential splicing test (Figure 1.4i-k; Supplementary Figure 1.9). It is interesting to note that in the Chen dataset the baseline (ten reads) filter was stricter than the combined filter, retaining 66 vs. 198 exons and thus achieving high precision at the expense of a lower recall. Indeed, due to the high mRNA recovery and low amplification in this dataset, our combined filter required only seven splice junction reads, rather than ten as in the baseline scheme. Conversely, in the Song dataset, which is predicted to have low mRNA recovery yet a large number of reads, the baseline filter retained almost all exons, while the combined filter retained only 104 of them, leading to significant over- representation of exons that are differentially spliced between clusters. The results show how differences in mRNA recovery and amplification between datasets may change the interpretation of an absolute number of reads. We corroborated these results using spatial autocorrelation as a way of identifying informative exons, rather than the clustering and differential splicing analysis. The autocorrelation test builds on the work of [76] to identify exons with a significantly high level of consistency in $\hat{\Psi}$ among transcriptionally similar cells (Supplementary Figure 1.10).

Taken together, these results indicate that careful filtering of exons can help identify observations that are consistent with the gene expression space and the stratification of cells into types, and are thus likely more biologically meaningful (e.g., capturing

known forms of regulation during differentiation) (Figure 1.4l, Supplementary Figure 1.11) [77]. The concept of "meaningful" is ascribed here to the observations, not the exons themselves, as there may be many additional exons with relevant variation (e.g., stable differences between cell types) that is not reliably measured due to low coverage. Indeed, the effects of the two filtering schemes and the differences between them exemplify how technical factors may distort our view of the data and consequently, our understanding of variation and regulation of splicing.

1.5 Discussion

The surprising result that alternatively splicing is bimodal among single cells provoked curiosity and speculation. Bimodal outcomes might reflect hidden cell subtypes, but the bimodality was seen even among apparently homogeneous cells. Did splicing outcomes reflect some unknown, stochastic cell state?

We have shown here that the bimodal patterns could have an entirely different explanation: profound technical limitations of single cell RNA sequencing. A crucial limit on biologically meaningful splicing observations in a single cell is the number of mRNAs available to inform the measurement. This is determined both by the expression of the genes that contain the exons of interest, and by the capture efficiency of the experiment. It is important to note that the depth of a sequencing library does not necessarily reflect its quality. Along with low mRNA numbers, splicing observations are also distorted by uneven amplification efficiency and cDNA overamplification. Increasing PCR amplification cycles in an attempt to compensate for low capture efficiency has the risk of worsening the technical distortion. Indeed, in our analysis, the dataset with the highest read count per cell actually had quite low mRNA recovery and large technical distortion, creating an appearance of bimodal splicing [56]. Moreover, a qualitative change in the observed $\hat{\Psi}$ distribution of an exon between single cell subpopulations does not necessarily reflect a change in the underlying splicing rate, as changes in gene expression and mRNA recovery between samples can create the illusion of a splicing change.

Further developments in statistical analysis that carefully account for both missing and redundant information due to low capture efficiency could make splicing observations in single cells more reliable. We set the foundation for such analysis by proposing a probabilistic process that describes the biological and technical steps that generate single cell splicing data. We also introduced a simple approach that builds on the Census normalization [63] to estimate the number of mRNAs recovered and the extent of artificial duplication of splicing information. This metric provides a practical filter for identifying exons with sufficient information to analyze. On the experimental side, improving the capture efficiency of scRNA-seq methods while moderating the extent of overamplification is crucial for increasing the subset of exons for which reliable observation can be made.

True biological insight into alternative splicing can indeed be found from high-quality scRNA-seq data, and we hope that new methods will allow better understanding of splicing regulation, cell-to-cell variation, and the importance of alternative splicing in defining cell fate [78, 79]. However, some limitations are inherent to the situation. Single cells express a limited number of mRNAs per gene; splicing observations in single cells will always be inherently noisy reflections of the underlying biology.

1.6 Methods and Materials

Analysis code is available at

https://github.com/lareaulab/sc_binary_splicing

1.6.1 Analysis of single cell RNA-seq datasets

Datasets Six publicly available single cell RNA sequencing datasets were analyzed (Table 1.1). These datasets are referenced with the first author’s name throughout this paper. For the Chen dataset, we selected the four cell types used to represent developmental stages in [69]: mouse embryonic stem cells cultured in 2i (ES2i) and LIF (ES), mouse EpiStem cells (Epi) and neurons. For the Lescroart dataset, we limited the analysis to the cells derived from mouse wild-type strains. For the Trapnell dataset, we only selected the runs that are annotated to have one cell per well.

Alignment, TPM quantification and Ψ estimation We aligned the reads of each dataset using STAR 2.5.3 [80] with two-pass mode and index overhang length adapted to the read length of each dataset. We used the hg38 genome annotation for the human RNA-seq datasets, and the mm10 annotation for the mouse datasets. Gene expression levels in transcripts per million (TPM) were calculated by running RSEM [37] on the BAM files produced by the STAR alignment. We ran rMATS 3.2.5 [39] on bulk human and mouse RNA-seq datasets from cell types matching the scRNA-seq datasets [45, 69, 81, 82] to find all annotated cassette exon alternative splicing events in each cell type. Then we used the SJ.out.tab files obtained from the scRNA-seq STAR alignment to obtain the splice junction reads compatible with the list of cassette exons found by rMATS. For each cell i , we calculated the observed $\hat{\Psi}$ of the cassette exon j as:

$$\hat{\Psi}_{ij} = \frac{SJ_{A_{ij}}}{SJ_{A_{ij}} + 2SJ_{B_{ij}}}$$

where $SJ_{A_{ij}}$ correspond to the number of reads that cover the two splice junctions compatible with cassette exon inclusion, and $SJ_{B_{ij}}$ are the reads that cover the splice junction compatible with its exclusion. We also determined the coverage of an exon j in i as $SJ_{ij} = SJ_{A_{ij}} + SJ_{B_{ij}}$. We used SJ_{ij} and $\hat{\Psi}_{ij}$ for the analyses shown in Figure 1.2.

Gene expression normalization and pseudotime inference For the purpose of visualization and clustering of cells, we normalized the gene expression data. First we selected the genes with TPM ≥ 20 in at least 20 cells. After filtering, we used SCONE 1.6.1 [83] to select the best normalization approach for the data. For improving the normalization of the data, we used additional information for each cell, including the annotated cell type and batch, total number of reads, housekeeping genes and genes that are expected to change in the biological process that the dataset covers. We applied principal component analysis (PCA) over the log-counts from the best SCONE normalization, and used the first two principal components to infer pseudotime using Slingshot 1.0.0 [46]. We used the cell type annotation as the cluster input for Slingshot, and manually indicated the direction of the biological process. Single cells can be highly heterogeneous, even within cells labeled to be in the same biological condition. To account for this variation, instead of relying on the annotated cell type provided with each dataset, we used agglomerative clustering over the PCA projection of the matrix of normalized gene expression to identify groups of similar cells in the Chen, Trapnell and Song datasets. For the Trapnell and the Song dataset we used the number of reported cell types by the authors. In the Chen dataset, we noticed that the cells annotated as neurons could form two well-defined clusters, with some of the cells presenting expression profiles consistent with mature neurons. For this reason, we used five instead of four clusters in the Chen dataset. In the Lescroart dataset we used the original labels provided by the authors because PCA over the normalized expression data separates the cells into two groups. This analysis was not performed in the Shalek dataset due to its small number of cells. Instead, we treated all the cells in this dataset as one single cluster. The Fletcher dataset was not considered for the analysis that required clustering due to the low number of exon observations that pass the minimum mRNA requirements described below.

mRNA counts estimation with the Census approach We performed our own implementation of the mRNA count estimation proposed in Census [63]. The total number of transcript mRNAs in cell i is estimated as:

$$M_i = \frac{n_i}{F_{X_i}(x_i^*) - F_{X_i}(0.1)}$$

where x_i^* is the mode of the log-transformed distribution of TPM values in cell i . As in Census, we found x_i^* by fitting a Gaussian kernel density estimation to each distribution and finding its peak. We also set 0.1 as is the minimum TPM below which it is assumed that no mRNA is present. n_i is the number of genes in cell i with an estimated TPM in the interval $(0.1, x_i^*)$. F_{X_i} is the cumulative distribution function of the TPM values in cell i . The original Census implementation also adjusts the mRNA estimation by multiplying by $\frac{1}{\theta}$, where θ is the capture efficiency of the dataset estimated with RNA spike-ins. Since for most datasets we do not have a reliable way of estimating the capture efficiency, we removed this adjustment from

the equation, so that M_i in our estimation is not an estimation of the amount of mRNAs present in the cell lysate as it is in Census, but an estimate of the mRNAs successfully captured into cDNA.

We found that some datasets contained outlier cells with M_i much higher than the median estimate (more than ten-fold increases). These outliers generally correspond to cells with a multimodal TPM distribution. An inflation in very low TPM values distorts the normalization by shrinking the values of $F_{X_i}(x_i^*) - F_{X_i}(\epsilon)$, thus inflating the M_i values in these cells. Because the Census method relies on a Gaussian kernel density estimation that performs inaccurately for multimodal distributions, we excluded this handful of outliers from further analysis. In the Trapnell dataset, we found that several cells had an unusually small number of recovered reads and genes observed. We reasoned that this had the potential to skew the TPM estimations on which the Census normalization depends (Supplementary Figure 1.5b), so we excluded the cells in the bottom 25% quantile of reads from the Census normalization and all downstream analyses that depend on it.

Finally, the number of mRNA transcripts of gene g in cell i is calculated as:

$$Y_{ig} = X_{ig} \cdot \frac{M_i}{10^6}$$

where X_{ig} is the expression of gene g in cell i expressed in TPM.

Nucleotide coverage and expected splice junction reads Amplification in short-read library preparation can lead to multiple reads from the same sequence fragment of a single mRNA molecule. To filter out exons with anomalously low read coverage, we wanted to know the number of splice junction reads expected to originate from one exon junction in a single mRNA, a number which differs in each cell and each experiment. We estimated the splice junction coverage rate of each cell as the expected number of reads covering the splice junction of a mRNA molecule:

$$\begin{aligned} C_j &= \text{read coverage at each splice junction of mRNAs in our sample} \\ &= \frac{\text{constitutive splice junction reads in cell } j}{\text{constitutive splice junctions of mRNAs in cell } j} \\ &= \frac{\sum_k r_{jk}}{\sum_k j_k \cdot m_{jk}} \end{aligned}$$

where C_j is the splice junction coverage rate, which is the number of reads expected to cover each splice junction in cell j ; r_{jk} is the number of reads that map to all constitutive splice junctions of gene k in cell j , as reported by STAR; j_k is the number of constitutive splice junctions in a mRNA molecule of gene k ; and m_{jk} is the Census estimate of captured mRNAs of gene k in cell j . To calculate the C_j of each cell, we only considered the constitutive splice junctions of genes that were estimated

to be expressed in at least 1 mRNA molecule by the Census normalization. This calculation captures the slight under-counting of splice junctions (relative to other positions) because of factors including ambiguous read mapping.

We identified the constitutive splice junctions of the human and mouse transcriptome using the Gencode hg38 and mm10 annotations respectively. For each organism, we identified the splice junctions that occur in all protein coding isoforms of each protein coding gene. This resulted in a total of 59,477 constitutive splice junctions in the human genome, and 108,481 in the mouse genome.

Then, based on the overall recovery of splice junction reads in a cell, the total expected splice junction reads for a particular cassette exon i in cell j is estimated as:

$$\begin{aligned} \text{SJ}_{E_{ij}} &= m_{ij} \cdot (\text{expected splice junctions per mRNA}) \cdot C_j \\ &= m_{ij} \cdot (1 + \hat{\Psi}_{ij}) \cdot C_j \end{aligned}$$

where $\text{SJ}_{E_{ij}}$ is the expected number of splice junction reads covering the splicing of exon i in cell j (both for mRNAs that splice-in or skip the exon). m_{ij} is the estimated number of mRNAs from the gene containing the cassette exon i in cell j ; $\hat{\Psi}_{ij}$ is the observed splicing rate of exon i in cell j . The expected number of splice junctions per mRNA is $1 + \hat{\Psi}_{ij}$ because 1 splice junction read is present in mRNA molecules that skip the exon, and 2 in those that include it.

Filtering to select good quality observations Simulations of the effect of the initial number of mRNA molecules of a gene and the underlying Ψ suggest that an average of 44 mRNA molecules are necessary to have a 50% chance of making an intermediate Ψ observation when the underlying Ψ is 0.5. This number goes up to 65 if the Ψ is 0.2 or 0.8, and to 127 if the Ψ is 0.1 or 0.9 (Supplementary Figure 1.8f). Assuming a capture efficiency of 10%, we rounded at 10 captured mRNA molecules as the lower threshold for a quality Ψ observation.

In some cases, the number of observed splice junction reads is discordant with the estimated number of mRNAs recovered. Therefore we set a additional filter based on the number of reads expected to come from 10 mRNA molecules that are informative about the splicing of a cassette exon:

$$\text{SJ}_{m_{ij}} = 10 \cdot (1 + \hat{\Psi}_{ij}) \cdot C_j$$

Therefore, for every observation, we required at least 10 mRNAs of the gene captured, and at least the number of reads that we expect if 10 mRNAs are informative. Notice that this minimum will be unique to each observation (combination of cassette exon and cell), as it depends on the cell-specific coverage rate, and the cell and exon specific observed $\hat{\Psi}$.

Kruskal-Wallis test and filter evaluation We evaluate the significance of splicing changes between different cell types, using the clusters defined above, for the Chen, Trapnell, and Song datasets. For each exon, we asked if its median $\hat{\Psi}$ is different between clusters. In order to do this, we grouped the exon’s $\hat{\Psi}$ observations by cluster and ran the Kruskal–Wallis test, which is a non-parametric one-way analysis of variance across all clusters at once. We reported the p-value from each exon. A significant result indicates that the exon has a significantly different median $\hat{\Psi}$ in at least one cluster relative to at least one other cluster, indicating cell type-associated differential splicing. A non-significant result from the test suggests that the exon’s median $\hat{\Psi}$ is not significantly different between any pairs of clusters. The strictness of the differential splicing test was determined by setting different p-value thresholds.

Autocorrelation test We reasoned that exons with reliable estimation of their splicing rate would tend to have similar observations ($\hat{\Psi}$) in cells that are transcriptionally similar to each other. To quantify this, we adapted the autocorrelation test described in [76] to compute, for every exon, the similarity in $\hat{\Psi}$ amongst neighboring cells in the space of the top two principal components of the gene expression space.

To calculate the autocorrelation score of one exon, first we normalized the $\hat{\Psi}$ as follows:

$$\hat{\Psi}'_{ij} = \frac{\hat{\Psi}_{ij} - \bar{\Psi}_j}{\text{Var}(\hat{\Psi}_j)}$$

where $\hat{\Psi}'_{ij}$ is the normalized $\hat{\Psi}_{ij}$ of exon i in cell j ; $\bar{\Psi}_j$ is the average observed $\hat{\Psi}$ of all the tested exons in cell j , and $\text{Var}(\hat{\Psi}_j)$ is the variance of all observed $\hat{\Psi}$ in cell j .

For each cell j , we identified all its k -nearest neighbors in the PCA projection of the normalized gene expression. For each neighbor k , we calculated a similarity score as follows:

$$w_{jk} = \exp\left(\frac{-d_{jk}^2}{\sigma_j^2}\right)$$

where d_{jk} is the Euclidean distance in the PCA projection between cell j and its neighbor k ; σ_j is the Euclidean distance from cell j to its farthest K -nearest neighbor. For all cells k that are not neighbors of j , we set $w_{jk} = 0$. For each analyzed dataset, we set K to be half the total number of cells. The smart-seq and smart-seq2 datasets that we analyzed here contain orders of magnitude fewer cells than the UMI-based datasets analyzed by DeTomaso et al., and we reasoned that the default setting of $K = \sqrt{N}$ might result in neighborhood sizes too small to ensure stability in some datasets. In the datasets with fewer cells, $K = \frac{N}{2}$ agreed better than $K = \sqrt{N}$ with the results of the Kruskal-Wallis test (Jaccard indexes: 0.52 and 0.45 respectively in

the Song dataset, 0.43 and 0.28 in the Trapnell dataset, and 0.51 and 0.54 in the Chen dataset).

For each exon i , we calculate the autocorrelation score as a variation of the Geary's C statistic as:

$$C' = 1 - \frac{(N-1) \sum_j \sum_k w_{jk} (\hat{\Psi}'_{ij} - \hat{\Psi}'_{ik})^2}{2W \sum_j (\hat{\Psi}'_{ij} - \bar{\Psi}'_i)^2}$$

where N is the total number of cells in the dataset, and W is the sum of all w_{ij} .

For each exon, we calculated an empirical p-value. To avoid bias from missing data and different splicing rates, we binned all the exons by average $\hat{\Psi}$ and percent of missing values. We grouped exons with average $\hat{\Psi}$ of: $\{(0.05 - 0.1 \& 0.9 - 0.95), (0.1 - 0.2 \& 0.8 - 0.9), (0.2 - 0.3 \& 0.7 - 0.8), (0.3 - 0.4 \& 0.6 - 0.7), (0.4 - 0.6)\}$. We also binned exons with missing observations between: $\{(50 - 60\%), (60 - 70\%), (70 - 80\%), (80 - 90\%), (90 - 100\%)\}$.

For each bin, we randomized the values of the exons across the cells 20,000 times. We then calculated the autocorrelation score for all the randomized exons.

For each exon, we calculated its p-value as $p = \frac{x+1}{20,001}$, where x are all the randomized exons of the same bin with a higher autocorrelation score.

Splicing change analysis and filter evaluation We evaluated if a subset of exons is enriched for exons with low p-values indicating significant change. We repeated this analysis for the p-values from the Kruskal-Wallis test and from the autocorrelation test. For each x as a threshold of low p-values, we define:

1. M as the set of all exons.
2. m as the set of exons with p-value $\leq x$.
3. P as the exons in the selected subset.
4. p as the exons in the subset with p-value $\leq x$.

We then calculate the fold enrichment as follows:

$$\text{enrichment} = \frac{|M| \cdot |p|}{|P| \cdot |m|}$$

We then use the hypergeometric test to determine if the subset of selected exons is significantly enriched for exons with p-values below the threshold. In our analysis, we tested x in a range between 0.1 and 0.00001, and corrected the p-values of the hypergeometric test for multiple testing using the Benjamini-Hochberg correction.

In addition, we define:

1. p as true positives.
2. $\neg m$ AND $\neg P$ as true negatives.
3. $\neg m$ AND P as false positives.
4. m AND $\neg P$ as false negatives.

Using these definitions, we calculate the precision, recall, specificity, accuracy and F1 score of the filter for every x .

1.6.2 Linear regression on transcriptional burst kinetics

Transcriptional burst kinetics parameters (burst frequency and burst size) of mouse embryonic stem cells were obtained from [71]. The authors modeled these parameters from cells from the Chen dataset [69]. We selected 619 intermediate exons observed in the ES2i cells that have binary observations (i.e., $\hat{\Psi} = 0$ or 1) in between 1% and 99% of the cells. For each exon, we calculated the logit of the proportion of cells that present binary observations as the target variable. Additionally, we matched each exon to three predictive features: the transcriptional burst size of its gene, the transcriptional burst frequency of its gene, and its expression, represented as the average number of informative splice junction reads that cover the exon. Each predictive feature was transformed by $\log_{10} + 1$, and all variables were scaled to its standard score. We trained a linear regression model to predict the logit of the proportion of binary exons, using several combinations of the previously described predictive features:

1. $\text{logit}(\Psi) \sim \text{burst size}$
2. $\text{logit}(\Psi) \sim \text{burst frequency}$
3. $\text{logit}(\Psi) \sim \text{size} + \text{freq} + \text{size} \cdot \text{freq}$
4. $\text{logit}(\Psi) \sim \text{expression}$
5. $\text{logit}(\Psi) \sim \text{size} + \text{freq} + \text{exp} + \text{size} \cdot \text{freq} + \text{size} \cdot \text{exp} + \text{freq} \cdot \text{exp}$

We evaluated each model by calculating the R^2 score between the $\text{logit}(\Psi)$ and the regression prediction.

1.6.3 Theoretical analysis of the observed $\hat{\Psi}$ with limited capture rate

mRNA molecules are captured at a limited rate, approximated in some instances as 10% of the molecules in the cell. Under the assumption of uniform sampling of transcripts and isoforms, and assuming the only nuisance factor is the limited capture rate, we formalize the probability for observing a splicing ratio $\hat{\Psi}$. We start

by specifying this probability, assuming that we know the total number of transcripts from the respective gene in the cell (m), the real splicing rate Ψ and the number of captured molecules r (assuming that for any capture molecule we know if it includes the exon or not). In that case:

$$Pr(\hat{\Psi} \mid \Psi, r, m) = \frac{\binom{m\Psi}{r\hat{\Psi}} \binom{m(1-\Psi)}{r(1-\hat{\Psi})}}{\binom{m}{r}}$$

Note that for this calculation, the capture efficiency (c) is not needed, since we assume that we know m and r . For a more useful analysis, we will next assume that only one of these variables is not known (starting with m and then r).

In a more realistic scenario, r and c can be estimated (e.g., using Census), while m remains unknown. We can therefore marginalize m to calculate:

$$\begin{aligned} Pr(\hat{\Psi} \mid \Psi, r, c) &= \sum_{m=0}^{\infty} Pr(\hat{\Psi}, m \mid \Psi, r, c) \\ &= \sum_{m=r}^{\infty} Pr(\hat{\Psi} \mid \Psi, r, c, m) \cdot Pr(m \mid r, c) \\ &= \sum_{m=r}^{\infty} Pr(\hat{\Psi} \mid \Psi, r, m) \cdot Pr(m \mid r, c) \end{aligned}$$

To estimate $Pr(m \mid r, c)$ we note the following:

$$\begin{aligned} Pr(m \mid r, c) &= \frac{Pr(r \mid c, m) \cdot Pr(m)}{\sum_{m'=0}^{\infty} Pr(r \mid c, m') \cdot Pr(m')} \\ &= \frac{Pr(r \mid c, m)}{\sum_{m'=0}^{\infty} Pr(r \mid c, m')} \\ &= \frac{\binom{m}{r} c^r (1-c)^{m-r}}{\sum_{m'=0}^{\infty} Pr(r \mid c, m')} \end{aligned}$$

where we model the probability of capturing r mRNA molecules as a binomial sample from m with probability c . Note that the third transition is done under the assumption of a uniform prior on m .

To compute the denominator, we expand:

$$\begin{aligned}
\sum_{m'=0}^{\infty} \Pr(r \mid c, m') &= \sum_{m'=r}^{\infty} \binom{m'}{r} c^r (1-c)^{m'-r} \\
&= c^r \sum_{k=0}^{\infty} \binom{r+k}{r} (1-c)^k = c^r \sum_{k=0}^{\infty} \frac{(r+k)!}{k!r!} (1-c)^k \\
&= c^r \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} (r+1)(r+2) \dots (r+1+(k-1))(c-1)^k \\
&= c^r \sum_{k=0}^{\infty} \frac{1}{k!} (-(r+1))(-(r+2)) \dots (-(r+k)) 1^{r+k+1} (c-1)^k \\
&\quad \text{by Taylor series centered in } 1 = c^r \frac{1}{c^{r+1}} = \frac{1}{c}
\end{aligned}$$

Thus

$$\begin{aligned}
\Pr(\hat{\Psi} \mid \Psi, r, c) &= \sum_{m=r}^{\infty} \Pr(\hat{\Psi} \mid \Psi, r, c, m) \cdot \Pr(m \mid r, c) \\
&= \sum_{m=r}^{\infty} \frac{\binom{m\Psi}{r\hat{\Psi}} \binom{m(1-\Psi)}{r(1-\hat{\Psi})}}{\binom{m}{r}} \frac{\binom{m}{r} c^r (1-c)^{m-r}}{\frac{1}{c}} \\
&\approx \sum_{m=r}^{10r/c} \binom{m\Psi}{r\hat{\Psi}} \binom{m(1-\Psi)}{r(1-\hat{\Psi})} \cdot c^{r+1} (1-c)^{m-r}
\end{aligned}$$

In the last equation, we estimate the sum going only up to a large value of m , since its posterior probability diminishes. In expectation $m \approx r/c$. We use ten times this value as the maximum.

We can use this equation to estimate the expected proportion of binary $\hat{\Psi}$ observations ($\hat{\Psi} = 0$ or 1) that is expected when we observe only r junctions from a splicing event with a given true rate Ψ (Supplementary Figure 1.6a). We can also estimate the chance to have an empirical $\hat{\Psi}$ that is at least within a certain delta (in absolute terms) from the real Ψ . Namely we can estimate $\Pr(|\hat{\Psi} - \Psi| < \delta \mid \Psi, r, c)$ (Supplementary Figure 1.6d).

In another calculation of interest, one can ask how many mRNA molecules should a gene have in a cell in order to correctly estimate the splicing rate, under a limited capture efficiency c . To estimate it, we denote by d a binary variable indicating that the gene has been detected (i.e., $r > 0$) and marginalize r in the following way:

$$\Pr(\hat{\Psi} \mid \Psi, c, m, d) = \frac{\Pr(\hat{\Psi}, d \mid \Psi, c, m)}{\Pr(d \mid c, m)}$$

Where

$$\begin{aligned}
\Pr(\hat{\Psi}, d \mid \Psi, c, m) &= \sum_{r=1}^m \Pr(\hat{\Psi}, r \mid \Psi, c, m) \\
&= \sum_{r=1}^m \Pr(\hat{\Psi} \mid \Psi, c, m, r) \cdot \Pr(r \mid c, m) \\
&= \sum_{r=1}^m \frac{\binom{m\Psi}{r\hat{\Psi}} \binom{m(1-\Psi)}{r(1-\hat{\Psi})}}{\binom{m}{r}} \cdot \binom{m}{r} c^r (1-c)^{m-r} \\
&= \sum_{r=1}^m \binom{m\Psi}{r\hat{\Psi}} \binom{m(1-\Psi)}{r(1-\hat{\Psi})} \cdot c^r (1-c)^{m-r}
\end{aligned}$$

and

$$\Pr(d \mid c, m) = 1 - (1 - c)^m$$

We use this in Supplementary Figure 1.6b to plot the chances to see only one isoform (binary $\hat{\Psi}$) for a fixed Ψ (set to 0.5) as a function of the number of molecules present in the cell (m).

1.6.4 Probabilistic simulator of splicing in single cell data

1) Biological process. We simulate the expression of 1500 hypothetical genes in 300 cells using SymSim, an *in silico* simulator of gene expression in single cells [75]; the expression of gene g in cell i is annotated as X_i . We simulate one cassette exon j for each gene g . For each cassette exon j in cell i , we simulate an underlying splicing distribution Ψ_{ij} as a Beta distribution with exon-specific parameters α_j and β_j . The splicing of j in i is simulated as a binomial sampling from X_{ig} with probability Ψ_{ij} .
2) Technical process. We simulate the capture, fragmentation and sequencing of each transcript using a modified version of SymSim's `True2ObservedCounts` function and a random vector of transcript lengths. Finally, we subsample the obtained reads based on the transcript length in order to simulate the coverage of informative splice junctions.

We simulate the splicing of cassette exons in a set of genes G expressed in a population of cells N . For each gene $g \in G$, we simulate the splicing of one cassette exon j . The inclusion of j forms the isoform j_A , while the exclusion of the exons forms the isoform j_B . The production of mRNA molecules from j_A in a single cell $i \in N$ is determined by the total expression of g , and by the action of the splicing machinery of i .

1. Biological process

Splicing from pre-mRNA transcripts in individual cells For each alternatively spliced gene g in a cell i , we simulate the expression of g and the splicing of its cassette exon.

$$\begin{aligned}
X_{ig} & \quad \text{expression of gene } g \text{ in } n. \\
\alpha_j, \beta_j & \quad \text{splicing rate distribution parameters.} \\
\Psi_{ij} & \sim \text{Beta}(\alpha_j, \beta_j) \\
X_{A_{ij}} & \sim \text{Binomial}(X_{ig}, \Psi_{ij}) \\
X_{B_{ij}} & = X_{ig} - X_{A_{ij}} \\
\Psi_{T_{ij}} & = X_{A_{ij}} / X_{ig}
\end{aligned}$$

X_{ig} represents the total number of pre-mRNA transcribed from each gene across all cells. We simulate the total counts using SymSim, an *in-silico* approach for simulation of single cell gene expression by accounting for the biological sources of variation [75]. Ψ_{ij} referred to as the underlying splicing rate, is the probability of splicing-in the cassette exon j of gene g in cell i . Notice that in this simulation, each gene g only has one cassette exon j . In a biological context, Ψ_{ij} would be determined by intrinsic attributes of the cassette exon inherent of g (e.g., sequence, secondary structure, binding sites), and by the profile of splicing factors expressed in n . $X_{A_{ij}}$ and $X_{B_{ij}}$ are respectively the counts of mRNA molecules from isoforms g_A and g_B in i . Notice that $X_{A_{ij}}$ is a random binomial sample from the total number of expressed pre-mRNA molecules of g in i with a probability Ψ_{ij} , as it has been modeled before [39, 66, 84, 85]. $\Psi_{T_{ij}}$ is the true isoform ratio that include cassette exon j in cell i , obtained as the proportion of molecules of gene g that include the cassette exon j .

The distribution of Ψ_{ij} across all cells $i \in N$ is modeled as a Beta distribution, which has been used in previous studies of single cell splicing [56, 57, 86]. In this model, the distribution is determined by the parameters $\alpha_j, \beta_j \in (0, \infty)$. The values of these parameter determine the distribution of Ψ_{ij} across all cells as follows:

- Unimodal with intermediate mode if $\alpha_j, \beta_j > 1$.
- Unimodal with mode 1 if $0 < \alpha_j < 1 \leq \beta_j$.
- Unimodal with mode 0 if $0 < \beta_j < 1 \leq \alpha_j$.
- Bimodal with modes 0 and 1 if $0 < \alpha_j, \beta_j < 1$.
- Uniform if $\alpha_j = \beta_j = 1$.

Notice that $\frac{\alpha_j}{\alpha_j + \beta_j} = \mu(\Psi_{ij})$. By controlling the α_j, β_j parameters we can compare the biological underlying distribution of the exon splicing rate with the observed distribution of Ψ inferred from single cell RNA-seq data.

To compare the results of our simulations under the binary-bimodal (alternative isoforms are present in the population, but rarely in the same cell; Figure 1.3a,c,d) and non-binary unimodal (both isoforms regularly appear in the same cell; Figure 1.3b,e,f) models of splicing, we simulated 500 alternatively spliced exons for each model. For the first model, we simulated the underlying splicing distribution of each exon as Bimodal Beta distributions. For the second model, we simulated the underlying splicing distributions of each exon with Unimodal Beta distributions with intermediate mode. For each cassette exon j in cell i :

- Binary-bimodal splicing:

$$\begin{aligned}\Psi_{B_{ij}} &\sim \text{Beta}(\alpha_{B_j}, \beta_{B_j}) \\ \alpha_{B_j}, \beta_{B_j} &\sim \frac{1}{\text{Uniform}(1, 30)}\end{aligned}$$

- Non-binary unimodal splicing:

$$\begin{aligned}\Psi_{U_{ij}} &\sim \text{Beta}(\alpha_{U_j}, \beta_{U_j}) \\ \alpha_{U_j}, \beta_{U_j} &\sim \text{Uniform}(1, 30)\end{aligned}$$

To simulate a realistic scenario in both models, we simulated 500 additional exons that are consistently included, and 500 that are consistently excluded. For the consistently included exons, we sampled the parameters for Unimodal Beta distributions with mode 1 as

$$\alpha_j \sim \text{Uniform}(1, 30); \beta_j \sim \frac{1}{\text{Uniform}(1, 5)}$$

For the consistently excluded exons, we sampled the parameters for Unimodal Beta distributions with mode 0 as

$$\alpha_j \sim \frac{1}{\text{Uniform}(1, 5)}; \beta_j \sim \text{Uniform}(1, 30)$$

2. Technical process

mRNA capture into cDNA After simulating the production of mRNAs of distinct isoforms in single cells, we simulate the process of capture and sequencing of mRNA molecules from $X_{A_{ij}}$ and $X_{B_{ij}}$.

- c expected capture efficiency
- $c_{I_{ng}}$ drawn from truncated normal with mean c and variance 0.002
- $C_{I_{ng}} \sim \text{Binomial}(X_{I_{ng}}, c_{I_{ng}})$

The process of mRNA capture is simulated using SymSim. $C_{I_{ng}}$ is the number of mRNA molecules of isoform $I \in \{g_A, g_B\}$ for gene g on cell n . This number is sampled from the total number of molecules for isoform that are present in the cell. c is a parameter that determines the expected capture efficiency. $c_{I_{ng}}$ is the specific probability of capture of isoform I of gene g in cell n , which is drawn from a truncated normal distribution with mean c . The variance of the truncated normal distribution is set at 0.002, which is the default variance in SymSim. The distribution is truncated at 0 and at 1.

RNA sequencing Sequencing is also simulated using SymSim’s approach, which includes a length-dependent PCR amplification bias modeled from experimental data.

$$\begin{aligned} l_{g_B} & \text{ sampled from SymSim’s database} \\ l_{e_A} & \text{ sampled from exon database} \\ l_{g_A} &= l_{g_B} + l_{e_A} \\ R_{I_{ng}} &= \text{True2ObservedCounts}(C_{I_{ng}}, l_{g_I}) \end{aligned}$$

l_{g_B} is the length of the mRNA transcript of isoform g_B of gene g , which represents the isoform that skips the alternative exon. l_{g_A} is the length of the mRNA transcript of the isoform g_A , which includes the exon. We assigned the lengths to the excluded isoform by drawing without replacement from SymSim’s transcript length database. l_{e_A} is the length of the alternative exon included in isoform g_A ; it is sampled without replacement from a database of skipped exon lengths from the human genome. $R_{I_{ng}}$ is the total number of observed reads from isoform I of gene g in cell n obtained from single cell RNA sequencing. These are obtained using the True2ObservedCounts function from SymSim with the modification previously described.

Splice junction coverage and observed Ψ calculation We also simulate the down-sampling from observing only reads that overlap the splice junctions that are informative about the splicing of the cassette exon.

$$\begin{aligned} l_r &= \text{read length (constant)} \\ j_{A_g} &= \frac{4(l_r - 1)}{l_{A_g}} \\ j_{B_g} &= \frac{2(l_r - 1)}{l_{B_g}} \\ \text{SJ}_{I_{ng}} &\sim \text{Binomial}(R_{I_{ng}}, j_{I_g}) \end{aligned}$$

l_r corresponds to the constant read length from the sequencing process. $\text{SJ}_{I_{ng}}$ is the number of reads that cover informative splice junctions for isoform $I \in \{g_A, g_B\}$ for gene g in cell n , which are sampled from the total number of reads covering the isoform. In order to account for variation in read density along transcripts, we sample

the splice junction reads of each event from a binomial distribution with probabilities j_{A_g} and j_{B_g} , which are respectively the probabilities of a given read to cover the splice junctions informative with isoform g_A and g_B . We derive these probabilities from the length of the simulated reads, and the length of the transcript. Each read can be mapped to $2(l_r - 1)$ positions in the transcript that overlap one splice junction. Thus, the probability of covering one given splice junction is defined as the number of possible positions in the transcript that are informative for the splice junction, divided by the length of the transcript. j_{A_g} is the probability to map to any of the two splice junctions that are informative for isoform g_A . j_{B_g} is the probability to map to one single splice junction, since there is only one junction informative for isoform g_B . Coverage depth variance might sometimes exceed the variance of the binomial distribution, such that the informative reads for one isoform might be much more likely to be sequenced than the informative reads of the other isoform in a given cell. In these situations, the technical variance of $\hat{\Psi}$ and the number of spurious binary observations could be higher than modeled in these simulations.

Finally, the observed Ψ is calculated as:

$$\hat{\Psi}_{ij} = \frac{\text{SJ}_{A_{ij}}}{\text{SJ}_{A_{ij}} + \text{SJ}_{B_{ij}}}$$

Simulator variants for studying sources of variation

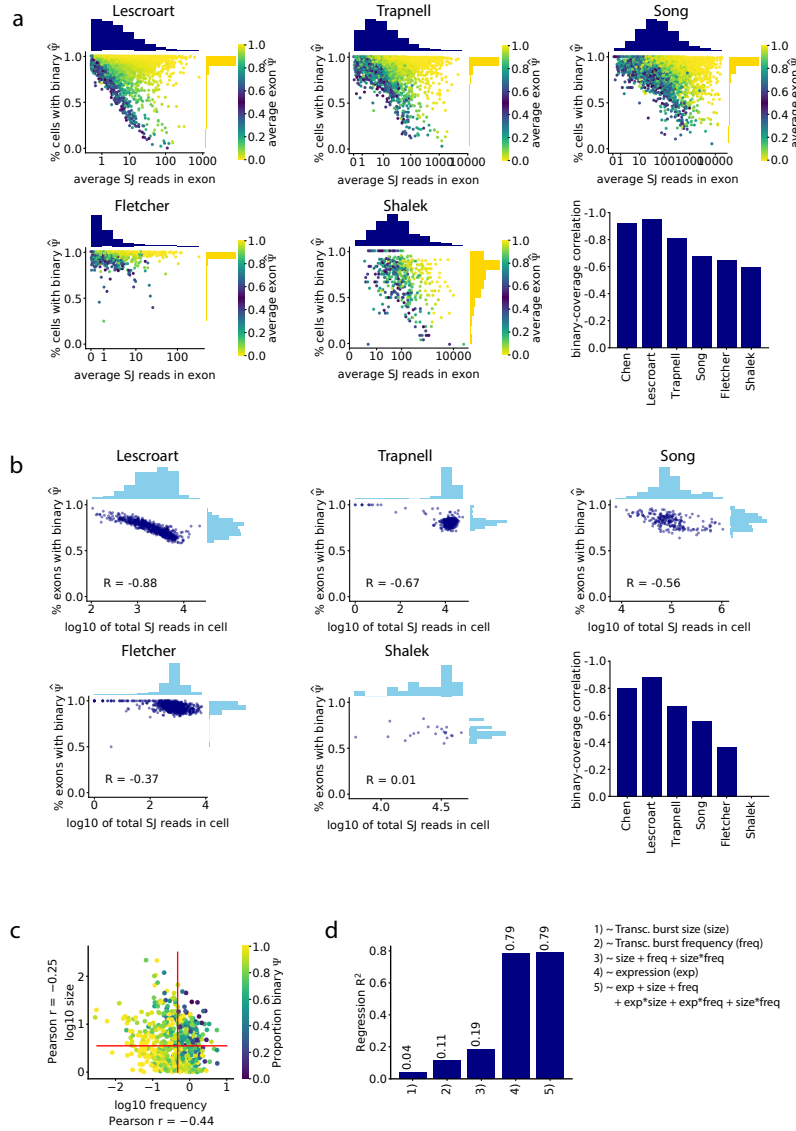
Gene expression and underlying Ψ We tested the effect of the interplay between gene expression and the ratio of isoforms that contain the cassette exon on the observed distribution of Ψ . For this test, we simulated a population of 300 single cells with 500 genes, indexed 1 to 500. For every cell i , the expression of gene g is fixed as $X_{ig} = g$, where $g \in \{1, 2, \dots, 500\}$. That is, every gene had a different level of expression, and the expression of every individual gene was constant across all cells. For each simulation, we fixed the underlying splicing rate of all cassette exons across all cells. That is, for each cassette exon j of gene g , in every cell, we set $\Psi_{ij} = \text{constant}$. We ran the simulator with different underlying splicing rates, with $\Psi_{ij} \in \{0.01, 0.02, \dots, 0.5\}$. For every simulation we used an average capture efficiency $c = 0.1$. We ran 30 simulations for every fixed Ψ_{ij} value. For every fixed Ψ_{ij} and for every fixed expression level g , we took the average proportion of cells with binary values ($\hat{\Psi} = 0$ or 1) for the observed $\hat{\Psi}$. That is, we reported:

$$\frac{1}{1500} \sum_{\text{sim}=1}^{30} \sum_{i=1}^{300} \mathbb{I}(\hat{\Psi}_{ij} = 1) + \mathbb{I}(\hat{\Psi}_{ij} = 0)$$

Gene expression and capture efficiency We tested the effect of capture efficiency in Ψ observations. To minimize the effect of the underlying Ψ in the simulations, in this analysis we fixed the true splicing rate of all exons to $\Psi_{Tng} = 0.5$ (we

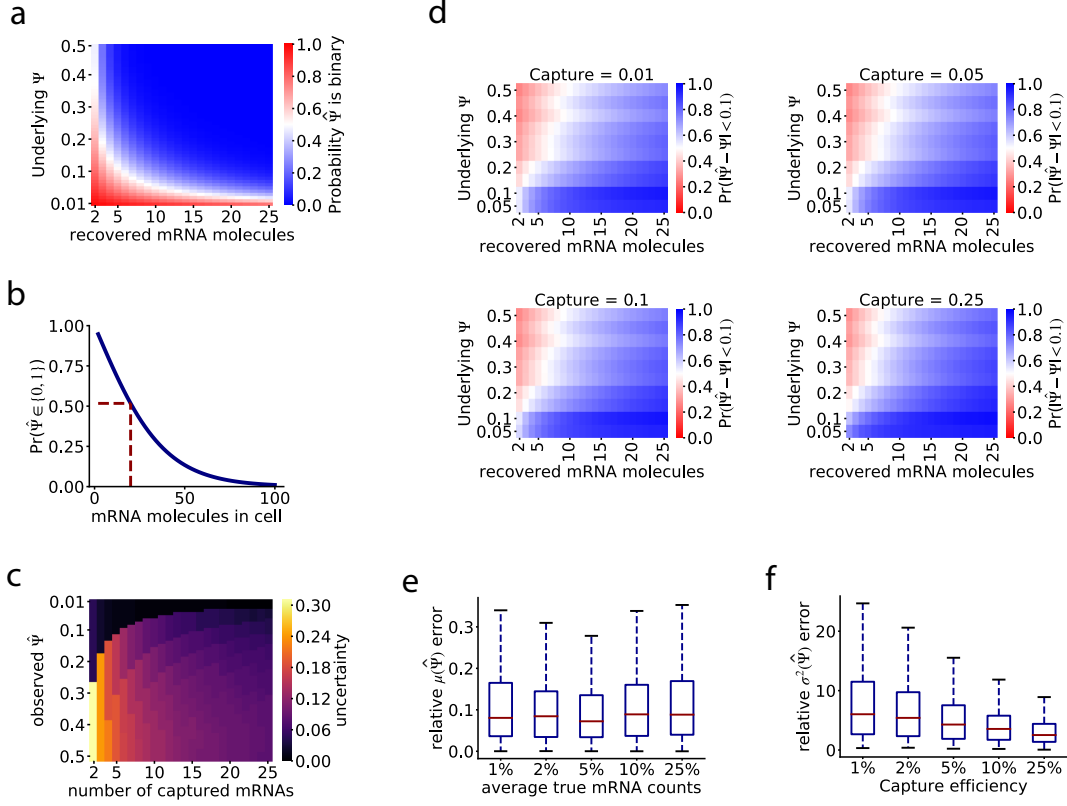
achieved this by setting $X_{Aij} = X_{Bij} = \frac{1}{2}X_{ig}$). We ran simulations for each possible value for the average capture efficiency in $c \in \{0.01, 0.011, \dots, 0.1\}$. For each tested average capture efficiency rate, we ranked the alternative splicing events by the number of reads that cover the informative splice junctions. For each alternative event, we observed the proportion of cells that present only one type of isoform (either including the cassette exon or excluding it, but not both).

1.7 Chapter appendix: supplementary tables and figures

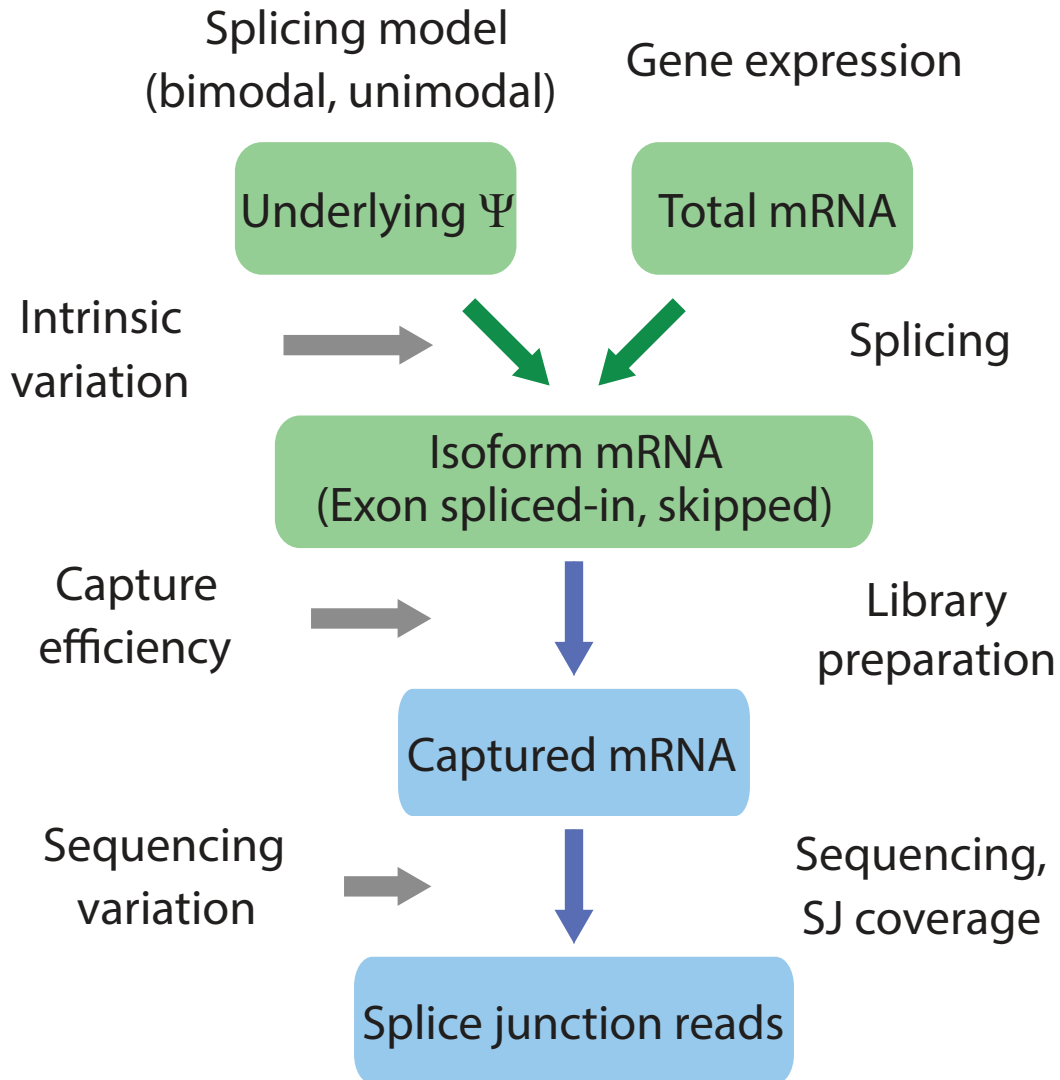


Supplementary Figure 1.5: Junction read coverage determines the proportion of binary observations in all analyzed datasets. a) As in Figure 1.2c, the splice junction read coverage of an intermediate exon was anti-correlated with the proportion of cells in which it shows binary $\hat{\Psi}$ (only one isoform is observed) in all datasets. Barplot: Pearson correlation between read coverage and proportion of binary observations of each cassette exon. b) As in Figure 1.2d, the total number of splice junction reads in a cell was inversely proportional to the fraction of exons that have binary values in the cell in all datasets. Barplot: Pearson correlation between the total number of captured reads in each cell and the proportion of cassette exons with binary $\hat{\Psi}$.

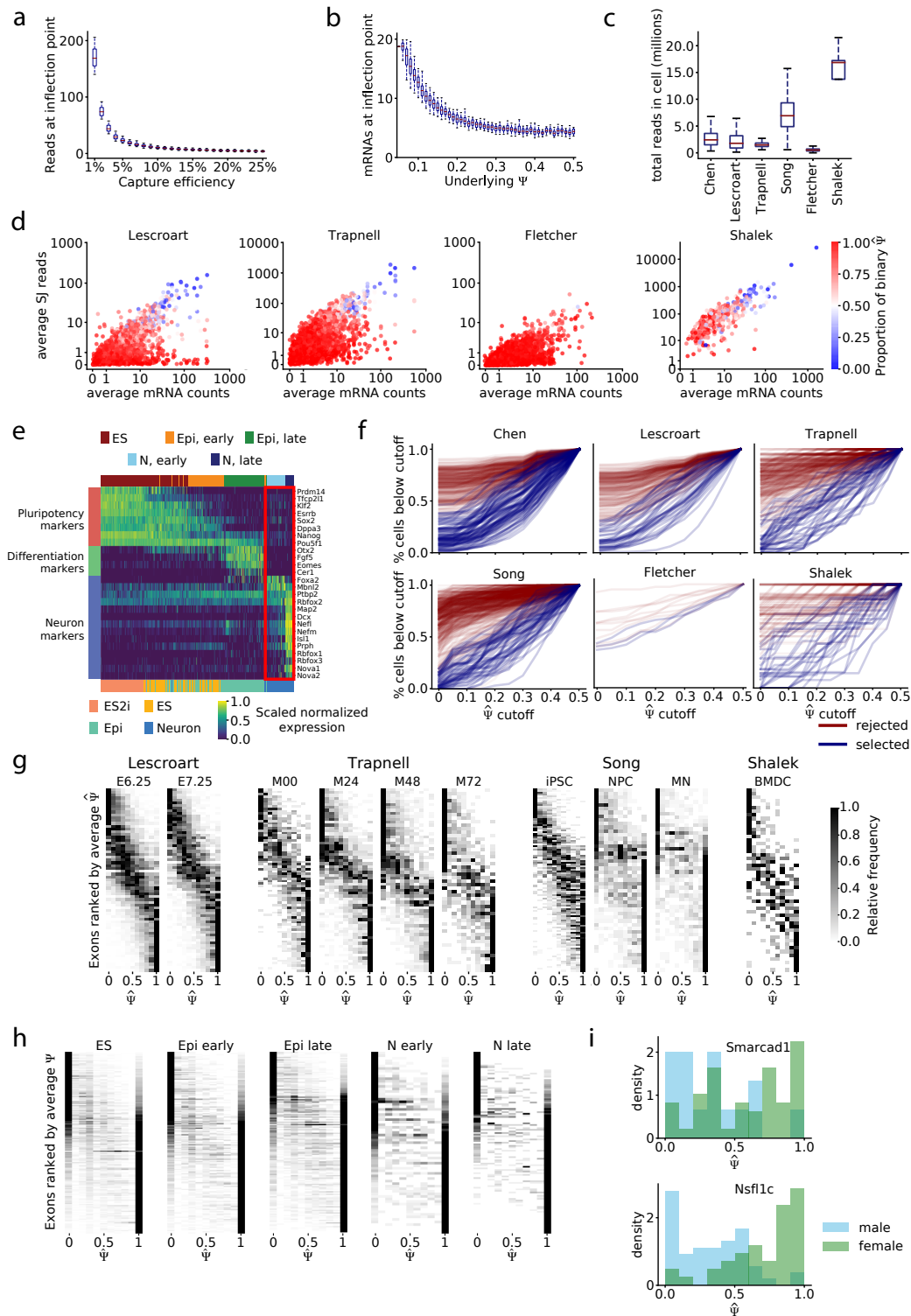
Supplementary Figure 1.5: c) Relationship between transcriptional burst frequency and size, and the proportion of binary observations of an exon. d) R^2 of the linear regression prediction of the proportion of binary observations per exon based on burst frequency, size, number of reads, and a combination of these features.



Supplementary Figure 1.6: Theoretical calculations and simulations of the effect of biological and technical factors in splicing observations. a) Theoretical likelihood of capturing only mRNAs representing one isoform of an alternatively spliced gene in a single cell, determined by the total number of mRNAs in the cell, the Ψ of the isoforms, and a 10% capture efficiency. b) Probability of observing a binary $\hat{\Psi}$ (only one isoform observed) given the number of mRNA molecules of that gene present in the cell, when the underlying Ψ is fixed at 0.5. c) Effect of the number of captured mRNA molecules on the uncertainty of $\hat{\Psi}$ observations. $\hat{\Psi}$ is understood as the maximum likelihood estimate of the underlying Ψ . d) The posterior probability of having an observed $\hat{\Psi}$ within 0.1 of the underlying Ψ . e) Distribution of the relative error between the average observed $\hat{\Psi}$, and the average true Ψ , in our simulations under different capture efficiency rates. The relative error is calculated as the difference between the average $\hat{\Psi}$ and the average true Ψ , scaled by the average true Ψ . f) Distribution of the relative error between the variance in the observed $\hat{\Psi}$, and the variance in the true Ψ , in our simulations under different capture efficiency rates.

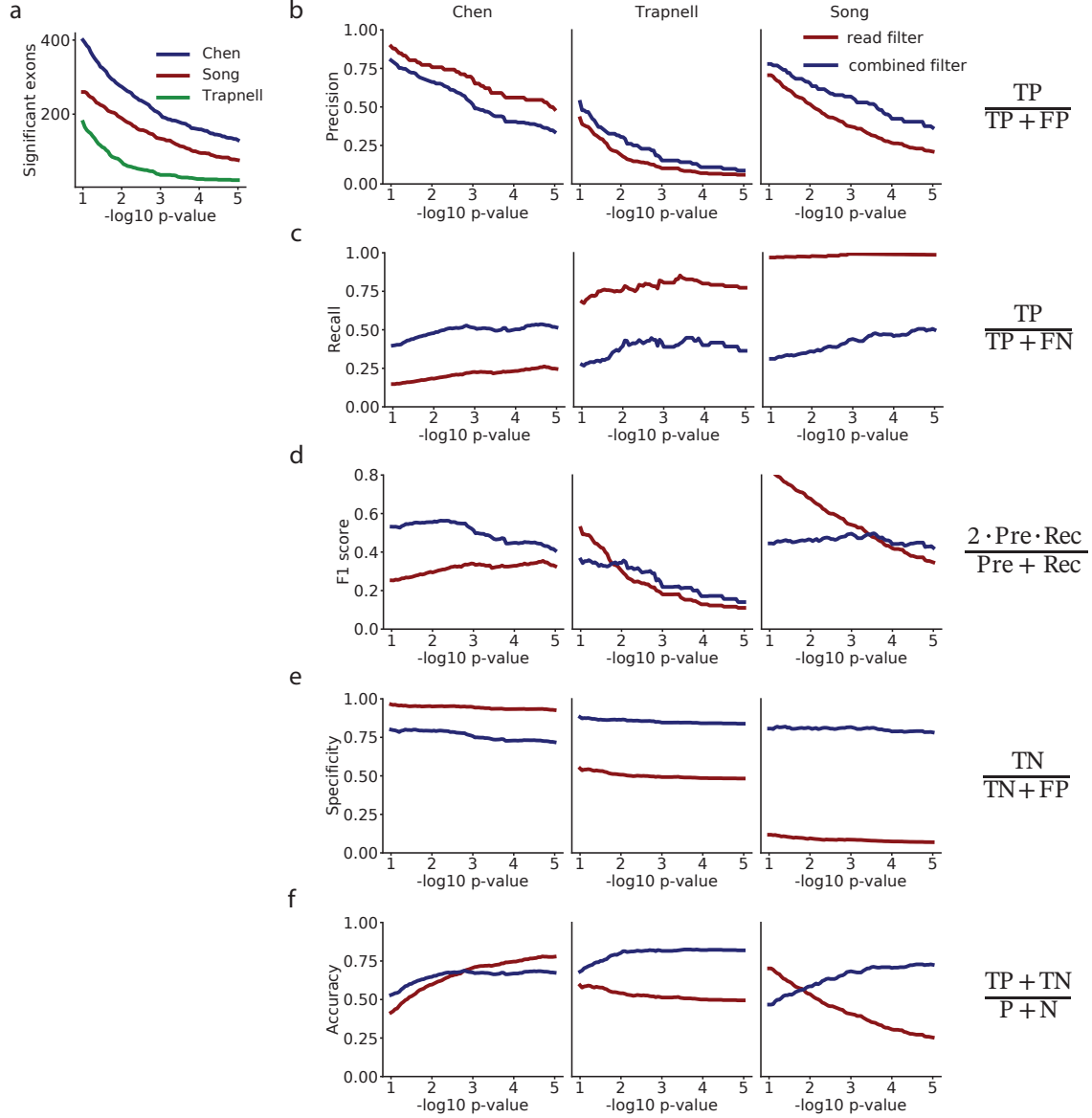


Supplementary Figure 1.7: Schematic of scRNA-seq splicing simulator. Simulator of scRNA-seq splicing data. Green elements represent biological variables; blue elements represent technical processes. The underlying Ψ is drawn from a Beta distribution with either a bimodal or unimodal shape. Individual gene expression determines the total number of mRNAs, and these mRNAs are spliced stochastically according to Ψ , producing isoforms that splice in or skip the exon. mRNAs are captured with a probability drawn from a normal distribution. Sequencing produces splice junction reads, which determine the observed $\hat{\Psi}$.

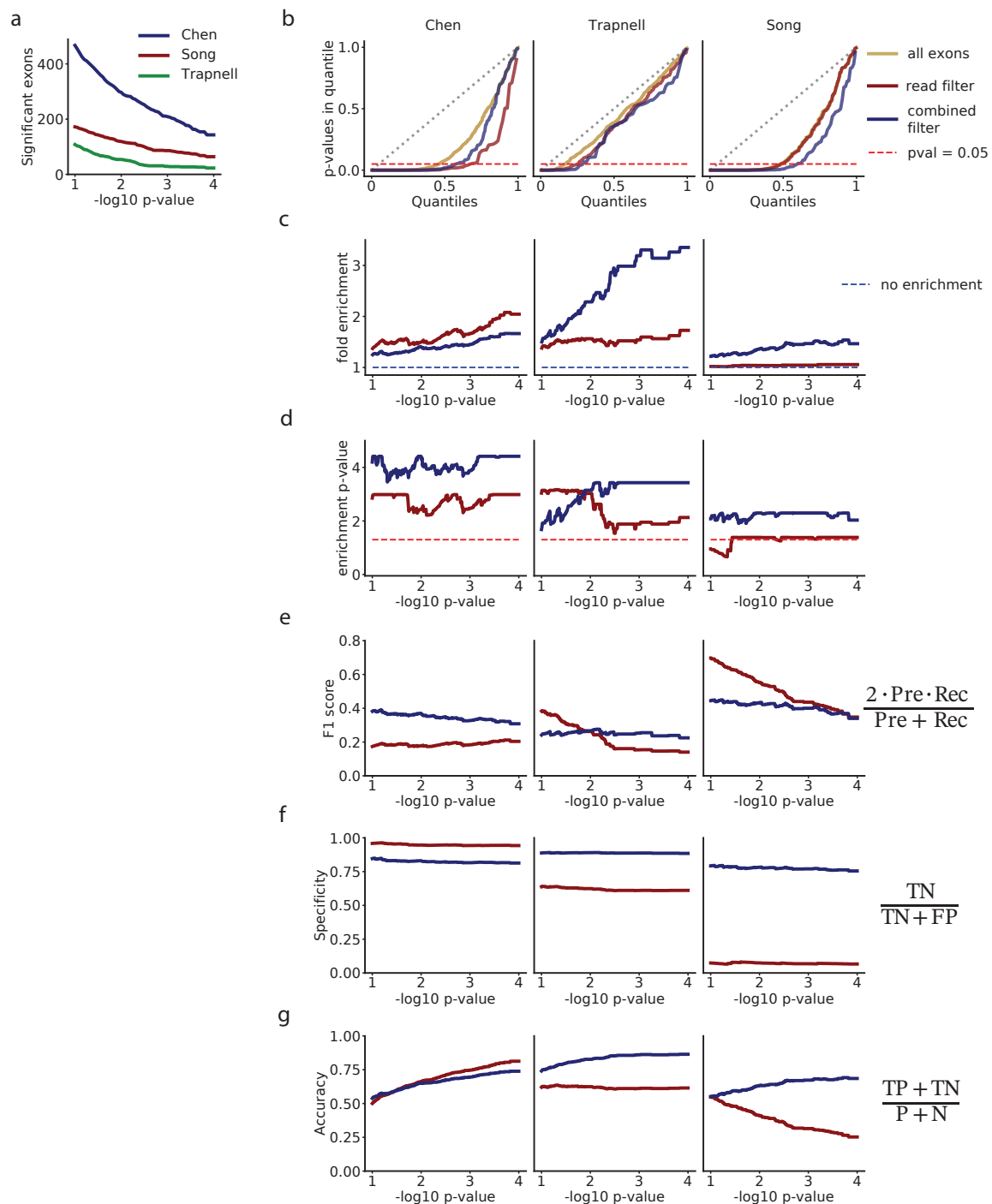


Supplementary Figure 1.8: Relationship between read coverage, captured mRNAs, and binarity in single cell datasets.

Supplementary Figure 1.8: a) Average number of splice junction reads required to have a 50% likelihood of observing both isoforms, from simulations in Figure 1.3j. b) Total mRNAs required to have a 50% likelihood of observing both isoforms, from simulations in Figure 1.3k. c) Total mapped reads per cell for each dataset. d) Estimated average number of recovered mRNAs vs. average number of splice junction reads for each exon. e) Expression of pluripotency and neuron differentiation marker genes in the Chen dataset; expression is scaled by each row's maximum and minimum. Cells are ordered by inferred pseudotime. Top color key, clusters from agglomerative clustering; bottom color key, labels from Chen et al. The red box highlights two sub-groups of neurons identified by clustering. f) Visualization of the distribution of $\hat{\Psi}$ observations towards extreme values in the first cluster of cells of each dataset. For each x cutoff in the x-axis, the y-axis corresponds to the proportion of cells that have $\hat{\Psi} \leq x$ or $\hat{\Psi} \geq (1 - x)$. Blue lines, exons passing the mRNA-based filter; red lines, exons removed by filter. g) Stacked histograms, as in Figure 1.4h, for other datasets. The Fletcher dataset is excluded because only five exons pass the mRNA filter. h) Stacked histograms of exons that do not pass the filter in the Chen dataset; exons were subsampled to match the number that do pass the filter in each cell type cluster. i) Sex-dependent distribution of $\hat{\Psi}$ of the two bimodal exons of the Chen dataset after cell-type stratification: Smarcd1 exon 3 (chr6: 65043836–65044108) in early Epi cells, Nsf1c exon 4 (chr2: 151502455–151502460) in late Epi cells.

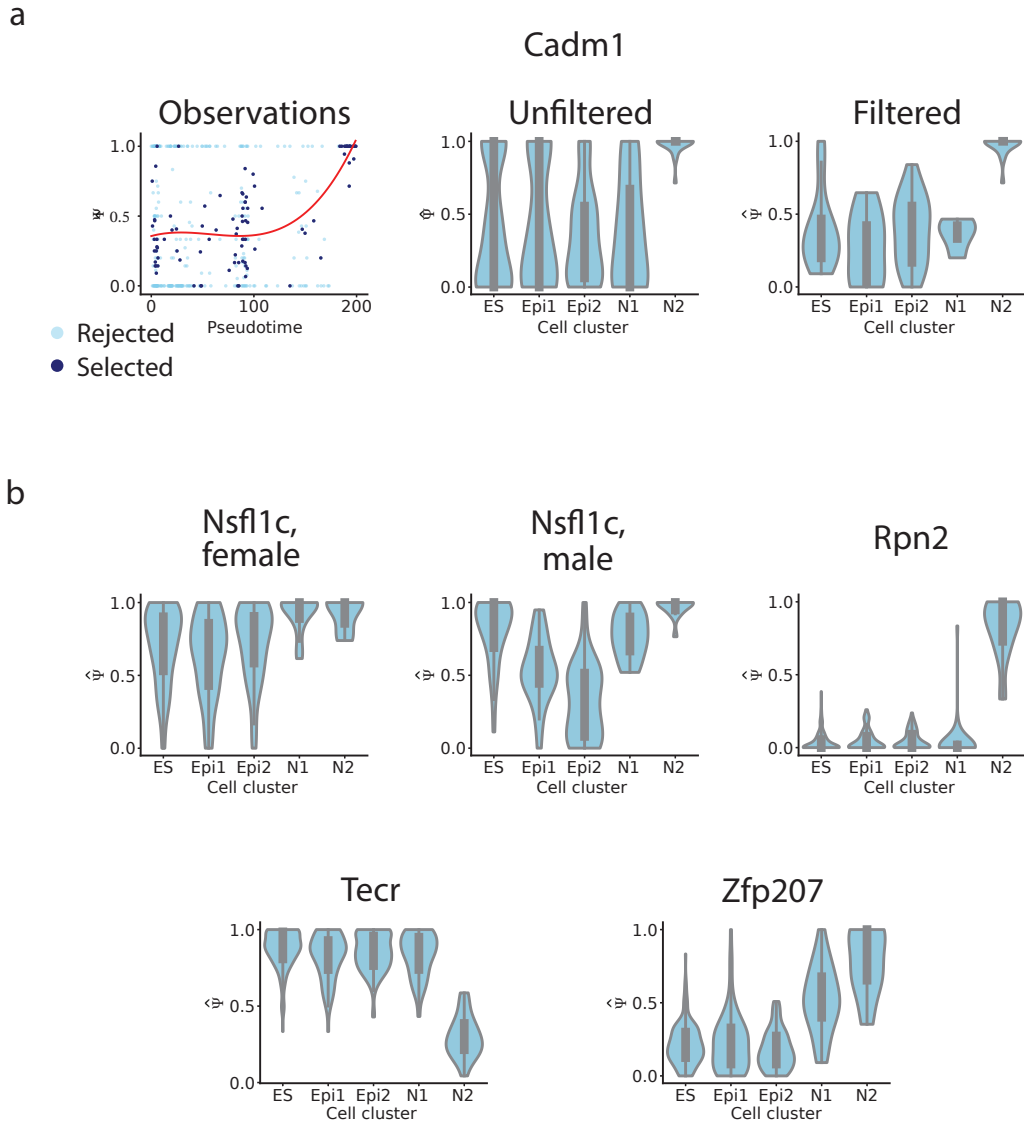


Supplementary Figure 1.9: Analysis of differential splicing among selected exons with the Kruskal-Wallis analysis of variance. a) Comparison of the precision of selecting exons with low p-values, for the mRNA-based filter versus selecting exons on a flat minimum of 10 reads. b) Recall. c) F1-score. d) Specificity. e) Accuracy. f) Number of exons with $p \leq x$ in each dataset.



Supplementary Figure 1.10: Analysis of differential splicing among selected exons with the autocorrelation test. a) Number of exons with $p \leq x$ in each dataset. b) QQ-plot comparing the quantiles of a uniform distribution, with the quantiles of the distributions of p-values from the autocorrelation test.

Supplementary Figure 1.10: c) Fold enrichment of exons with an autocorrelation $p \leq x$ in the set of exons selected with the mRNA-based filter (blue), and exons selected with a flat read minimum filter (red). d) Significance p-value of the enrichment shown in c). For each p-value limit x , we estimate the significance with the hypergeometric test, and correct for multiple testing using the Benjamini-Hochberg FDR adjustment. e) F1-score. f) Specificity. g) Accuracy.



Supplementary Figure 1.11: Distribution of $\hat{\Psi}$ of example exons. a) Skipped *Cadm1* exon 8 (chr9: 47829377–47829409). We show the relationship between pseudotime and observed $\hat{\Psi}$; as well as violinplots of the distribution of $\hat{\Psi}$ in each cluster before filtering, and after filtering. b) Violinplots of the distribution of $\hat{\Psi}$ in each cluster after filtering for the exons in Figure 4k: *Nsfl1c* exon 4 (chr2: 151502455–151502460), *Rpn2* exon 16 (chr2: 157323223–157323270), *Tecr* exon 4 (chr8: 83573411–83573455), *Zfp207* exon 8 (chr11: 80393084–80393176).

Table 1.1: Single cell RNA-seq datasets.

Dataset	Organism	Biological process	Cells	No. ASE	Mean reads/event	Protocol	Accession	Reference
Chen	mouse	mES neuron differentiation	488	3276	3.1	Smart-seq2	GSE74155	[69]
Lescroart	mouse	cardiomyogenesis	598	3007	3.2	Smart-seq2	GSE100471	[87]
Trapnell	human	skeletal myogenesis	314	4457	14.4	Smart-seq	GSE52529	[45]
Song	human	iPS motor neuron differentiation	206	5355	88.8	Smart-seq	GSE85908	[56]
Fletcher	mouse	olfactory neurogenesis	849	684	2.7	Smart-seq	GSE95601	[88]
Shalek	mouse	bone-marrow-derived dendritic cells	18	380	213.4	Smart-seq	GSE41265	[40]

Table 1.2: Prevalence of bimodal splicing distributions before and after filtering.

dataset	cell type	cluster	cells	exons	bimodal	% bimodal	selected	selected bimodal	% bimodal selected	p-val (adj)
Chen	ES		217	446	118	26%	94	0	0%	3.3e-14
Chen	Epi, early		98	402	107	27%	98	1	1%	9.3e-14
Chen	Epi, late		104	516	136	26%	76	1	1%	1.1e-09
Chen	Neuron, early		47	364	117	32%	43	0	0%	3.6e-08
Chen	Neuron, late		22	517	146	28%	61	0	0%	1.1e-09
Lescroart	Heart E6.75		172	286	77	27%	33	0	0%	2.0e-05
Lescroart	Heart E7.25		341	291	78	27%	36	0	0%	8.0e-06
Trapnell	Myoblast 00h		35	400	142	36%	41	0	0%	1.2e-08
Trapnell	Myoblast 24h		89	251	97	39%	27	0	0%	1.2e-06
Trapnell	Myoblast 48h		72	242	97	40%	32	1	3%	9.1e-07
Trapnell	Myoblast 72h		35	252	101	40%	24	1	4%	5.1e-05
Song	iPSC		62	616	269	44%	55	0	0%	3.3e-14
Song	NPC		73	212	92	43%	28	1	4%	1.2e-06
Song	MN		67	168	82	49%	19	4	21%	9.4e-03
Shalek	BMDc		13	149	51	34%	27	1	4%	6.6e-05

Chapter 2

Identifying cell-state associated alternative splicing events and their co-regulation

Publication and authorship permissions

This work will be submitted for publication in 2021 as:

Buen Abad Najar Carlos F., Burra Prakruthi, Yosef Nir, Lareau Liana F. *Identifying cell-state associated alternative splicing events and their co-regulation*. 2021.

The co-authors of this paper granted me their authorization for its reproduction in this dissertation.

Alternative splicing shapes the transcriptome and contributes to each cell's unique identity, but single-cell RNA sequencing has struggled to capture the impact of alternative splicing. We previously showed that low recovery of mRNAs from single cells led to erroneous conclusions about the cell-to-cell variability of alternative splicing [59]. Here, we present a new method, Psix, to confidently identify splicing that changes across a landscape of single cells, using a probabilistic model that is robust against the data limitations of scRNA-seq. Its autocorrelation-inspired approach finds patterns of alternative splicing that correspond to patterns of cell identity, such as cell type or developmental stage, without the need for explicit cell clustering, labeling, or trajectory inference. Applying Psix to data that follow the trajectory of mouse brain development, we identify genes whose alternative splicing patterns cluster into modules of co-regulation. We show that the exons in these modules are enriched for binding by distinct neuronal splicing factors, and that their changes in splicing cor-

respond to changes in expression of these splicing factors. Thus, Psix reveals cell-type-dependent splicing patterns and the wiring of the splicing regulatory networks that control them. Our new method will enable scRNA-seq analysis to go beyond transcription to understand layers of post-transcriptional regulation in determining cell identity.

2.1 Cell-type associated alternative splicing

Transcriptome profiling at a single cell level has revolutionized our understanding of the continuous biological variation in gene expression that determines a cell’s unique identity [41, 42]. Alternative mRNA splicing is a major source of transcriptome variability that plays an important role in determining the identity of a cell. The fine-tuned regulation of splicing contributes to many continuous biological processes such as neurogenesis [27, 28], while its misregulation is associated with complex diseases [11–17]. Thus, a full understanding of cell identity at a single-cell level requires measuring changes in alternative splicing and decoding their regulation.

Despite the enormous progress in computational modeling of cell identity from single-cell gene expression studies [48–50, 89], formidable challenges remain in capturing the impact of alternative splicing [53]. Alternative isoforms are distinguished by only a few specific regions of the transcript, even when using methods that capture reads along the full length of a transcript. On top of this major data limitation, low capture efficiency in single cell RNA-sequencing adds enormous technical variation. As a result, the observed distribution of an exon’s inclusion (described as percent spliced in, or Ψ) across a population of single cells is greatly distorted, with an inflation of spurious extreme values [59]. This distortion is particularly severe in exons from genes with relatively low expression. Because of these limitations, few computational methods for identifying informative alternative splicing events in single cells have been developed. Splicing changes have been characterized as qualitative changes in the empirical distribution of observed splicing rates [56, 57]; however, such qualitative descriptions can be misleading due to the aforementioned coverage-dependent bias [59]. Other methods detect differences in splicing between pre-defined groups of single cells, clustered by condition or sample, and thus cannot take full advantage of the high-resolution insight that single cell data provides for continuous biological processes [55, 64, 65, 90]. A method that analyzes splicing across a continuous cell population, while considering the impact of low mRNA capture efficiency, has yet to be developed.

2.2 Psix: a probabilistic approach to cell-state associated alternative splicing

Our goal in this work was to create a method that could identify alternative splicing events that vary in an informative way over a landscape of single cells. We defined this broadly to mean that the alternative splicing followed some pattern related to underlying cell state; for instance, an alternative exon might be spliced into mRNAs at low levels in stem cells and at increasingly high levels in differentiating cells (Figure 2.1a). Importantly, this definition does not require dividing the cells into explicit groups by clustering or by labeling cell types, nor does it require an explicit trajectory of cell progression. Rather, we wanted to find all exons whose splicing varied as the underlying cell state varied. One of the main challenges that we described in our previous work is that, if few mRNA molecules are captured, the observed $\hat{\Psi}$ will largely deviate from the underlying Ψ [59]. The solution that we presented to this problem in our earlier work was to rely only on the $\hat{\Psi}$ observations that we estimated to come from a large number of captured mRNA molecules, and which were therefore less prone to suffer from these biases. Here we aim to instead retain the information available from these low-coverage observations, taking into consideration technical distortion by using a probabilistic modeling approach.

To achieve this, we developed Psix, a robust yet simple probabilistic method inspired by autocorrelation models that identifies informative splicing events in a continuous single cell space. Psix estimates the likelihood of an alternative exon’s observed inclusion in each cell and determines if the variance in the observed values are the result of biological changes across a continuous landscape by comparing each cell to its neighbors. Our method requires two things: knowledge of which cells are more similar (in transcriptome space), which we will infer from their relative positions in a low-dimensional space derived from overall gene expression [48–50], and a model of the distribution of observed $\hat{\Psi}$ values given an underlying unknown Ψ . Based on the conclusions of our previous work [59], we reasoned that large differences in $\hat{\Psi}$ observations between similar cells should be mostly the result of technical variance, and how much a cell appears to differ from its neighbors largely depends primarily on how many mRNA molecules were captured during sequencing. Meanwhile the differences between phenotypically dissimilar cells could reflect a combination of different splicing programs and technical variance (Figure 2.1a).

Psix estimates the likelihood of an exon’s $\hat{\Psi}$ observation in each cell given two models: one in which the original Ψ of the cell, before the data loss imposed by sequencing, was similar to the distance-weighted average Ψ of the neighboring cells, and another in which it was similar to the Ψ average across the entire dataset. The likelihood of each cell’s observation is estimated based on the number of mRNA molecules captured in the cell, given a binomial sample without replacement. If the exon is autocorrelated, then the first model is would be more likely than the later (Figure 2.1b). We formalize

this comparison as a likelihood ratio, Ψ_ξ , and assign it an empirical p -value based on randomly shuffled cell labels to determine a background distribution.

2.3 Validation of Psix

We validated our approach using simulated single cell splicing data from continuous cell models. The simulation allows us to test Psix performance on exons with different degrees of change (reflecting differences in simulated cell states), or on exons with no change in splicing (i.e., reflecting noise properties of this data [59]) (Figure 2.2a). We tested the advantage of Psix’s cluster-free approach by comparing the results with a simple Kruskal-Wallis test that detects differences in the median Ψ between defined clusters [59, 90]. We also tested the advantage of considering the number of mRNA molecules captured for each cell by comparing the results with estimating a Geary’s C, an autocorrelation test statistic that does not model the number of mRNAs. We found that Psix outperforms these simple approaches in detecting splicing changes on data simulated at a range of different capture efficiencies (Figure 2.2b). Importantly, unlike other approaches, Psix was not enriched for low p -values in exons simulated without splicing changes (Figure 2.2c). This showed that modeling the technical distortion due to low mRNA recovery successfully enabled Psix to avoid excessive false positives. Psix also outperformed these approaches when simulating a complex single cell population with multiple lineages (Supplementary Figure 2.5). Our simulations showed that a k -nearest neighborhood of $\sim 10\%$ of the single cell population is optimized the performance of Psix when simulating 1,000 cells (Supplementary Figure 2.6).

Next, we sought to test our approach on real single-cell RNA-seq data. We applied Psix to three different published neurogenesis scRNA-seq datasets. Since we do not have a controlled ground truth as in the case of simulations, we compared our predictions against differentially spliced exons identified in bulk RNA-seq datasets that closely matched the scRNA-seq datasets, e.g., a single-cell dataset of embryonic mouse neurons was matched with a bulk RNA-seq dataset taken from similar time points (Figure 2.2d). We selected exons that showed no change in the bulk data as the negative groups, while we kept exons with a high change in the bulk data as positives. We compared Psix against the simple Kruskal-Wallis and Geary’s C tests as described above. In contrast with our numerical simulations, this analysis starts directly from the sequencing data. For this reason, we were also able to compare Psix with an existing computational single-cell approach, BRIE2 [65], which requires aligned reads as its input. In all datasets, Psix overperformed all methods in classifying the exons (Figure 2.2e).

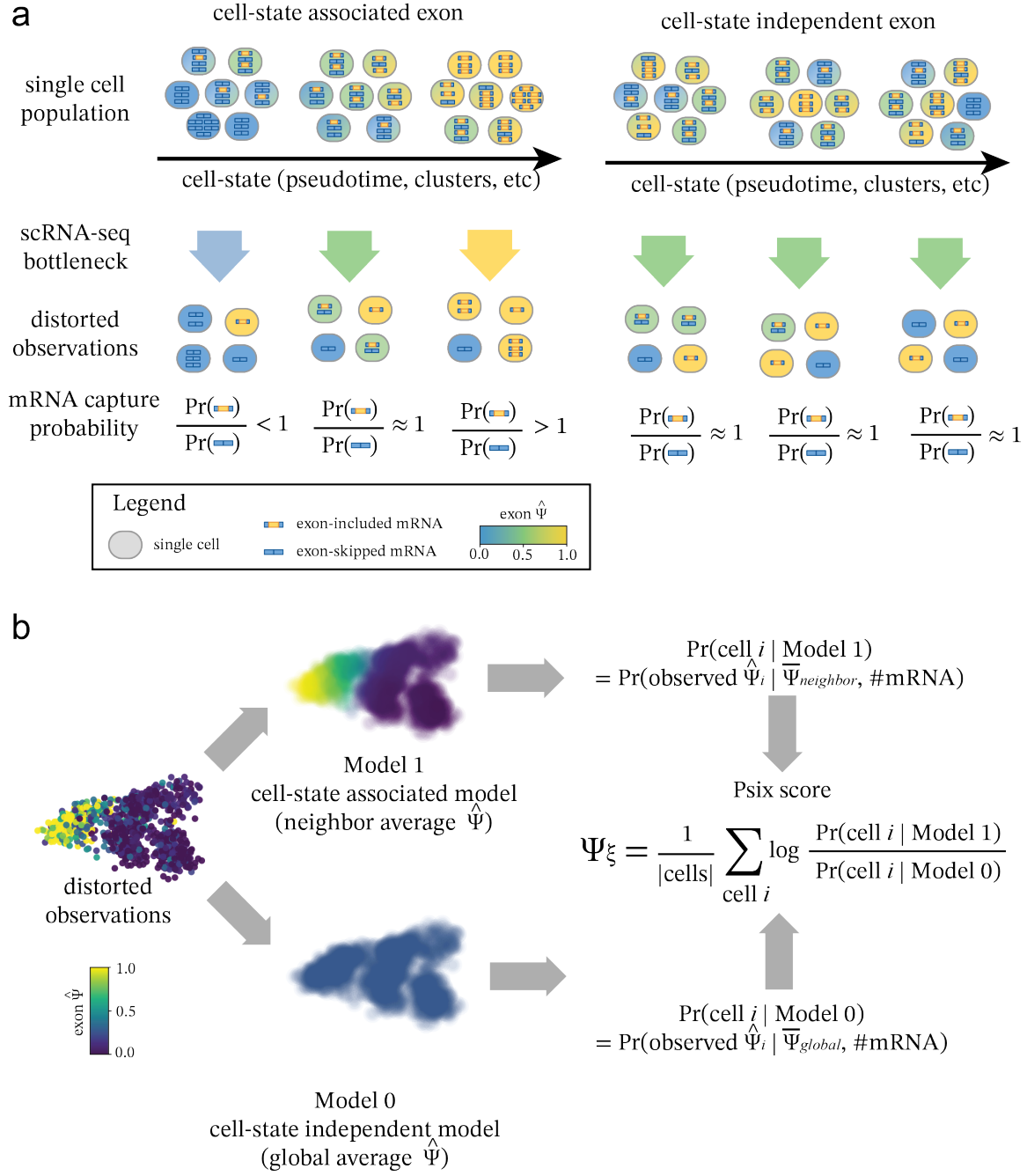


Figure 2.1: A) Cell-state associated exons change across the phenotypic landscape of a single cell population. Cell-state independent exons do not change across the phenotypic landscape. Low-capture efficiency in scRNA-seq experiments adds additional technical variance depending on the number of captured mRNA molecules. The probability of capturing each alternative isoform depends on the underlying distribution of exons in the single cell population.

Figure 2.1: B) Psix compares the likelihood of each single cell observation given two models: Model 1 in which the exon is cell-state associated (probability of the observation given the neighbor average Ψ), vs Model 0 in which the exon is cell-state independent (probability of the observation given the global average Ψ). Model 1 is more likely for a cell-state associated exon will be consistently, while the likelihood of both models is similar for a cell-state independent exon.

2.4 Finding modules of co-regulated splicing in single cells

Because scRNA-seq reports on the biological state of hundreds or thousands of single cells, it has great potential for uncovering regulatory networks that control the changes in individual gene expression. This approach has been successfully implemented to explore the regulation of transcription at the single-cell level [47]. However, its implementation to study the regulation of splicing with single cell data has been limited, and usually involve pooling pre-defined clusters of cells [28, 65]. Alternative splicing is regulated by a network of splicing factors, some of which are highly cell-type specific, and we posited that scRNA-seq data could reveal coordinated changes in co-regulated exons as well as the corresponding changes in expression of the splicing factors that regulate these exons.

To begin reconstructing splicing regulatory networks, we first applied Psix to a dataset consisting of mid-brain neurons across different stages of mouse development [91] (Figure 2.4a). Psix identified 798 cell-state associated exons over the brain development landscape, including many exons that have been extensively reported to change in neuron development [27, 92–94] (Figure 2.4b). A gene ontology enrichment analysis confirmed that the set of alternatively spliced genes identified by Psix was enriched for genes associated with neuron and synaptic development (Supplementary Figure 2.7).

Next we set out to identify co-regulated splicing among the autocorrelated exons. To avoid the high local variance of individual splicing observations, we calculated the neighborhood average Ψ of autocorrelated exons identified by Psix, and clustered these observations into ten modules of co-spliced exons using a UPGMA algorithm (Figure 2.4c). These modules of exons showed distinct patterns of change across brain development, and across different neuron types in late stages of development, suggesting that the modules corresponded to patterns of splicing co-regulation (Figure 2.4d). To find potential splicing factor regulators associated with these modules, we analyzed the binding of RNA binding proteins using published CLIP-seq data [95]. We identified 20 splicing factors that were enriched in binding to cell-state associated exons. Binding of Nova and Rbfox family splicing factors was associated to specific clusters of correlated exons in neurogenesis, consistent with previous knowledge of their role in regulation of splicing in neurogenesis [22, 27, 29]. We found that the

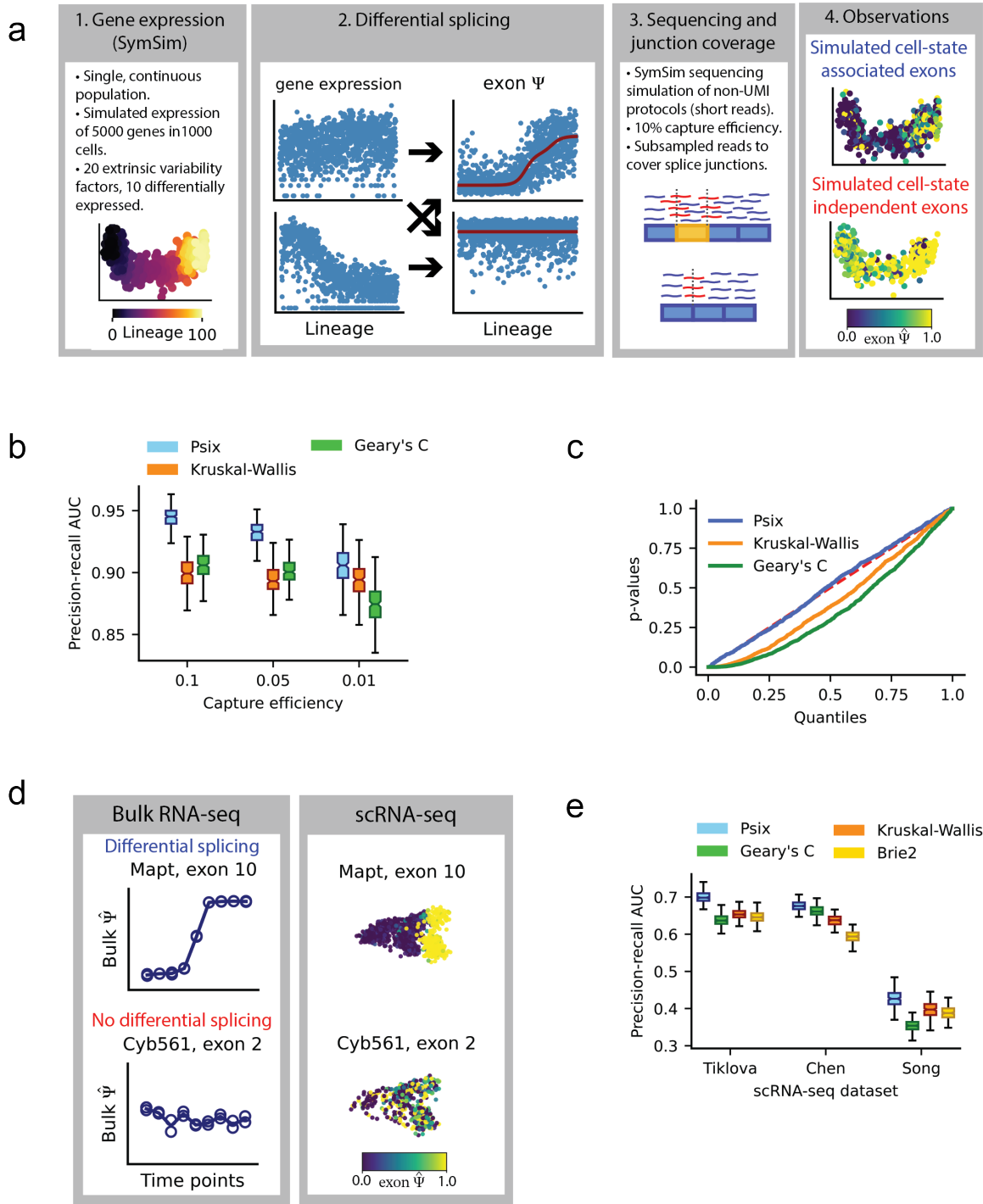


Figure 2.2: A) Pipeline for simulation single cell splicing. Variance in the observations can come from a real change in splicing (positives), or from technical variation or purely changes in gene expression (negatives). B) Area under the precision-recall curve of Psix, the Kruskal-Wallis test, and Geary's C, indentifying exons with a $\Delta\Psi \geq 0.2$, under three different capture efficiencies. C) P-value distributions of the negative exons when tested with Psix, Kruskal-Wallis, and Geary's C. The p-values of Psix do not deviate significantly from the uniform distribution.

Figure 2.2: D) Validation of single cell splicing changes with bulk RNA-seq data. E) Area under the precision-recall representing the overlap of the exons selected by different methods on real scRNA-seq data, with the exons from corresponding bulk datasets.

expression of these splicing factors was strongly correlated with the inclusion levels of the exons from the different modules, further supporting regulatory roles (Figure 2.4f). These results show that the unsupervised exon selection and module inference of Psix discovers relevant switches in alternative splicing during neurogenesis.

Surprisingly, we also found that the splicing factors of the SR family were strongly enriched to binding into one module of exons associated to an early increase in Ψ (Figure 2.4f). Interestingly, this module included, among other exons, the poison exons of four genes of the SR family, which have been previously proposed to regulate their own protein production through the splicing of poison exons [4, 96–100], but have not been suggested to have a specific role in neurogenesis.

A strength of single-cell data is its insight into hidden heterogeneity within cell subtypes. Indeed, the neurons in the final time point of this dataset, collected 90 days post natal, formed distinct sub-populations. These cells showed cell-state associated expression of marker genes for different lineages of midbrain neurons [91], primarily across the third principal component (Supplementary Figure 4). Interestingly, several splicing modules that were otherwise similar showed distinct patterns over these sub-populations (Figure 3d). We searched for the exons associated to the different lineages in midbrain neurons. To achieve this, we implemented Psix in the postnatal day 90 cells only (to which we refer as p90 cells) in order to find exons associated to the cell-state diversity of late midbrain neurons. Psix identified 78 cell-state associated exons, which group in two modules (Module 1 and 2). Principal component analysis (PCA) of these exons showed the p90 cells forming roughly two groups (Figure 4a). We designed the first principal component of this decomposition as the splicing component.

Next we set to find the regulatory mechanisms that control the emergence of this association during neurogenesis. We implemented Psix individually on the postnatal day 1 cells (p1), and in the embryonic day 13.5 cells (e13.5). We then applied the PCA decomposition on the 78 exons of the p90 cells, into the same exons in the p1 and e13.5 cells, in order to find the variance of these cells across the splicing component. Both these sub-groups of cells showed less variance across the splicing component than p90 cells, with the p1 cells being an intermediate state. We observed that Module 1 is starting to be clearly formed in p1 cells, while Module 2 seems still underdeveloped. This is reflected in the variance of p1 cells across the splicing component.

We identified 14 splicing factors that are enriched in the 78 exons associated cell-state in late midbrain neurons using the hypergeometric test over CLIP-seq data. Of these,

Nova1, Rbfox1 and Mbnl2 had a significant correlation between expression and the the splicing component in p90 cells, indicating a positive correlation with Module 1 (Figure 4b,c). On the other hand, expression of Ptbp2 showed a positive correlation with the splicing component in p90 cells, indicating a correlation with Module 2. Interestingly, expression of Nova1 and Rbfox1 are also significantly negatively correlated to the splicing component in p1 cells, but not in e13.5. This suggests that Nova1 and Rbfox1 might be regulators of the early formation of Module 1. In contrast, the correlations of Mbnl2 and Ptbp2 with the splicing component are not significant in p1, and Mbnl2 has the opposite correlation in e13.5 cells. This could indicate that these splicing factors are associated to the late formation of Module 2 in midbrain neurons.

These results suggest that the well-known families of splicing factors that regulate exon usage in neurogenesis also regulate cell-state associated splicing between different lineages in midbrain neurons. The definition of these exon modules arises from differential expression of these splicing factors as midbrain neural lineages emerge.

Discussion

Single-cell data allows to study continuous biological functions at a higher resolution than bulk studies. However, coverage-dependent biases have limited the study of single-cell splicing to discrete pre-defined clusters of cells, hindering its potential [59]. Psix aims to solve these problems by combining cell identity information from the transcriptome space, with bias-aware probabilistic modeling. Our approach finds cassette exons that vary continuously across the cell phenotypic landscape, without mistaking gene expression variance for changes in splicing. Psix also identifies groups of exons with similar patterns of biological variation, which are indicative of potential splicing co-regulation. Our cluster-free and bias-aware approach sets the ground for further studies of alternative splicing regulation that take full advantage of single-cell information.

Methods

Code and data availability

Psix is available at <https://github.com/lareaulab/Psix>. The analysis of simulations and publicly available data is documented at

https://github.com/lareaulab/analysis_psix

The Tiklova [91], Chen [69], and Song [56] datasets were downloaded from accessions GSE116138, GSE74155, and GSE85908, respectively.

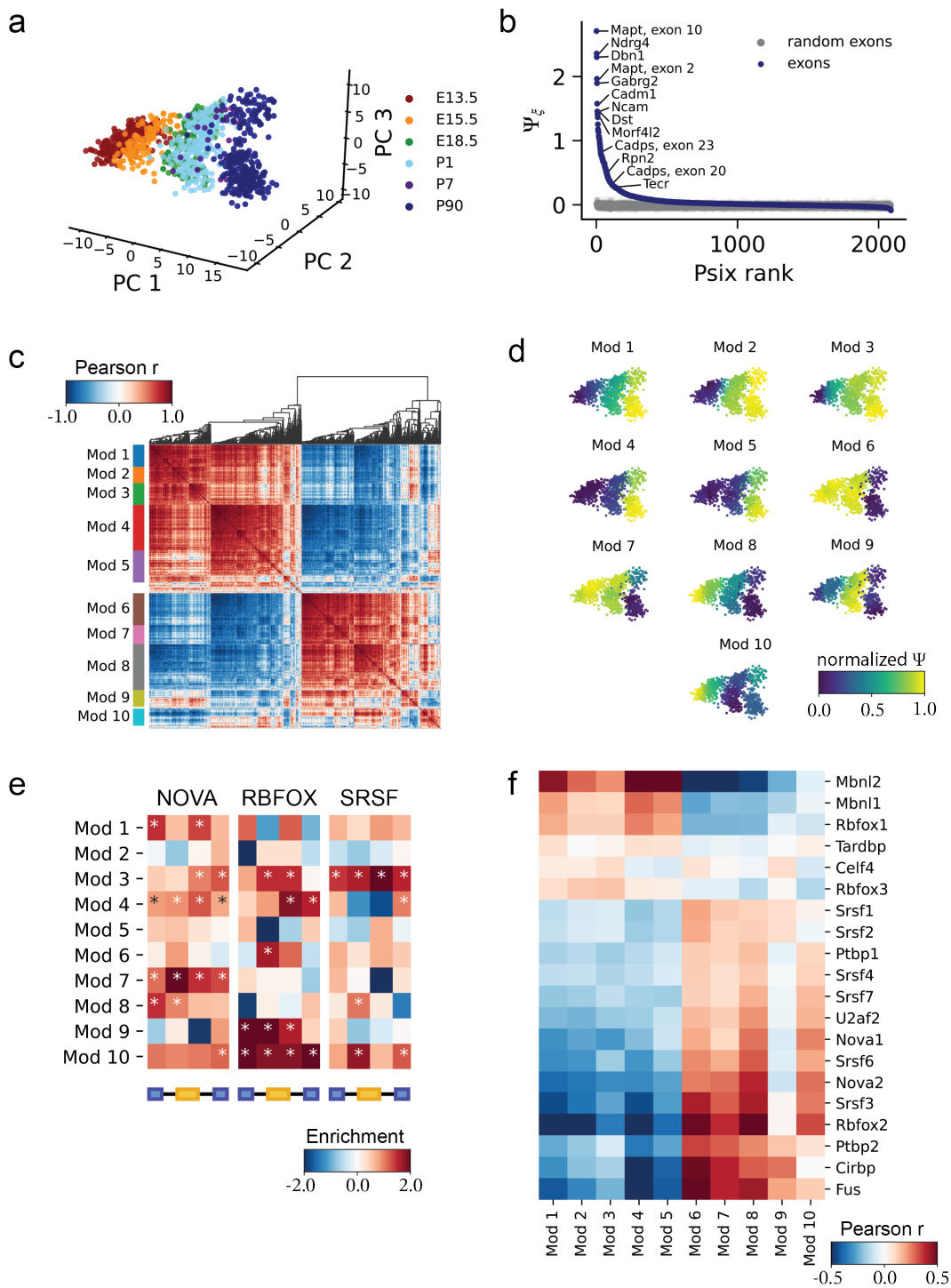


Figure 2.3: a) PCA of the normalized counts of mouse brain neurons at different stages of development. b) Psix scores of cassette exons, compared to the scores of randomized exons. Some exons known to be regulated in neurogenesis are highlighted.

Figure 2.3: c) Correlation map of the neighbor average $\bar{\Psi}$ of the informative exons. Modules were identified with a modification of the UPGMA algorithm. d) Neighbor average normalized $\bar{\Psi}$ of the exons in each module. e) Enrichment of CLIP-tag clusters in the exons of the seven modules of informative exons. A * marks significant enrichment in a hypergeometric test after Benjamini-Hochberg correction. f) Correlation of module splicing with gene expression of 20 splicing factors enriched in splicing modules.

Conceptual overview

To determine if a splicing event is associated with cell identity or cell state, we observe how its Ψ varies given a cell-cell similarity metric. If a splicing event is informative about a cell’s identity, we expect that similar cells will have closer Ψ than cells with different identities across the similarity metric.

For a population of n single cells, and for m alternatively spliced exons, Psix takes as input three matrices: one (n, m) matrix of $\hat{\Psi}_{ij}$ observations for each exon i in each cell j ; another (n, m) matrix of the estimated number of informative mRNA molecules for each exon in each cell; and a (n, n) similarity map between the cells (Figure 2.1b). Psix estimates the likelihood of the $\hat{\Psi}_{ij}$ observation of an exon in each cell given the local average Ψ of its closest neighbors, and the number of captured molecules x_{ij} . It then contrasts this likelihood with the likelihood of the same observations, given the global average Ψ of all cells. If the likelihood of the observations given the local average is significantly higher than the likelihood given the global average, the exon is considered to be informative of the biological state.

In our implementation, we estimate $\hat{\Psi}_{ij}$ using splicing junction reads of each analysed cassette exon. Because smart-seq2 and similar methods do not directly report the number of mRNAs recovered per gene, for smart-seq2 data we estimate the number of mRNAs, x_{ij} , with a modification of the Census normalization [63] as described in [59]. The similarity matrix is estimated from the Euclidean distance in the gene expression space of the single cell populations, e.g.: the principal components of a normalized matrix of gene expression. The method could be generalized to use other similarity metrics for single cells, such as spatial positions [101] or single cell lineage [102].

The Psix model

Psix is built over the likelihood of $\hat{\Psi}$ observations based on underlying Ψ and r captured mRNA molecules with a capture efficiency c that we developed on a previous

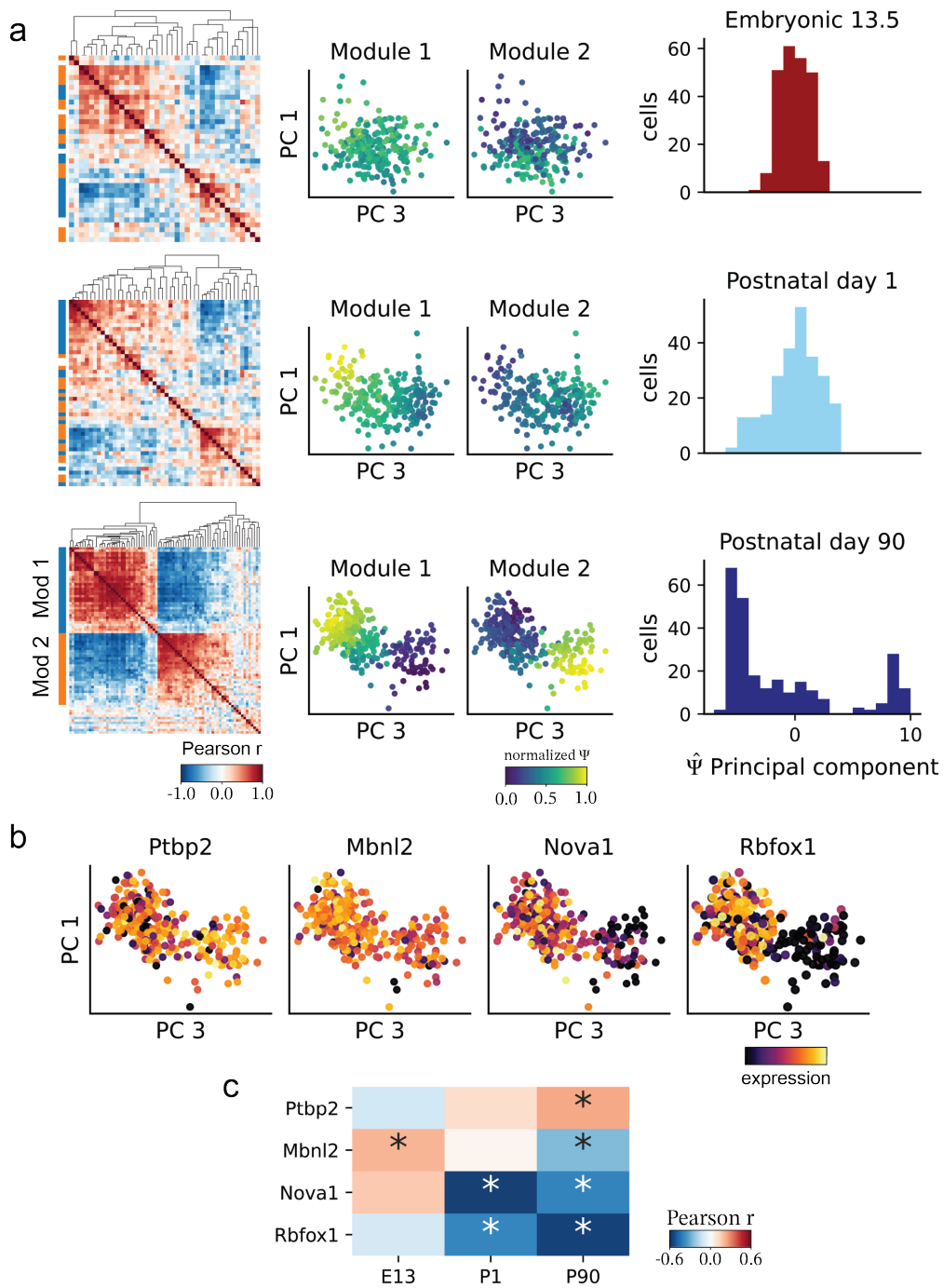


Figure 2.4: a) Cell-state associated exon modules of postnatal day 90 cells, at different stages of development. Heatmaps show the correlation between exons in the modules. Scatterplots show the splicing patterns of the exons of the module in sub groups of cells. The histograms show the distribution of the cells across the first principal component of the $\hat{\Psi}$ of the cell-state associated exons.

Figure 2.4: b) Cell-state associated expression of splicing factors. c) Correlation of splicing factor expression with the first principal component of the $\hat{\Psi}$ of the cell-state associated exons in embryonic day 13.5, postnatal day 1 and postnatal day 90 cells. Asterisks mark significant correlation.

study [59]:

$$\Pr(\hat{\Psi} \mid \Psi, r, c) \approx \sum_{m=r}^{10r/c} \binom{m\Psi}{r\hat{\Psi}} \binom{m(1-\Psi)}{r(1-\hat{\Psi})} \cdot c^{r+1}(1-c)^{m-r} \quad (2.1)$$

where m is the number of molecules present in the cell before sequencing.

Using this likelihood equation, we can test the likelihood of an observation given its underlying splicing context. In order to find informative exons, given a cell-cell similarity metric, we define the Psix score of an exon as follows:

$$\Psi_{\xi} = \frac{1}{\sum_{i \in \text{cells}} I(i)} \sum_{i \in \text{cells}} \left(\log \Pr(\hat{\Psi}_i \mid \bar{\Psi}_i, r_i, c) - \log \Pr(\hat{\Psi}_i \mid \bar{\Psi}, r_i, c) \right) \quad (2.2)$$

where $\hat{\Psi}_i$ correspond to the observed exon Ψ in cell i , and r_i are the number of captured mRNA molecules that are informative for the exon in cell i . $\bar{\Psi}_i$ corresponds to the weighted average observed Ψ of the k -nearest neighbors of cell i in the cell-cell similarity metric, and $\bar{\Psi}$ is the global average Ψ . $I(i)$ is a function that determines if an observation was made in cell i .

Given a cell-cell similarity metric, for cell i we estimate the weighted average Ψ of its neighbors as follows:

$$\bar{\Psi}_i = \frac{\sum_{k \in K_i} (w_{ik} \hat{\Psi}_k)}{\sum_{k \in K_i} w_{ik}} \quad (2.3)$$

where K_i is the set of k -nearest neighbors of cell i (not including itself), and w_{ik} is a similarity score between cells i and k in the cell-cell metric.

Identifying modules of co-spliced exons

We are interested in identifying exons with similar patterns of autocorrelation, since they could potentially share regulatory splicing factors. Due to the high technical noise in single cell splicing observations, clustering of exons is difficult in the raw

data. Instead, for each pair of exons, we obtain the Pearson correlations of the k-nearest neighbors averages. We cluster the exons in a bottom-up procedure similar as described by DeTomaso & Yosef [89]: we merge iteratively the two exons or modules that have the highest Pearson correlation. Then we update the correlation score of the module with all other exons and modules using the UPGMA approach. When a module hits a minimum of 20 exons, we assign a label to it. We stop clustering when the maximal correlation score is lower than 0.3, and return the labeled modules. Exons that do not belong to a labeled module are returned as “unassigned”.

Practical considerations for the application of the Psix model

Estimating captured mRNA molecules in smart-seq2 datasets

In practical usage, non-UMI based scRNA-seq methods such as smart-seq2 do not directly quantify the number of captured mRNA molecules of a gene per cell. As a result, it is difficult to estimate the parameter r_i . To approximate r , we use a modification of the Census normalization [63] as reported previously:

$$M_i = \frac{n_i}{F_{X_i}(x_i^*) - F_{X_i}(0.1)}$$

x_i^* is the mode of the distribution of log-transformed gene TPM counts in cell i . x_i^* is the argmax of a Gaussian kernel density fit to the log TPM distribution. n_i is the number of genes in cell i with a TPM between 0.1 and x_i^* . F_{X_i} is the cumulative distribution of TPM counts in cell i . As in Qiu et al., we assume that genes with TPM below 0.1 are not represented by any mRNA molecules.

Continuous approximation to probability of observations

Due to variance in sequencing depth and potential overamplification, the number of informative splice junction reads for an event will rarely match the number of informative estimated mRNA molecules captured. For this reason, it is often impossible to calculate $\Pr(\hat{\Psi} \mid \Psi, r, c)$ directly due to the combinations part of the equations. In Psix, we use a continuous approximation of this probability by approximating the combinations in Equation 1 as:

$$\binom{n}{k} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \quad (2.4)$$

Cell-cell metric

To determine if an exon is informative of a cell’s identity, we need a cell-cell metric that describes the similarity between individual cells. In practical terms, from smart-seq2 data this metric can be an interpretable low-dimensional manifold representation

of the normalized gene expression profiles of the cells. Our approach to normalization and dimensionality reduction depended on the dataset we analysed. For the Tiklova [91], Chen [69] and Song [56] datasets, we normalized the TPM counts per gene using SCONE [83] due to the relatively small number of cells, and performed dimensionality reduction with PCA over the top 1000 fano factor genes. In the three cases, the first two principal componets captured most of the variance, and thus we used these components as the cell-cell metric.

Once we obtained a cell-cell metric for each dataset, for each cell i , we used the Euclidian distance in the low-dimensional manifold projection to identify its k -nearest neighbors. For each neighbor k , we estimated a similarity score as follows:

$$w_{ik} = \exp\left(\frac{-d_{ik}^2}{\sigma_i^2}\right)$$

where d_{ik} corresponds to the Euclidean distance between cells i and k ; σ_i is the Euclidean distance between i and its most distant neighbor in the cell-cell metric. We set $w_{ik} = 0$ for $k = i$, and for all cells k that are not neighbours of i .

Optionally, the user can choose to use the unweighted neighborhood average. To calculate $\bar{\Psi}_i$ as the k -nearest neighbor average without weights, we simply set $w_{ik} = 1$ for all k neighbors of i , and $w_{ik} = 0$ for all other cells.

Neighborhood size selection

Our simulation analysis showed that Psix performance was optimal when the size of the neighborhood used for estimating neighbor average $\hat{\Psi}$ is around $\sim 10\%$ of the total number of cells in the dataset (Supplementary Figure 2.6a,b)

Capture efficiency in single cell datasets

We set a capture efficiency of 0.1, as supported by single-cell studies [44, 58, 63, 72]. The choice of capture efficiency has very little effect on Psix scores (Supplementary Figure 2.6c).

Empirical p-value estimation

Psix compares the probability of two models. However, the null model (global average $\bar{\Psi}$) is not embedded into the alternative model (neighbor average $\hat{\Psi}$). For this reason, we obtained an empirical p-value by randomizing the $\hat{\Psi}$ observations and their respective mRNA counts and obtaining a random distribution of Psix scores. We observed that the mean mRNA counts and the total $\hat{\Psi}$ variance had a slight effect on the random distribution of the Psix scores. For this reason, we divided the exons into bins according to their ranks in mRNA counts and $\hat{\Psi}$ variance (by default, 5 bins

for each parameter, totalling 25 bins). For each bin, we randomized the exons of the bin, and for each exon we estimated an empirical p-value as follows:

$$p = \frac{x + 1}{n + 1} \quad (2.5)$$

where x is the number of random exons in the bin to which the exon belongs that have a higher Psix score than the non-randomized exon, and n is the total number of randomized exons in the bin (2000 by default). By default, Psix uses the Benjamini-Hochberg procedure to correct for multiple tests.

Processing of smart-seq2 datasets

For the mouse brain development dataset[91] we aligned the reads using STAR to the mouse genome (annotation mm10) with added sequences for the Illumina spike-ins and the enhanced green fluorescent protein. TPM was estimated using RSEM. After quality control and removing a substantial number of outlier cells expressing glial or microglial gene markers, we were left with a total of 1067 cells that we used for downstream analysis. For dimensionality reduction, we first selected genes that are expressed in the dataset (average normalized counts ≥ 0.1), and selected the top 1000 genes with high Fano factor. We then applied PCA on the resulting genes, which revealed a clear trajectory of neuron differentiation.

We selected 2087 cassette exons that had observations in at least 25% of cells (i.e., a maximum dropout rate of 75%). We applied Psix over these exons using a neighborhood size of 100 cells and a capture efficiency rate of 0.1.

The Chen [69] and Song [56] datasets were processed as we described in our previous work [59]. These datasets are smaller than the Tiklova dataset. A neighborhood of size 30 was used in both cases.

Comparison of Psix with other single cell splicing methods

We compared the performance of Psix against several approaches that have been proposed for addressing single cell splicing.

Kruskal-Wallis test

The Kruskal-Wallis test is a non-parametric version of ANOVA. It has been previously used to identify if the median $\hat{\Psi}$ of an exon is different between pre-defined clusters of cells [59, 90]. We applied this test for all the exons we evaluated with Psix in all datasets. We clustered the cells according to the labels provided by the authors of each dataset.

Geary’s C

Geary’s C is a statistic that measures the spatial autocorrelation of a variable. It has been used for computing the autocorrelation of gene signatures with cell identity [76], and we previously used an adaptation of this test for finding autocorrelation of exon observations [59].

For each exon, we normalize the observed $\hat{\Psi}$ as follows:

$$\hat{\Psi}'_{ij} = \frac{\hat{\Psi}_{ij} - \bar{\Psi}_j}{\text{Var}(\hat{\Psi}_j)}$$

where $\hat{\Psi}_{ij}$ is the observed splicing of exon i in cell j . $\bar{\Psi}_j$ is the mean $\hat{\Psi}$ of all the observed exons in cell j , and $\text{Var}(\hat{\Psi}_j)$ is the variance of the observed $\hat{\Psi}$ of all exons in cell j .

We calculate a variant of the Geary’s C statistic of each exon as:

$$C = \frac{(N - 1) \sum_j \sum_k w_{jk} (\hat{\Psi}'_{ij} - \hat{\Psi}'_{ik})^2}{2W \sum_j (\hat{\Psi}'_{ij} - \bar{\Psi}'_i)^2}$$

where N is the number of cells in the dataset; $w_{i,j}$ is the weight between cell i and cell j in the cell-cell metric, and W is the sum of all $w_{i,j}$. We use the same similarity metric that we used in Psix.

In order to make the statistic more intuitive, in which a positive score close to 1 indicates high autocorrelation, we transform our statistic as:

$$C' = 1 - C$$

BRIE2

BRIE2 is a computational tool that regresses single-cell splicing data against cell-level features, and can detect differential splicing between groups of cells [65]. We applied BRIE2 directly to the STAR-aligned BAM files [80] of the Tiklova, Chen and Song datasets. We used the author-provided labels for each dataset. In a second approach, we also ran BRIE2 using the pseudotime derived by Slingshot [46] in each of the datasets as the input for the cell feature coefficients.

Alternative splicing in bulk RNA-seq data

We compared each method’s ability to detect alternative splicing events identified in bulk RNA-seq datasets using rMATS 3.2.5 [39]. We compared the Tiklova dataset [91] with a bulk mouse brain development dataset [27]; the Chen dataset [69] with

a bulk dataset of mouse embryonic stem cells induced to neurogenesis [81]; and the Song dataset with the bulk RNA-seq data included in the same study [56].

RNA binding-protein analysis

CLIP-seq data compilation and analysis

CLIP-seq data was obtained from multiple public sources (Table 2.1). Some datasets were downloaded from the CLIPdb dataset [95], while others were obtained from their respective studies depending on availability. For iCLIP-seq data that was not already pre-processed by CLIPdb, we used Piranha version 1.2.1 [103] to call peaks from CLIP-tags. Peaks from different datasets of the same RBP were combined into a single file per RBP. We next reported the overlap of the RBPs into four regions per cassette exon:

- E1 region: 100 nt upstream and i_1 nt downstream from the splice junction of the upstream flanking exon. Where i_1 is the minimum of 100 and half the distance between the upstream flanking exon and the cassette exon.
- S1 region: i_1 nt upstream and e nt downstream from the 5' splice junction of the cassette exon. Where e is the minimum of 100 and half the distance between the two splice junctions of the cassette exon.
- S2 region: e nt upstream and i_2 nt downstream from the 3' splice junction of the cassette exon. Where i_2 is the minimum of 100 and half the distance between the cassette exon and the downstream flanking exon.
- E2 region: i_2 nt upstream and 100 nt downstream from the splice junction of the downstream flanking exon.

Analysis of postnatal day 90 modules

We implemented Psix on the 290 cells collected from the postnatal day 90 brain, using the transcriptomic space in Figure 3a. Due to the smaller sample size (compared with the entire dataset), we used a neighborhood of 30 cells ($\sim 10\%$ of the total number of cells). We implemented principal component analysis of the 78 cell-state associated exons by applying the decomposition over the neighbor average Ψ used in the Psix model, and selected the first component for analysis (splicing component). Using the neighbor average functions as an imputation method for the missing data.

We also implemented Psix in the postnatal day 1 and in the embryonic day 13.5 subgroups individually. To analyze the emergence of the variance of the 78 exons in these cell groups, we used the principal component analysis model trained in the postnatal day 90 cells, and implemented this decomposition in the neighbor average Ψ of these two sub-groups. This allowed us to observe the variance of the cells in these sub-groups across the splicing component.

In order to find the potential regulators that control the formation of these patterns during neurogenesis, we identified the RNA binding proteins that are significantly enriched in binding to the 78 exons in these modules using CLIP-seq data (hypergeometric test, $\text{FDR} \leq 0.05$). Due to the small number of exons in this module, we tested for enrichment of binding to any of the four regions of the exon previously described.

After selecting the RBPs that are significantly enriched in these exons, we ran a Pearson correlation test between the normalized expression (SCONE normalized TPM counts) of these RBPs, and the splicing component. We implemented this for each of the three sub-groups individually, and selected the RBPs with significant correlations in at least one timepoint ($\text{FDR} \leq 0.05$).

Simulations of alternative splicing in single cells

We simulate alternative splicing of exons in single cells. This consists of three main steps: First, we simulate the single cell population and its gene expression. Second, we simulate alternative splicing in some of the genes of this population across a continuous trajectory. Third, we simulate the sequencing and splice junction coverage of the exons.

Simulating a single cell population

We used SymSim [75] to simulate the expression of 5000 genes in a continuous population of 1000 single cells with default parameters. To simulate a single developmental trajectory, we simulated single cells across a phylogenetic tree with two branches with SymSim, and then we merged the two lineages into a single lineage. Each cell’s relative position in the tree corresponded to their position in the developmental trajectory, starting from the tip of one branch to the other. We normalized the relative positions of the lineage to span from 0 (first cell) to 100 (last cell).

Simulating splicing in single cells

We simulate splicing in the single cells as three sub-steps. First, we model the expected Ψ as a sigmoid or impulse function [104]. For each exon, we simulate a “Platonic” Ψ that will be dependent on the cell’s position in the lineage:

$$z(t_i) = \frac{1}{h_1} \left(h_0 + (h_1 - h_0) \frac{1}{1 + e^{-\beta(t_i - t_1)}} \right) \left(h_2 + (h_1 - h_2) \frac{1}{1 + e^{\beta(t_i - t_2)}} \right) \quad (2.6)$$

where $z(t_i) = \text{logit}(\Psi(t_i))$. $t_i \in [0, 100]$ is the position of a given cell in the simulated lineage. h_0 , h_1 and h_2 are amplitude parameters of the curve (first, second and third plateau), t_1 and t_2 are the state transition times; i.e., when the inflection points of

the impulse happen. β is a slope parameter that determines how steep the impulse is.

We sample the impulse parameters from the following probabilistic distributions:

- $t_1 \sim \text{Uniform}(0, 70)$
- $t_2 \sim \text{Uniform}(t_1, 99)$
- $\beta \sim \text{Uniform}(0.05, 0.25)$
- $h_x \sim \text{Uniform}(-3, \text{logit}(0.7))$
- $h_y \sim \text{Uniform}(\text{logit}(\text{expit}(h_x) + 0.2), 3)$
- $h_z \sim \text{Uniform}(h_x, h_y)$
- To simulate different impulse shapes, we set $[h_0, h_1, h_2]$ as a permutation of $[h_x, h_y, h_z]$
- To simulate exons that do not change across the lineage, we randomly select a value x from $[h_x, h_y, h_z]$ and set $h_0 = h_1 = h_2 = x$.

We add noise to simulate the biological variability that affects the true splicing profile of a cell in a manner independent of the lineage:

$$z_i \sim \mathcal{N}(z(t_i), \sigma^2) \quad (2.7)$$

where z_i is the *logit* transformation of the true Ψ of cell i . t_i is the temporal stage of cell i ; i.e., its position in the continuous trajectory (lineage). σ^2 is a variance parameter that is specific of each exon. For each exon, we randomly sample $\sigma^2 \sim \text{Uniform}(0.5, 1)$. Finally, we simulate splicing of the mRNA molecules of each gene as a stochastic process as follows:

$$I_i \sim \text{Binomial}(X_i, \text{expit}(z_i)) \quad (2.8)$$

$$E_i = X_i - I_i \quad (2.9)$$

where I_i is the number of mRNA molecules of the gene in cell i that include the alternatively spliced exon. E_i is the number of mRNA molecules that exclude the exon. X_i is the total number of mRNA molecules of the gene in cell i . We merge E and I into a single matrix of mRNA molecules M , in which each isoform corresponds to an independent row.

Sequencing and splice junction coverage

Once we have simulated the expression and splicing of each gene, we simulate the sequencing of the mRNA molecules. For this, we first assign a transcript length to each molecule as follows: For each gene g with isoforms g_E (excludes the exon) and g_I (includes the exon), we randomly sample without replacement l_{g_E} as the length of g_E from SymSim’s transcript length database. We then sample without replacement l_{g_i} , the length of the alternative exon of g , from a database of lengths of cassette exons in the human genome. Finally, we set the length of g_I as $l_{g_I} = l_{g_E} + l_{g_i}$.

We use SymSim’s `True2ObservedCounts` function to simulate non-UMI sequencing of the mRNA molecules in M and lengths l_{g_E} and l_{g_I} . We used the following parameters: mean capture efficiency: 0.1, capture efficiency standard deviation: 0.05, depth sequencing mean: 1e5, depth sequencing standard deviation: 1e4. To simulate a dataset with bad capture efficiency, we set mean capture efficiency: 0.05, capture efficiency standard deviation: 0.02; and for a dataset with very poor capture efficiency we set mean capture efficiency: 0.01, capture efficiency standard deviation: 0.01.

Finally, we simulate the splice junction coverage of isoforms as follows:

$$\begin{aligned} l_r &= \text{read length (constant)} \\ j_{g_I} &= \frac{4(l_r - 1)}{l_{g_I}} \\ j_{g_E} &= \frac{2(l_r - 1)}{l_{g_E}} \\ \text{SJ}_{g_I} &\sim \text{Binomial}(R_{g_I}, j_{g_I}) \\ \text{SJ}_{g_E} &\sim \text{Binomial}(R_{g_E}, j_{g_E}) \end{aligned}$$

where l_r corresponds to the constant read length from the sequencing process (set as default to 50). SJ_{g_I} and SJ_{g_E} are the number of reads that cover informative splice junctions for isoforms g_I and g_E respectively; and R_{g_I} and R_{g_E} are the total number of reads simulated to map to isoforms g_I and g_E respectively.

For each cell, the observed $\hat{\Psi}$ of each exon is calculated as:

$$\hat{\Psi} = \frac{\text{SJ}_{g_I}}{\text{SJ}_{g_I} + 2 \cdot \text{SJ}_{g_E}} \quad (2.10)$$

Simulation of a branching lineage of single cells

We simulated a diverging lineage of single cells, using SymSim with the following three-branched phylogenetic tree (Supplementary Figure X):

(A:1,(B:0.5,C:0.5):0.5)

We divided the single cell population into three lineages:

- Lineage 1: Root to A
- Lineage 2: Root to B
- Lineage 3: B-C splitting point to C

We simulated the “Platonic” Ψ of each lineage as follows:

1. For Lineage 1, sample h_0, h_1, t_1 and β as shown for the single-lineage impulse parameters. With a probability of 0.25, simulate differential splicing with a sigmoid function as follows:

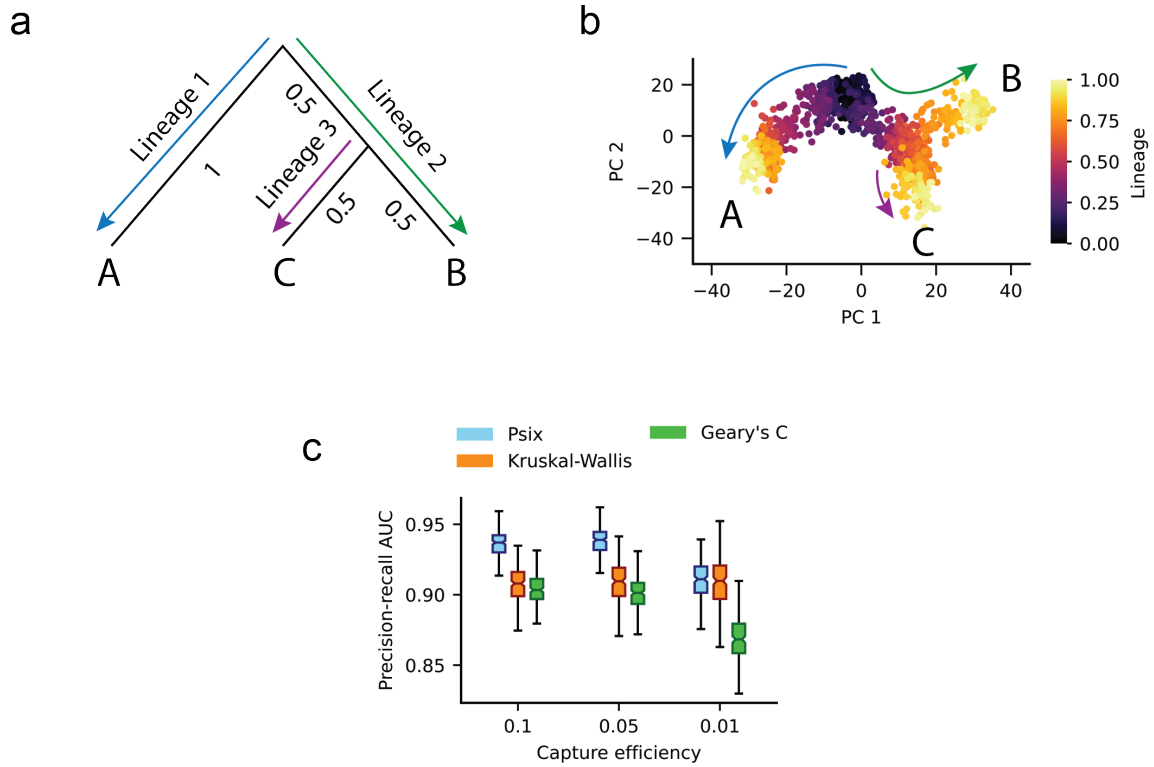
$$z_1(t_i) = \left(h_0 + (h_1 - h_0) \frac{1}{1 + e^{-\beta(t_i - t_1)}} \right)$$

Else, we fit a flat line $z_1(t_i) = h_0$.

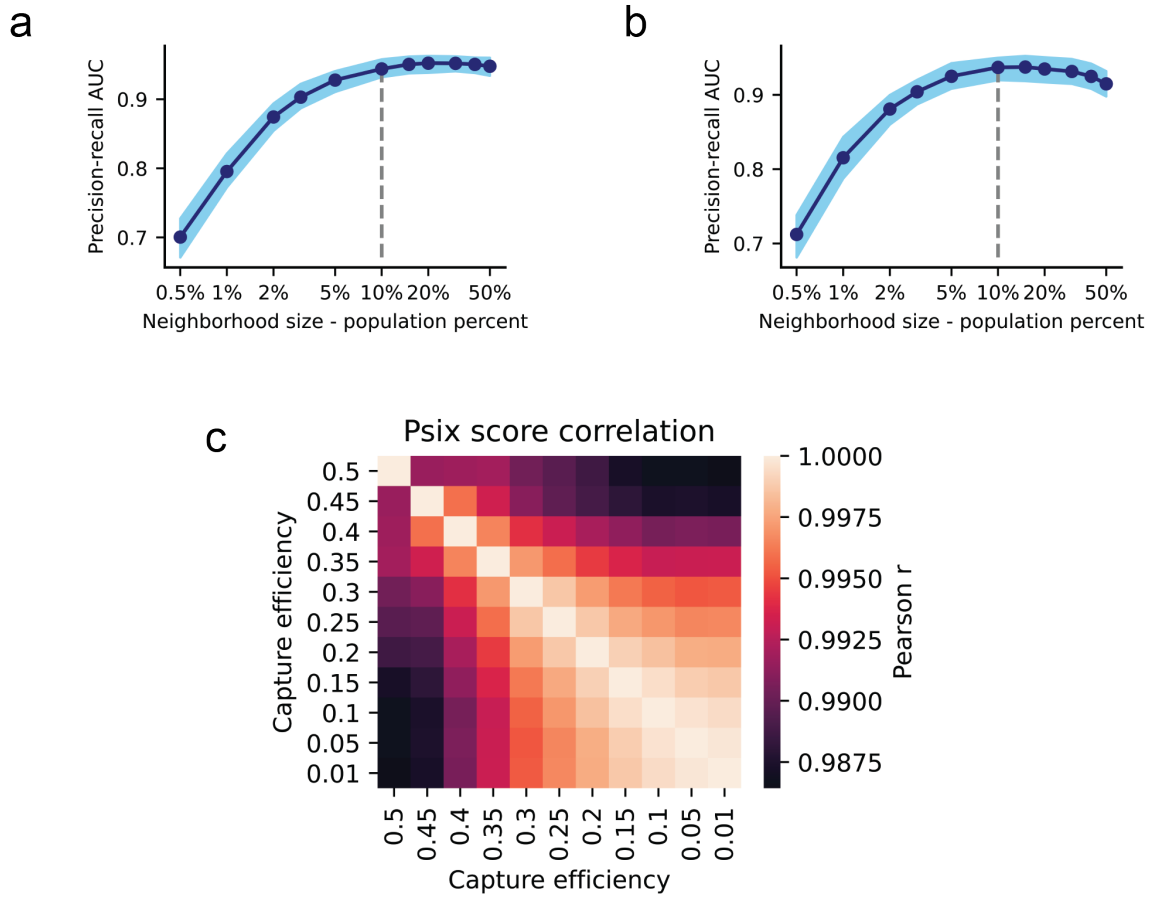
2. For Lineage 2, we use the same h_0 as in Lineage 1, and randomly sample the other parameters. We simulate differential splicing with a sigmoid z_2 with a probability of 0.25, or fit a flat line $z_2(t_i) = h_0$ otherwise.
3. For Lineage 3, we set $h_0 = z_2(t_{BC})$, where z_2 is the function (sigmoid or flat line) used for Lineage 2, and t_{BC} is the timepoint at which Lineage 2 and Lineage 3 split. We sample the other parameters randomly. We simulate splicing with a sigmoid z_3 with a probability of 0.25, or fit a flat line $z_3(t_i) = h_0$ otherwise.

Each simulated gene is marked as differentially spliced if the splicing of at least one lineage was simulated with a sigmoid. Otherwise the gene is marked as non-differentially spliced. Once we’ve simulated the “Platonic” Ψ of each gene, we repeat the simulation steps for the single lineage simulations.

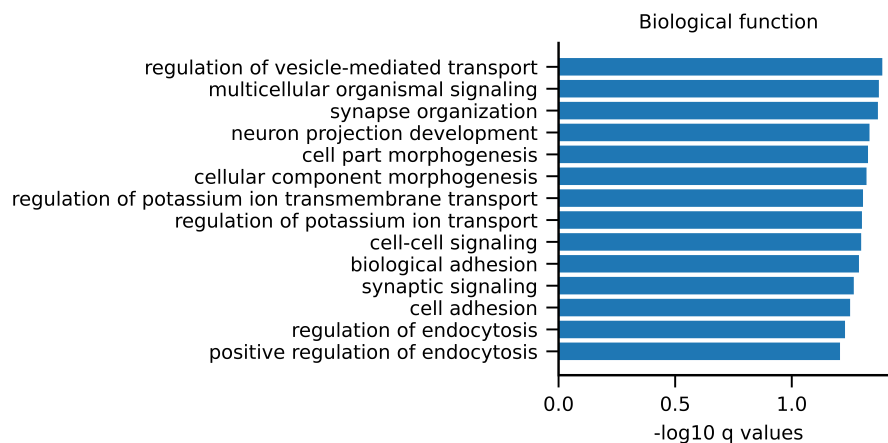
2.5 Chapter appendix: supplementary tables and figures



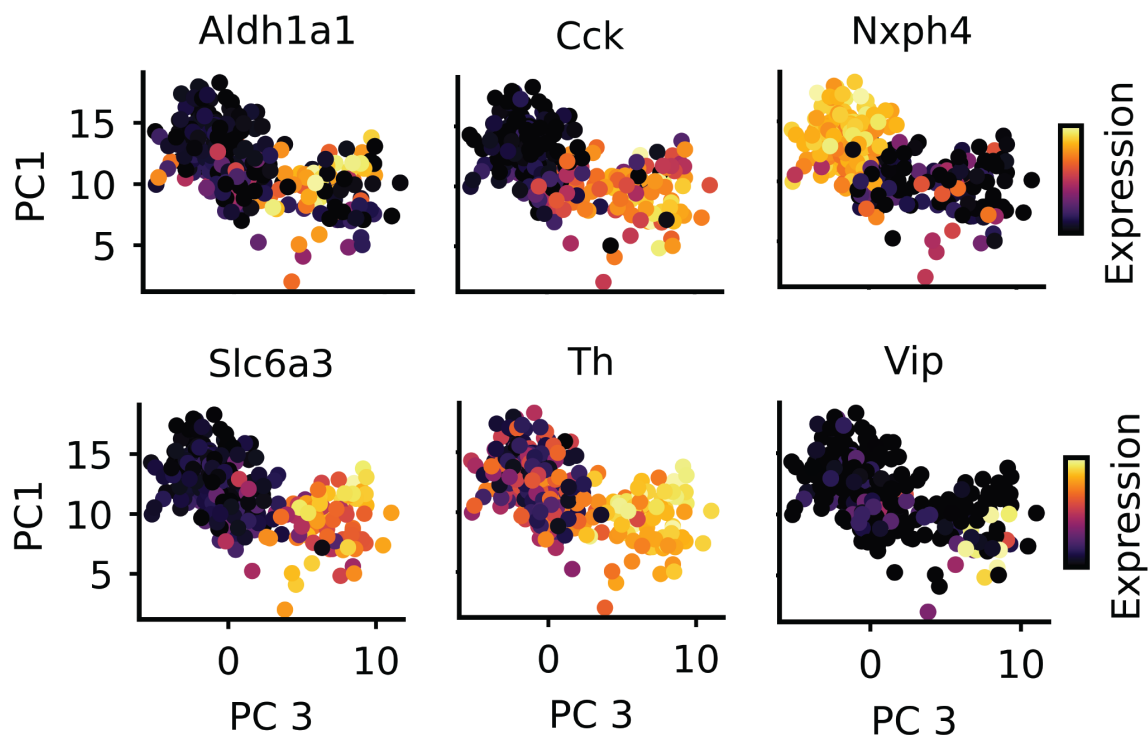
Supplementary Figure 2.5: Simulation of alternative splicing in a complex single cell population. a) Phylogenetic tree used as input for SymSim. b) Principal component analysis of the transcriptome landscape of the simulated population. c) Area under the precision-recall curve of Psix, the Kruskal-Wallis test, and Geary's C, identifying exons with a $\Delta\Psi \geq 0.2$ under different capture efficiencies.



Supplementary Figure 2.6: a) Area under the precision-recall curve of Psix in simulated data with different k sizes. Blue area represents 0.05 to 0.95 quantiles. Data is shown from the single lineage simulation, and from b) the three lineages simulation. c) Pearson correlation of Psix score in the Tiklova dataset [91] using different capture efficiency rates in the model.



Supplementary Figure 2.7: Gene Ontology enrichment analysis of the Psix score in the Tiklova dataset (biological function terms).



Supplementary Figure 2.8: Principal component 3 is associated to neuron diversity in postnatal day 90 midbrain neurons. a) Expression of genes characteristic of midbrain neuron subtypes. b) Cell-state associated expression of splicing factors.

Table 2.1: Mouse CLIP-seq data.

RBP	Experiment	Tissue	Preprocessing	Source/Accession
AGO2	HITS-CLIP	mESC	CLIPdb	GSE25310
CELF1	HITS-CLIP	C2C12	CLIPdb	PRJEB2614
CELF4	iCLIP	Mouse brain	CLIPdb	PRJEB3066
CIRBP	HITS-CLIP	NIH/3T3 fibroblasts	CLIPdb	GSE37685
CIRBP	PAR-CLIP	MEF	CLIPdb	GSE40468
ELAVL1	HITS-CLIP	Mouse brain	CLIPdb	GSE45148
EZH2	PAR-CLIP	mESC	CLIPdb	GSE49433
FMR1	HITS-CLIP	Mouse brain	CLIPdb	GSE45148
FUS	iCLIP	Mouse brain	CLIPdb	E-MTAB-1223
FUS	HITS-CLIP	Mouse brain	CLIPdb	GSE40653
FUS	HITS-CLIP	mES-derived neurons	CLIPdb	GSE43308
LIN28A	HITS-CLIP	mESC	CLIPdb	GSE37114
MBNL1	HITS-CLIP	C2C12	CLIPdb	E-MTAB-414
MBNL1/2	HITS-CLIP	Brain, heart, muscle, and C2C12	CLIPdb	GSE39911
MBNL2	HITS-CLIP	Hippocampus	CLIPdb	GSE38497
NOVA1	HITS-CLIP	Mouse brain		PMID: 20558669
NOVA1	HITS-CLIP	Mouse cortex		GSE69710
NOVA2	HITS-CLIP	Mouse cortex		GSE69710
PABPC1	PAR-CLIP	MEL	CLIPdb	GSE69755
PTBP1	HITS-CLIP	C2C12	CLIPdb	E-MTAB-414
RBFOX1	HITS-CLIP	Whole mouse brain	CLIPdb	SRP030031
RBFOX2	iCLIP	mESC	CLIPdb	GSE54794
RBFOX2	HITS-CLIP	Whole mouse brain	CLIPdb	SRP030031
RBFOX3	PAR-CLIP	P19-derived neurons	CLIPdb	SRP032810
RBM3	PAR-CLIP	MEF	CLIPdb	GSE40468
SRRM4	PAR-iCLIP	N2A cells	CLIPdb	GSE57278
SRSF1	iCLIP	P19 cells	Piranha	GSE69734
SRSF2	iCLIP	P19 cells	Piranha	GSE69734
SRSF3	iCLIP	P19 cells	Piranha	GSE69734
SRSF4	iCLIP	P19 cells	Piranha	GSE69734
SRSF5	iCLIP	P19 cells	Piranha	GSE69734
SRSF6	iCLIP	P19 cells	Piranha	GSE69734
SRSF7	iCLIP	P19 cells	Piranha	GSE69734

RBP	Experiment	Tissue	Preprocessing	Source/Accession
TAF15	HITS-CLIP	mES-derived neurons	CLIPdb	GSE43294
TARDBP	HITS-CLIP	Whole mouse brain	CLIPdb	GSE27394
TARDBP	HITS-CLIP	Whole mouse brain	CLIPdb	GSE40653
TARDBP	iCLIP	Mouse brain	CLIPdb	E-MTAB-1223
U2AF2	PAR-iCLIP	N2A cells	CLIPdb	GSE57278
U2AF2	iCLIP	Mouse brain	CLIPdb	E-MTAB-1223
UPF1	HITS-CLIP	mESC	CLIPdb	GSE41785

Chapter 3

Regulation of SR poison exons in neurogenesis

3.1 Background: Nonsense mediated decay and SR poison exons in neurogenesis

Nonsense mediated decay (NMD) is a mRNA degradation mechanism that targets transcripts containing premature termination codons (PTC) [6] (Figure 3.1). NMD was initially thought to be a quality control mechanism, but since then, several studies have found functional links between NMD and biological processes such as the promotion of cell proliferation and suppression of differentiation [5]. It is also clear now that NMD is tightly linked to pre-mRNA splicing through the alternative splicing of exons bearing PTC, also known as poison exons. Other forms of alternative splicing can also result in NMD-targeted isoforms. Termination codons between 50 and 55 nucleotides upstream of the last exon-exon junction of a mRNA molecule are usually recognized as PTC and trigger NMD [3, 6]. NMD-associated alternative splicing events are known to play a regulatory role in key splicing factors during neurogenesis. For example, PTBP1 inhibits PTBP2 activity by promoting the inclusion of PTBP2's exon 10, which induces NMD [23, 105]. For this reason, a decrease in exon 10 inclusion is necessary for increased PTBP2 activity and neuron differentiation.

Interestingly, some mammalian splicing factors have been reported to present highly conserved poison exons. A well known example are the genes in the SR family, all of which contain alternatively spliced poison exons that, if included in the mRNA molecule during splicing, mark it for degradation via NMD, decreasing the protein expression levels of these splicing factors. Strikingly, the poison exons and part of their surrounding intronic sequences in some of the SR genes are ultraconserved, meaning that they present completely identical sequences in humans, mice and rats [4, 106].

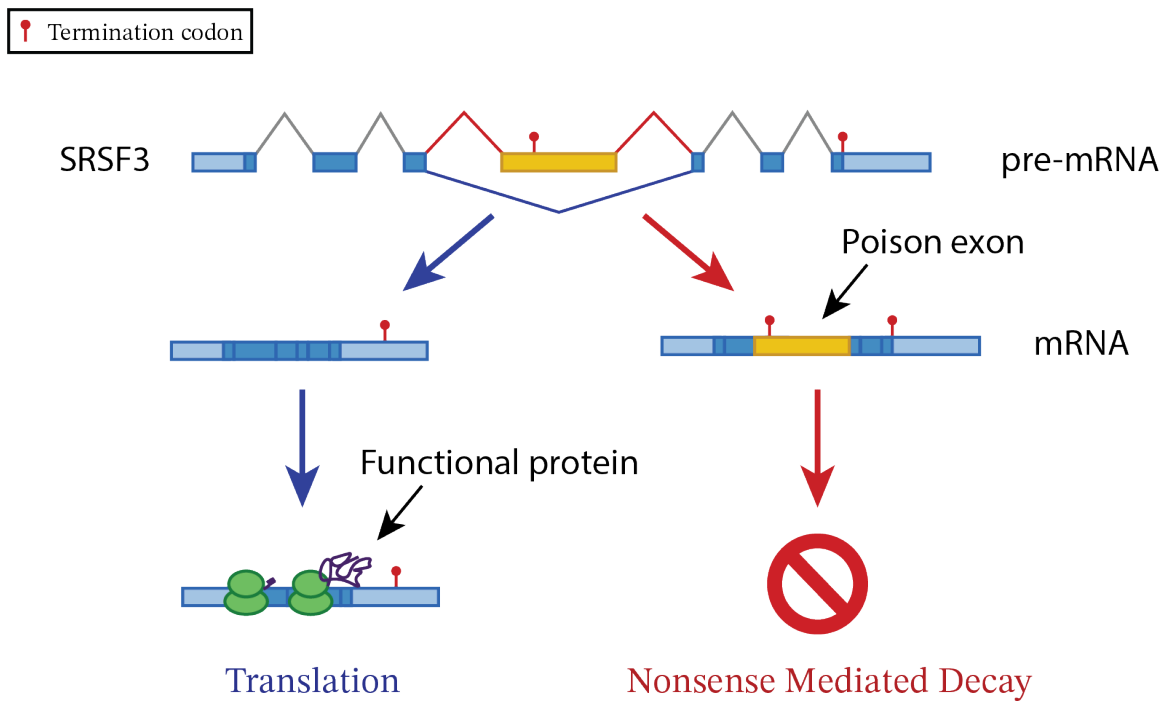


Figure 3.1: Alternative splicing of SRSF3 gene in humans. Inclusion of a poison exon (exon with an early termination codon), SRSF3 forms an isoform that is degraded through nonsense mediated decay. Skipping this poison exon results in a mRNA molecule that can be translated into a functional SRSF3 protein.

The extraordinarily high sequence conservation of these exons suggests that they might have a necessary regulatory role in the proper function of alternative splicing mechanisms. Consistent with this hypothesis, some studies have found a connection between misregulation of poison exon splicing and complex diseases such as cancer [11–17]. To the best of our knowledge, a clear link between these exons and neurogenesis is yet to be made. However, previous studies have found that misregulation of SR protein production has a deleterious effect in neuron development. Overexpression of the splicing repressor Serine/Arginine Rich Splicing Factor 10 (SRSF10) blocks inhibits the primary neuronal differentiation, while its depletion results in dysregulation of neurogenesis [107]; suggesting that a finely tuned regulation of SRSF10 expression is required for the proper development of the nervous system. Experimental research has found links between NMD and neurogenesis. A study in mice found that NMD is downregulated in developing embryonic brain and in neural stem cells (NSC) lines that undergo maturation. These changes in the rate of NMD were associated with a decrease in the expression of the core elements of the NMD pathway [5].

Our analysis in Chapter 2 showed several SR poison exons, including the poison exon of SRSF10, increase their Ψ during neurogenesis [91]. We asked if increased SR poison exon usage is a necessary regulatory mechanism in neurogenesis. In this Chapter we explore the alternative splicing of poison exons of SR genes in neurogenesis, and its potential auto-regulation. In addition, we apply machine learning approaches to CLIP-seq data in order to identify binding sites of each SR protein to the poison exon of its own pre-mRNA. Consistent with the results from the previous chapter, we observe a regulated increase in poison exon usage during neurogenesis. Our machine learning approaches also suggest strong binding preference of several SR proteins to their own poison exons.

3.2 Increased poison exon usage in SR genes during neurogenesis

All genes from the SR family that produce functional proteins can form an alternative isoform that is targeted for NMD. In most cases, the NMD target is formed by the inclusion of a poison exon (Table 3.1). Some cases such as SRSF2 produce NMD targeted isoforms through other types of alternative splicing events. We observed the change in SR poison exon usage across two time-series bulk RNA-seq datasets in both human [82] and mouse samples [81]. We use bulk RNA-seq data to reduce the noise that we previously reported in single-cell RNA-seq datasets [59]. For each dataset, we tested the significance of differential splicing by comparing each timepoint with the first of the series using rMATS [39]. In multiple dataset we observed a clear increase in poison exon usage during neurogenesis (Supplementary Figure 3.5). In contrast, poison exons across the transcriptome change in both directions in all datasets during neurogenesis. This observation rules out the possibility that the increase in SR poison

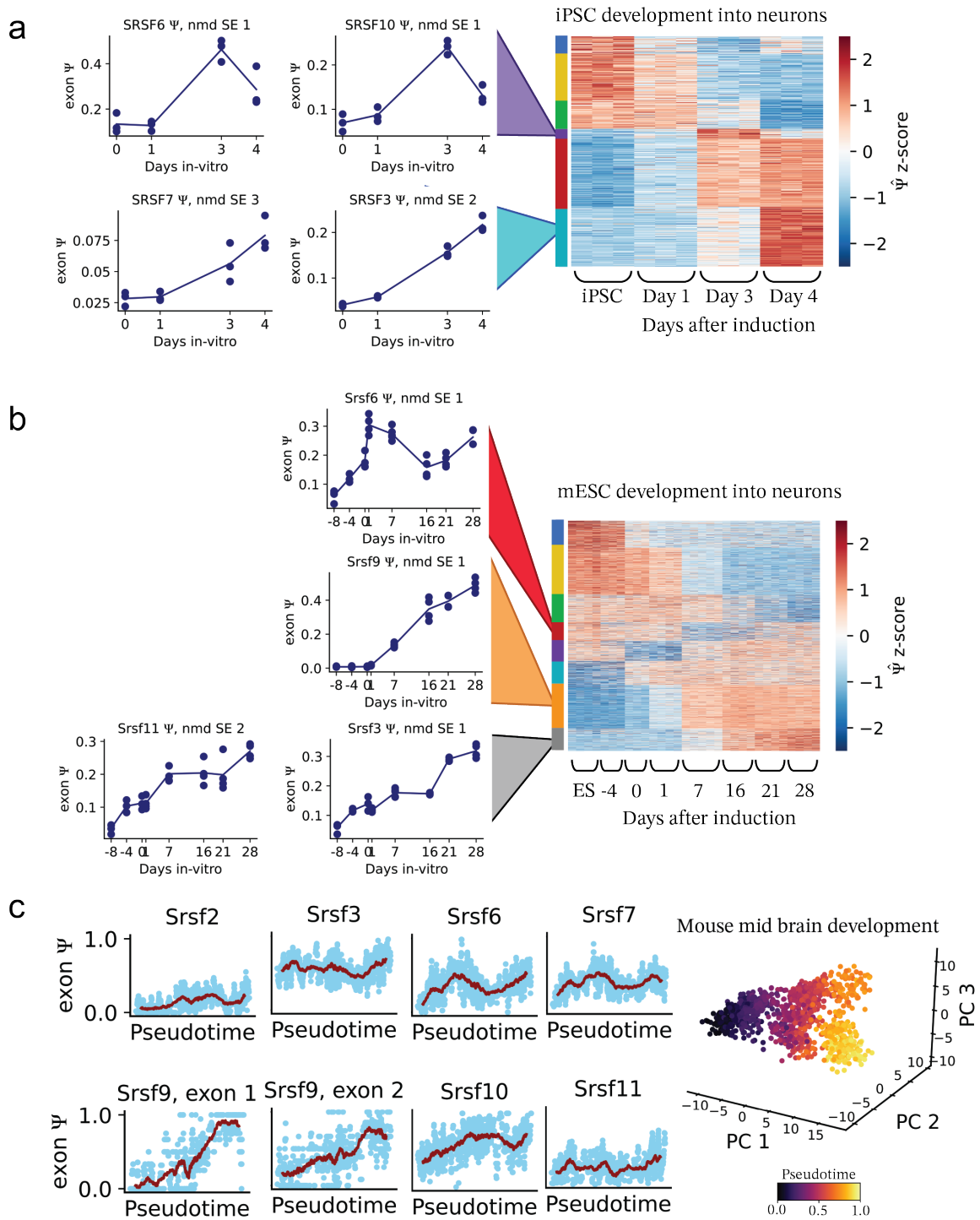


Figure 3.2: Alternative splicing of poison exons of the SR family in neurogenesis. SR family genes increase their poison exon usage in neurogenesis. a) Human induced pluripotent stem cells induced into neurogenesis. b) Mouse embryonic stem cells induced into neurogenesis. c) Mouse brain development.

exons is merely effect being part of a global change in NMD rates, which is known to occur in neurogenesis [5, 108, 109]. Interestingly, we found that while SR poison exons show an upward trend in most datasets, the timepoints at which they change varies between genes. Poison exons in SRSF6 and SRSF10 genes have an early change, compared with poison exons in SRSF2, SRSF3, SRSF7, SRSF9 and SRSF11, which have a late increase. This suggests that the poison exons of SR genes might have a diverse set of regulatory mechanisms.

Poison exons have the potential to act as a post-transcriptional mechanism of regulation of protein production through degradation of mRNA molecules. Under this hypothesis, we would expect that an increase in poison exon usage corresponds to a decrease in productive mRNA abundances and in protein production. We set to discover if SR genes present a significant change in mRNA abundance during neurogenesis. We found that the majority of SR genes that exhibited an increase in poison exon retention, also showed a significant decrease in mRNA abundance when tested for differential expression with DESeq2 [110] (Supplementary Figure 3.6). Consistently with the changes in poison exon splicing, the temporal dynamics of these changes in mRNA abundances were different between different SR genes. Overall, these results indicate that an increase in SR poison exon usage during neurogenesis results in a decrease of functional mRNA production of these genes.

3.3 A network of SR splicing factor auto and cross-regulation through poison exon splicing

We sought to reconstruct the network of regulatory factors that regulate exon usage of the poison exons, especially in the context of neurogenesis. SR proteins bind to their own poison exons [4, 96–100]. SR proteins tend to promote exon inclusion [96, 98, 111], which implies that binding to their own poison exons could act as a auto-regulatory negative feedback mechanism. In this model, SR proteins regulate their own production at a post-transcriptional level by promoting poison exon inclusion in their own mRNA molecules. In addition, four families of splicing factors have been extensively documented as key regulators of differential splicing during neurogenesis [27]: NOVA [22], PTBP [23, 24], RBFOX [25] and MBNL [26]. A splicing factor of this group, RBFOX2, regulates the alternative splicing of SR poison exons [112]. Based on this prior knowledge, we set to find the potential regulatory interactions between splicing factors and poison exons of the SR family in the context of neurogenesis.

3.3.1 A CLIP-seq data metaclassifier

In order to reconstruct the network of SR poison exon regulation, we collected CLIP-seq data from both human and mouse sources [95, 113]. We used the CLIP-peaks defined by the authors of each dataset. CLIP-seq protocols usually target mRNA

molecules, which brings forth two complications. Firstly, its peaks will only fall in genes that are expressed in the tissue or cell-line that is used in the experiment. Secondly, CLIP sequencing targets mature mRNAs, thus intronic binding sites are less likely to be captured. Thirdly, unstable mRNA molecules, such as those targeted for NMD, are less likely to be captured in CLIP-seq experiments, than their stable (non-NMD) counterparts. To ameliorate the effect of these biases, we trained a machine learning classifier to test the sequence affinity of each splicing factor.

Our classifier trains on peak sequences from CLIP-seq data, and predicts the affinity of a splicing factor to any given sequence (Supplementary Figure 3.7a). For each dataset, we obtained a set of negative sequences by randomly shuffling the CLIP-seq peaks in the genome. In order to avoid unwanted biases, we limited our shuffling to the sequences of the genes that showed CLIP-seq peaks. We also kept the same proportion of exonic and intronic peaks in the negative dataset, as in the CLIP-seq (positive) dataset. Once we had a set of positive and negative sequences, we randomly selected 80% of the sequences for training (training set), and withheld the other 20% for validation (test set). We further divided the peaks of the training set in two equally sized sets (each representing 40% of the total peaks of the dataset): training set 1, and training set 2.

We used training set 1 to train two SVM classifiers: Graphprot [114] and LS-GKM [115]. We then scored all the peaks in training set 2 with Graphprot, LSGKM, and DeepBind [116]: a neural-network classifier that is pre-trained for RBP affinity in RNACompete data and therefore does not require training on CLIP-seq peaks. We then normalized the scores of training set 2 from each of the three classifiers with isotonic regression. Finally, we used the normalized scores of the three classifiers in training set 2 to train a logistic regression model, which we used as a metaclassifier of splicing factor sequence affinity. Our metaclassifier overperformed each of the three individual classifiers, even when these were trained in the whole training set (both training set 1 and training set 2; Supplementary Figure 3.7b).

3.3.2 Poison exon binding maps

Once we had the metaclassifier of splicing factor sequence affinity, we sought to evaluate the binding affinity of each splicing factor to the poison exons of the SR genes in both humans and mice. In order to do that, we scored with the metaclassifier the sequence within a sliding window of 50 nucleotides extending from the upstream exon of the poison exon, to the downstream exon. For each nucleotide, we averaged the classifier score of the 25 windows that include it, and we used this score as a predictor of splicing factor affinity using a 0.15 false discovery rate.

Our results show a strong affinity of several SR proteins to the poison exons of their own genes (Figure 3.3). Moreover, we also found that many SR proteins bind to the poison exons of other SR genes. This allowed us to reconstruct a network of regulatory

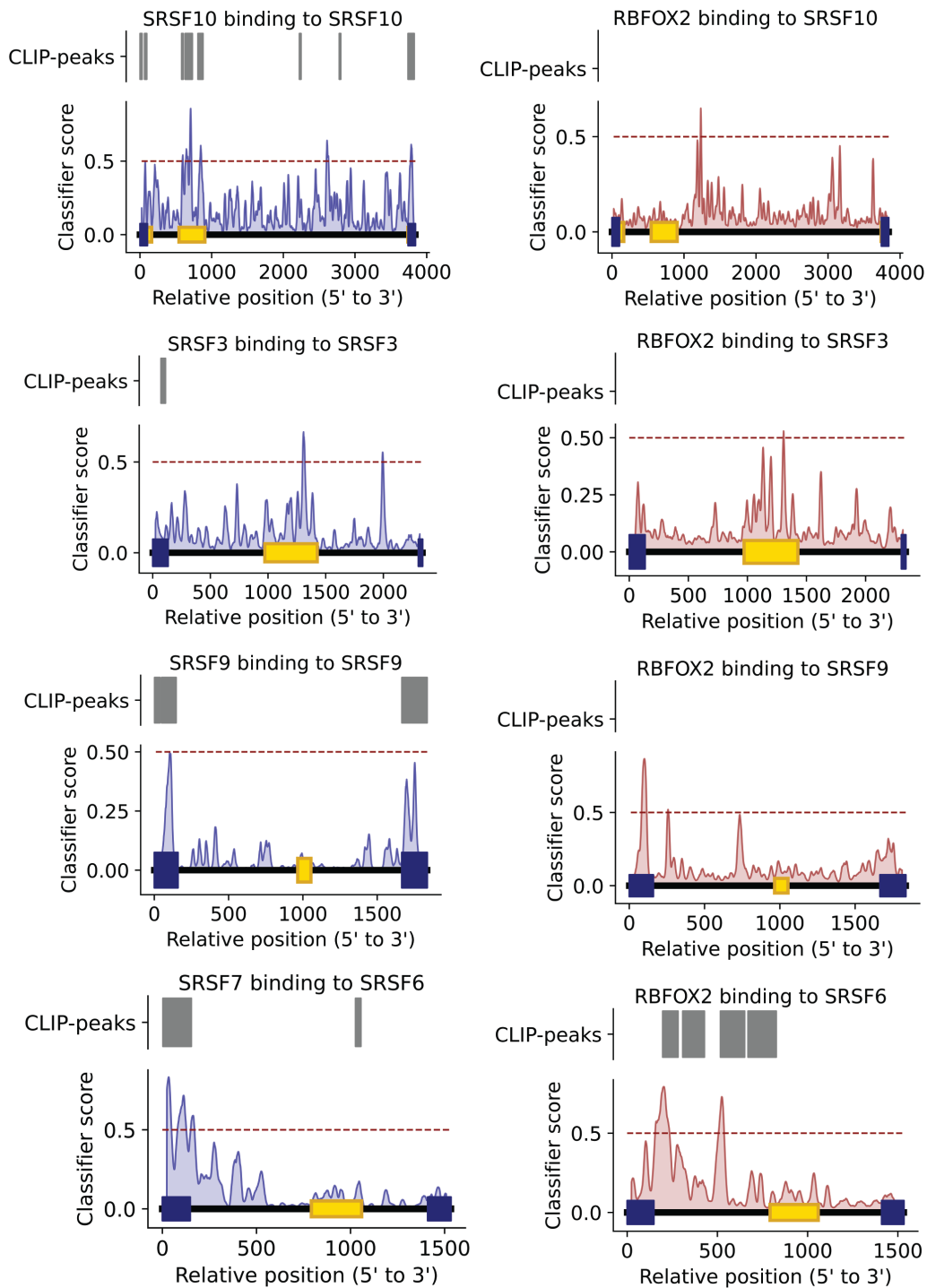


Figure 3.3: Binding maps show the probability of RBP binding to the proximity of poison exons of human SR genes, as predicted by the metaclassifier. SR proteins tend to bind to their own poison exons, or to one of their flanking exons. RBFOX2 also binds to the poison exons of SR genes.

interactions between SR proteins (Figure 3.4). When data was available, we assessed the direction of these interactions by observing the direction of change in poison exon inclusion in knockdown and knockout data of SR genes from ENCODE [113]. An SR protein can repress the production of another SR protein (or itself) by promoting poison exon inclusion in the latter SR gene. In this case, knockdown/knockout of the first SR gene would result in decreased poison exon usage in the second SR gene. An SR protein could also promote the production of another SR protein (or itself) by suppressing poison exon inclusion. In this case, knockdown/knockout of the first SR gene would result in increased poison exon usage in the second SR gene. The majority of the SR protein interactions in the regulatory network were repressive.

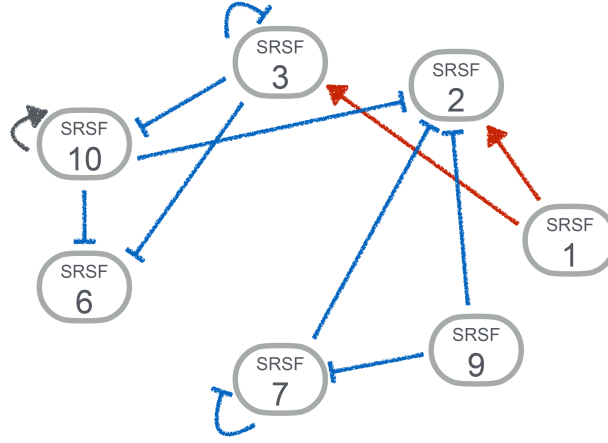
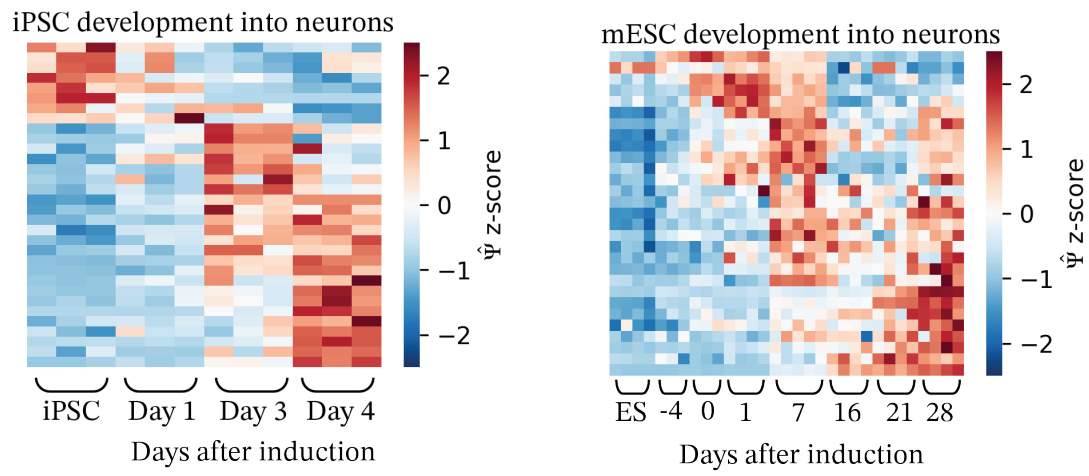
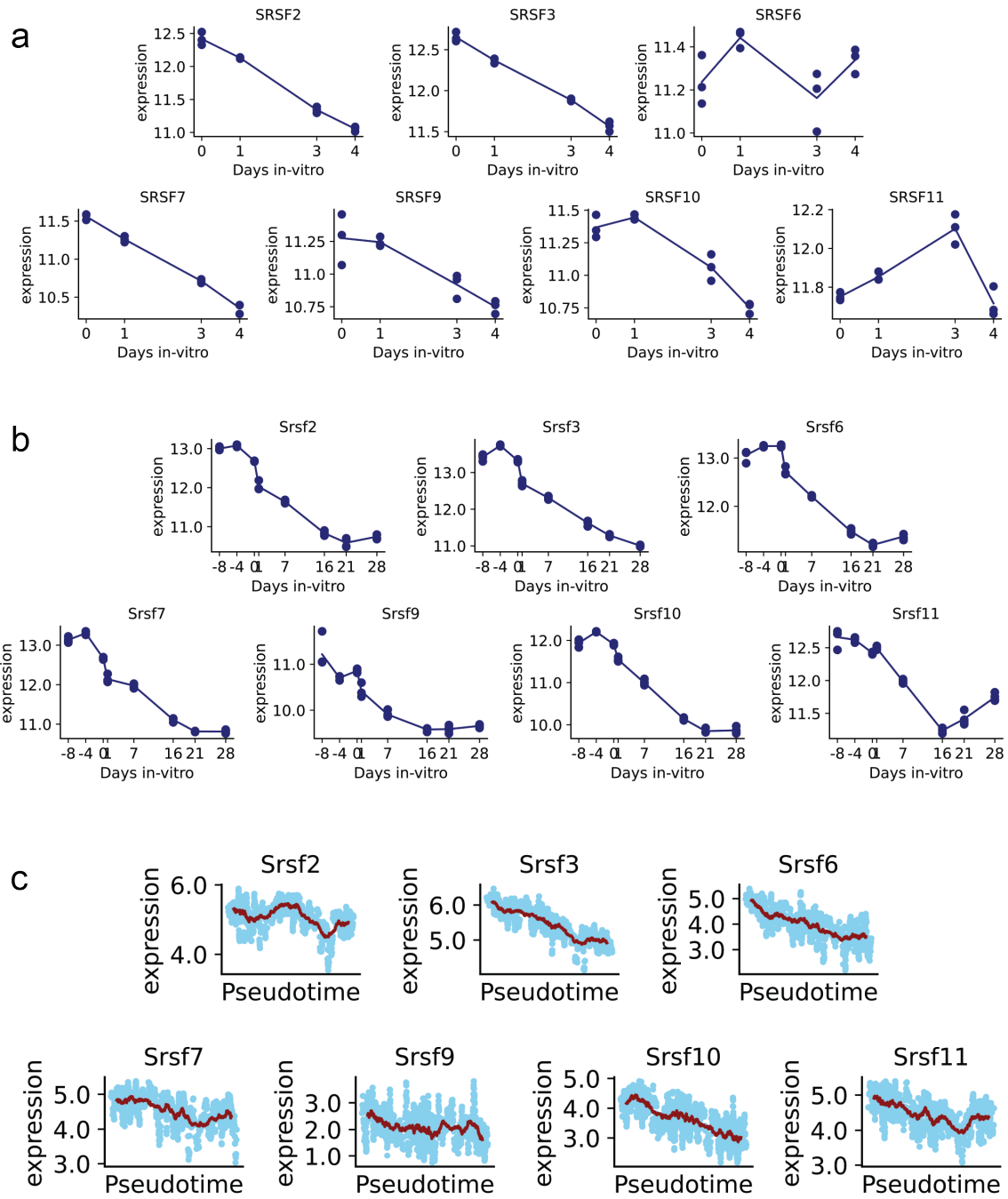


Figure 3.4: Interaction of SR proteins through binding to poison exons, as determined by the metaclassifier. Blue arrows indicate promotion of poison exon usage (repression of SR protein), while red arrows indicate a suppression of poison exon usage (promotion of SR protein).

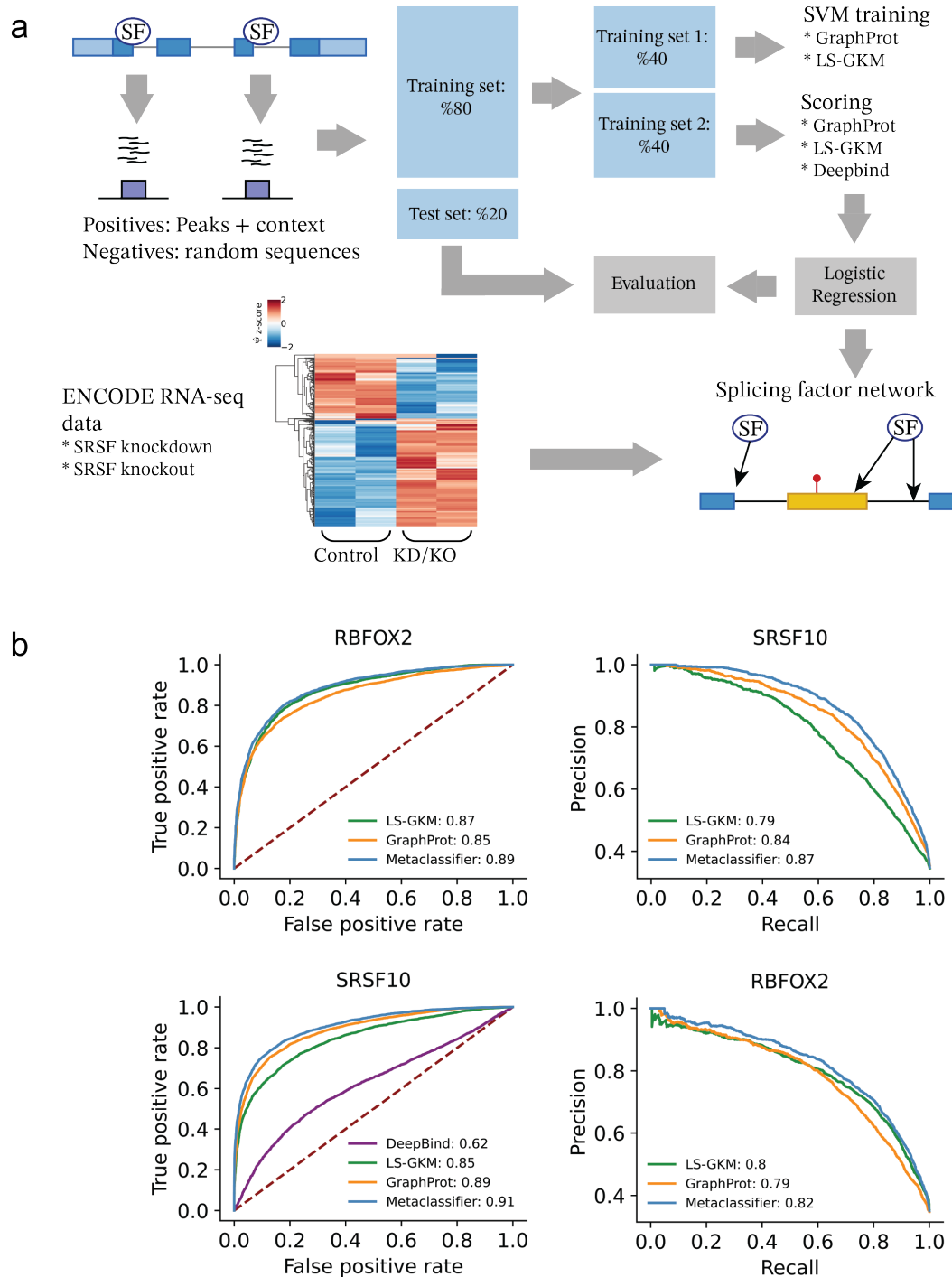
3.4 Supplementary figures and tables



Supplementary Figure 3.5: Alternative splicing of poison exons in neurogenesis.



Supplementary Figure 3.6: Most SR genes decrease their expression in neurogenesis. a) Human induced pluripotent stem cells induced into neurogenesis. b) Mouse embryonic stem cells induced into neurogenesis. c) Mouse brain development.



Supplementary Figure 3.7: Analysis of RBP binding to RNA. a) Workflow of the metaclassifier of RBP binding. CLIP-seq peaks are scored with multiple classifiers, and then merged into one single metaclassifier that determines the probability of an RBP to bind to a sequence of RNA. b) The metaclassifier outperforms each classifier on its own in predicting CLIP-seq peaks.

Table 3.1: SR poison exon coordinates.

SR gene	Human (hg38)	Mouse (mm10)
SRSF2	chr17:76735772-76735875:-	chr11:116851648-116851751:-
SRSF3	chr6:36599821-36600276:+	chr17:29039453-29039909:+
SRSF4	chr1:29159864-29160091	chr4:131873637-131901700:+
SRSF5	chr14:69769511-69770540	chr12:80948129-80948504:+
SRSF6	chr20:43459153-43459420:+	chr2:162932814-162933082:+
SRSF7	chr2:38748898-38749346:-	chr17:80204621-80205081:-
SRSF9	chr12:120464733-120464822:-	chr5:115328347-115328704:+; chr5:115331450-115331539:+
SRSF10	chr1:23977911-23978273:-	chr4:135858841-135859201:+
SRSF11	chr1:70231095-70231203:+	chr3:158023243-158023359:-

References

- [1] Brenton R. Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2):100–107, 2001.
- [2] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, Dec 2008.
- [3] Eszter Nagy and Lynne E. Maquat. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci*, 23(6):198–199, Jun 1998.
- [4] Liana F Lareau, Maki Inada, Richard E Green, Jordan C Wengrod, and Steven E Brenner. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, 446(7138):926–929, Apr 2007.
- [5] Chih H Lou, Ada Shao, Eleen Y Shum, Josh L Espinoza, Lulu Huang, Rachid Karam, and Miles F Wilkinson. Posttranscriptional control of the stem cell and neurogenic programs by the nonsense-mediated RNA decay pathway. *Cell Rep*, 6(4):748–764, Feb 2014.
- [6] Nele Hug, Dasa Longman, and Javier F Cáceres. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res*, 44(4):1483–1495, Feb 2016.
- [7] Yeon Lee and Donald C Rio. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, 84:291–323, March 2015.
- [8] Hong Han, Ulrich Braunschweig, Thomas Gonatopoulos-Pournatzis, Robert J. Weatheritt, Calley L. Hirsch, Kevin C.H. Ha, Ernest Radovani, Syed Nabeel-Shah, Tim Sterne-Weiler, Juli Wang, Dave O’Hanlon, Qun Pan, Debashish Ray, Hong Zheng, Frederick Vizeacoumar, Alessandro Datti, Lilia Magomedova, Carolyn L. Cummins, Timothy R. Hughes, Jack F. Greenblatt, Jeffrey L. Wrana, Jason Moffat, and Benjamin J. Blencowe. Multilayered control of alternative splicing regulatory networks by transcription factors. *Molecular Cell*, 65(3):539–553.e7, 2017.

- [9] Enrique Lara-Pezzi, Manuel Desco, Alberto Gatto, and María. Gómez-Gaviro. Neurogenesis: Regulation by Alternative Splicing and Related Posttranscriptional Processes. *Neuroscientist*, 23(5):466–477, 10 2017.
- [10] Elisabetta Furlanis and Peter Scheiffele. Regulation of Neuronal Differentiation, Function, and Plasticity by Alternative Splicing. *Annu Rev Cell Dev Biol*, 34:451–469, 10 2018.
- [11] N. Agrebi, I. Ben-Mustapha, N. Matoussi, N. Dhouib, M. Ben-Ali, N. Mekki, M. Ben-Ahmed, B. Larguèche, S. Ben Becher, M. Béjaoui, and M.R. Barbouche. Rare splicing defects of fas underly severe recessive autoimmune lymphoproliferative syndrome. *Clinical Immunology*, 183:17–23, 2017.
- [12] Michael Hecker, Annelen Rüge, Elena Putscher, Nina Boxberger, Paulus Stefan Rommer, Brit Fitzner, and Uwe Klaus Zettl. Aberrant expression of alternative splicing variants in multiple sclerosis – a systematic review. *Autoimmunity Reviews*, 18(7):721–732, 2019.
- [13] Benedetta Perrone, Valentina La Cognata, Teresa Sprovieri, Carmine Ungaro, Francesca Luisa Conforti, Sebastiano Andò, and Sebastiano Cavallaro. Alternative splicing of als genes: Misregulation and potential therapies. *Cellular and Molecular Neurobiology*, 40(1):1–14, Jan 2020.
- [14] James D. Thomas, Jacob T. Polaski, Qing Feng, Emma J. De Neef, Emma R. Hoppe, Maria V. McSharry, Joseph Pangallo, Austin M. Gabel, Andrea E. Belleville, Jacqueline Watson, Naomi T. Nkinsi, Alice H. Berger, and Robert K. Bradley. Rna isoform screens uncover the essentiality and tumor-suppressor activity of ultraconserved poison exons. *Nature Genetics*, 52(1):84–94, Jan 2020.
- [15] Nathan K. Leclair, Mattia Brugiolo, Laura Urbanski, Shane C. Lawson, Ketan Thakar, Marina Yurieva, Joshy George, John T. Hinson, Albert Cheng, Brenton R. Graveley, and Olga Anczuków. Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. *Mol Cell*, 80(4):648–665, 11 2020.
- [16] Wei Jiang and Liang Chen. Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Computational and Structural Biotechnology Journal*, 19:183–195, 2021.
- [17] Yuanjiao Zhang, Jinjun Qian, Chunyan Gu, and Ye Yang. Alternative splicing and cancer: a systematic review. *Signal Transduction and Targeted Therapy*, 6(1):78, Feb 2021.
- [18] Bushra Raj and Blencowe Benjamin J. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron*, 87(1):14–27, July 2015.

- [19] Nuno L Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y Xiong, Serge Guernoussov, Leo J Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Colak, TaeHyung Kim, Christine M Misquitta-Ali, Michael D Wilson, Philip M Kim, Duncan T Odom, Brendan J Frey, and Benjamin J Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, 2012.
- [20] Jason Merkin, Caitlin Russell, Chen Ping, and Christopher B Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, 2012.
- [21] Celine K Vuong, Black Douglas L, and Zheng Sika. The neurogenetics of alternative splicing. *Nature Reviews. Neuroscience*, 17(5):265–281, 2016.
- [22] Chaolin Zhang, Maria A Frias, Aldo Mele, Matteo Ruggiu, Taesun Eom, Christina B Marney, Huidong Wang, Donny D Licatalosi, John J Fak, and Robert B Darnell. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science*, 329(5990):439–443, 2010.
- [23] Eugene V Makeyev, Jiangwen Zhang, Monica A Carrasco, and Maniatis Tom. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Molecular Cell*, 27(3):435–448, 2007.
- [24] Xiaochang Zhang, Ming Hui Chen, Xuebing Wu, Andrew Kodani, Jean Fan, Ryan Doan, Manabu Ozawa, Jacqueline Ma, Nobuaki Yoshida, Jeremy F Reiter, Douglas L Black, Peter V Kharchenko, Phillip A Sharp, and Christopher A Walsh. Cell-type-specific alternative splicing governs cell fate in the developing cerebral cortex. *Cell*, 166(5):1147–1162, 2016.
- [25] Bushra Raj, Manuel Irimia, Ulrich Braunschweig, Timothy Sterne-Weiler, O’Hanlon Dave, Zhen-Yuan Lin, Chen Ginny I, Laura E Easton, Jernej Ule, Anne-Claude Gingras, Eduardo Eyras, and Blencowe Benjamin J. Global regulatory mechanism underlying the activation of an exon network required for neurogenesis. *Molecular Cell*, 56(1):90–103, 2014.
- [26] Konstantinos Charizanis, Kuang-Yung Lee, Ranjan Batra, Marianne Goodwin, Chaolin Zhang, Yuan Yuan, Lily Shiue, Melissa Cline, Marina M Scotti, Guangbin Xia, Ashok Kumar, Tetsuo Ashizawa, H Brent Clark, Takashi Kimura, Masanori P Takahashi, Harutoshi Fujimura, Kenji Jinnai, Hiroo Yoshikawa, Mário Gomes-Pereira, Geneviève Gourdon, Noriaki Sakai, Seiji Nishino, Thomas C Foster, Manuel Ares Jr, Robert B Darnell, and Maurice S Swanson. Muscleblind-like 2-mediated alternative splicing in the developing brain and dysregulation in myotonic dystrophy. *Neuron*, 75(3):437–450, Aug 2012.
- [27] Sebastien M Weyn-Vanhentenryck, Huijuan Feng, Dmytro Ustianenko, Rachel

- Duffié, Qinghong Yan, Martin Jacko, Jose C Martinez, Marianne Goodwin, Xuegong Zhang, Ulrich Hengst, Stavros Lomvardas, Maurice S Swanson, and Chaolin Zhang. Precise temporal regulation of alternative splicing during neural development. *Nature Communications*, 9(1):2189–17, June 2018.
- [28] Huijuan Feng, Daniel F Moakley, Shuonan Chen, Melissa G McKenzie, Vilas Menon, and Chaolin Zhang. Complexity and graded regulation of neuronal cell-type-specific alternative splicing revealed by single-cell RNA sequencing. *PNAS*, 118(10):e2013056118, March 2021.
- [29] Sebastien M Weyn-Vanhentenryck, Aldo Mele, Qinghong Yan, Shuying Sun, Natalie Farny, Zuo Zhang, Chenghai Xue, Margaret Herre, Pamela A Silver, Michael Q Zhang, Adrian R Krainer, Robert B Darnell, and Chaolin Zhang. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Reports*, 6(6):1139–1152, March 2014.
- [30] Mathieu Quesnel-Vallières, Manuel Irimia, Sabine P Cordes, and Benjamin J Blencowe. Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes & Development*, 29(7):746–759, 2015.
- [31] Ryan Lister, Ronan C O’Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, Harvey Millar, and Joseph R Ecker. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, 113(3):523–536, 2008.
- [32] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.
- [33] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Beren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, 2010.
- [34] Jingyi Jessica Li, Ci-Ren Jiang, James B Brown, Haiyan Huang, and Peter J Bickel. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *PNAS*, 108(50):19867–19872, 2011.
- [35] Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, The RGASP Consortium, Tim J Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10:1177–1184, 2013.

- [36] Yarden Katz, Eric T. Wang, Edoardo M. Airolidi, and Christopher B. Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, Dec 2010.
- [37] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323–16, August 2011.
- [38] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Nature Methods*, 10:1177–1184, 2013.
- [39] Shihao Shen, Juw Won Park, Zhi-Xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. 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(51):E5593–601, December 2014.
- [40] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaubblomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z Levin, Hongkun Park, and Aviv Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, June 2013.
- [41] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160, Nov 2016.
- [42] Amos Tanay and Aviv Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338, 01 2017.
- [43] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, Mar 2015.
- [44] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinus, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell*, 65(4):631–643, February 2017.
- [45] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, Apr 2014.
- [46] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef,

- Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477–16, June 2018.
- [47] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. Scenic: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, Nov 2017.
- [48] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(284), 2018.
- [49] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018.
- [50] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(390), 2019.
- [51] Ángeles Arzalluz-Luque and Ana Conesa. Single-cell RNAseq for the study of isoforms-how is that possible? *Genome Biology*, 19(1):110, August 2018.
- [52] Jennifer Westoby, Marcela Sjöberg Herrera, Anne C Ferguson-Smith, and Martin Hemberg. Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. *Genome Biology*, 19(1):191–14, November 2018.
- [53] Jennifer Westoby, Pavel Artemov, Martin Hemberg, and Anne Ferguson-Smith. Obstacles to detecting isoforms using full-length scRNA-seq data. *Genome Biology*, 21(74), March 2020.
- [54] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*, 11(2):163–166, Feb 2014.
- [55] Joshua D. Welch, Yin Hu, and Jan F. Prins. Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research*, 44(8):e73–e73, 01 2016.
- [56] Yan Song, Olga B. Botvinnik, Michael T. Lovci, Boyko Kakaradov, Patrick Liu, Jia L. Xu, and Gene W. Yeo. Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. *Mol Cell*, 67(1):148–161, Jul 2017.
- [57] Stephanie M. Linker, Lara Urban, Stephen J. Clark, Mariya Chhatriwala, Shradha Amatya, Davis J. McCarthy, Ingo Ebersberger, Ludovic Vallier, Wolf

- Reik, Oliver Stegle, and Marc Jan Bonder. Combined single-cell profiling of expression and dna methylation reveals splicing regulation and heterogeneity. *Genome Biology*, 20(1):30, Feb 2019.
- [58] Georgi K. Marinov, Brian A. Williams, Ken McCue, Gary P. Schroth, Jason Gertz, Richard M. Myers, and Barbara J. Wold. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Research*, 24(3):496–510, Mar 2014.
- [59] Carlos F. Buen Abad Najar, Nir Yosef, and Liana F. Lareau. Coverage-dependent bias creates the appearance of binary splicing in single cells. *eLife*, 9:e54603 DOI: 10.7554/eLife.54603, June 2020.
- [60] Nir Yosef, Alex K. Shalek, Jellert T. Gaublomme, Hulin Jin, Youjin Lee, Amit Awasthi, Chuan Wu, Katarzyna Karwacz, Sheng Xiao, Marsela Jorgolli, David Gennert, Rahul Satija, Arvind Shakya, Diana Y. Lu, John J. Trombetta, Meenu R. Pillai, Peter J. Ratcliffe, Mathew L. Coleman, Mark Bix, Dean Tantin, Hongkun Park, Vijay K. Kuchroo, and Aviv Regev. Dynamic regulatory network controlling TH17 cell differentiation. *Nature*, 496(7446):461–468, Apr 2013.
- [61] Jellert T Gaublomme, Nir Yosef, Youjin Lee, Rona S Gertner, Li V Yang, Chuan Wu, Pier Paolo Pandolfi, Tak Mak, Rahul Satija, Alex K Shalek, Vijay K Kuchroo, Hongkun Park, and Aviv Regev. Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell*, 163(6):1400–1412, Dec 2015.
- [62] D. Donny Licatalosi and Robert B. Darnell. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*, 11(1):75–87, Jan 2010.
- [63] Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mrna quantification and differential analysis with census. *Nature Methods*, 14(3):309–315, Mar 2017.
- [64] Yuanhua Huang and Guido Sanguinetti. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology*, 18(1):123, June 2017.
- [65] Yuanhua Huang and Guido Sanguinetti. Computational identification of splicing phenotypes from single cell transcriptomic experiments. *bioRxiv*, November 2020.
- [66] Zeev Waks, Allon M. Klein, and Pamela A. Silver. Cell-to-cell variability of alternative RNA splicing. *Mol Syst Biol*, 7:506, Jul 2011.
- [67] Hédia Maamar, Moran N Cabili, John Rinn, and Arjun Raj. linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis. *Genes & development*, 27(11):1260–1271, June 2013.

- [68] Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Tomislav Ilicic, Sarah A Teichmann, and John C Marioni. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, 6(1):8687, October 2015.
- [69] Geng Chen, John Paul Schell, Julio Aguila Benitez, Sophie Petropoulos, Marlene Yilmaz, Björn Reinius, Zhanna Alekseenko, Leming Shi, Eva Hedlund, Fredrik Lanner, Rickard Sandberg, and Qiaolin Deng. Single-cell analyses of X Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Research*, 26(10):1342–1354, October 2016.
- [70] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–226, October 2008.
- [71] Anton JM Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, January 2019.
- [72] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, April 2014.
- [73] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, September 2013.
- [74] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271, December 2008.
- [75] Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell RNA sequencing. *Nature Communications*, 10:2611–27, June 2019.
- [76] David DeTomaso, Matthew G Jones, Meena Subramaniam, Tal Ashuach, Chun J Ye, and Nir Yosef. Functional interpretation of single cell similarity maps. *Nature Communications*, 10(4376), September 2019.
- [77] Jiancheng Liu, Anqi Geng, Xiwei Wu, Ren-Jang Lin, and Qiang Lu. Alternative RNA splicing associated with mammalian neuronal differentiation. *Cerebral Cortex*, 28(8):2810–2816, August 2018.
- [78] Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton J M Larsson, Omid R Faridani, and

- Rickard Sandberg. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology*, May 2020.
- [79] Ishaan Gupta, Paul G. Collier, Bettina Haase, Ahmed Mahfouz, Anoushka Joglekar, Taylor Floyd, Frank Koopmans, Ben Barres, August B. Smit, Steven A. Sloan, Wenjie Luo, Olivier Fedrigo, M. Elizabeth Ross, and Hagen U. Tilgner. Single-cell isoform rna sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*, 36(12):1197–1202, Dec 2018.
 - [80] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 10 2012.
 - [81] Kyle S Hubbard, Ian M Gut, Megan E Lyman, and Patrick M McNutt. Longitudinal RNA sequencing of the deep transcriptome during neurogenesis of cortical glutamatergic neurons from murine ESCs. *F1000Research*, 2(35):35, 2013.
 - [82] Volker Busskamp, Nathan E Lewis, Patrick Guye, Alex H M Ng, Seth L Shipman, Susan M Byrne, Neville E Sanjana, Jernej Murn, Yinqing Li, Shangzhong Li, Michael Stadler, Ron Weiss, and George M Church. Rapid neurogenesis through transcriptional activation in human stem cells. *Molecular Systems Biology*, 10(11):760–760, 2014.
 - [83] Michael B. Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. Performance assessment and selection of normalization procedures for single-cell rna-seq. *Cell Systems*, 8(4):315–328.e8, 2019.
 - [84] Hui Yuan Xiong, Yoseph Barash, and Brendan J Frey. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics*, 27(18):2554–2562, September 2011.
 - [85] Lior Faigenbloom, Nimrod D Rubinstein, Yoel Kloog, Itay Mayrose, Tal Pupko, and Reuven Stein. Regulation of alternative splicing at the single-cell level. *Molecular Systems Biology*, 11(12):845–845, December 2015.
 - [86] Yosef Barash, Benjamin Blencowe, and Brendan Frey. Model-based detection of alternative splicing signals. *Bioinformatics*, 26(12):i325–i333, June 2010.
 - [87] Fabienne Lescroart, Xiaonan Wang, Xionghui Lin, Benjamin Swedlund, Souhir Gargouri, Adriana Sánchez-Dànes, Victoria Moignard, Christine Dubois, Catherine Paulissen, Sarah Kinston, Berthold Göttgens, and Cédric Blanpain. Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science*, 359(6380):1177–1181, March 2018.

- [88] Russell B. Fletcher, Diya Das, Levi Gadye, Kelly N. Street, Ariane Baudhuin, Allon Wagner, Michael B. Cole, Quetzal Flores, Yoon G. Choi, Nir Yosef, Elizabeth Purdom, Sandrine Dudoit, Davide Risso, and John Ngai. Deconstructing Olfactory Stem Cell Trajectories at Single-Cell Resolution. *Cell Stem Cell*, 20(6):817–830, 06 2017.
- [89] David DeTomaso and Nir Yosef. Identifying Informative Gene Modules Across Modalities of Single Cell Genomics. *bioRxiv*, February 2020.
- [90] Wei Xiong Wen, Adam J Mead, and Supat Thongjuea. VALERIE: Visual-based inspection of alternative splicing events at single-cell resolution. *PLoS Computational Biology*, 16(9):e1008195, September 2020.
- [91] Katarína Tiklová, Asa K Björklund, Laura Lahti, Alessandro Fiorenzano, Sara Nolbrant, Linda Gillberg, Nikolaos Volakakis, Chika Yokota, Markus M Hilscher, Thomas Hauling, Fredrik Holmström, Eliza Joodmardi, Mats Nilsson, Malin Parmar, and Thomas Perlmann. Single-cell RNA sequencing reveals midbrain dopamine neuron diversity emerging during mouse brain development. *Nature Communications*, 10(581), February 2019.
- [92] Pascal Y Smith, Charlotte Delay, Johanne Girard, Marie-Amélie Papon, Emmanuel Planel, Nicolas Sargeant, Luc Buée, and Sébastien S Hébert. Microrna-132 loss is associated with tau exon 10 inclusion in progressive supranuclear palsy. *Human Molecular Genetics*, 20(20):4016–4024, August 2011.
- [93] Jia Bei Wang and David R Burt. Differential expression of two forms of gabaa receptor gamma 2-subunit in mice. *Brain Research Bulletin*, 27(5):731–735, November 1991.
- [94] Dina Speidel, Frederique Varoqueaux, Carsten Enk, Mari Nojiri, Ruslan N Grishanin, Thomas FJ Martin, Kay Hoffman, Nils Brose, and Kerstin Reim. A family of ca2+-dependent activator proteins for secretion: comparative analysis of structure, expression, localization, and function. *Journal of Biological Chemistry*, 278(52):52802–52809, December 2003.
- [95] Yu-Cheng T Yang, Chao Di, Boqin Hu, Meifeng Zhou, Yifang Liu, Nanxi Song, Yang Li, Jumpei Umetsu, and Zhi J Lu. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, 16(51), February 2015.
- [96] Hassan Jumaa and Peter J. Nielsen. The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *EMBO J*, 16(16):5077–5085, Aug 1997.
- [97] A. Sureau, R. Gattoni, Y. Dooghe, J. Stévenin, and J. Soret. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J*, 20(7):1785–1796, Apr 2001.

- [98] Fabrice Lejeune, Xiaojie Li, and Lynne E. Maquat. Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. *Mol Cell*, 12(3):675–687, Sep 2003.
- [99] Minna-Liisa Änkö, Michaela Müller-McNicoll, Holger Brandl, Tomaz Curk, Crtomir Gorup, Ian Henry, Jernej Ule, and Karla M Neugebauer. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol*, 13(3):R17, 2012.
- [100] Michaela Müller-McNicoll, Valentina Botti, Antonio M. de Jesus Domingues, Holger Brandl, Oliver D. Schwich, Michaela C. Steiner, Tomaz Curk, Ina Poser, Kathi Zarnack, and Karla M. Neugebauer. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev*, 30(5):553–566, Mar 2016.
- [101] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C. Devlin, Maayan Baron, Cristina H. Hajdu, Diane M. Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38(3):333–342, Mar 2020.
- [102] Matthew G Jones, Alex Khodaverdian, Jeffrey J Quinn, Michelle M Chan, Jeffrey A Hussmann, Robert Wang, Chenling Xu, Jonathan S Weissman, and Nir Yosef. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biology*, 21(92), April 2020.
- [103] Philip J Uren, Emad Bahrami-Samani, Suzanne C Burns, Mei Qiao, Fedor V Karginov, Emily Hodges, Gregory J Hannon, Jeremy R Sanford, Luiz OF Penalva, and Andrew D Smith. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, 28(23):3013–3020, December 2012.
- [104] David Fischer, Fabian Theis, and Nir Yosef. Impulse model-based differential expression analysis of time course sequencing data. *Nucleic Acids Research*, 46:e119, 2018.
- [105] Paul L. Boutz, Peter Stoilov, Qin Li, Chia-Ho Lin, Geetanjali Chawla, Kristin Ostrow, Lily Shiue, Manuel Ares, and Douglas L. Black. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev*, 21(13):1636–1652, Jul 2007.
- [106] Julie Z. Ni, Leslie Grate, John P. Donohue, Christine Preston, Naomi. Nobida, Georgeann O’Brien, Lily Shiue, Tyson A. Clark, John E. Blume, and Manuel Ares. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev*, 21(6):708–718, Mar 2007.
- [107] Karen J. Liu and Richard M. Harland. Inhibition of neurogenesis by SRp38, a

- neuroD-regulated RNA-binding protein. *Development*, 132(7):1511–1523, Apr 2005.
- [108] Tahani Alrahbeni, Francesca Sartor, Jihan Anderson, Zosia Miedzybrodzka, Colin McCaig, and Berndt Müller. Full UPF3B function is critical for neuronal differentiation of neural stem cells. *Mol Brain*, 8:33, May 2015.
 - [109] Samie R. Jaffrey and Miles F. Wilkinson. Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. *Nat Rev Neurosci*, 19(12):715–728, 12 2018.
 - [110] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12):550, 2014.
 - [111] Zhihong Zhou and Xiang-Dong Fu. Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma*, 122(3):191–207, Jun 2013.
 - [112] Mohini Jangi, Paul L. Boutz, Prakriti Paul, and Phillip A. Sharp. Rbfox2 controls autoregulation in RNA-binding protein networks. *Genes Dev*, 28(6):637–651, Mar 2014.
 - [113] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.
 - [114] Daniel Maticzka, Sita J. Lange, Fabrizio Costa, and Rolf Backofen. Graph-Prot: modeling binding preferences of RNA-binding proteins. *Genome Biol*, 15(1):R17, Jan 2014.
 - [115] Dongwon Lee. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics*, 32(14):2196–2198, 07 2016.
 - [116] Babak Alipanahi, Andrew DeLong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 33(8):831–838, Aug 2015.