UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational analysis of single-cell alternative splicing

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

in

Bioinformatics and Systems Biology

by

Olga Borisovna Botvinnik

Committee in charge:

Professor Gene Yeo, Chair
Professor Sheng Zhong, Co-Chair
Professor C. Titus Brown
Professor Amy Pasquinelli
Professor Sam Pfaff
Professor Kun Zhang

2017

The dissertation of Olga Borisovna Botvinnik is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2017

DEDICATION

To my family, my parents, and Kwasi.

EPIGRAPH

Always stay gracious, best revenge is your paper – Beyoncé Giselle Knowles Carter

TABLE OF CONTENTS

List of Figures

ACKNOWLEDGEMENTS

There are too many people to thank but I will do my best.

Thank you to Dr. Yan Song for being a patient mentor, keeping me on track, and giving key criticisms of the internal workings of all the algorithms, and for making sure everything is explained as clearly as possible. I couldn't have done any of this without you.

Thank you to Dr. Gene Yeo for giving me freedom and opportunities to teach my passions of open-source software, open science to unsuspecting biologists.

Thank you to Dr. Sheng Zhong for key computational feedback on `anchor`'s classifier.

Thank you to Dr. C. Titus Brown for believing in me and sponsoring me for the NumFOCUS John Hunter Technical Fellowship.

Thank you to Dr. Amy Pasquinelli for keeping the broader biological picture in mind, to Dr. Kun Zhang for knowing what's out there in the single-cell world, and to Dr. Sam Pfaff for neurobiological insights.

Thank you to Dr. Mike Lovci for teaching me about alternative splicing, to Dr. Boyko Kakaradov for teaching me about machine learning, to Gabriel Pratt for being my Python partner in crime, to Dr. Andrew Gross for answering my annoying `pandas` questions and to the entire Yeo Lab for being supportive of my open source endeavors.

Thank you especially to especially Pantea Khodami who will always me my Kappa Alpha Theta big sister and has always been there for me and has supported me every step of the way. Thank you to my friends Dr. Anne-Ruxandra Carvunis, Dr. Charisse Crenshaw, Colleen Stoyas, Cynthia Hsu, Kara Gordon, Emily Wheeler, David Nelles, Ron Batra, En-Ching Luo, Alain Domissy, Alice

Ho, Rose Hurwitz, Alexa Robinson, Wendi Zhang, Nazita Lejevardi, Michelle Ma, Liam Fedus .... Thank you to the *Lean In and Stay Classy San Diego crew*: Dianna Cowern, Shari Haynes, Kristen Peña Breanna Berry, Daryl Fairweather, Shelly Wanamaker, and ... Thank you to my current and former roommates and affiliates Cailey Bromer, Daniella Bardalez Gagliuffi, Robert Schwartz, Jennifer Hammond, Patrick Wise, Sara Glass, Keawe Kolohe, Sharon and Matt Scott, and Andie Rotner for being a kind, supportive and welcoming environment to come home to, and reminding me there's so much more to life than just cells.

Thank you to my family, especially my mother and father for leaving the Soviet Union and bringing me to the United States to give me the opportunity to write a dissertation like this. Thank you to my brothers for being understanding of my mad scientist ways, thank you to my stepmother Ira whose family warmly accepted me as one of their own, especially Alya and Max Zolotorev.

Huge thank you to Kwasi Nti for surviving the Herculean task of dating a PhD student, for listening so hard and making sure I always made the best decision for me. I love you.

Chapter 1, in part, is currently being prepared for submission for publication of the material. Botvinnik, Olga; Song, Yan; Yeo, Gene W. The dissertation author was the primary investigator and author of this material.

Chapter 2, in part, has been accepted for publication as the supplementary material as it may appear in Molecular Cell, 2017, Yan Song*, Olga B Botvinnik*, Michael T Lovci, Boyko Kakaradov, Patrick Liu, Jia L. Xu and Gene W Yeo (* These authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3, in full, has been accepted for publication as it may appear in Molecular Cell, 2017, Yan Song*, Olga B Botvinnik*, Michael T Lovci, Boyko

Kakaradov, Patrick Liu, Jia L. Xu and Gene W Yeo (* These authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this paper.

VITA

2010	S. B. in Mathematics, Massachusetts Institute of Technology
2010	S. B. in Biological Engineering, Massachusetts Institute of Technology
2012	M. S. in Biomolecular Engineering and Bioinformatics, University of California, Santa Cruz
2017	Ph. D. in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

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. *Accepted*. * These authors contributed equally to this work.

Curtis A Nutter, Elizabeth A Jaworski, Sunil K Verma, Vaibhav Deshmukh, Qiongling Wang, **Olga B Botvinnik**, Mario J Lozano, Ismail J Abass, Talha Ijaz, Allan R Brasier, Nisha J Garg, Xander H T Wehrens, Gene W Yeo, and Muge N Kuyumcu-Martinez. Dysregulation of RBFOX2 Is an Early Event in Cardiac Pathogenesis of Diabetes. *Cell Reports*, 15(10):2200-2213, 2016.

Jong Wook Kim*, **Olga B Botvinnik***, Omar Abudayyeh, Chet Birger, Joseph Rosen- bluh, Yashaswi Shrestha, Mohamed E Abazeed, Peter S Hammerman, Daniel DiCara, David J Konieczkowski, et al. Characterizing genomic alterations in cancer by complementary functional associations. *Nature Biotechnology, 2016*. * These authors contributed equally to this work.

P Compeau and P Pevzner. *Bioinformatics Algorithms* Volume 1, volume 1 of An Active Learning Approach. Active Learning Publishers LLC, 2 edition, 2015. Contributed text, figures, problems and code solutions, primarily to "Chapter 4: How Do We Sequence Antibiotics?".

Kris C Wood, David J Konieczkowski, Cory M Johannessen, Jesse S Boehm, Pablo Tamayo, **Olga B Botvinnik**, Jill P Mesirov, William C Hahn, David E Root, Levi A Garraway, et al. MicroSCALE screening reveals genetic modifiers of therapeutic response in melanoma. *Science Signaling*, 5(224):rs4, 2012.

A Goncearenco, P Grynberg, **Olga B Botvinnik**, Geoff Macintyre, and Thomas Abeel. Highlights from the Eighth International Society for Computational Biology (ISCB) Student Council Symposium 2012. *BMC Bioinformatics*, 2012.

Naomi Galili, Pablo Tamayo, **Olga B Botvinnik**, Jill P Mesirov, Margarita R Brooks, Gail Brown, and Azra Raza. Prediction of response to therapy with ezatiostat in lower risk myelodysplastic syndrome. *Journal of Hematology & Oncology*, 5(1):1, 2012.

Naomi Galili, Pablo Tamayo, **Olga B Botvinnik**, Jill P Mesirov, Jennifer Zikria, Gail Brown, and Azra Raza. Gene Expression Studies May Identify Lower Risk Myelodys- plastic Syndrome Patients Likely to Respond to Therapy with Ezatiostat Hydrochloride (TLK199). *Blood*, 118(21):2779-2779, 2011.

Michael F Berger, Gwenael Badis, Andrew R Gehrke, Shaheynoor Talukder, Anthony A Philippakis, Lourdes Pena-Castillo, Trevis M Alleyne, Sanie Mnaimneh, **Olga B Botvinnik**, Esther T Chan, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133(7):1266-1276, 2008.

ABSTRACT OF THE DISSERTATION

Computational analysis of single-cell alternative splicing

by

Olga Borisovna Botvinnik

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2017

Professor Gene Yeo, Chair
Professor Sheng Zhong, Co-Chair

Alternative splicing (AS) generates isoform diversity critical for cellular identity and homeostasis in multicellular life. Although AS variation has been observed among single cells for a few events, little is known about the biological significance of such variation. We developed Expedition, a computational framework consisting of outrigger, a *de novo* splice graph transversal algorithm to detect AS; anchor, a Bayesian approach to assign modalities and bonvoyage, a visualization tool using non-negative matrix factorization to display modality changes. Applying Expedition to single iPSCs undergoing neuronal differentiation, we discover up to 20% of AS exons exhibit bimodality and are flanked by more conserved introns harboring distinct cis-regulatory motifs. Bimodal exons constitute the majority of cell-type specific splicing, are highly dynamic during cellular transitions, preserve translatability and reveal intricacy of cell states invisible to global gene expression analysis. Systematic AS characterization in single cells redefines our understanding of AS complexity in cell biology.

Chapter 1

Single-cell mRNA processing: If you liked it, you should have put a Seq on it

1.1 Introduction

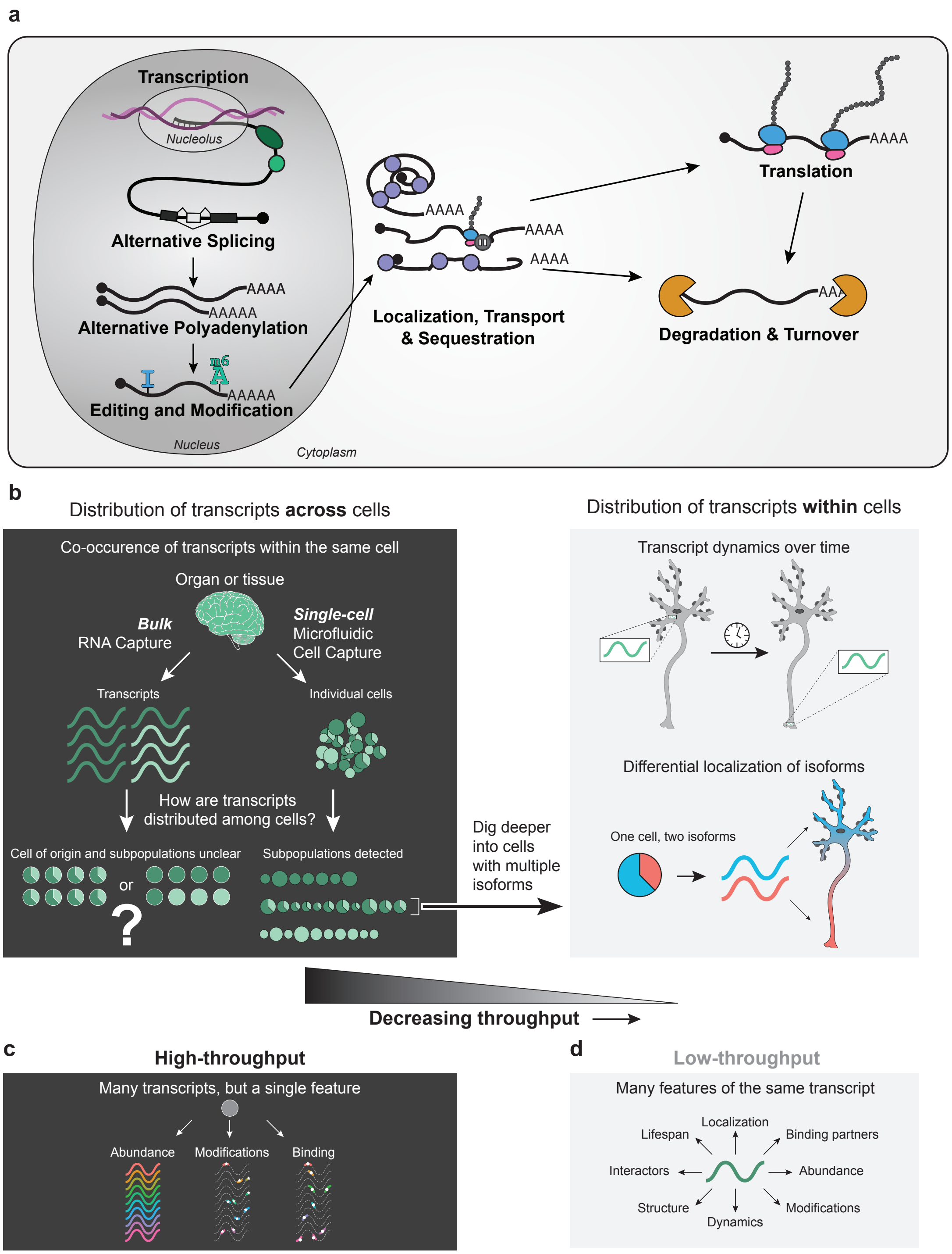
The human body contains an estimated 3.72×10^{13} cells[1], all of which are highly specialized in form and function, and yet despite their incredible diversity in phenotypes, each cell contains nearly identical genotypes. These cells are heterogeneous because of their different RNA, protein, and metabolite molecules, which coordinately regulate the cell to express precise phenotypes. To study the variation between cells, we turn to single-cell analysis.

The original tool for single-cell analysis is the microscope[2;3], which can visualize structural differences between individual cells, but the molecules that create these differences are too small to resolve in live cells by current microscope technology. To compare the molecules of single cells, recent advances in microfluidics have allowed for capture of one cell at a time, which can be coupled with modern high-throughput technology to measure many messenger RNA (mRNA) molecules per cell, and together these are combined to create single-cell RNA-sequencing (scRNA-Seq)[4;5]. Computational analysis of these high-dimensional data can identify distinct cellular states or delineate cellular trajectories (reviewed by Bacher and Kendziorski[6]; Cannoodt et al.[7]; Liu and Trapnell[8]; Trapnell[9]; Stegle et al.[10]).

While single-cell capture has enabled probing of cellular state measured through mRNA abundances, the study of an mRNA molecule's rich life (**Figure 1.1**a) from birth (transcription) to death (degradation), the collection of actions known as mRNA processing[11–15], has only started to be addressed at the single-cell level. As in bulk RNA-seq[16–23], scRNA-seq has enabled the investigation of RNA processing features that are measureable by sequencing, such as alternative splicing, RNA editing, and alternative polyadenylation[24–29]. However, the high-throughput nature of scRNA-seq captures only the abundance of RNA transcripts in a snapshot in time and loses the information of RNA modifications, dynamics, localization, binding partners, and secondary structure. Thus, these features must be measured a different way.

Ideally, we would capture the entire cellular and molecular context of an RNA molecule. To accomplish this, we turn back to the microscope, a tried and true tool. While even the highest resolution microscopes cannot discern individual molecules without significant amplification[30;31], microscopy captures cellular context including morphology and subcellular localization, and in the case of live-cell imaging, dynamics. Microscopy is limited by the ability to design fluorescent constructs to visualize RNA and protein molecules, and as a result, can only be performed for a few targets a time. Middle-ground technologies that are relatively high-throughput but also measure several aspects of the same cell or same transcript[32;33] have highest potential for discovery. We will review the available methods to probe RNA processing at the single cell level, and highlight the current limitations, showing opportunities for novel technology to make breakthroughs in the knowledge of RNA processing.

Figure 1.1 *(next page)*: **Overview of open questions in single-cell RNA processing.**
a. Overview of the processing steps in an RNA's life cycle: transcription (biogenesis), alternative splicing, poly-adenylation, modification, export, localization, translation, and degradation.
b. Dichotomy of investigating distribution of transcripts across cells with high-throughput methods, and distribution of transcripts within cells using high-resolution methods.
c. Examples of high-throughput measurements, where many transcripts can be measured at once, but only one feature of them may be measured.
d. Examples of high-resolution measurements, where only a few transcripts can be measured at once, but many features of them can be profiled.


a
Transcription
Nucleolus
Alternative Splicing
AAAA
AAAAA
Alternative Polyadenylation
m6
A
I
AAAAA
Editing and Modification
Nucleus
Cytoplasm
AAAA
AAAA
AAAA
Localization, Transport
& Sequestration
AAAA
Translation
AAA
Degradation & Turnover
b
Distribution of transcripts across cells
Co-occurence of transcripts within the same cell
Organ or tissue
Bulk
RNA Capture
Single-cell
Microfluidic
Cell Capture
Transcripts
Individual cells
How are transcripts
distributed among cells?
Cell of origin and subpopulations unclear
or
?
Subpopulations detected
Dig deeper
into cells
with multiple
isoforms
Distribution of transcripts within cells
Transcript dynamics over time
Differential localization of isoforms
One cell, two isoforms
Decreasing throughput
c
High-throughput
Many transcripts, but a single feature
Abundance
Modifications
Binding
d
Low-throughput
Many features of the same transcript
Localization
Lifespan
Binding partners
Interactors
Abundance
Structure
Modifications
Dynamics

1.2 Balancing high-throughput and high-resolution single-cell technologies

A complex tissue such as the human brain contains many different transcripts, but by measuring them at the bulk level, the cell of origin for each transcript is unknown (**Figure 1.1**b, left). Using single-cell technologies, we quantify an RNA processing event either as presence or absence (e.g. m6A or splicing) or a continuous quantity (e.g. abundance or poly-A tail length). With these quantifications in hand, we want to be able to capture individual cells and measure each cell's transcripts to understand two separate questions (**Figure 1.1**): How are transcripts distributed (1) across cells, and (2) within cells?

The questions of distributions of transcripts across cells and within cells represent the ends of a spectrum, each with their own advantages and limitations. Where on the one extreme there are high-throughput methods which can measure many transcripts per cell, but are low-resolution and can only measure one aspect, abundance, and on the other extreme are high-resolution methods which can measure a limited number of transcripts (low-throughput) but can measure many aspects beyond abundance, such as dynamics and localization.

To measure RNA processing across cells, we use high-throughput single-cell technologies to study cellular biology and answer the question, is a particular RNA processing event found only within certain subpopulations of cells, or does it co-occur within individual cells? If it's found within the same cell, are these on the same transcript, or on different transcripts in individual cells? Since these technologies only extract a snapshot in cellular time, akin to a still frame in a movie, we don't know how these transcripts change over time or how they are physically used within an individual cell, and thus high-throughput technologies are hypothesis-generating experiments.

To test these hypotheses, we turn to low-throughput, high resolution technologies in molecular biology, which can answer the question, within cells, are the different transcripts differentially localized in different populations? Does the transcript have different temporal dynamics? Does it have different interactors, binding partners, or three-dimensional structure? To truly understand this, we would need to follow up with an experiment that turns the process of or over expresses it to see how it affects cellular fate.

1.2.1 High-throughput

We define "throughput" as the number of different molecules that can be measured at once from a single cell (**Figure 1.1**c). High-throughput methods such as scRNA-Seq can measure 18,000 transcripts per cell (2000 genes/cell) for up to one million cells[34]. These high-dimensional datasets can be used to study two main questions regarding the distribution of transcripts across cells (**Figure 1.1**b): (1) Do these processes co-occur on the same transcript, or on different transcripts within the same cell? (2) If they occur in different cells, do these cells comprise distinct population sub-structures? Digging into the population sub-structures can especially elucidate novel cell states or types, and understanding of cellular biology. Thus, high-throughput methods allow for measurement of one feature across many targets, and are especially enable the deep study of cellular biology and cell state.

High-throughput technologies come at the cost of resolution: many scRNA-seq techniques measures only the abundance of the 5′ end of the mature, poly-adenylated mRNA[4;5;35], thus missing large portions of the transcripts, the immature pre-mRNA and due to sequencing technology limitations, these methods cannot measure nucleotide modifications. Additionally, these measurements are destructive and the original cells cannot be restored to re-analyze to observe how they would respond to perturbations. If we develop a hypothesis from the high throughput data, we must test it on completely new and independent cells. Thus, high-throughput data such as scRNA-seq represents only a snapshot in time and loses the dynamics of the transcript's biogenesis, the localization of the transcript in the cell, its interactors, and any modifications or secondary structure.

The digital measurements of high-throughput data are necessarily lossy in part because the technology itself defines what can be observed, and all other features remain undetected. This echoes Jaron Lanier's "You are Not a Gadget"[36] which discusses how the digital representation of an object inevitably removes all unmeasured features, and this can be problematic as they are still a part of the object. As an analogy, digitizing an impressionist painting as a photo does not capture the time of day the flax seeds were pressed to create the oil paints, the tautness of the muslin cloth on the frame, or the names of every person who has ever viewed the painting, but all of these are part of the history of the painting. These are examples of incidental measurements that are lost as soon as the object is digitized, because by digitizing, you've made decisions about what you think is important, and lose information about what you've chosen not to measure. Thus, as soon as you measure the abundance of a cell's transcripts through RNA-seq, you lose all other features of the transcript, such as its structure and nucleotide modifications, its localization and binding partners, its lifespan, and even more unmentioned features which have not yet been observed but could contribute to the RNA molecule's biography.

Finally, high dimensional data requires many computational manipulations[6;7;10], which could retreat from the biology. Due to the lossy nature of the digital measurements, computational methods may not retain the biology as the algorithm may latch onto signals that are artefacts of the technology, rather than true biology. To many, the enormous amounts of data may feel like staring into "tea leaves" and trying to draw conclusions, and rigorous follow up and investigation across multiple algorithms is required to ensure the analysis is biologically correct. Thus, while high-throughput single-cell measurements such as RNA-Seq allow for exploration of huge biological datasets and digging into cellular populations, they are limited in their ability to study multiple features of an mRNA transcript's life cycle at once.

1.2.2 Low-throughput

To become closer to the biology, we turn to lower-throughput techniques such as microscopy, which allows for observation of dynamics and localization, trends that are not currently visible in high-throughput data (**Figure 1.1**d). While almost an antediluvian tool, microscopy, especially fluorescence and confocal microscopy heavily used in single-cell analyses, has undergone many advances in resolution and throughput, allowing for visualization of RNA and protein molecules in thick (millimeter) tissue slices[37–40]. To further investigate the subcellular characteristics such as time scales and localization, of RNA processing, we turn to low throughput analyses to answer two main questions: (1) Do the processes exist at the same or different, time and place? (2) What is the fate of transcripts with different RNA processing features? For example, if a gene's transcript with distinct RNA features co-occur in the same cell, how does the cell use the diverse transcripts differently? By studying RNA processing in both time and space, we will become ever closer to a deep understanding of the regulation of RNA.

1.2.3 "Goldilocks" balance of throughput and resolution

Technologies that balance both throughput and resolution that are "just right," as in the children's fairy tale of where the Goldilocks character finds the perfect porridge that isn't too salty or too sweet. These in-between methods are currently limited in the still-growing field of technological development in single cells and are a major need. Most technological innovation has focused on increasing either throughput or resolution, but we argue that the large leaps will be made by combining the two to show several aspects of a single RNA transcript molecule, in a way that has never been seen before. For example, in situ sequencing[38;41–46] combines localization of transcripts with high-throughput microscopy or sequencing to spatially resolve transcripts within, and across cells. Other combination technologies include single-cell multiomics methods which measure several aspects of a cell at once, to answer the questions of how DNA mosaicism or epigenetics contribute to gene expression[33;47–49], or how RNA levels influence protein levels with Seq-well. The development of methods which maximize the "bang for the buck" – i.e. high-throughput while retaining cellular context – will revolutionize single-cell biology and create opportunities for novel biological questions.

1.3 Insights into RNA processing through single-cell technologies

Here we discuss the aspects of RNA processing that are measurable by current technologies, and summarized in **Figure 1.2**.

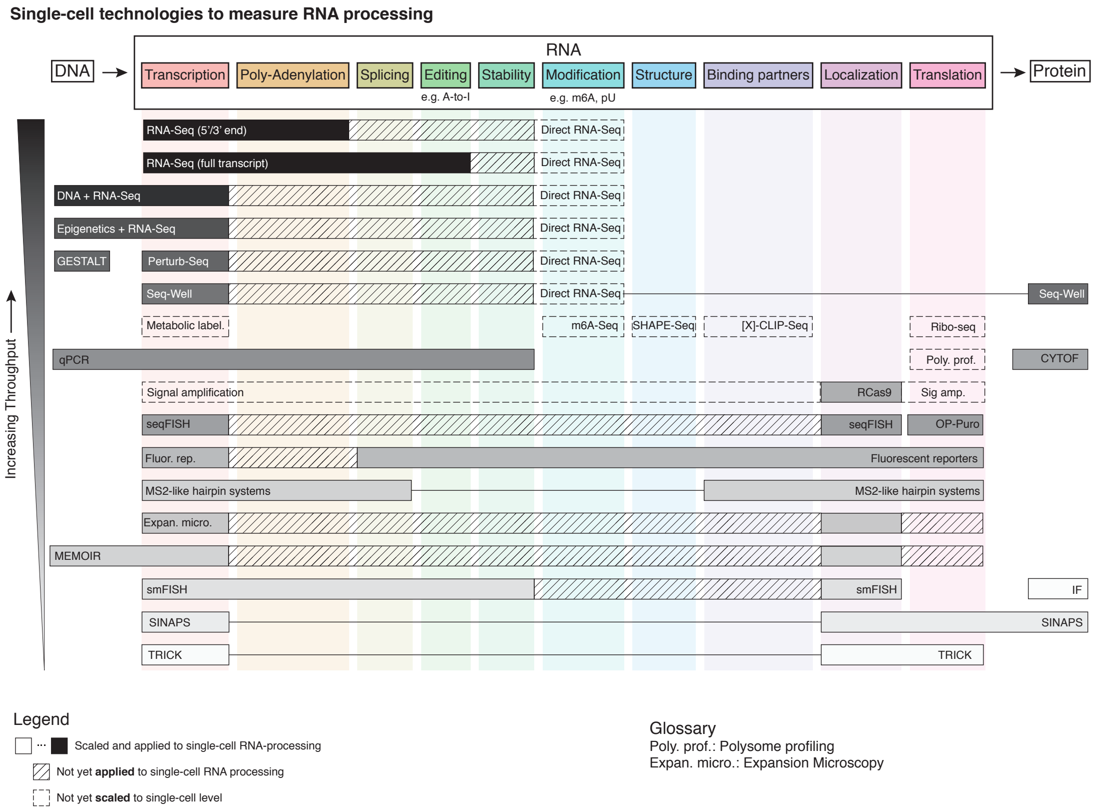


Figure 1.2: Overview of what can be measured at different steps of the central dogma in terms of RNA processing.

1.3.1 Chaotic, "bursty" transcription is harmonized by slow nuclear export

Single-cell analyses have long showed that cells do not constantly transcribe genes, but rather do it in a, stochastic "bursty" fashion[50–53]. Single-molecule imaging paved the way to show the stochasticity of gene expression within genetically homogeneous cell populations[53;54] (, and was confirmed by single-cell RNA-Seq, which also showed bursty and cyclical transcriptional kinetics that would otherwise be masked in bulk sequencing data[27;55;56]. How does the cell deal with what appears to be such chaotic creation of RNA? Two papers showed that while sudden bursts of expression in the nucleus are common, mRNAs aren't immediately exported to the cytoplasm[57;58], suggesting the mRNAs are first sequestered in storage facility before they are exported. Thus, bursty transcription is tempered with a slow drip of RNA export from the nucleus.

1.3.2 Variability in isoforms is nonrandom but functional implications remain unclear

Alternative splicing (AS) is a co- and post-transcriptional modification of mRNA that is a mechanism for proteomic diversity[59–64]. AS removes introns, sequences from the immature mRNA which are not contained in the mature, poly-adenylated mRNA. Since the outcome of AS is the presence or absence of an intron, this can be readily measured using RNA-sequencing (RNA-Seq)[23]. RNA-seq is a readily available technology for single cells and thus AS analysis can be directly applied. We discuss below the applications of single-cell analyses to studying alternative splicing.

Overall variation of AS in single cells can be studied using high-throughput methods such as RNA-sequencing. While bulk measurements show that individual isoforms may vary within a population, this doesn't show how individual cells use different transcripts. Understanding how individual cells choose one transcript or another has been challenging to measure. Do cells tend to have only one isoform of a gene, or do they contain many? One question that RNA-seq AS analysis can answer is, unbiasedly, which splicing events are changing within a cell population, or across cell populations? One early study found variability in isoforms using RNA fluorescence in situ hybridization (FISH)[65]. The earliest scRNA-seq study found that single-cell AS was more "all or nothing" – each cell tended to have only one isoform, compared to bulk samples, which showed many isoforms[27]. This shows that individual cells tend to only have a single isoform, and that variation in isoform composition, such as having multiple isoforms, is likely observed in bulk samples because of the heterogeneity of cells, rather than heterogeneity of transcripts within cells. Completeness of splicing is also associated with higher conservation of introns and exons[66]. Another study used full-transcript scRNA-seq to find that single cells had higher percentages of novel splice junctions than bulk data[25]. Another study looked at alternative splicing in single cells captured from the mouse visual cortex and found changes in alternative splicing throughout the different cortical layers[67]. Another study developed statistical models to find variable splicing events across single cells, and applied this to find splicing events that change with cell cycle[29]. Contrary to the initial study that most splicing events are "all or nothing," another study used UMIs coupled with long-read sequencing to extract poly-adenylated mRNAs from from mouse oligodendrocytes and vascular and leptomeningeal cells[24]. They found a purifying selection of exon splice sites in protein coding genes, with very few junctions mapping to mis-spliced exons. They found up to 25 isoforms per cell, with most isoform differences occurring due to alternative transcription start and end sites, rather than cassette exons or 3′ or 5′ end positional differences. Earlier studies also showed that the choice of 3′ isoform is highly variable, more variable than random selection[28], a finding that was confirmed by RNA-FISH. More work is needed to analyze the choice of final exons, which can be aided by advances in computational methods for end-sequence analysis[68]. These results show that overall, alternative transcript architectures are highly variable in a statistically significant manner, but the purpose of this variation is still unknown. The ability to decipher whether there is function in the variability[69–72] is limited by the capture methods of transcripts, and is limited to highly-expressed transcripts.

The dynamics of splicing can be studied using microscopy of fluorescently labeled transcripts. For example, the competition between transcript release and splicing of the final intron of human beta-globin was found to favor transcript release, then splicing. Interestingly, splicing of diffuse RNA occurred rapidly, faster than diffusion[73] and alternative splicing is known to occur primarily after transcription[54]. These studies have shed light onto how the transcription of RNA is decoupled from the creation of alternative transcripts.

The differential usage of the same gene's isoforms remains an interesting question. Why would a cell contain multiple isoforms? What are their different functions[72]? For example, the cell polarity gene Cdc42 has two possible terminal exons, exon 6 and exon 7, but in non-neuronal cells, exon 6 is strongly suppressed by polypyrimidine tract binding protein (PTB/Ptbp1) and only exon 7 is included[74]. However, in neurons, Ptbp1 is not expressed, and both exon 6 and exon 7 transcripts are expressed in equimolar ratios, disrupting this equimolar concentration resulted in defects in neuronal development. Too much of the exon 6 isoform led to insufficient axonogenesis and too much of the exon 7 isoform led to insufficient dendritic maturation. This suggests the exact ratios and distributions are critical for normal neuronal development.

1.3.3 Adenosine to Inosine RNA Editing

Another method of transcriptome and proteome diversification is through RNA editing, the most commonly occurring one being the deamination of adenosine to inosine (A-to-I) editing[19;75]. Inosine has three positions for hydrogen bonds and performs base-pairing like guanine, and thus by sequencing can be detected by an A to G transition. However, true identification of editing sites is difficult, as the negative control of knockout of the A-to-I editing enzyme family ADAR is embryonic lethal in mammals (though not in *Caenorhabditis elegans*). How can true editing sites be identified in mammals? Again, the questions we are interested in are, how is A-to-I editing distributed (1) across and (2) within cells? Across cells (1) could theoretically be answered with single-cell full-transcript sequencing, but to our knowledge has not yet been performed. Within cells (2) can be answered using microscopy based methods, for example visualizing adenosine-to-inosine edited transcripts using inoFISH[76]. RNA in situ hybridization methods can also be applied to RNA editing. There were differences in variability between transcripts and between cells: editing of GRIA2 was highly variable from cell to cell but editing of NUP43 was fairly constant between cells, suggesting that an individual gene's editing is regulated separately. By using microscopy, the authors were able to address questions of localization, co-/post-transcriptionality, and variability within and between cells. A limitation of this method is the need to design probes against all possible flanking sequences of edited transcripts, and the cells must be fixed. Technology that can use live cells and/or resolve tens or hundreds of edited sites at a time will allow for interrogation of broader trends.

Other forms of RNA editing, such as G-to-A[77], C-to-U[11], and U-to-C[78] editing, and their across-cell distributions, and within-cell localization and dynamics have yet to be explored at the single cell level.

1.3.4 Spatial organization of transcripts

The physical location of an RNA molecule informs its position in the mRNA life cycle, where immature transcripts are in the nucleus and mature in the cytoplasm. For example, single-molecule RNA-FISH (smFISH) amplifies the signal of individual RNA molecules using multiple probes or branching[31;79;80] can visualize individual transcripts with high resolution and has shown that nonsense mediated decay doesn't occur in the nucleoplasm but almost immediately upon nuclear export[81]. Visualization of individual RNA molecules has been multiplexed across many different RNA species in methods termed FISSEQ and seqFISH[41;45;46]. Together with whole body clearing such as CLARITY[40], seqFISH can spatially resolve RNA molecules even in millimeter-thick brain tissues[38;46]. Another method of increasing resolution of smFISH is "expansion microscopy," which links RNA molecules to an expandable polymer, and after expansion, individual RNA molecules can be visualized without the need for signal amplification. smFISH has also been applied simultaneously to RNA and DNA, such as in the simultaneous measurement of DNA methylation and gene expression[82]. Expanding on this work, the combination of CRISPR/Cas9 genome editing "scratchpads," with smFISH can visualize both the lineage histories of individual cells and their gene expression in a method termed MEMOIR[83]. Finally, the microscopy-free method of "spatial transcriptomics" uses location barcodes in fixed tissues to mark RNA molecules by position, then perform scRNA-seq and reconstruct the two-dimensional position of the RNA[84]. Beyond transcripton, the spatial resolution of RNA transcripts has only scratched the surface of RNA processing and many opportunities for discoveries remain.

1.3.5 Translation

The penultimate step in an mRNA's life is translation of its encoded information into proteins. Exact regulation of translation rates is critical for hematopoietic stem cells[85;86], shown by marking nascent polypeptides with fluorescent marker and measure incorporation using fluorescently-activated cell sorting (FACS). This study found that modulating the translation rate by knocking down a ribosomal subunit prevented proper hematopoietic development. Spatial analysis of the "pioneer" round of translation using TRICK[87] showed that endoplasmic reticulum proteins are translated almost immediately upon contact with the ER. Live tracking of translation in neurons using SINAPS[88] showed a translation rate of approximately 5 amino acids per second, and translation occurred throughout the neuron, including while the ribosome was transported along the axon. Insights into the conversion of RNA to protein will close the gap in understanding how the transcriptome is converted to the proteome.

1.3.6 Computational challenges and considerations

Single-cell RNA-seq is inherently noisy and requires careful consideration of experimental design[6], cell capture methods[5], library preparation and transcript barcoding[5;89], and downstream computational analysis[10].

1.4 Future technologies necessary to measure the entire life cycle of RNA

Using current technologies, tracking every step of the entire lifecycle of a single RNA transcript, from biogenesis, binding to proteins, editing, modifications, localization, translation, and degradation is not possible. It is not known what the prediction model is that takes an RNA sequence and can predict the outcome of each step of RNA processing. For example, RNA transcripts are specifically exported from the nucleus independently of transcription[90–93], and may be sequestered in one of many RNA granules[94–98] – how long is the typical residence time of an RNA molecule in these granules? How does this vary for different transcripts or cell types? A universal framework for understanding the contributions of RNA sequence to RNA processing is still needed.

Many steps of RNA processing are performed by RNA binding proteins (RBPs), but many are seemingly redundant due to their high homology such as in helicases and splicing factors, indicating a lack of understanding the specific functions of individual proteins. Beyond redundancy, does a single RBP have different activities based on cellular localization (nucleus vs cytoplasm) or transcript region (intron versus UTR)? For example, the RBFOX1/2/3 family of proteins all bind (U)GCAUG[99–103], but are expressed in different cell types, suggesting untapped complexity in understanding their activities. Recent papers have shown that the proteins have differential functions based on localization in the cytoplasm or nucleus[100] or binding partners[99], but a general rule explaining the activities of different RBPs based on their local molecular context has not been established. An experiment that profiles the localization and transcript binding preferences of related RBPs in cellular context could reveal how related RBPs have subtly different functions, or how a single RBP performs multiple functions based on its neighbors. A general mapping of RBP amino acid sequence to function and binding partners has yet to be established.

To deeply understand RNA processing events, we propose the creation of new technologies, both high-throughput to establish distributions of RNA features across a cell population, and high-resolution to demonstrate the within-cell localization and dynamics of RNA processing. Many of these technologies are extensions of existing methods which have not yet been adopted to the single cell level due to scalability. The high-throughput methods that are still only able to be performed at the bulk level tend to require high numbers of input material as the purification methods are too lossy to capture a substantial amount of molecules from every cell. For example, antibody-based methods such as m6A sequencing[104–106] or the many methods for probing protein-RNA interactions[107–110] would lose too much RNA and must be optimized to be more sensitive to RNA detection. To study individual molecules within cells, high-resolution measurements need strong signal amplification as microscopes are not generally capable of resolving single molecules. For example, tracking of RNA molecules using RNA-targeted CRISPR[111] has weak signal per molecule, but the molecules are detectable in aggregation. Thus, sensitivity of molecular detection, signal amplification, microscope resolution will be critical paths for innovation in understanding RNA processing.

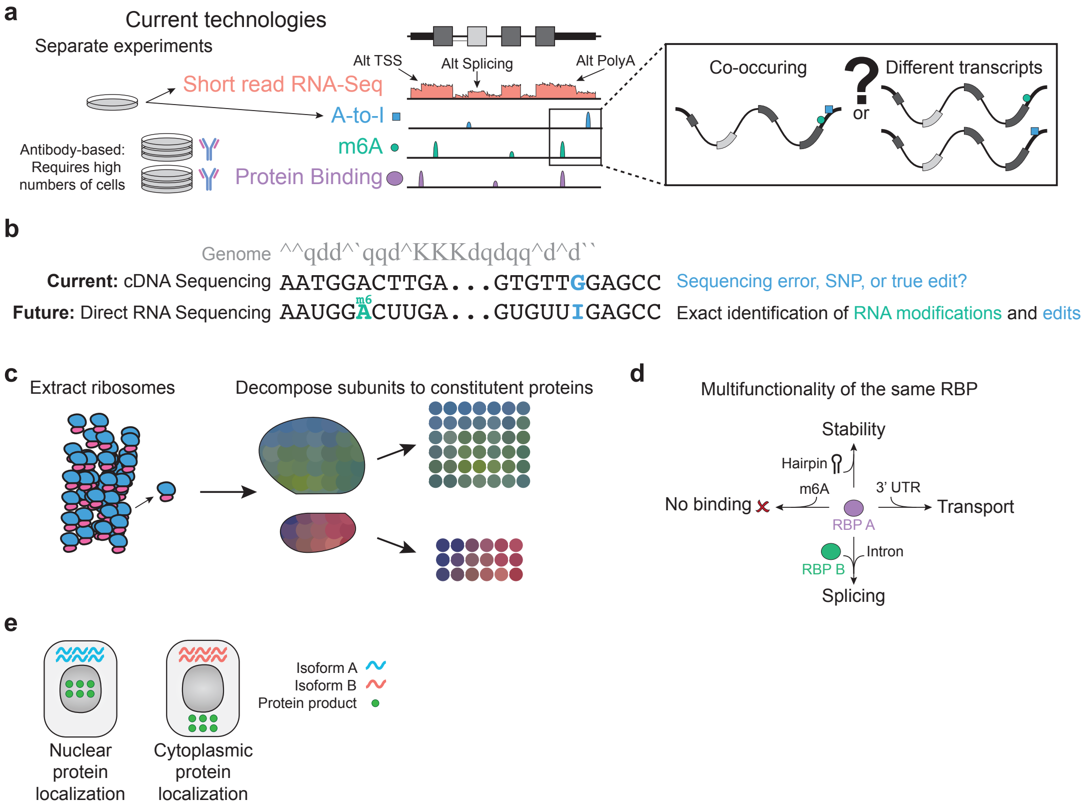


Figure 1.3: Unmet needs in RNA processing and potential future technologies. **a.** Left, example RNA transcript with RBPs (purple), A-to-I editing (blue), and m6A (green). Right, possible inclusion and exclusion isoforms with different post-transcriptional modifications of splicing, m6A, RBP binding, and m6A modification.
b. Current technologies allow for visualization of RNA abundance, A-to-I editing, m6A methylation, and protein binding, but cannot determine whether these occur on physically different or the same molecules.
c. Technologies are needed to investigate multifunctionality of the same protein, e.g. if it performs different functions based on where it binds in the transcripts. Additionally, this would be interesting to study redundancy of different proteins such as splicing factors.
d. Future direct RNA sequencing technologies would allow for direct identification of multiple RNA modifications and edits on the same transcript, unlike currently available technologies.

1.4.1 Across-cell distributions of RNA and protein with high-throughput technologies

Beyond model organisms such as *Caenorhabditis elegans*, the number of different cell types and niches in the body are unknown at the organism level[112;113]. For example, in the neurodegenerative disease amyotrophic lateral sclerosis (ALS), protein aggregation in motor neurons of the spinal cord are thought to be the main cause of disease[114;115], but there are many other supporting cells in the spinal cord such as glia and astrocytes whose functions are are unknown. As a result, researchers either remove the entire spinal cord or laser-capture micro-dissect only the motor neurons from donors[116–118]. But the spinal cord is an integrated system of many cell types – how do glia and astrocytes contribute to the progression of the disease? If instead researchers could perform single-cell capture of the entire spinal cord, and then spatially reconstruct the cell types and locations without a priori knowledge[119–122], they then could focus on analyzing the differences in RNA and protein processing within and across, cell types and individuals, rather than making inferences based on incomplete information. Such a "cell atlas" would greatly inform the understanding of disease.

Indeed, a Human Cell Atlas (HCA)[123–125] project is currently underway, where researchers are working towards establishing molecular and physical markers of cell types in the human body. In the future, it will be possible to "align" a transcriptome of interest to the HCA and obtain the closest cell types and cell locations, helping to provide a more complete understanding of human cellular biology. This will reducing the burden for researchers to painstakingly capture spatial locations of cells in the body when harvesting tissue, thus lowering the barrier for biological research and paving the way for discovery.

Mutually exclusivity and co-occurrence of RNA features

Future understanding of the interdependence of RNA features will require the measurement of multiple aspects of RNA processing at once. Current technologies to measure transcript abundance and over 120 possible nucleotide modifications[126–130], must be performed in separate, bulk experiments, and beyond correlations, the co-occurrence of these features on the same transcript is unknown (**Figure 1.3**a). One potential method for measuring multiple features at once is direct sequencing of RNA and its modified bases, without creation of a cDNA template such as by the Oxford Nanopore MinION[131–134] or Pacific Biosciences Single Molecule Real-Time (SMRT) Sequencing[135]. These technologies can directly detect RNA modifications (**Figure 1.3**b) such as m6A and inosine[129;136;137], exon structure driven by alternative splicing in transcripts[24], and help to answer the question, what percentage of cells have a modification or transcript structure? Unfortunately, these technologies are plagued with high error rates and this challenging problem of accuracy for single molecule sequencing will need to be addressed. Additionally, many of the library preparation methods for capturing entire transcripts measure only mature, poly-adenylated mRNA and not any byproducts of mRNA processing such as mirtrons or circular RNAs[138], which can be cell-type specific and for which other capture methods must be developed. Nonetheless, if the entire transcript is measured, these technologies would reveal the co-occurrence of relationships such as between alternative splicing and nucleotide modifications, shedding light on the co-dependence (or independence) of RNA processing elements.

Translation of mRNAs can vary from tissue to tissue, but has not yet been shown to vary from cell to cell. There are several methods for bulk samples to answer the question, which transcripts are actively translating?[139;140]. BacTrap and RiboTag fluorescently label ribosomal subunits and then capture the ribosomal-bound transcripts to perform full-length transcript sequencing[140]. Ribosome profiling (also called “ribo-seq” or “ribosome footprinting”) measures protected RNA fragments[141–143], and can pinpoint the paused locations of ribosomes. All of these are possible to scale to the single-cell level but will have high rRNA contamination, and at the single-cell level, every nucleotide counts, so protocols must be optimized to be sensitive only to the molecules of interest.

Beyond coding sequences, there are many disease-associated mutations that occur in non-coding regions such as introns and 5′/3′ UTRs, which are likely to influence its ability to form three-dimensional structures and can vary between cells. For example, Thus, measuring RNA structure and binding partners at the single-cell level is a critical problem, but due to technical limitations, remains unsolved. Current methods for measuring RNA secondary structure, typically by selectively measuring only single- or double-stranded RNA, require high amounts of starting material, meaning, many thousands of cells as input[144–150], thus averaging out the signal across many cells. Scaling these protocols down to single cells will be challenging, it will require tiny amounts of each reaction occurring in nanoliter volumes of captured cells. Double-stranded structure is also important in lncRNAs, as their functions range from inhibiting entire chromosomes (XIST) to sequestering RBPs (MALAT1)[151]. Combining the measurement of RNA structure with RNA modifications will inform how individual nucleotides promote or inhibit certain RNA structures. Capturing the three-dimensional structure of RNA in single cells will be a challenging problem to solve, but will pay great dividends in understanding RNA biology.

Even if there existed the perfect high-throughput method of interrogating all possible nucleotide modifications, RNA structures, and translating transcripts, it is likely the subcellular context would be lost. Need to perform follow-up experiments showing the localization and dynamics of the different transcripts. These experiments would answer questions such as, what are the time-scales of RNA modifications and translational pausing? How transient are RNA structures and how do they affect localization of the mRNA? While High-throughput experiments can create millions of data points, they are merely a starting point, creating a scaffold upon which to build knowledge of RNA processing.

RBP specialization in single cells

While high-throughput transcriptomics from single cells is possible though imperfect, high-throughput proteomics in single cells is a challenging problem that remains to be solved. It is not yet possible to perform a massive "sequencing-style" experiment on proteins as their building blocks, amino acids, do not have Watson-Crick base-pairing rules. Instead, mass cytometry[152–156] using antibodies conjugated to heavy metal isotopes has been applied primarily to immune and cancer genes, and could be applied to RBPs, specifically to ribosomes[157]. Ribosomes have been shown to be specialized to certain tissues, and have been shown to require specific subunits for proper function[86;158–162]. Ribosomal dysfunction has been implicated in a wide variety of neurological disorders such as Alzheimer's disease and Charcot-Marie-Tooth disorder, as well as viral infections such as foot-and-mouth disease[163;164]. Performing mass cytometry of ribosomal subunits could be used to investigate, what is the exact composition of ribosomal subunits of each ribosome in each cell (**Figure 1.3**c)? Follow-up experiments showing how differently composed ribosomes function differently in binding preferences, translation rates, or localization would be needed. For true discovery, it is important to have the ability to shatter all proteins and reassemble them as in sequencing using technologies such as mass spectrometry[165;166], but these have not yet been scaled to single cells and would require a dramatic reduction in cost to be accessible to researchers.

The concurrent measurement of RNA and protein would inform how RBPs specialize to different aspects of RNA processing in different contexts (**Figure 1.3**d). An intermediate technology could be to measure RNA abundance and protein levels simultaneously. Seq-well, RNA-seq coupled with antibody-based markers has applied to immune genes[32]. Seq-well could be extended to RBPs, for example, to dissect families of related proteins, such as ribosomal subunits or splicing factors, relative to transcript abundance. However, a multiplexed approach measuring the sequence specificities of proteins in addition to the RNA abundance would be more informative.

For example, heart, brain and muscle all express members of the homologous RBFOX family at different levels[99–103]. RBFOX proteins have been implicated in splicing and in transcript stability[167;168], in some cases dependent on the localization of the protein[100]. Understanding exactly which of RBFOX1/2/3 bind at different locations in the transcript would inform how these highly similar proteins have very specialized functions. There are existing methods for measuring RNA-protein interactions is cross linking and immunoprecipitation (CLIP)-Seq, which now has many variants (eCLIP/iCLIP/irCLIP)[107–110] However, few RBP-RNA interaction sites exist relative to the total amount of RNA and due to the inefficiency of antibodies, these techniques require high numbers of cells as input to be able to detect even a few interaction sites. Scaling *CLIP-Seq to single cells will require substantial optimization of the protocol. Ideally, detecting several RBPs, their RBP-RNA interaction sites and the entire transcript could be measured, to provide insight into how homologous RBPs differentially bind in different cellular contexts and ultimately create a wide variety of distinct phenotypes.

1.4.2 Molecular organization in space and time

How do different RNA isoforms influence the protein product? To capture the RNA-protein transformation, need technologies that combine RNA and protein measurements. For example, coupling of RNA-FISH with antibody-based immunofluorescence of the protein product would allow for visualization of differential isoform usage and protein localization. For example, if one isoform appears in cells where the protein is cytoplasmic while the other appears with a nuclear protein localization (**Figure 1.3**e), this suggests that the different mRNA isoforms influence the localization signals of the protein. However, this only examines cells fixed in one state, and can only garner correlations but not causes – need perturbations and/or dynamics to visualize how isoforms are differentially processed. For example, do the isoforms have different transcription or splicing kinetics? How do their differential sequence, modifications, editing, binding partners, and structure, contribute to localization and translation?

Moonshot target: Live cell imaging of both dynamics and spatial organization of RNA processing

Ideally, one could measure the phenotype of a cell, then perturb that same cell and observe the result. However, most single-cell technologies are destructive – once you sample the cell, you can't put it back and see how it responds in a new situation. Live cell imaging of RNA and protein enables encoding information of both localization and time scales (reviewed by Buxbaum et al.[37]). However, imaging-based methods are low throughput as they cannot yet resolve individual molecules without significant signal amplification, or discover novel RNA molecules. Nonetheless, as smFISH began by measuring a single RNA and can now be multiplexed to visualize hundreds of molecules[46], live cell imaging has started with a few RNAs at a time and will eventually expand to visualize poly-adenylation sites, editing, nucleotide modifications, RNA structure, and binding partners such as DNA, RNA, protein, and metabolites.

Just as transcription factors come together in pulses[169–178], RBPs that facilitate transcription and perform co-transcriptional tasks must also be aggregated in pulses, and are especially important for the assembly and disassembly of membrane-free organelles used in RNA storage and splicing such as paraspeckles, stress granules and the spliceosome. How does the cell "know" it is time to transcribe and "tell" the molecular components to convene in one place? How exactly do these molecules come together? Current methods for live cell imaging of transcripts are difficult because they require design of RNAs containing bulky hairpins, then adding the hairpin coat protein, creating large structure on the RNA molecule and possibly disrupting normal RNA processing[73;179;180].

Technologies to measure the assembly of these organelles in live cells will be necessary, labeling both RNA and protein with fluorophores. RNA could be labeled in cellulo fluorescently without adding much molecular weight[181], with molecular beacons[182], or with novel RNA-targeted CRISPR technologies[111], provided the technique scaled to the single molecule level. The biggest challenges in optimizing the protocol to single molecules is signal amplification and multiplexing across multiple RNA molecules, but techniques from in situ sequencing could be applied to increase the number of fluorophores and increase the number of different transcripts measured. Given the ultimate system that images single

RNA molecules in real time, would allow for real-time visualization of an RNA molecule's life.

"Goldilocks" balance: Single-cell multiomics

Previously, it was thought that there are certain unalienable "uncertainty principles" in biology, such that one could not know both both the genotype and phenotype of a living cell[183] or both the cellular "position" (current cell state) and "momentum" (a cell's past or future, i.e. its lineage or differentiation trajectories)[184]. However, recent work in single-cell multiomics has turned these principles on their head [reviewed in Macaulay et al.[33]]. Both the genotype and phenotype can be measured from a single cell by capturing both DNA and RNA[49;185], and even coupling with measuring epigenomics and RNA[47;48]. A cell's "position" and "momentum" can be inferred through algorithms that delineate cellular trajectories from phenotypic measurements such as RNA-seq, reviewed thoroughly by Cannoodt et al.[7]. While simultaneous capture of the (epi)genome and transcriptome have been major breakthroughs, there are many more aspects of cellular state that are still invisible to the sequencing eye. We expect more technologies to upend traditional thinking of what is possible at the single cell level.

The ability to simultaneously measure RNA and DNA has not yet been applied to questions of RNA processing, and there are many "low-hanging fruit" opportunities to study the question, are RNA features marked (epi)genomically in the DNA? With the simultaneous measurement of both the (epi)genetic context and transcriptome, researchers can observe mutual exclusivity/co-occurrence of DNA features and RNA processing, as with within RNA features (**Figure 1.3**a). For example, there is evidence that alternative splicing regulation is influenced by polymerase speed, GC content and epigenetic marks[63;186;187]. These simultaneous measurements would help to to answer the questions, how do chromatin modifications[188;189], four-dimensional genome structure[190], and single nucleotide polymorphisms influence alternative splicing? In a perfect world, to study the interplay between DNA and alternative splicing, one could observe genomic features, transcription speed, and alternative splicing simultaneously, for all transcripts. Alternative splicing is but one facet of RNA processing, and these multiomics measurements would help to answer, Is an RNA's fate encoded in the (epi)genome?

Another method to create additional context for each individual cell is to combine high-throughput measurements with genome editing such as with CRISPR/Cas9, allowing for dissection of complex phenotypes in mammalian cells at large scales. For example, Perturb-seq is a method that combines knockdown of genes using CRISPRi with single-cell RNA-seq, and was used to study the unfolded protein response[191] and effect of lipopolysaccharides on dendritic cells[192]. This created a computational scientist's dream dataset, as for each gene that was knocked down, there was a control dataset, and thus for developing algorithms, one could always have a negative control to check with. Perturb-Seq could be applied to study any aspect of RNA processing, e.g. systematically knocking down ribosomal subunits or splicing factors, enabling fine-tuned dissection of high-level regulatory units.

So far, the technologies we have discussed have focused on observing a cell's present, but recent developments have enabled encoding of a cell's history or lineage in its genome using CRISPR/Cas9[83;193]. Coupling phenotypic measurements such as RNA-Seq or smFISH with lineage tracing, would allow for comparison of present cell state while considering cellular time. Using phylogenetic techniques, cell lineages could be reconstructed and even the times at which cells asymmetrically divided to change fates could be found. If RNA-seq encodes a cell's present, then its traced lineage encodes its past. This lineage tracing method, coupled with direct RNA sequencing, would help to understand how developmentally regulated RNA processing events such as RNA editing, m6A, and alternative splicing, are finely tuned in different lineages. Do all cells that were committed to a particular lineage also have certain RNA processing events? This could indicate inheritability of the event, either encoded through the genome or by asymmetrically dividing the RNA content of a mother cell. The simultaneous measurement of a cell's past and present will illuminate a deeper understanding of cellular processing and will enable the computational prediction of cellular futures.

1.5 Conclusions

Each RNA molecule lives a rich, fulfilled life, and while advances single-cell technologies have greatly expanded our understanding of RNA processing, many questions remain. How do the many transient aspects of RNA, such as nucleotide modifications, binding partners, 3D structure, and localization, affect each other? What is the effect of non-RNA, such as DNA, protein, or metabolites, on RNA? Where in the cell does an RNA molecule travel? What does it interact with on the way? Finally, when a feature is variable, is it noise or is it functional? Maybe the noise itself is functional[69–71]? In conclusion, there are ample opportunities for application of current and future technologies to understanding the entirety of an RNA's existence, and ultimately the molecular processes that drive diversity in cell types across human life.

1.6 Acknowledgements

Chapter 1, in full, is currently being prepared for submission for publication of the material. Botvinnik, Olga; Song, Yan; Yeo, Gene W. The dissertation author was the primary investigator and author of this material.

Chapter 2

The *Expedition* software suite: Computational tools for transcriptome analysis

In this paper, we developed the *Expedition* suite, consisting of software packages that addressed three key deficiencies in single-cell alternative splicing analysis:

1. **Detect and quantify alternative splicing quickly, with minimum false positives: `outrigger`, Section 2.1**
   In single-cell analysis, absolute quantitation of gene expression or "percent spliced-in" (Psi/Ψ) is important and enable us to learn the distribution of these quantitations. Previously, relative quantitation for splicing ($\Delta\Psi$) is more commonly used to calculate the difference between groups. Such relative quantitation tolerates false positive better, as false positives may not vary between groups, $\Delta\Psi \sim 0$ and are thus not noticeable in pairwise comparisons. However, when studying distribution of absolute quantitation,

such false positives obscure the observation in unpredictable way and hinder biological interpretation. The second main problem of previous splicing algorithm is the inflexible definitions of alternative exons. The same alternative exons may utilize different flanking exons in different cells/samples, thus leading to different biological interpretation. To address these problems, we create `outrigger`, which uses junction reads to find de novo exons, creates a splice graph to define junction-based alternative events, filters for conserved splice sites, and strictly rejectes cases of alternative events incompatible with the data at hand. Finally, we discuss and compare to the popular MISO[194] algorithm.

2. **Classify modalities of alternative splicing events, including bimodal: `anchor`, Section 2.2**

   The power of single-cell analyses rises from the ability to study the distribution of a parameter-of-interest. There are a few statistical methods for finding bimodal distributions, but none are sufficient because they are either not sensitive enough, or not robust enough to noise. Additionally, these methods only deal with bimodal distribution and do not classify other distributions, such as unimodal or multimodal. To create a sensitive distribution classifier for all modalities, we used Bayesian methods to create `anchor`, and compare our method to a simple binning method, the bimodality index[195], and the bimodal dip test[196].

3. **Quantify and visualize dynamics in distributions: `bonvoyage`, Section 2.3**

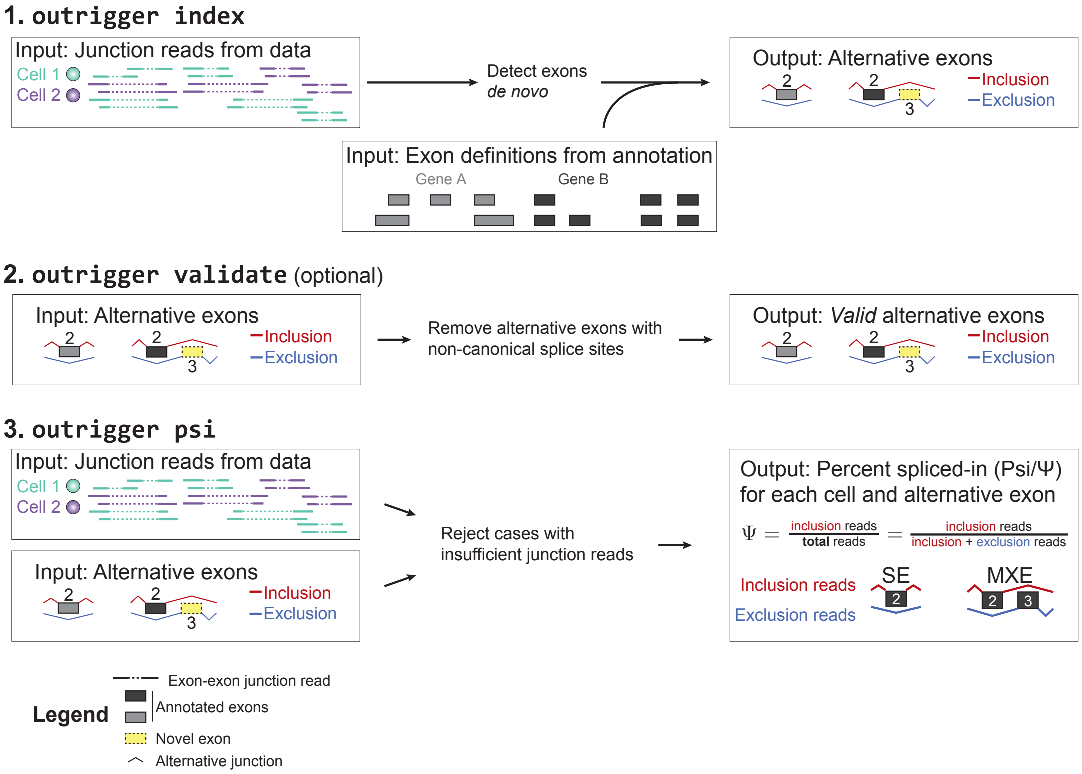
   While there are many statistical tests to compare changes in distributions, few of them is coupled with visual tools to present changes in distribution with both magnitude and direction. For the specific question of alternative splicing changes, we are interested in observing a event becomes more included or more excluded. Thus we have employed machine learning methods to create a visualizable, interpretable 2d space with "included" and "excluded" axes. This method is compared to the quantification offered by the Jensen-Shannon Divergence (JSD)[197].

2.1 `outrigger`: Splicing estimation with *de novo* annotation and graph traversal

Currently available tools for AS detection and quanitification have two major problems: (1) inflexible definitions that cannot handle different configurations of flanking exons for the same alternative junctions, and (2) lack of rejection of an alternative event even if its definition is incompatible with the data-at-hand. The first problem is solved with `outrigger index`, which defines all potential alternative events based on the junctions and alternative exons from the aggregate of entire sample sets in a given project, and enumerates all biologically possible flanking exon combinations. This step maximize the likelihood to identify all possible alternative events. To ensure only valid alternative events were generated, we added `outrigger validate` to remove alternative events with introns lacking conserved splice sites. The second prolbem is solved with `outrigger psi`, which applies strict rules to only permit junctions with sufficient coverage for an event in a given sample. All the parameters in the rules can be user-defined. Thus, outrigger addresses key issues with current alternative splicing software.

2.1.1 Algorithm overview

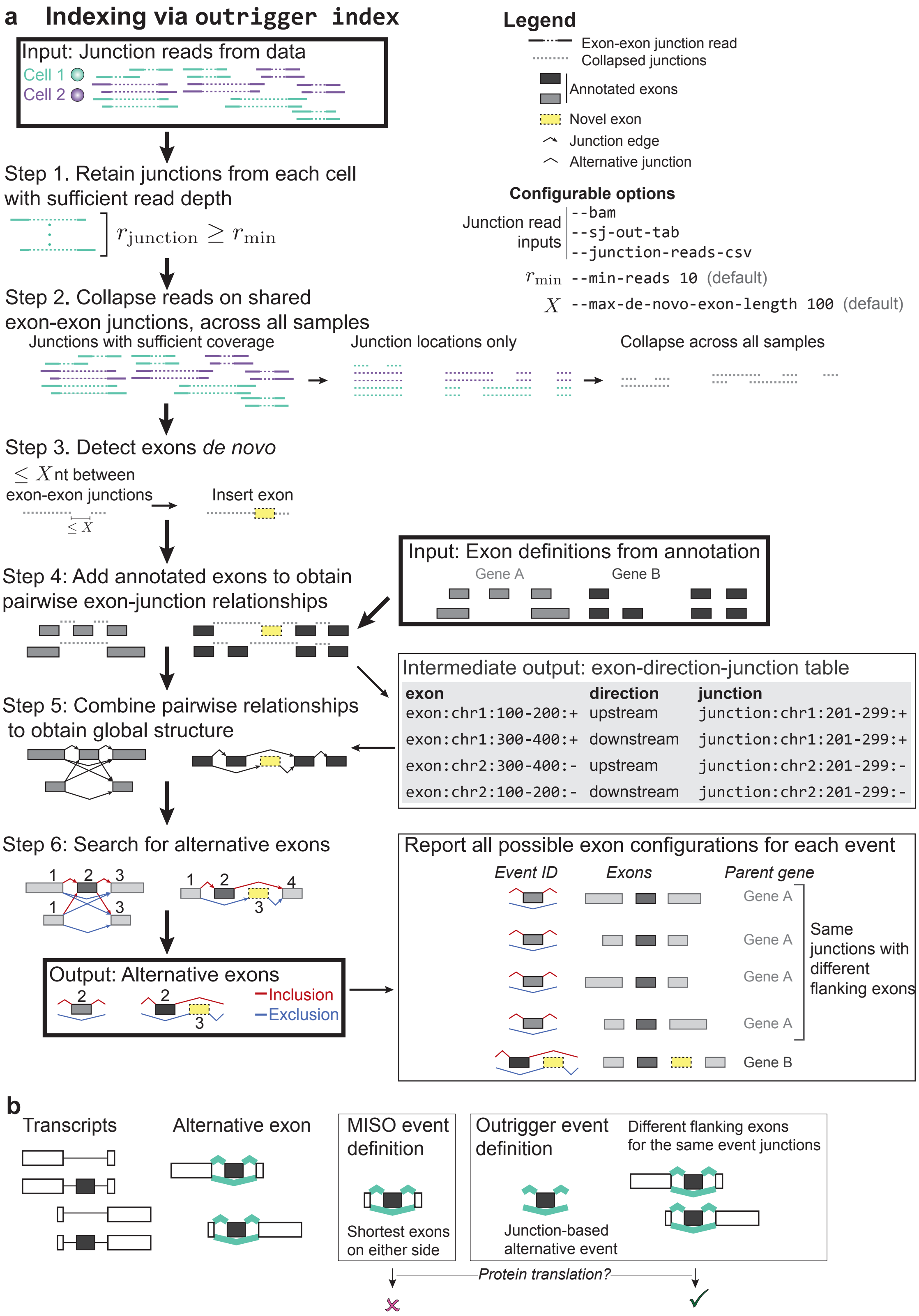
Broadly, the goal of `outrigger` is to create a custom, *de novo* alternative splicing annotation by using junction reads and exon definitions to create a exon-junction graph, traversing the graph to find alternative events, and calculate percent spliced-in (Psi/Ψ) of the alternative exons.



1. outrigger index
Input: Junction reads from data
Cell 1
Cell 2
Detect exons
de novo
Output: Alternative exons
2
2
3
Inclusion
Exclusion
Input: Exon definitions from annotation
Gene A
Gene B
2. outrigger validate (optional)
Input: Alternative exons
Remove alternative exons with non-canonical splice sites
Output: Valid alternative exons
3. outrigger psi
Input: Junction reads from data
Input: Alternative exons
Reject cases with insufficient junction reads
Output: Percent spliced-in (Psi/Ψ) for each cell and alternative exon
Ψ = inclusion reads / total reads = inclusion reads / inclusion + exclusion reads
Inclusion reads
Exclusion reads
SE
MXE
Legend
Exon-exon junction read
Annotated exons
Novel exon
Alternative junction

Figure 2.1: Overview of `outrigger`'s three steps and associated commands: indexing (`outrigger index`), validation (`outrigger validate`) and percent spliced-in (Psi/Ψ) calculation (`outrigger psi`). In the first step of building an index, `outrigger` considers the entirety of junction reads from the user-input dataset to detect exons *de novo*, adds annotated exons, then searches for alternative exons. In the second, optional, step of validating the detected events, `outrigger` removes alternative exons with flanking introns lacking consensus splice sites. For the third step of calculating Psi/Ψ, `outrigger` utilizes junction reads together with alternative exons defined in the indexing step and calculates Ψ for each sufficiently covered event. Only junction reads are used to represent inclusion or exclusion reads. SE, Skipped Exon; MXE, Mutually Exclusive Exons.

Figure 2.2 *(next page)*: Internal steps of indexing via `outrigger index`: Exons identification and defining alternative events.
a. Internal workings of the indexing step via `outrigger index`. User-provided inputs junction reads can be either genome-aligned `.bam` files, the `.SJ.out.tab` splice junction files from the STAR aligner, or a compiled table in `.csv` of all junction reads from all samples for the project. Step 1, only junction reads with sufficient depth in a cell/sample are retained. By default, the minimum number of reads is 10 per cell/sample, which can be modified with the flag `--min-reads`. Step 2, junction reads are used to identify junction locations, and reads are aggregated across all cells/samples regardless of which cell/sample it came from. Step 3, if there is a "gap" between two junctions that is smaller than certain length X (by default, $X = 100$ nucleotides but can be modified with the flag `--max-de-novo-exon-length`), then an exon is inserted. Step 4, the identified exons are compared with the annotated exons to obtain the pairwise relationships between exons and junctions. Step 4 outputs a table of "triples:" of `(exon, direction, junction)` encoding the directional relationship between exons and junctions. Step 5, the output tables from step 4 are utilized to connect exons through junctions and creates a graph database. Finally, in Step 6, alternative exons are identified by traversing the graph database. The output of the indexing step run by the command `outrigger index`, is junction-based, outputting the alternative exon and all possible configurations of flanking exons for each event. For example, on the bottom right, the same skipped exon event using the same alternative junctions, have four possible configurations of flanking exons. They are considered to be the same event, but are reported with all four configurations for the ease-to-use in downstream analysis.
b. Defining alternative events and comparison of biological interpretability of events found by MISO and `outrigger`. For a given alternative exon (black box), there can be multiple transcripts corresponding to the alternative exon but with different flanking exons. MISO chooses to define the alternative event using the shortest exons on both sides. Yet, this MISO-defined alternative event may not actually exist as a transcript in the dataset and will be misleading to interpret. For example, attempts to translate such non-existing transcript(s) will be inappropriate. In contrast, `outrigger` defines the event based on the junctions, and outputs all corresponding flanking exon configurations, thus enabling broader use of the outputs and more relevant biological interpretation.


a
Indexing via outrigger index
Input: Junction reads from data
Cell 1
Cell 2
Step 1. Retain junctions from each cell with sufficient read depth
$r_{\text{junction}} \geq r_{\min}$
Step 2. Collapse reads on shared exon-exon junctions, across all samples
Junctions with sufficient coverage
Junction locations only
Collapse across all samples
Step 3. Detect exons *de novo*
$\leq X$ nt between exon-exon junctions
$\leq X$
Insert exon
Step 4: Add annotated exons to obtain pairwise exon-junction relationships
Input: Exon definitions from annotation
Gene A
Gene B
Step 5: Combine pairwise relationships to obtain global structure
Step 6: Search for alternative exons
1
2
3
1
3
1
2
4
3
Output: Alternative exons
2
2
3
Inclusion
Exclusion
Legend
Exon-exon junction read
Collapsed junctions
Annotated exons
Novel exon
Junction edge
Alternative junction
Configurable options
Junction read inputs
$r_{\min}$
X
Intermediate output: exon-direction-junction table
Report all possible exon configurations for each event
Event ID
Exons
Parent gene
Gene A
Gene A
Gene A
Gene A
Same junctions with different flanking exons
Gene B
b
Transcripts
Alternative exon
MISO event definition
Shortest exons on either side
Outrigger event definition
Junction-based alternative event
Different flanking exons for the same event junctions
Protein translation?
✗
✓

```
--bam
--sj-out-tab
--junction-reads-csv
--min-reads 10 (default)
--max-de-novo-exon-length 100 (default)
```

exon	direction	junction
exon:chr1:100-200:+	upstream	junction:chr1:201-299:+
exon:chr1:300-400:+	downstream	junction:chr1:201-299:+
exon:chr2:300-400:-	upstream	junction:chr2:201-299:-
exon:chr2:100-200:-	downstream	junction:chr2:201-299:-

`outrigger index`: Create custom alternative splicing annotation. The following is a narrative describing **Figure 2.2a**.

Inputs. Two inputs are required for `outrigger index`: junction counts and gene annotations. The junction counts can be provided in many forms: either `.bam`[198] genome alignment files, splice junction count `.SJ.out.tab` files created by the STAR aligner[199], or a pre-compiled table of samples' junction reads in a `.csv` format. The gene annotations can be provided in `.gtf` or `.gff` format.

Step 1: Retain junctions from each cell with sufficient read depth. Junctions with reads in an individual sample less than the minimum number of reads, r_{min} are removed. By default, $r_{\text{min}} = 10$, and can be adjusted by the user, for example to a minimum of 88 reads, with `--min-reads 88` on the command line. To illustrate, if one junction is observed with two (2) reads in 100 samples, although there were a total of 200 reads observed on the junction, it will be discarded at this step. Because, there is not sufficient evidence to suggest that this junction is well-covered in any sample.

Step 2: Collapse reads on shared exon-exon junctions, across all samples. The aggregate of all junctions from all samples in a given project are create to maximize the likelihood of identifing all potential alternative events.

Step 3: Detect exons de novo. If the gap between two junctions is under X nucleotides, an exon will be inserted at the gap. This maximum X is necessary, because otherwise we could insert "exons" that are many kilobases long, but aren't true exons -- they are the intergeneic space between genes. By default, $X = 100$, and this can be adjusted by the user, for example to 157 nucleotides, with the command line flag, `--max-de-novo-exon-length 157`.

Step 4: Integrate exon annotation to obtain pairwise exon-junction relationships. Annotated exons are integrated with the *de novo* exons and create a table of the pairwise relationships of each exon to each junction. We do this by creating a database of genes, transcripts, and exons from a GTF gene annotation file using `gffutils`[200], and observing which junctions are adjacent to each exon. This outputs an *"exon-direction-junction"* table which is used in Step 5.

Step 5: Combine pairwise relationships to obtain global structure. We then use the adjacencies to build a directional graph which connects exons to each other via junctions. This graph database was built using `graphlite`[201], a Python program that provides a lightweight graph wrapper over SQLite.

Step 6: Search for alternative exons. To find alternative events, all exons in the graph database were transversed to test, if starting from that exon, it could be a first exon of an skipped exon (SE) or mutually exclusive exon (MXE) event.

Outputs. The output of `outrigger index` is a folder containing the following. The `events.csv` file contains the event definitions will be used by `outrigger psi`. The `exonN.bed` files, where `N` is an exon number, will be used by `outrigger validate` to check for canonical or non-canonical splice sites.

The splicing event definitions in the `events.csv` files are specified by the junctions and the alternative exon. As there may be multiple potential flanking exons with the same junctions, rather than choosing a single version (as is done by MISO, **Figure 2.2b**), we output all possible flanking exon configurations. Thus, while the critical alternative exons are exon 2 for SE events and exons 2 and 3

```
outrigger_output/
├── index
│   ├── gtf ..................................................... Added by Step 3
│   │   ├── gencode.vM10.annotation.gtf ......................... Added by Step 4
│   │   ├── gencode.vM10.annotation.gtf.db ...................... Added by Step 4
│   │   └── novel_exons.gtf ..................................... Added by Step 3
│   ├── exon_direction_junction.csv ............................. Added by Step 4
│   ├── mxe ..................................................... Added by Step 6
│   │   ├── event.bed ........................................... Added by Step 6
│   │   ├── events.csv .......................................... Added by Step 6
│   │   ├── exon1.bed ........................................... Added by Step 6
│   │   ├── exon2.bed ........................................... Added by Step 6
│   │   ├── exon3.bed ........................................... Added by Step 6
│   │   ├── exon4.bed ........................................... Added by Step 6
│   │   └── intron.bed .......................................... Added by Step 6
│   └── se ...................................................... Added by Step 6
│       ├── event.bed ........................................... Added by Step 6
│       ├── events.csv .......................................... Added by Step 6
│       ├── exon1.bed ........................................... Added by Step 6
│       ├── exon2.bed ........................................... Added by Step 6
│       ├── exon3.bed ........................................... Added by Step 6
│       └── intron.bed .......................................... Added by Step 6
└── junctions ................................................... Added by Step 1
    ├── metadata.csv ............................................ Added by Step 2
    └── reads.csv ............................................... Added by Step 1
```

Figure 2.3: Example output of `outrigger index` command.

for MXE events, we show all possible exon flanking exon 1s and exon 3s for SE, and all possible flanking exon 1s and exon 4s for MXE events (**Figure 2.2a**, lower right).

Below is an example command using `outrigger index`:

```
outrigger index --bam *sorted.bam \
    --gtf gencode.vM10.annotation.gtf
```

This creates a folder called `outrigger_output` with the following contents:

Besides outputting the relevant `events.csv` which is used in `outrigger psi` to define events, we also output `.bed` files for the entire event, the alternative intron, and each exon, facilitating downstream sequence analysis.

`outrigger validate`: Remove alterantive splicing lacking conserved splice sites. The following describes the biological intuition behind **Figure 2.5a**. Major (U2) splicesome recognize splice-sites as (5′ end of intron/3′ end of intron) GT/AG and GC/AG the Minor (U12) spliceosome recognizes splice-sites as AT/AC[202;203]. By default, these combinations of splice-sties are allowed. But the valid splice sites can be user-specified and changed for example to AA/AA and GG/GG with --valid-splice-sites AA/AA,GG/GG.

The output of outrigger validate is a splice_sites.csv folder containing the splice sites, and an additional folder in the splice type folder, called validated, containing filtered events.csv which only contain alternative events with valid splice sites. For example, as a follow up on our previous outrigger index command, we validate the alternative exons with the command,

```
outrigger validate -{}-genome mm10 \
    -{}-fasta GRCm38.primary_assembly.genome.fa
```

This creates the following additions to the outrigger_output folder:

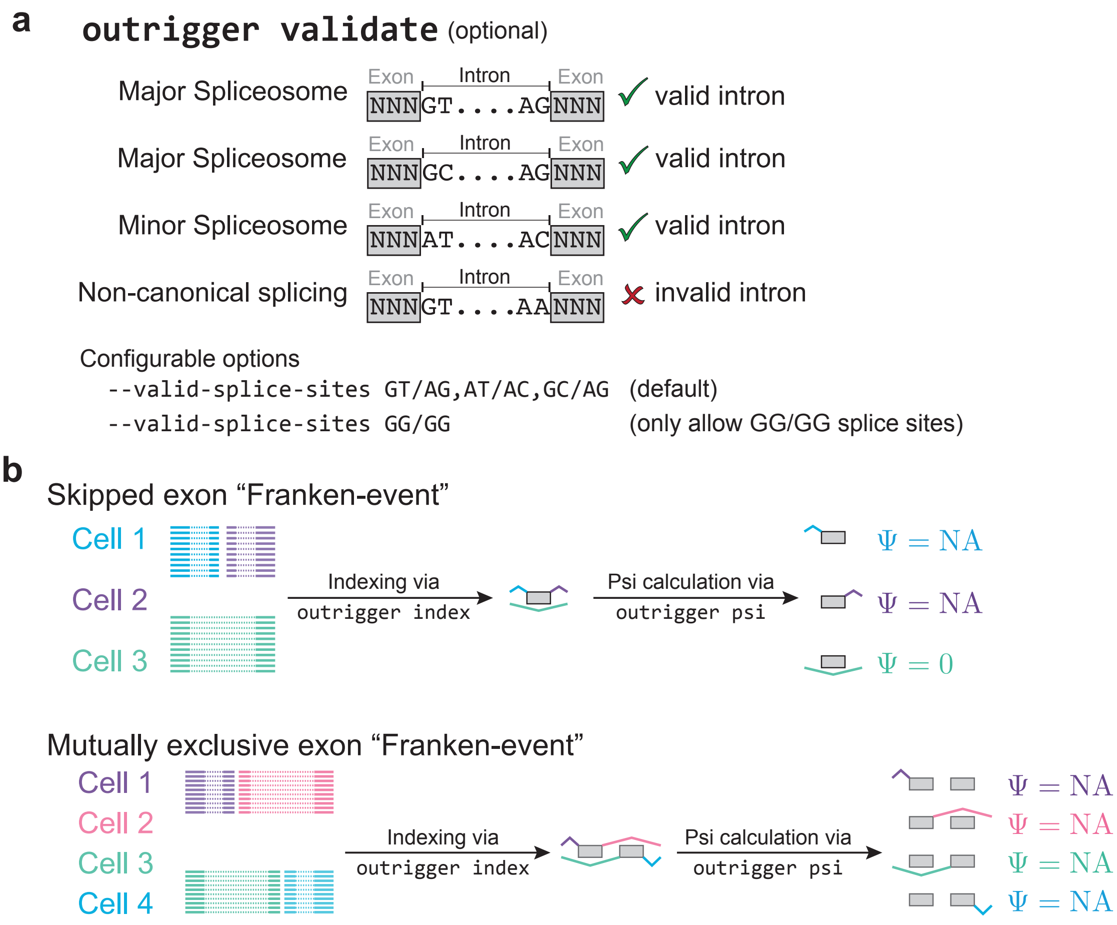
Potential "Franken-events" created by combining junctions over multiple datasets. As many junctions may occur spuriously in a single cell (sample), aggregating all junctions across all cells (sample) may create events that were not observed in any individual cell (**Figure 2.5**b). We wanted to ensure we strictly defined when events were valid or not in these cases.

In the case of SE events, the exon will have Ψ = NA for the cell with the observed inclusion junctions, since they don't have sufficient reads on both sides of the exon. For the cell with the exclusion junction, it will have $\Psi = 0$ since no inclusion reads were observed.

For MXE events, if each of the four junctions was observed independently in a different cell, then all of the cells will have Ψ = NA for that splicing event since there are no cells which have sufficient reads on all junctions of either isoform.

```
outrigger_output/
├── index
│   ├── gtf
│   │   ├── gencode.vM10.annotation.gtf
│   │   ├── gencode.vM10.annotation.gtf.db
│   │   └── novel_exons.gtf
│   ├── exon_direction_junction.csv
│   ├── mxe
│   │   ├── event.bed
│   │   ├── events.csv
│   │   ├── exon1.bed
│   │   ├── exon2.bed
│   │   ├── exon3.bed
│   │   ├── exon4.bed
│   │   ├── intron.bed
│   │   ├── splice_sites.csv ............................ Added by outrigger validate
│   │   └── validated.................................... Added by outrigger validate
│   │       └── events.csv............................... Added by outrigger validate
│   └── se
│       ├── event.bed
│       ├── events.csv
│       ├── exon1.bed
│       ├── exon2.bed
│       ├── exon3.bed
│       ├── intron.bed
│       ├── splice_sites.csv ............................ Added by outrigger validate
│       └── validated.................................... Added by outrigger validate
│           └── events.csv............................... Added by outrigger validate
└── junctions
    ├── metadata.csv
    └── reads.csv
```

Figure 2.4: Example output of `outrigger validate` command.



```
--valid-splice-sites GT/AG,AT/AC,GC/AG  (default)
--valid-splice-sites GG/GG              (only allow GG/GG splice sites)
```

Figure 2.5: `outrigger` validation and pathological cases.
a. Validation via `outrigger validate`: Removal of alternative events with introns lacking consensus splice sites. In this optional step, exons with flanking introns lacking known splice site motifs are removed. This is configurable. By default, the valid splice sites are specified as, `--valid-splice-sites GT/AG,GC/AG,AT/AC`, but can be any pair of two nucleotides.
b. Possible pathological cases of `outrigger`. These "Franken-events" consist of junctions that were observed in independent samples. At the indexing step, aggregated reads from multiple cells/samples are considered to construct an index of all junctions to maximize the number of AS events. Yet, at the Psi/Ψ calculation step, in each individual cell/sample, insufficient reads may be observed for certain junction resulting in Ψ = NA in some cells/samples for the same event. Top, skipped exons, if each junction is observed only in one cell, the cell with the exclusion junction is assigned a Ψ = 0 while the remaining cells are assigned as Ψ = NA. Bottom, mutually exclusive exons, Ψ = NA for all 4 cells, as there is insufficient evidence of exon inclusion or exclusion in any one cell. Thus, the number of detected events output by `outrigger index` can greatly overestimate the number of valid events in the dataset found by `outrigger psi`.

`outrigger psi`: Calculate percent spliced-in of alternative exons To calculate percent spliced-in (Psi/Ψ) of a potentially alternative exon identified in `outrigger index`, we use the equation for $\Psi = \frac{\text{inclusion reads}}{\text{total reads}}$ 23, with substantial checks for whether the event is valid (**Figure 2.6**). For SE, there is only one exclusion junction and thus the the exclusion junction is weighted by two to compensate (Eq. Equation (2.1)). For MXE, the calcluation is simply the inclusion reads divided by the total reads (Eq. Equation (2.2)). The junction reads between exon i and exon j are presented as $r_{i,j}$, displaying inclusion reads in red and exclusion reads in blue.

SE Ψ

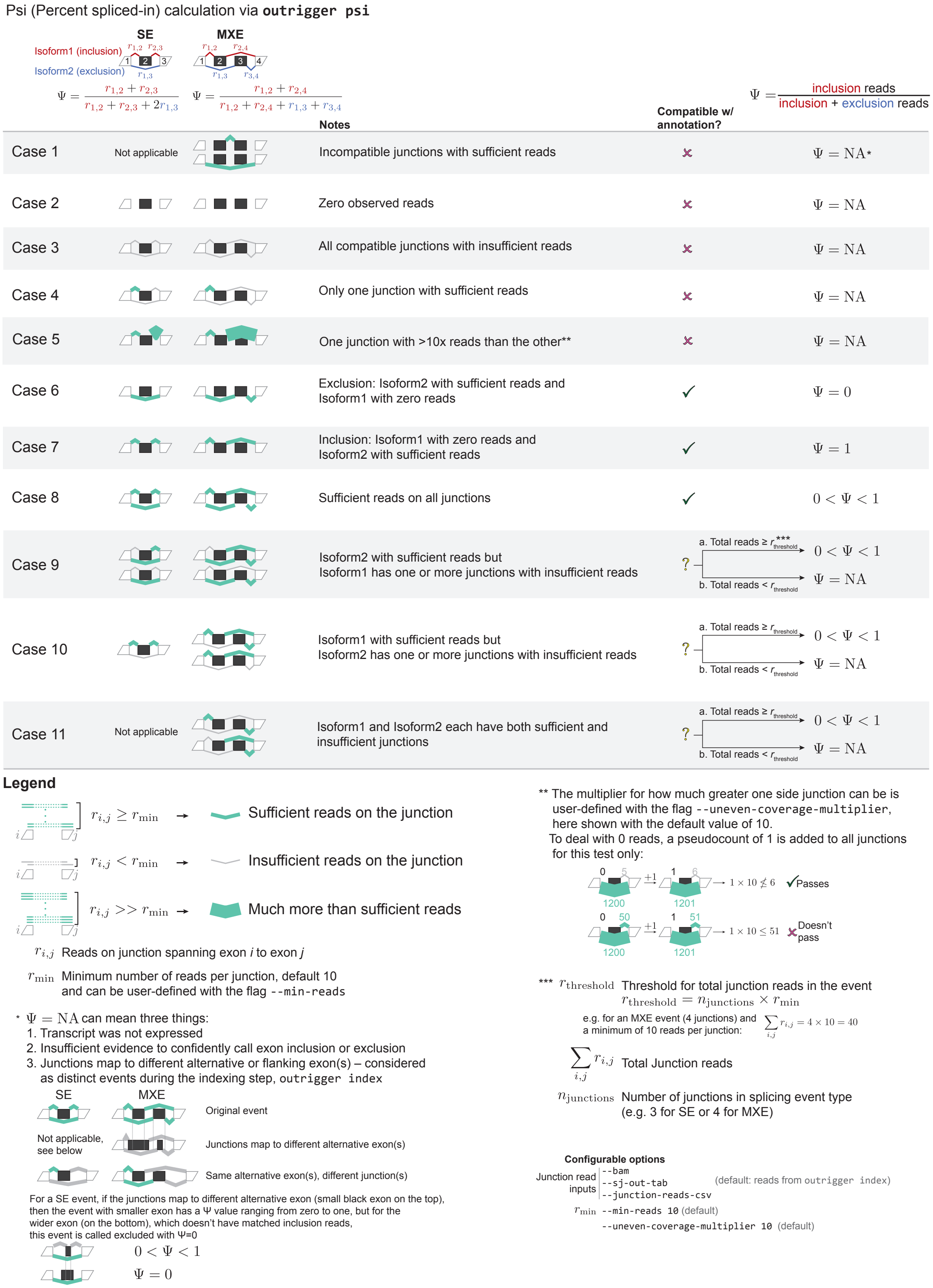
$$\Psi = \frac{r_{1,2} + r_{2,3}}{r_{1,2} + r_{2,3} + 2r_{1,3}} \tag{2.1}$$

MXE Ψ

$$\Psi = \frac{r_{1,2} + r_{2,4}}{r_{1,2} + r_{2,4} + r_{1,3} + r_{3,4}} \tag{2.2}$$

Multiple validation steps were incorporated to ensure that the junction reads observed in each sample are consistent with the type of splicing event annotated by `outrigger`. This process is described in **Supplementary Software . Figure 2.6**.

Figure 2.6 *(next page)*: Cases created by percent spliced-in calculation via the command `outrigger psi`.
The table describes the 11-step sequential logic of `outrigger` to reject an event in a cell/sample based on that cell/sample's junction reads. If an event reaches a Ψ = NA case, then it is rejected from that sample, otherwise, it continues through the cases. If the event is rejected, then it is assigned Ψ = NA, if it is not rejected, then it gets a $0 \leq \Psi \leq 1$ value based on the junction reads.
Strict evaluation of percent spliced-in (Psi/Ψ). To compute the percent spliced-in (Psi/Ψ) of skipped exon (SE) and mutually exclusive exons (MXE) alternative events during the execution of the command `outrigger psi`, we use $\Psi = \frac{\text{inclusion reads}}{\text{total reads}}$. We represent the number of reads spanning the junction between exon_i and exon_j as $r_{i,j}$.



Psi (Percent spliced-in) calculation via **`outrigger psi`**

$$\Psi = \frac{r_{1,2} + r_{2,3}}{r_{1,2} + r_{2,3} + 2r_{1,3}} \qquad \Psi = \frac{r_{1,2} + r_{2,4}}{r_{1,2} + r_{2,4} + r_{1,3} + r_{3,4}}$$

$$\Psi = \frac{\text{inclusion reads}}{\text{inclusion + exclusion reads}}$$

	SE	MXE	Notes	Compatible w/ annotation?	
Case 1	Not applicable		Incompatible junctions with sufficient reads	✗	$\Psi = \text{NA}*$
Case 2			Zero observed reads	✗	$\Psi = \text{NA}$
Case 3			All compatible junctions with insufficient reads	✗	$\Psi = \text{NA}$
Case 4			Only one junction with sufficient reads	✗	$\Psi = \text{NA}$
Case 5			One junction with >10x reads than the other**	✗	$\Psi = \text{NA}$
Case 6			Exclusion: Isoform2 with sufficient reads and Isoform1 with zero reads	✓	$\Psi = 0$
Case 7			Inclusion: Isoform1 with zero reads and Isoform2 with sufficient reads	✓	$\Psi = 1$
Case 8			Sufficient reads on all junctions	✓	$0 < \Psi < 1$
Case 9			Isoform2 with sufficient reads but Isoform1 has one or more junctions with insufficient reads	? a. Total reads ≥ $r_{\text{threshold}}$*** b. Total reads < $r_{\text{threshold}}$	$0 < \Psi < 1$ $\Psi = \text{NA}$
Case 10			Isoform1 with sufficient reads but Isoform2 has one or more junctions with insufficient reads	? a. Total reads ≥ $r_{\text{threshold}}$ b. Total reads < $r_{\text{threshold}}$	$0 < \Psi < 1$ $\Psi = \text{NA}$
Case 11	Not applicable		Isoform1 and Isoform2 each have both sufficient and insufficient junctions	? a. Total reads ≥ $r_{\text{threshold}}$ b. Total reads < $r_{\text{threshold}}$	$0 < \Psi < 1$ $\Psi = \text{NA}$

Legend

$r_{i,j} \geq r_{\text{min}}$ → Sufficient reads on the junction

$r_{i,j} < r_{\text{min}}$ → Insufficient reads on the junction

$r_{i,j} >> r_{\text{min}}$ → Much more than sufficient reads

$r_{i,j}$ Reads on junction spanning exon *i* to exon *j*

r_{min} Minimum number of reads per junction, default 10 and can be user-defined with the flag `--min-reads`

* $\Psi = \text{NA}$ can mean three things:

1. Transcript was not expressed
2. Insufficient evidence to confidently call exon inclusion or exclusion
3. Junctions map to different alternative or flanking exon(s) – considered as distinct events during the indexing step, `outrigger index`

Original event

Junctions map to different alternative exon(s)

Same alternative exon(s), different junction(s)

For a SE event, if the junctions map to different alternative exon (small black exon on the top), then the event with smaller exon has a Ψ value ranging from zero to one, but for the wider exon (on the bottom), which doesn't have matched inclusion reads, this event is called excluded with Ψ=0

$0 < \Psi < 1$

$\Psi = 0$

** The multiplier for how much greater one side junction can be is user-defined with the flag `--uneven-coverage-multiplier`, here shown with the default value of 10.

To deal with 0 reads, a pseudocount of 1 is added to all junctions for this test only:

$\xrightarrow{} 1 \times 10 \nleq 6$ ✓Passes

$\xrightarrow{} 1 \times 10 \leq 51$ ✗ Doesn't pass

*** $r_{\text{threshold}}$ Threshold for total junction reads in the event
$r_{\text{threshold}} = n_{\text{junctions}} \times r_{\text{min}}$

e.g. for an MXE event (4 junctions) and a minimum of 10 reads per junction: $\sum_{i,j} r_{i,j} = 4 \times 10 = 40$

$\sum_{i,j} r_{i,j}$ Total Junction reads

$n_{\text{junctions}}$ Number of junctions in splicing event type (e.g. 3 for SE or 4 for MXE)

Configurable options

Junction read inputs | `--bam` `--sj-out-tab` `--junction-reads-csv` (default: reads from `outrigger index`)

r_{min} `--min-reads 10` (default)

`--uneven-coverage-multiplier 10` (default)

Case 1: Incompatible junctions with sufficient reads. This step checks whether the junction reads are compatible with a MXE event, or rather a twin cassette event. Specifically, evidence of $r_{2,3} > r_{\text{min}}$ or $r_{1,4} > r_{\text{min}}$ suggests this junction is a twin cassette event but not an MXE event. In such cases, $\Psi = \text{NA}$. As described in `outrigger index`, the minimum number of reads is user-defined, for example to 37 with `--min-reads 37`.

Case 2: Zero observed reads. Given no reads is observed, this event is $\Psi = \text{NA}$, rather than $\Psi = 0$ since $\Psi = 0$ indicates exclusion.

Case 3: All compatible junctions with insufficient reads. No single junction has the minimum number of reads r_{min}, by default r_{min} is 10, and can be modifiable by the `--min-reads` flag. If this is the case, we assign $\Psi = \text{NA}$.

Case 4: Only one junction with sufficient reads. This applies to a single junction of two junctions per isoform, e.g. Isoform2 of either SE or MXE events, and Isoform1 of an MXE event, has sufficient reads. Since only one junction has the minimum number of reads, r_{min}, no sufficient evidence indicates inclusion of exon-of-interest, thus, we assign $\Psi = \text{NA}$.

Case 5: One junction with $> 10\times$ more reads than the other. When the alternative exon is covered on the two sides with junction reads of great disparity, there is insufficient evidence supporting the inclusion of alternative exon or suggests the exon may involved in a complex splicing, rather than a SE or MXE. Thus, $\Psi = \text{NA}$. The default multiplier is 10 and can be modified by the user, for example to 55 by `--uneven-coverage-multiplier 55`.

Case 6: Exclusion: Isoform2 with sufficient reads and Isoform1 with zero reads. All junctions on Isoform2 have greater than the minimum reads

r_{min}, and all junctions of Isoform1 have no observed reads, thus $\Psi = 0$.

Case 7: Inclusion: Isoform2 with zero reads and Isoform1 with sufficient reads. All junctions on Isoform2 have no observed reads and all junctions of Isoform1 have greater than the minimum reads r_{min}, thus $\Psi = 1$.

Case 8: Sufficient reads on all junctions. Both Isoform1 and Isoform2 have greater than the minimum reads on all their junctions. This is the best possible case for alternative splicing.

Case 9: Isoform2 with sufficient reads but Isoform1 has one or more junctions with insufficient reads. If the exclusion isoform, Isoform2 has sufficient reads, but the inclusion isoform (Isoform1) does not, then we assess whether the total read coverage of the event, $\sum_{i,j} r_{i,j}$exceeds $r_{\text{threshold}}$. If so, a Ψ is calculated; if not, $\Psi = \text{NA}$. We define $r_{\text{threshold}}$ as the number of junctions n times the minimum number of reads r_{min}. For example, with a minimum read count is 10 on an SE event, $r_{\text{threshold}} = 30$. For a minimum read count of 10 on an MXE event, $r_{\text{threshold}} = 40$.

Case 10: Isoform2 has one or more junctions with insufficient reads but Isoform1 has sufficient reads. Similar to Case 9, we again test if the total read coverage is sufficient to calculate Ψ, i.e. if $\sum_{i,j} r_{i,j} \geq r_{\text{threshold}}$. If so, we calculate Ψ, and if not, we assign $\Psi = \text{NA}$.

Case 11: Isoform1 and Isoform2 each have both sufficient and insufficient junctions. This case only applies to MXE events as SE events have as single Isoform2 junction, and cannot have both sufficient and insufficient junctions. If by the per-junction coverage, it is unclear whether the event has sufficient coverage, then we test if the total coverage of the event is sufficient. If so, we calculate Ψ, and if not, we assign Ψ = NA.

Outputs The output of `outrigger psi` is added into the `outrigger_output` folder by creating a `psi` folder for each splice type. `psi.csv` contains Ψ in a matrix, and the `summary.csv` produces a summary of all the events observed in all samples with their junction reads.

To follow up with our `outrigger index` and `outrigger validate` commands, we can run the below example command in the same directory:

```
outrigger psi
```

This command adds to the existing output folder `outrigger_output`. Therefore, we don't need to specify a genome location or reads or index location if this command is run from the same folder as the `outrigger index` command was run, and there exists in the directory a folder called `outrigger_output`.

Advantages and limitations of `outrigger`. The main advantages of `outrigger` are speed and conserved memory footprint. As `outrigger` operates only on junction reads, rather than resampling reads from a `.bam` alignment file, which can range in size from 500MB to 20GB and results in a high memory footprint, `outrigger` summarizes each `.bam` file to only its junction reads and uses that to estimate Psi/Ψ values. Additionally, employing three steps of `outrigger` `outrigger` is able to maximize the number of potential alternative events and subsequently apply strict validation rules in the step of outrigger psi calculation to eliminate false positive events from each sample. However, currently, `outrigger` can only deal with SE and MXE events. We are in the process of incorporating other alternative splice types.

```
outrigger_output/
├── index
│   ├── gtf
│   │   ├── gencode.vM10.annotation.gtf
│   │   ├── gencode.vM10.annotation.gtf.db
│   │   └── novel_exons.gtf
│   ├── exon_direction_junction.csv
│   ├── mxe
│   │   ├── event.bed
│   │   ├── events.csv
│   │   ├── exon1.bed
│   │   ├── exon2.bed
│   │   ├── exon3.bed
│   │   ├── exon4.bed
│   │   ├── intron.bed
│   │   ├── splice_sites.csv
│   │   └── validated
│   │       └── events.csv
│   └── se
│       ├── event.bed
│       ├── events.csv
│       ├── exon1.bed
│       ├── exon2.bed
│       ├── exon3.bed
│       ├── intron.bed
│       ├── splice_sites.csv
│       └── validated
│           └── events.csv
├── junctions
│   ├── metadata.csv
│   └── reads.csv
└── psi. ..................................................... Added by outrigger psi
    ├── mxe..................................................... Added by outrigger psi
    │   ├── psi.csv ............................................. Added by outrigger psi
    │   └── summary.csv ........................................ Added by outrigger psi
    ├── outrigger_psi.csv.................................... Added by outrigger psi
    └── se...................................................... Added by outrigger psi
        ├── psi.csv ............................................. Added by outrigger psi
        └── summary.csv ........................................ Added by outrigger psi
```

Figure 2.7: Example output of `outrigger psi` command.

2.1.2 Comparison to other methods

In comparison to the popular splicing program MISO[194], `outrigger` has three major advantages:

1. Ability to build de novo exon indexes (`outrigger index`)
2. Flexiblity of junction-based definitions of alternative exons, enumerating all possible flanking exons (`outrigger index`)
3. Ability to eliminate incompatible alternative events (`outrigger psi`)
4. Speed of evaluation. Instead of using the huge `.bam` alignment files directly, `outrigger` summarizes the files as junction reads, leading to much faster calculation of percent spliced-in. Once an index is built with `outrigger index` (24-48 hours), then calculation of Ψ/Psi takes 2-4 hours, even on hundreds of samples. With MISO, the calculation can take 8 hours per sample.

Ability to build de novo exon indexes. MISO provides pre-built alternative splicing indexes, which may not be incompatible with the data at hand. There is a program, GESS[204] to detect alternative exons from `.bam` files, which can only handle a handful files at a time and freeze when given hundreds of single-cell `.bam` files. In contrast, in the outrigger indexing step, `outrigger` builds indexes based on provided data, which will be integrated with provided exon annotation allowing identification of novel exons.

Flexiblity of junction-based definitions of alternative exons, enumerating all possible flanking exons. Multiple possible flanking exons can be associated with an alternative exon, most algorithms, including MISO and rMATS[22], choose a single set (often the shortest one), rather than being flexible and allowing the user to choose the relevant ones. The resulting "best guess" of the alternative event may not be biologically relavent and may be misleading to interprete. In such case, computational translation of alternative events, as demonstrated in Figure 4, will not be possible.

Ability to eliminate incompatible alternative events Comparing MISO Ψ values side-by-side with a corresponding `outrigger psi` calculation, we find that 46% of MISO Ψ values are rejected and assigned Ψ = NA by `outrigger` (**Figure 2.8**).

A large group of false positives that are correctly rejected by `outrigger` are Case 1, where only incompatible junctions present sufficient reads. For example, when twin cassette events are annotated as MXE events and the data indicates inclusion of both alternative exons, MISO will calculate Ψ as 0.5. Because MISO uses a prior of $\Psi = 0.5$ and resamples the data to calculate Ψ. In such a case, MISO is never convinced that Ψ should be towards 1 or 0 and remains at Ψ 0.5 (**Figure 2.8a**).

Figure 2.8 *(next page)*: Examples of inconsistencies in MISO's estimation with single-cell data.

a-c. Representative examples of SE and MXE AS events measured by MISO, but were unsupported with visual inspection on IGV browser, and were disqualified by `outrigger`. To identify SE and MXE events, `outrigger` constructs a *de novo* splicing index based on the junction reads in all libraries in the dataset (see details in **Figures figures 2.2, 2.5 and 2.6**). The following examples are not considered by `outrigger`as true SE or MXE events, therefore annotated as NA. Note, MISO does not estimate modality for each event, `anchor` (see details in **Figures 2.10, 2.11 and 3.4**) was used to estimate modality.

a. Top, a MISO-annotated MXE event in ARF4 with MISO estimated Ψs ~ 0.5 and classified as "middle" modality in each of iPSC, NPC, and MN by `anchor`. Yet, in the IGV browser (bottom), this event appears as a twin cassette event, where both exons 2 and 3 are included, indicating that at least in our dataset this event is not consistent with the MISO annotation. Outrigger disqualifies this event as a MXE and assign NA (top left).

b. Top, a MISO-annotated SE event in CLF1 with MISO estimated Ψs ranging from 0.1 to 0.6 and is classified as a "middle" modality event by `anchor` in each of iPSC, NPC, and MN. Yet, in the IGV browser (bottom), exon 1 for this annotation is not covered at all. Given the data, outrigger do not consider this as a bona fide SE event and assign NA to this event.

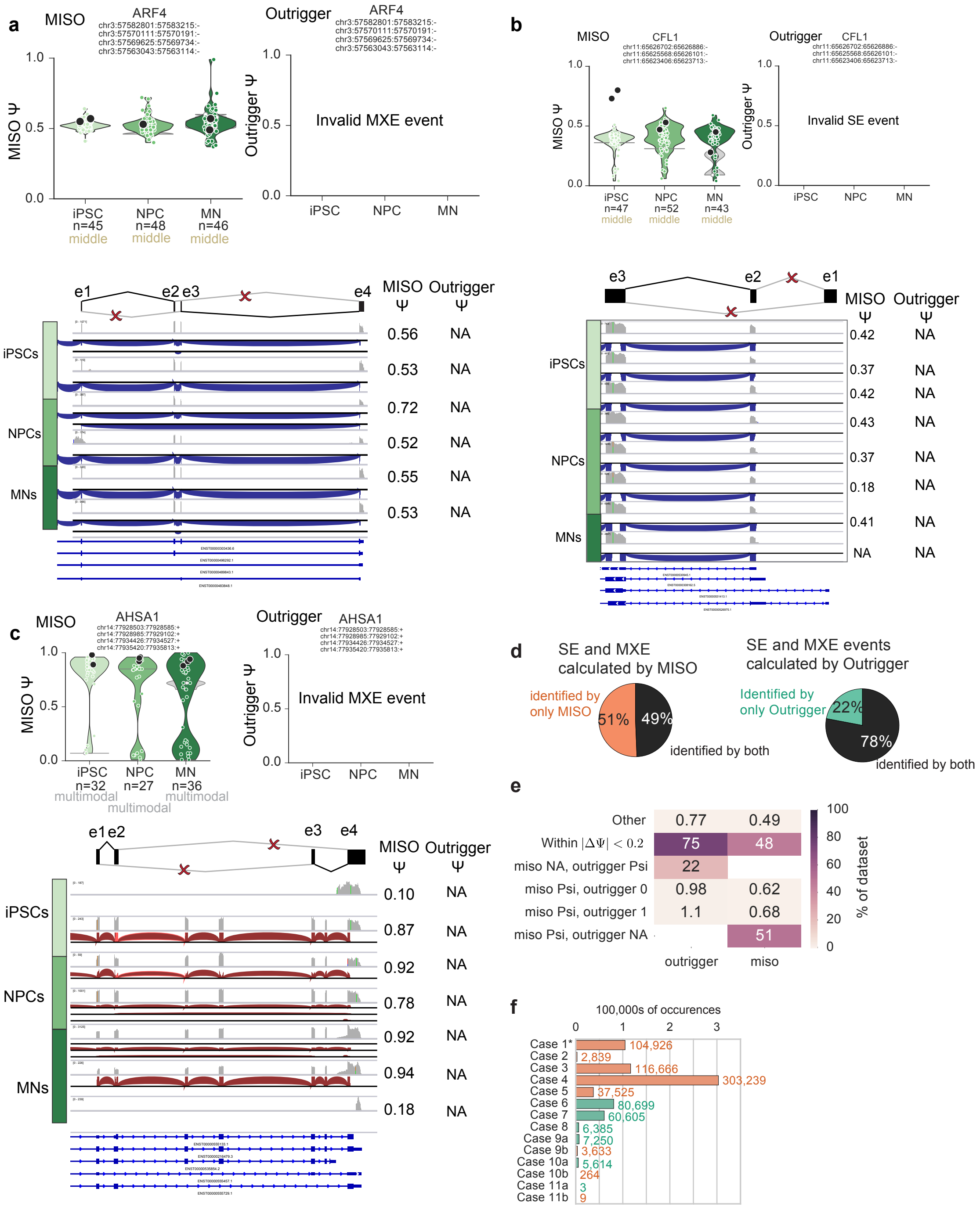
c. Top, a MISO-annotated MXE event in AHSA1 with a wide range of MISO calculated Ψs and is classified as the "multimodal" modality in each of iPSCs, NPC, and MN populations by `anchor`. Bottom, in the IGV browser. Exons 2 and 3 are the annotated alternative exons for MXE, however, another two well-covered exons between exon 2 and 3 were observed and one extra exon between exon 3 and 4, which disqualify this event as an MXE event. Furthermore, when both exon 2 and 3 are included, MISO estimated Ψ scores are closer to 1 instead of around 0.5, as was seen in (**a**). Thus, outrigger rejects this as MXE and assign NA.

d. Using `outrigger`'s strict rules on MISO annotations, the majority (51%) of the data generated by MISO was rejected by `outrigger` (left). Right, using the exact same annotation from MISO, `outrigger` 22% of events found by `outrigger` had too wide of a confidence interval (> 0.4) by MISO.

e. Heatmap comparing the numbers and percentages of alternative events that were within $|\Delta\Psi| < 0.2$, switched to exactly 1 or 0 in `outrigger`, were NA in either MISO or `outrigger`, or were in another case.

f. Barplot of the number of cases found only in MISO (orange) and rejected as NA by `outrigger`, and of the cases found only by `outrigger`(green) and considered to have too wide of a confidence interval by MISO.

To summarize, `outrigger` follows strict rules to identify alternative splicing (**Figures figures 2.2, 2.5 and 2.6**) and provides a Ψ distribution more localized at the extremes of $\Psi = 0$ and $\Psi = 1$. Although `outrigger`, may identify fewer events, they are true SE and MXE events.



* For a detailed explanation of cases, see Supplementary Software Figure 4

The majority of the false positives are Case 4, where only one junction has sufficient reads. As MISO counts both junctions to calculate Ψ, shown in **Figure 2.8b-c**, many of the events are not covered on both sides of the alternative exons, which may suggest the events are not true SE events, but rather alternative first exon events, for instance.

We used MISO's event definitions and found that as many as 50% of MISO events did not pass the stringent rules of `outrigger`, primarily due to the incompatibility with the annotation of SE and MXE and insufficient coverage (**Figure 2.8j-l**).

2.2 anchor: Modality estimation

2.2.1 Algorithm overview

Model modalities as beta distributions We define *modality* as a distinct type of distributions. Since Ψs are continuous value between $(0,1)$, distribution of Ψ can be modeled as Beta distribution. The probability density function for the Beta distribution, $\Pr(\alpha,\beta)$ is defined between $(0,1)$, with parameters $\alpha > 0$ and $\beta > 0$,

$$\Pr(\alpha,\beta) \sim \frac{1}{\mathrm{B}\left(\alpha,\beta\right)} x^{(\alpha-1)} (1-x)^{(\beta-1)}, \tag{2.3}$$

where $\mathrm{B}\left(\alpha,\beta\right)$ is the Beta function, defined by $\alpha > 0$ and $\beta > 0$. It may be easier to think about how the α and β parameters affect distribution by observing the mean and variance **Figure 2.8a**. The beta distributions can be described by four parameterizations: $1 \leq \alpha < \beta$, $\alpha = \beta > 1$, $\alpha > \beta \geq 1$, $\alpha = \beta < 1$ (**Figure 2.8b**). Conveniently, these four configurations correspond to the four modalities we are interested in: $1 \leq \alpha < \beta$ corresponds to *excluded*, $\alpha = \beta > 1$ to *middle*, $\alpha > \beta \geq 1$ to *included*, and $\alpha = \beta < 1$ to *bimodal* (**Figure 2.8c**). The final *multimodal* modality corresponds to $\alpha = \beta = 1$, which is equivalent to the uniform distribution used as null model.

Model parameterization To describe feature distribution as modalities, we parameterized the four parameterizable modalities and used Bayesian model selection to choose the best model to describe the distribution. Python package `scipy`[205;206] was used to implement Beta distribution. For included (excluded) modality, we fixed β (α) at 1 and linearly increased α (β) from 2 to 20 (**Figure 2.8**d). We chose 2 as a starting parameter since it is near the $\alpha = \beta = 1$ uniform distribution, as we wanted to allow excluded and included distributions with noise. For bimodal (middle) modality, we changed α and β simultaneously, monotonically decreasing (increasing) the parameters from $\alpha = \frac{1}{12}, \beta = \frac{1}{12}$ ($\alpha = 2, \beta = 2$) to $\alpha = \frac{1}{30}, \beta = \frac{1}{30}$ ($\alpha = 20, \beta = 20$). The parameters for bimodal start at $\frac{1}{12}$ rather than $\frac{1}{2}$ because starting the parameters from $\frac{1}{2}$ resulted in more false positive "bimodal" events, whereas starting the parameters from $\frac{1}{2}$ ensures any density near 0.5 is downweighted.

The fit of feature distribution is assessed to the four configurations using Bayes Factors, represented by K,

$$K^{(m)} = \frac{P(D|M_1^{(m)})}{P(D|M_0)} \tag{2.4}$$

$$= \frac{\sum_i P(\alpha_i^{(m)}, \beta_i^{(m)}|M_i^{(m)})P(D|\alpha_i^{(m)}, \beta_i^{(m)}, M_i^{(m)})}{\sum P(\alpha_0, \beta_0|M_0)P(D|\alpha_0, \beta_0, M_0)} \tag{2.5}$$

$$= \frac{\sum_i P(\alpha_i^{(m)}, \beta_i^{(m)}|M_i^{(m)})P(D|\alpha_i^{(m)}, \beta_i^{(m)}, M_i^{(m)})}{1} \tag{2.6}$$

$$= \sum_i P(\alpha_i^{(m)}, \beta_i^{(m)}|M_i^{(m)})P(D|\alpha_i^{(m)}, \beta_i^{(m)}, M_i^{(m)}) \tag{2.7}$$

Where $M_i^{(m)}$ is the model of interest (e.g. $M_i^{(\text{bimodal})}$) and $\alpha_i^{(m)}, \beta_i^{(m)}$ are the corresponding parameters from the parameterization shown in **Figure 2.8d**. The null model, M_0 is the uniform distribution, where $\alpha_0 = \beta_0 = 1$, and thus $P(D|M_0) = 1$ for all datasets. We use a Bayes Factor cutoff of K_{cutoff} to indicate the threshold where the model begins to explain the data reasonably well. In practice we set $K_{\text{cutoff}} = 2^5$ ($\log_2 K_{\text{cutoff}} = 5$).

The excluded and included modalities vary only one parameter at a time, whereas middle and bimodal modalities vary both α and β simoutanously. Models with more parameters are more likely to fit, thus we fit to the one-parameter models first, assessing whether $K > K_{\text{cutoff}}$ for either excluded or included. No distribution can fit both excluded and included modalities, thus it is assigned to the modality with highest K. Next, the distribution is fitted to the two-parameter bimodal and middle models, checking if $K > K_{\text{cutoff}}$. If neither modality applies, we assign the modality to *multimodal* (**2c**).

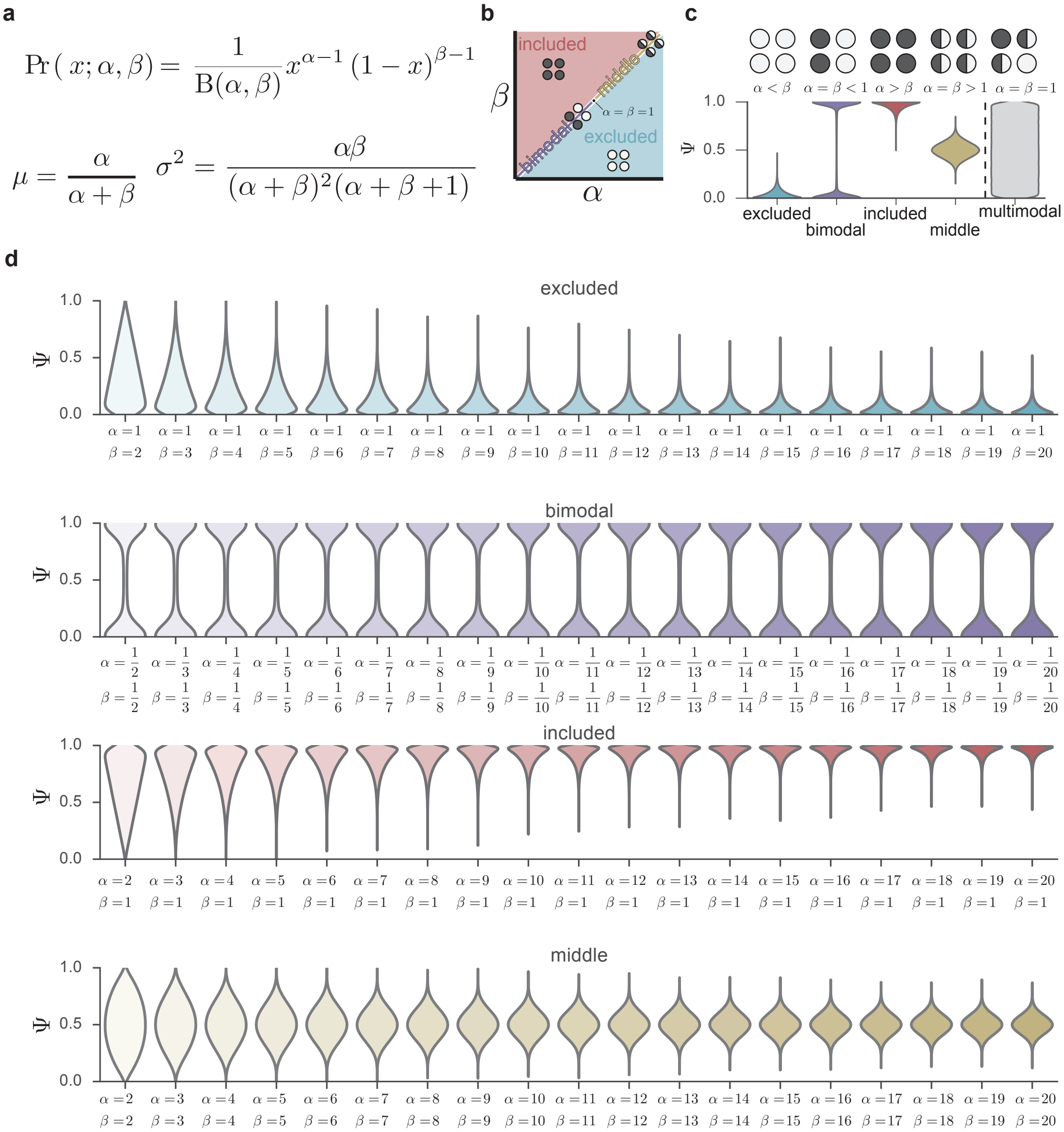
Figure 2.8 *(next page)*: Overview of `anchor` parameterization of the Beta distribution.

a. Top, equation for the Beta distribution of the random variable x with parameters $\alpha, \beta > 0$. Bottom left, equation for the mean (μ) of the Beta distribution as a function of its parameters. Bottom right, equation for the variance (σ^2) of the Beta distribution as a function of parameters.

b. Cartoon of valid values of α and β parameters of Beta distribution, showing how the space is partitioned by the modalities.

c. Violinplots representing the four ideal modalities, plus the null "multimodal" distribution. Each modality is annotated with examples of four cells representing within-cell distributions of included (dark grey) and excluded (light grey) transcripts, and the corresponding parameters of the Beta distribution.

d. Violinplots of 1 million random samples of the family of Beta distributions specified by the α and β (x tick labels) parameterization of the four modalities: excluded, bimodal, included, and middle.


a
$\Pr(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
$\mu = \frac{\alpha}{\alpha+\beta}$
$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
b
included
excluded
bimodal
middle
$\alpha=\beta=1$
β
α
c
$\alpha<\beta$
$\alpha=\beta<1$
$\alpha>\beta$
$\alpha=\beta>1$
$\alpha=\beta=1$
Ψ
1.0
0.5
0.0
excluded
bimodal
included
middle
multimodal
d
excluded
$\alpha=1$, $\beta=2$ … $\alpha=1$, $\beta=20$
bimodal
$\alpha=\frac{1}{2}$, $\beta=\frac{1}{2}$ … $\alpha=\frac{1}{20}$, $\beta=\frac{1}{20}$
included
$\alpha=2$, $\beta=1$ … $\alpha=20$, $\beta=1$
middle
$\alpha=2$, $\beta=2$ … $\alpha=20$, $\beta=20$

As exact 0 and 1 are not in the range of the Beta distribution, we implement this model selection by adding a small number (0.001) to 0 and subtracting this small number from 1. Thus, we approximate the data-derived distribution from the invalid closed interval [0, 1] to the valid open interval of (0, 1).

2.2.2 Simulations

We optimized the algorithm parameters using test datasets and visually inspecting random samples from both the best- and worst-fitting data and ensuring that the even the worst fitting data was still believably categorized as the modality (**Figure 2.9**).

Dataset 1: "Perfect Modalities" with noise To test the limits of `anchor`, we simulated perfectly excluded, middle, included, and bimodal distribution, added uniform random noise with 100 iterations, and estimated modality at each noise level with iteration (**Figure 2.10a**). As expected, the most frequently predicted modality was "multimodal," since the dataset was created from randomly added noise (**Figure 2.10b**). The next frequent modality was bimodal, followed by a tie with excluded and included, and the least frequent one is middle modality. We found that these parameterizations can accurately predict modality with up to 35% noise added to the middle modality, 50% noise added to excluded and included modalities, and up to 70% noise added to the bimodal modality(**Figure 2.10d**). By visual inspection of distributions fit best or worst to each modality (**Figure 2.9a**), we observed that the bimodal distributions are sufficiently different from other parameterizations, demonstrating the robustness of the algorithm.

Figure 2.9 *(next page)*: Best and worst fitting modality data using `anchor`. Left, 10 events with largest Bayes Factor, K (best fit) from the assigned modality. Right, 10 events with smallest Bayes Factor, K (worst fit) from their assigned modality. For multimodal, as there is no fit, this simply shows 20 random events.
a. Bayesian `anchor` method on "Perfect modalities" dataset.
b. Bayesian `anchor` method on "Maybe bimodals" dataset.

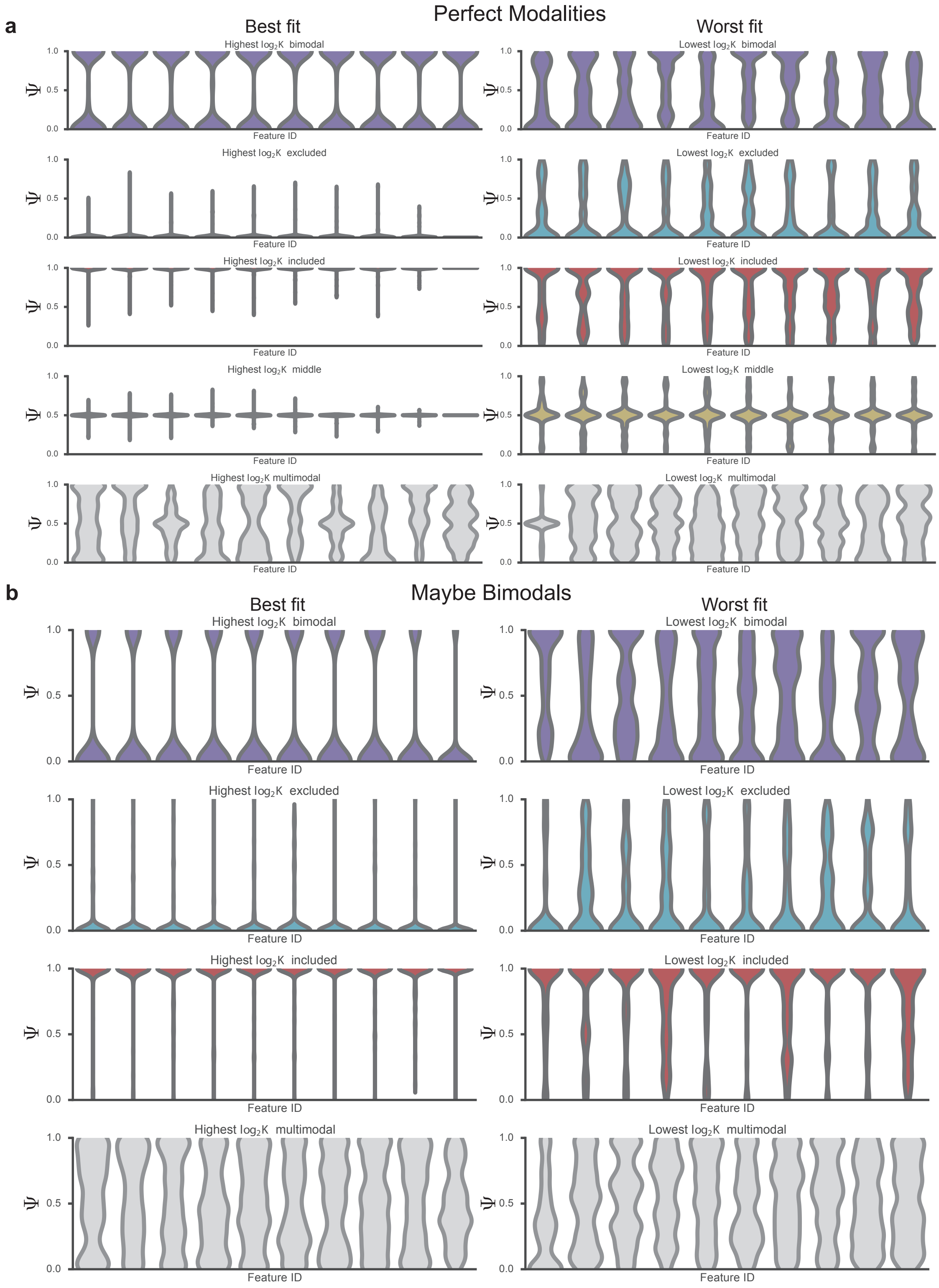
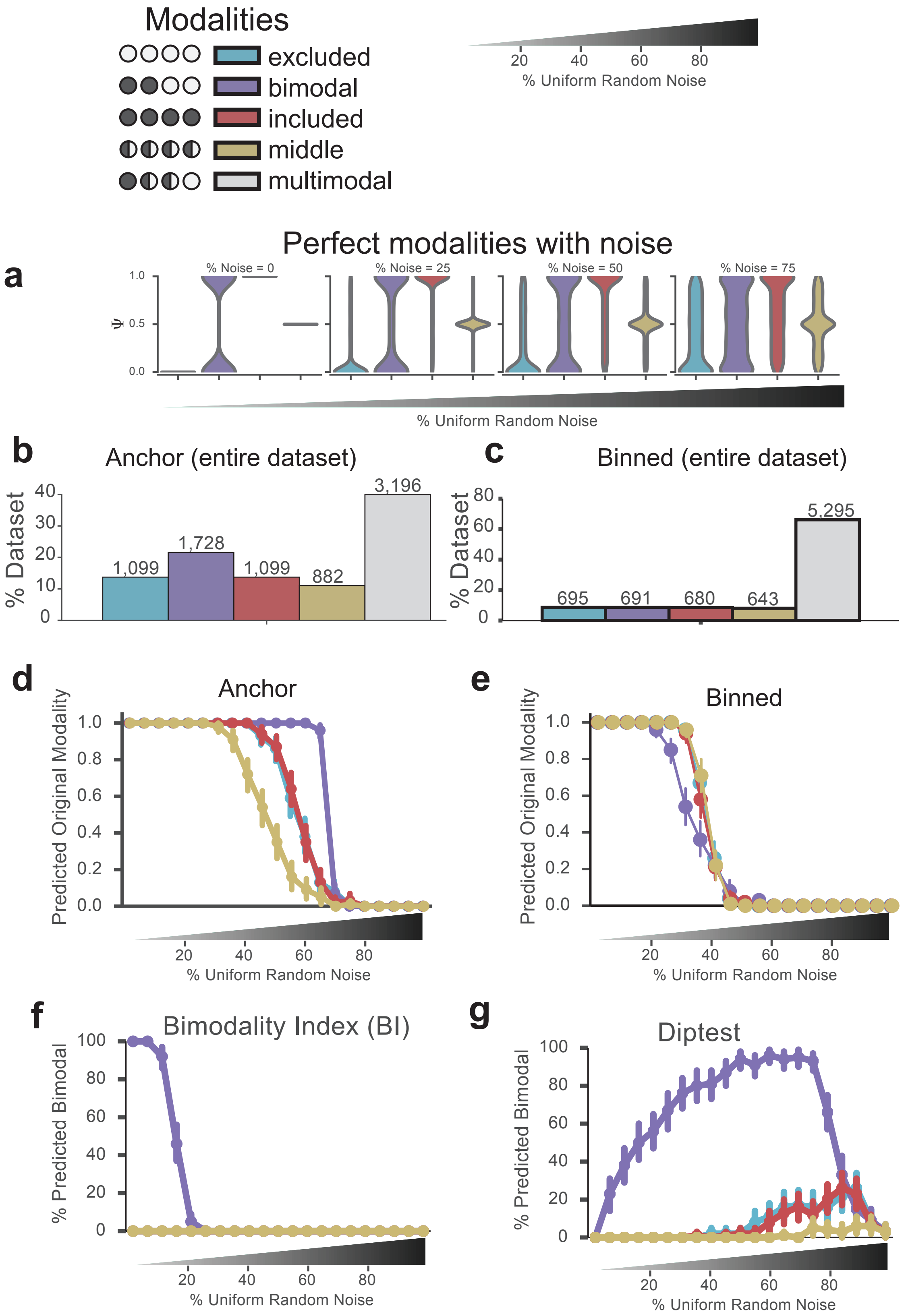

a
Perfect Modalities
Best fit
Worst fit
Highest log$_2$K bimodal
Lowest log$_2$K bimodal
Highest log$_2$K excluded
Lowest log$_2$K excluded
Highest log$_2$K included
Lowest log$_2$K included
Highest log$_2$K middle
Lowest log$_2$K middle
Highest log$_2$K multimodal
Lowest log$_2$K multimodal
Ψ
1.0
0.5
0.0
Feature ID
b
Maybe Bimodals
Best fit
Worst fit
Highest log$_2$K bimodal
Lowest log$_2$K bimodal
Highest log$_2$K excluded
Lowest log$_2$K excluded
Highest log$_2$K included
Lowest log$_2$K included
Highest log$_2$K multimodal
Lowest log$_2$K multimodal
Ψ
1.0
0.5
0.0
Feature ID

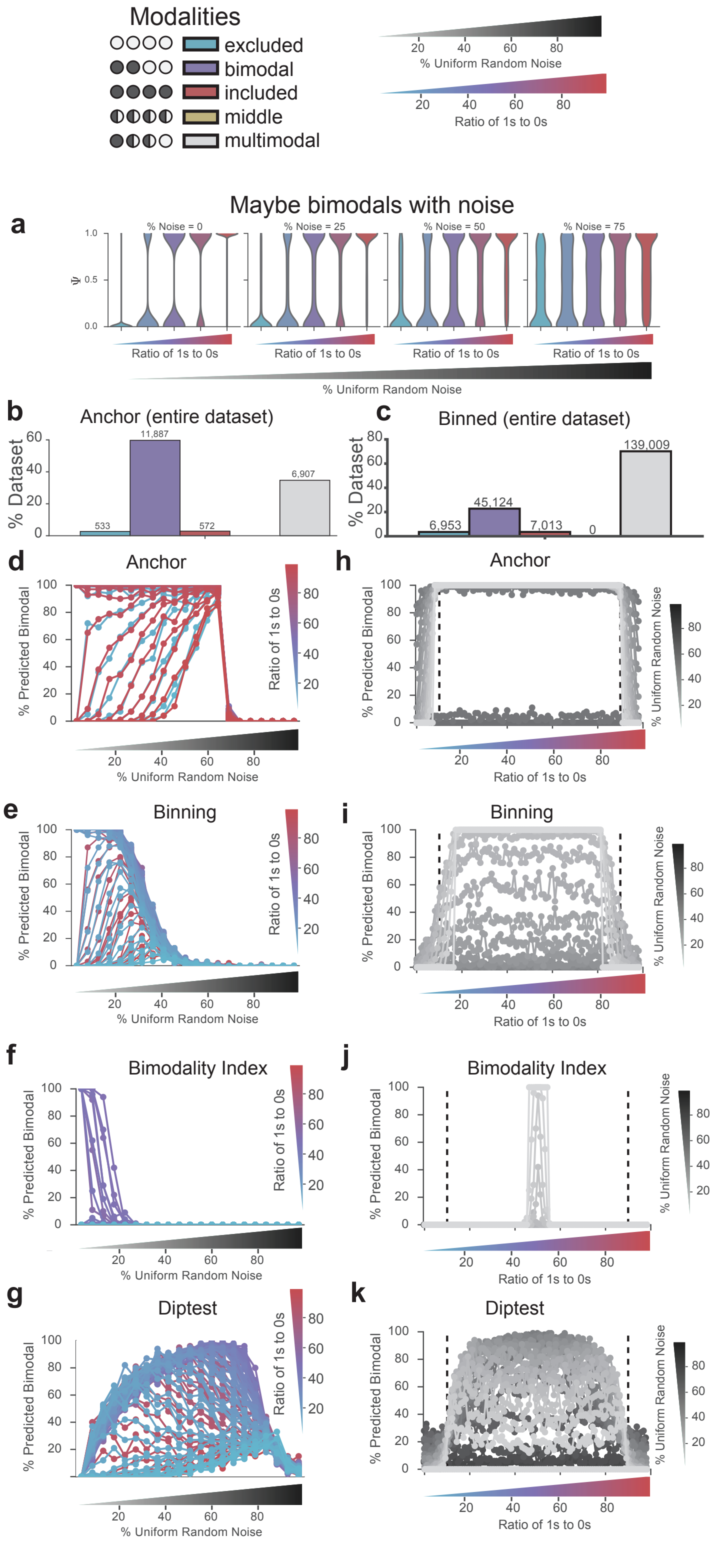
Figure 2.10 *(next page)*: Simulated dataset to test performance of `anchor`.
a. Violinplots depicting the creation of simulated modality datasets with increasing noise. The base dataset (% Noise = 0) consisted of 100 samples of either all zeros (excluded), half zeros and half ones (bimodal), all ones (included), or all 0.5s (middle), exactly representing the four modalities. Uniform random noise was added in 5% increments, with 100 iterations at each noise level.
b. Percentage of events categorized as different modalities by `anchor` in the randomly generated test datasets, across all noise levels, as illustrated in (**a**). Number of events for each modality is annotated on top of the barplots.
c. Percentage of events categorized as different modalities by binning in the randomly generated test datasets, across all noise levels, as illustrated in (**a**). Number of events for each modality is annotated on top of the barplots.
d-g. Specificity of modality estimation. Recapitulation of the original modality as a function of additional noise, using `anchor` (**d**), binning (**e**), Bimodality index (**f**), and diptest (**g**) methods. The x-axis depicts the percent of uniform random noise added (visualized as a triangle gradient), and the y-axis depicts the fraction of times a noisy feature was categorized into each modality. The hue of the line is the modality.


Modalities
excluded
bimodal
included
middle
multimodal
20 40 60 80
% Uniform Random Noise
a
Perfect modalities with noise
% Noise = 0
% Noise = 25
% Noise = 50
% Noise = 75
1.0
Ψ 0.5
0.0
% Uniform Random Noise
b
Anchor (entire dataset)
% Dataset
40
30
20
10
0
1,099
1,728
1,099
882
3,196
c
Binned (entire dataset)
% Dataset
80
60
40
20
0
695
691
680
643
5,295
d
Anchor
Predicted Original Modality
1.0
0.8
0.6
0.4
0.2
0.0
20 40 60 80
% Uniform Random Noise
e
Binned
Predicted Original Modality
1.0
0.8
0.6
0.4
0.2
0.0
20 40 60 80
% Uniform Random Noise
f
Bimodality Index (BI)
% Predicted Bimodal
100
80
60
40
20
0
20 40 60 80
% Uniform Random Noise
g
Diptest
% Predicted Bimodal
100
80
60
40
20
0
20 40 60 80
% Uniform Random Noise

Dataset 2: "Maybe Bimodals" with noise To test the proportions of zeros and ones that able to constitute "bimodal" distribution, we created another dataset comprised 100 samples of varying amounts of 0s and 1s, and adding random uniform noise (**Figure 2.11a**). The primary predicted modality was bimodal, then multimodal, and finally included and excluded (**Supplementary Figure 2.11b**). No distribution was predicted as the middle modality, indicating the bimodal and middle modalities are drastically different with little chance of mis-assignment. The falloff of correctly predicting bimodality is at adding 70% noise (**Supplementary Figure 2.11b**), consistent with the previous simulation with "Perfect Modalities" dataset (**Figure 2.10d**). We found that bimodality is determing with a 90:10 (10:90) proportion of samples of 0:1 (0:1) (**Supplementary Figure 2.11d**). Visual inspection of distributions fit best or worst to each modality confirmed the assignment of each modality(**Figure 2.9b**).

To summarize, simulation with two different datasets indicates that 1) bimodal modality can tolerate to up to 70% of uniform random noise, and middle modality is least tolerable to noise at only 30%, 2) included and excluded modalities are drastically different, so as the middle and bimodal modalities, thus the two step modality assignment procedure (**Figure 2**) is well-grounded, 3) anchoris able to determine a bimodal modality with up to 90:10 proportion of zeros and ones.

Figure 2.11 *(next page)*: Simulated bimodal dataset to test performance of `anchor`.
a. Violinplots depicting the creation of the "Maybe Bimodals" test set consists of potential bimodal events, each containing 100 samples of only zeros ($\Psi = 0$) and ones ($\Psi = 1$) in every combination, shown here as relative to the number of ones. We added uniform random noise in increasing 5% levels for 100 iterations at each level. While each combination of 1s and 0s was created, only a subset are shown for brevity – 1:99, 25:75, 50:50, 75:25, and 99:1 ratios of 1:0 are shown, with added uniform random noise of 0% (original), 25%, 50%, and 75%.
b. Percentage of events categorized in modalities by `anchor` in the randomly generated bimodal test datasets, across all noise levels, as illustrated in (**h**). Number of events for each modality is annotated on top of the barplots.
c. Percentage of events categorized in modalities by binning in the randomly generated bimodal test datasets, across all noise levels, as illustrated in (**h**). Number of events for each modality is annotated on top of the barplots.
d-k. Accuracy of bimodality prediction, as a function of the noise added to the dataset.
d-g. Specificity of bimodality estimation upon addition of uniform random noise. The x-axis shows the percent added uniform random noise (visualized as a triangle gradient), and the y-axis indicates the fraction of time features in each noise percentage and proportion of $1:0$ was categorized as bimodal. Overall, all but the very extremes of the $1:0$ proportions were consistently categorized as bimodal until 70% noise, after which point nearly all events became multimodal. Modality estimations are shown using `anchor` (**k**), binning (**l**), Bimodality Index (**m**), and Diptest (**n**).
h-k. Sensitivity of bimodality detection. Percentage of events predicted as bimodal given different proportions of 0s and 1s, and increasing uniform random noise. Events are called as bimodal with approximately 9:1 ratio of 0s and 1s (and vice versa), shown with a dotted line at 10% ones and 90% ones. Bottom triangle gradient shows increasing ratio of ones to zeros, i.e. from exclusion to bimodal, to inclusion. Bimodality estimations are shown using `anchor` (**o**), binning (**p**), Bimodality Index (**q**), and Diptest (**r**).


Modalities
excluded
bimodal
included
middle
multimodal
20 40 60 80
% Uniform Random Noise
20 40 60 80
Ratio of 1s to 0s
a
Maybe bimodals with noise
% Noise = 0
% Noise = 25
% Noise = 50
% Noise = 75
Ψ
1.0
0.5
0.0
Ratio of 1s to 0s
Ratio of 1s to 0s
Ratio of 1s to 0s
Ratio of 1s to 0s
% Uniform Random Noise
b
Anchor (entire dataset)
% Dataset
60
40
20
0
533
11,887
572
6,907
c
Binned (entire dataset)
% Dataset
80
60
40
20
0
6,953
45,124
7,013
0
139,009
d
Anchor
% Predicted Bimodal
100
80
60
40
20
0
Ratio of 1s to 0s
80
60
40
20
20 40 60 80
% Uniform Random Noise
h
Anchor
% Predicted Bimodal
% Uniform Random Noise
80
60
40
20
20 40 60 80
Ratio of 1s to 0s
e
Binning
% Predicted Bimodal
Ratio of 1s to 0s
20 40 60 80
% Uniform Random Noise
i
Binning
% Predicted Bimodal
% Uniform Random Noise
20 40 60 80
Ratio of 1s to 0s
f
Bimodality Index
% Predicted Bimodal
Ratio of 1s to 0s
20 40 60 80
% Uniform Random Noise
j
Bimodality Index
% Predicted Bimodal
% Uniform Random Noise
20 40 60 80
Ratio of 1s to 0s
g
Diptest
% Predicted Bimodal
Ratio of 1s to 0s
20 40 60 80
% Uniform Random Noise
k
Diptest
% Predicted Bimodal
% Uniform Random Noise
20 40 60 80
Ratio of 1s to 0s

2.2.3 Comparison to other methods

Simple binning We can compare this to other methods we attempted, such as fixing bins of $[0, 0.3, 0.7, 1]$ and using cutoffs for the densities, which does not account for the continuous nature of the underlying distributions. We found the modality whose binned distribution was the smallest distance (measured by Jensen-Shannon Divergence[197]) away from each binned event. In both the simulated modalities and simulated bimodal datasets, we found a sharp increase in multimodal distributions and by eye, poorer categorization of the bimodal modality, especially at the decision boundary of low JSD (**Figures figure 2.11c, e, j, l, p**).

Bimodality index Another test for bimodality is the Bimodality Index[195] (BI), which requires estimating each feature as a mixture of Gaussian models. We used the implementation of Generalized Mixture Models in `scikit-learn`[207] to estimate two Gaussian distributions for each model, and calculated the BI. For perfect bimodal featues, the value is large, for example, we found that for the zero-noise bimodal event, the BI = 402) and was the single bimodality index that was larger than 100 for any feature (**Figure 2.11**f, j). This shows that our method is more sensitive to finding bimodal features with the addition of noise, which BI cannot handle.

Hartigan's Dip test A commonly used test for unimodality is Hartigan's dip test[196]. If the distribution fails the unimodality test, then it is considered bimodal. To define a cutoff for when the dip statistic becomes reliable, we calculated the dip statistic using a Python implementation of the test, called `diptest`[208]. We used a p-value cutoff of $p < 0.05$ as our threshold for assigning an event as bimodal. We used the diptest statistic on the two datasets, and found that while the zero-noise bimodal event was not detected as bimodal, adding as small amount of noise *improved* the diptest's detection of bimodal events (**Figure 2.11g,k**), and the accuracy dropped off at a very high noise level - 90%. As expected, the excluded, included, and middle modalities weren't detected as bimodal, except at higher noise levels, which we also saw with `anchor`.

2.3 bonvoyage: Transformation of distributions to *waypoints* and *voyages*

2.3.1 Algorithm overview

The goal of `bonvoyage` is to be able to summarize the entire distribution of a feature into a single point in space, enabling visualization multiple distributions at a time with intuitive interpretation. To accomplish this, we will transform one-dimensional vectors into two-dimensional space. Specifically, the x-axis will represent the *excluded* dimension and the y-axis will represent the *included* dimension, and all points will be described as a sum of excluded and included components (**6a**, left). For example, for two distinct cell-types, we can imagine a feature that starts at a included modality in the first and changes to a excluded event in the second, or changes from middle to bimodal (**6a**, right).

Data discretization We will use a reduced representation of our splicing data by binning each feature on bins b of size 0.1, where b_n represents the nth bin. We represent the binned splicing matrix with B_Ψ, where $B_\Psi[k, j]$ represents the fraction of non-null samples in feature j with Ψ value contained in b_k. In practice, we pre-filter the data by using only features for which there are enough samples. In the main text for this paper, we used a minimum of 10 cells.

Dimensionality reduction via non-negative matrix factorization Non-negative matrix factorization (NMF) is a parts-based dimensionality reduction algorithm which results in meaningful, interpretable results[209]. It is an alternative to other dimensionality reduction methods such as principal- and independent-component analyses (PCA and ICA) because its features are both independent, and non-negative, and thus each feature is composed of a sum of the underlying structure of the data, without pesky negative terms.

Thus, for NMF, we will be reducing B_Ψ as such,

$$B_\Psi \approx W \times H, \tag{2.8}$$

Where W is a (features, 2)-size matrix of the composition of each feature as a sum of how many samples are excluded and included. We found that in the alternative splicing data, the primary components were the included and excluded values, but in other datasets, this may not be the case. Thus, as the components of NMF are the most prominent features, to ensure reproducibility of the axes across datasets, we seeded the NMF transformation with a matrix that is composed of features that are primarily excluded plus a single included feature. We used the Python package `scikit-learn`[207] for the Projected Gradient NMF implementation.

We call the projected distributions "waypoint space," and the distance between two points a "voyage," such as the voyage of the MXE event in PKM (**Figure 3.15**c).

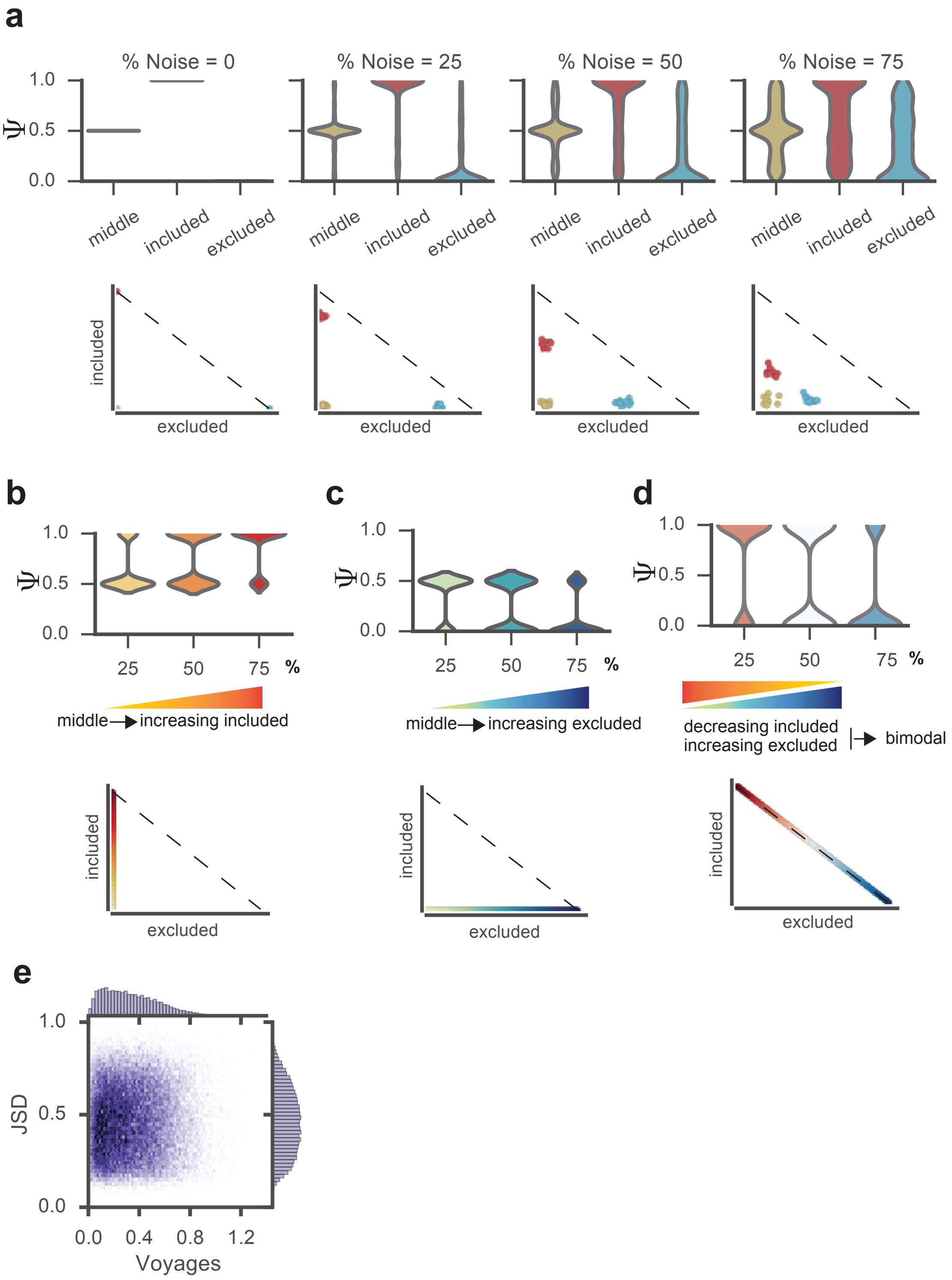
2.3.2 Simulations

Transformation of static distributions To demonstrate the ability of bonvoyage, we created a simulated dataset which we call "Maybe Everything" consisting of every combination of 0s, 1s, and 0.5s (**Figure 2.12**a-d), essentially incorporating both the "Perfect Modalities" (from **Section 2.2.2**) and "Maybe Bimodals" (from **Section 2.2.2**) into a single dataset. Again, we added uniform random noise at 5% intervals. We transformed the entire simulated dataset into the *"waypoint"* space.

To identifying features which change in distribution, we calculate the *"voyage"* between them in waypoint space. As a demonstration, we shuffle the simulated data to create two different *in silico* phenotypes. We will use each feature as a *"waypoint"* along the voyage, and calculate total travel distance of each feature between the phenotypes.

A key aspect of the waypoint space is that while changes from exclusion to inclusion are easy to spot by a change in means, the change from a middle to a bimodal is not, and requires a battery of other tests to find. Here, voyage space has a significant advantage as it gives both the magnitude of change and a directly interpretable direction.

Figure 2.12 *(next page)*: Visualization capabilities of `bonvoyage` shown with simulated data
a-d. Datasets used for testing `bonvoyage`. Uniform random noise was added in 5% intervals to all datasets, up to 95% noise, for 100 iterations at each noise level.
a. Perfect middle, included, and excluded modalities, with added noise. Only 0%, 25%, 50% and 75% noise levels are shown for brevity. Top, averaged violinplots for all features at a given level of noise. Bottom, waypoint space of all features at the specified noise level.
b. Maybe middle-included modalities, created with every combination of 0.5 and 1.0 values. Only the 0% noise dataset is shown for brevity. Top, violinplots, bottom, waypoint plots.
c. Maybe excluded-middle modalities, created with every combination of 0.0 and 0.5 values. Only the 0% noise dataset is shown for brevity. Top, violinplots, bottom, waypoint plots.
d. Maybe bimodal modalities, created with every combination of 0 and 1 values. Only the 0% noise dataset is shown for brevity. Top, violinplots, bottom, waypoint plots.
e. Comparison of voyage magnitude and JSD between "Maybe everything" data and a shuffled copy to show the entire distribution.


a
% Noise = 0
% Noise = 25
% Noise = 50
% Noise = 75
1.0
Ψ 0.5
0.0
middle
included
excluded
included
excluded
b
1.0
Ψ 0.5
0.0
25
50
75
%
middle → increasing included
included
excluded
c
1.0
Ψ 0.5
0.0
25
50
75
%
middle → increasing excluded
included
excluded
d
1.0
Ψ 0.5
0.0
25
50
75
%
decreasing included
increasing excluded
→ bimodal
included
excluded
e
1.0
JSD 0.5
0.0
0.0
0.4
0.8
1.2
Voyages

2.3.3 Comparison to other methods

As there exist many methods for comparing distributions, we will show that the magnitude of change obtained from `bonvoyage` is comparable to other metrics for assessing changes in distribution. In particular, we will show the metrics within each modality, and across modalities, compared to Jensen-Shannon Divergence[197] (JSD) in (**Figure 2.12**). While JSD is more sensitive to slight changes in distribution (their scatterplots are skewed towards the right), it does not also encode directionality of change. Thus, `bonvoyage` offers a unique perspective on how to interpret changes in distribution.

2.4 Acknowledgements

Chapter 2, in part, has been accepted for publication as the supplementary material as it may appear in Molecular Cell, 2017, Yan Song*, Olga B Botvinnik*, Michael T Lovci, Boyko Kakaradov, Patrick Liu, Jia L. Xu and Gene W Yeo (* These authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3

Single-cell alternative splicing analysis with *Expedition* reveals splicing dynamics during neuron differentiation

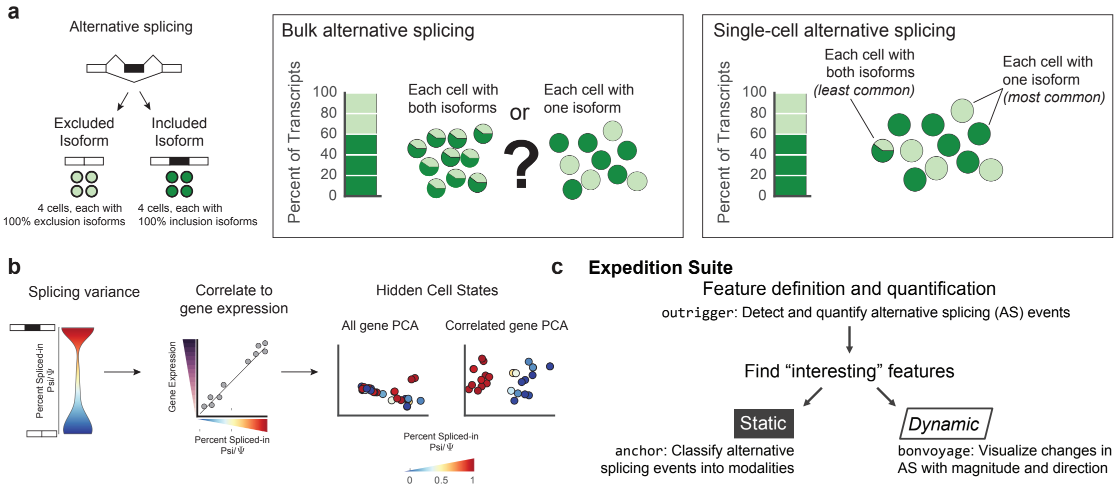

a
Alternative splicing
Excluded Isoform
Included Isoform
4 cells, each with 100% exclusion isoforms
4 cells, each with 100% inclusion isoforms
Bulk alternative splicing
Percent of Transcripts
100
80
60
40
20
0
Each cell with both isoforms
or
Each cell with one isoform
?
Single-cell alternative splicing
Percent of Transcripts
100
80
60
40
20
0
Each cell with both isoforms (least common)
Each cell with one isoform (most common)
b
Splicing variance
Percent Spliced-in Psi/Ψ
Correlate to gene expression
Gene Expression
Percent Spliced-in Psi/Ψ
Hidden Cell States
All gene PCA
Correlated gene PCA
Percent Spliced-in Psi/Ψ
0
0.5
1
c
Expedition Suite
Feature definition and quantification
outrigger: Detect and quantify alternative splicing (AS) events
Find "interesting" features
Static
Dynamic
anchor: Classify alternative splicing events into modalities
bonvoyage: Visualize changes in AS with magnitude and direction

Figure 3.1: Graphical abstract of biological findings from this chapter.

3.1 Introduction

Alternative splicing (AS) generates protein diversity in human cells as over 90% of multi-exon genes are alternatively spliced[23;210–212]. Transcriptome profiling by sequencing (RNA-seq) has emerged as a powerful technology to detect and quantify AS in tissue or cell populations[23;213;214]. Neural tissues have especially high levels of alternative splicing, though it is unclear whether it is a result of high levels splicing within each cell or heterogeneity of cells, impeding precise understanding of AS regulation and dynamics. While single-cell technologies (scRNA-seq) can, in principle, address the issue of heterogeneity, and AS variation has been observed in single-cells[25;27;29], we still do not know if variable AS events are evolutionarily or biologically distinct from less variable events. Robust computational methods are needed to fully characterize the complexity of AS at the whole transcriptome level in single cells.

Previous studies that investigated AS in single cells were limited to only a few examples[27;65] or simply discovered novel splice junctions[25]. However, the key challenge in single-cell AS analysis is not only to measure, but to describe variation in AS within a group of single cells, enabling the discovery of differential AS distribution between populations. Most computational tools for AS were developed for bulk RNA-sequencing and were designed for pairwise comparisons to compute relative differences, such as DEXSeq[16] and rMATs[22]. Yet, for single cells, calculating all pairwise comparisons are impractical. Additionally, many algorithms do not consider the compatibility of splicing annotation with the observed data. Algorithms, such as MISO[194], utilize probabilistic priors which can assign AS events percent-spliced-in (Psi) values near the prior (**Figure 2.8**), resulting in false positive AS events and also prevent meaningful estimation of splicing variation. Other available methods that reconstruct isoforms or estimate read dispersion (Cufflinks, TIGAR2, WemIQ)[215–217] are not appropriate due to the current low molecular capture rate and uneven transcript coverage in single cell RNA-seq datasets. Thus, the lack of computational tools to describe the distribution of AS limits single cell AS analysis to only a few cells or a few events and prevents us from applying systems biology methods to understand AS complexity on a global scale. Similarly, inability to visualize distribution changes from one cell-type/state to another impedes identification of dynamic AS events subjected to specific regulation.

Three key concepts need to be addressed in single-cell AS analyses: (1) implementation of strict rules to identify AS events and ensure compatibility of the annotation and observed data, (2) description of variation and distribution of AS events and (3) visualization of AS distribution and its dynamics from one cell-type or state to another. Therefore, we developed *Expedition*, a suite of algorithms integrated in a complete software package. Expedition can identify and quantify AS events in scRNA-seq data (`outrigger`), categorize splicing modalities (`anchor`) and visualize modality dynamics (`bonvoyage`). To illustrate its utility we sequenced and analyzed single cells from induced pluripotent stem cells (iPSCs), in vitro differentiated neural progenitor cells (NPCs) and motor neurons (MNs). AS events were quantitated and classified into five distinct modalities. Up to 75% of AS events exhibit unimodality, where exons are primarily included or excluded with low variance in each cell population. Only 20% of AS events are highly varying, composed primarily by bimodal AS events. Interestingly, these bimodal AS events account for essentially all AS events that change modalities during neuronal differentiation, thus representing cell-type specific splicing. Furthermore, we demonstrate that individual bimodal and multimodal events are able to reveal the substructure of a cell population that was undetected by global gene expression analysis. Finally, our study revealed that highly variance AS events exhibit evolutionary and sequence characteristics distinct from unimodal events, illustrating the importance of single-cell analysis of RNA processing.

3.2 Results

3.2.1 Identification of alternative splicing events in single cells with `outrigger`

To study alternative splicing in a neural differentiation system, human iPSCs were differentiated towards neural progenitor cells (NPCs) and motor neurons (MNs), as supported by immunofluorescence staining and qRT-PCR of known markers (**Figure 3.2**a, **Figure 3.3**a). We prepared scRNA-seq libraries[218] which were sequenced to an average depth of 15-25 million, 100 bp paired-end (PE) reads per cell (**Figure 3.3**b). Bulk sequencing libraries were also generated from 1,000 cells. We mapped reads to the hg19 genome using RNA-STAR[199] and estimated gene expression as transcripts per million (TPM) using sailfish[219]. Genes detected in at least 10 cells were retained and 4,000-11,000 genes were identified per cell in each population (**Figure 3.3**c-d). Downstream analyses were performed on scRNA-seq datasets from 62 iPSCs, 69 NPCs and 60 MNs that satisfied stringent quality control metrics, after excluding outliers detected by k-means clustering (**Figure 3.3**e). Lineage-specific transcription factors (POU5F1, PAX6 and ISL1) and RNA binding proteins (LIN28A, MSI1 and RBFOX1) that distinguished each cell-type were observed (**Figure 3.3**f). Principal and independent component analysis (PCA and ICA) confirmed that iPSCs, NPCs and MNs were homogenous, yet distinct populations (**Figure 3.3**g, h).

Figure 3.2 *(next page)*: Cell-type specific alternative splicing is an independent feature of cell identity.

a. Human iPSCs are directly differentiated into neuron progenitor cells (NPC) or motor neurons (MN) in vitro. Cell identity is verified by immunofluorescence staining. 63 iPSCs (light green), 73 NPCs (medium green) and 70 MNs (dark green) passed QC and were retained for splicing analysis. Bulk samples are independent samples of 1000 cells.

b. Pyruvate kinase M (PKM) is consistently expressed in iPSCs, NPCs and MNs, shown by log2(TPM+1) in single cells by cell-types.

c. Differential inclusion of a mutually exclusive exon (MXE) alternative splicing (AS) event in PKM is observed in the three cell-types from single cell RNA-seq. top, Schematic of the MXE composed by exon 10 (e10) and exon 9 (e9). bottom, distribution of Ψ for exon 9 in single cells is illustrated by cell-types. Ψ score is estimated by `outrigger` (see Methods). Each green dot in the violin plots represents one cell. Black dots represent measurements in bulk samples.

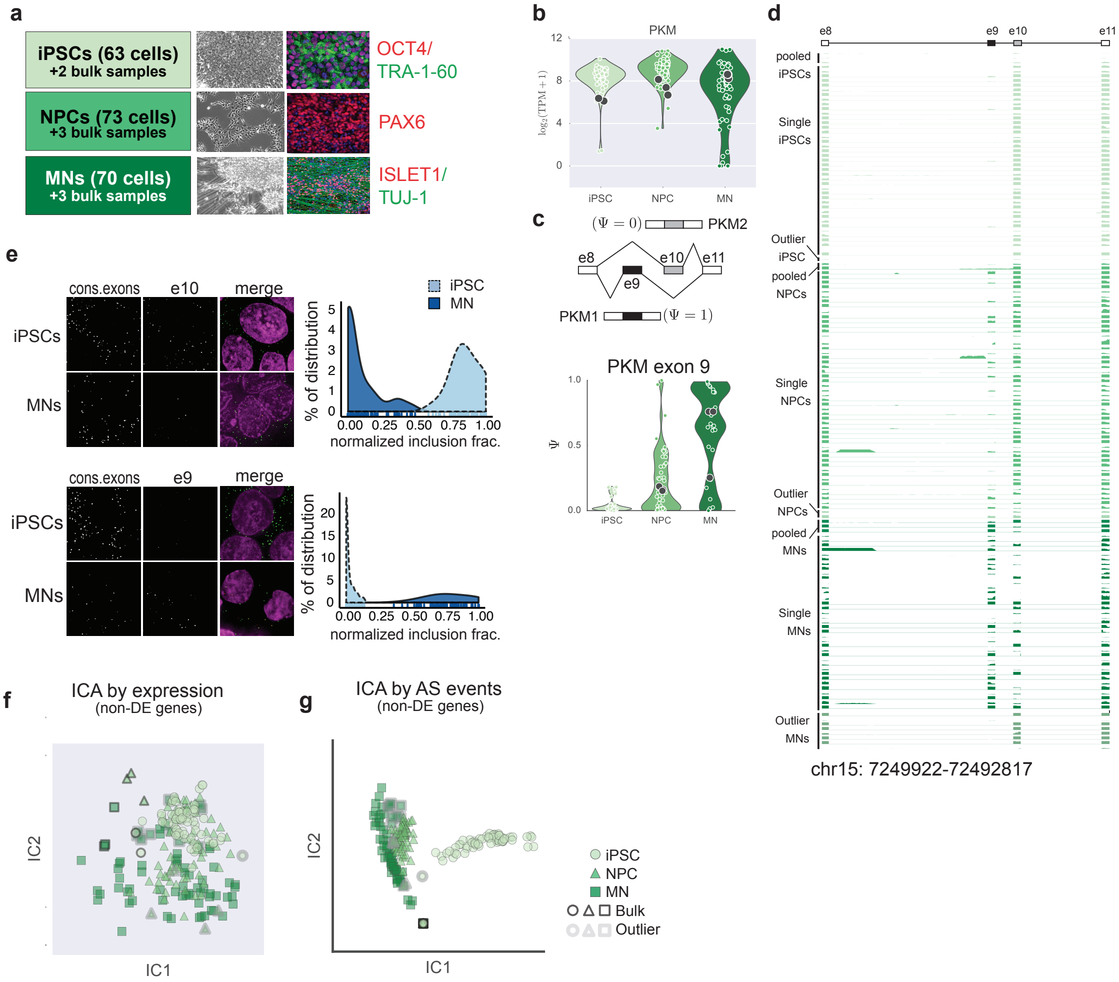
d. Coverage track of MXE exons in pyruvate kinase M (PKM) gene. Each row represents a single cell/sample.

e. Preferential inclusion of e10 and e9 in iPSCs and MNs, respectively, were demonstrated in single cells by smRNA-FISH. Probe sets against constitutive exons (green in merge images) and either exon 10 or exon 9 (red in merge images) were designed in PKM gene. Representative smRNA-FISH images for exon 10 (upper) and exon 9 (lower) (left panel). Distribution of normalized exon inclusion is depicted in iPSCs (light blue with dashed outline) and MNs (dark blue with solid outline; right panel). 74 iPSCs and 101 MNs were counted for e10 inclusion; 125 iPSCs and 67 MNs were counted for e9 inclusion. Normalized inclusion fraction is determined by the percentage of exon specific probes co-localized with constitutive probes/constitutive probes, and resulting percentage is normalized by the 95 percentage of the maximal inclusion.

f-g. AS profile is an independent feature of cell-types. 12,685 Non-differentially expressed (non-DE) genes were identified by non-parametric Kruskal-Wallis test with Bonferonni-corrected q-values > 1.

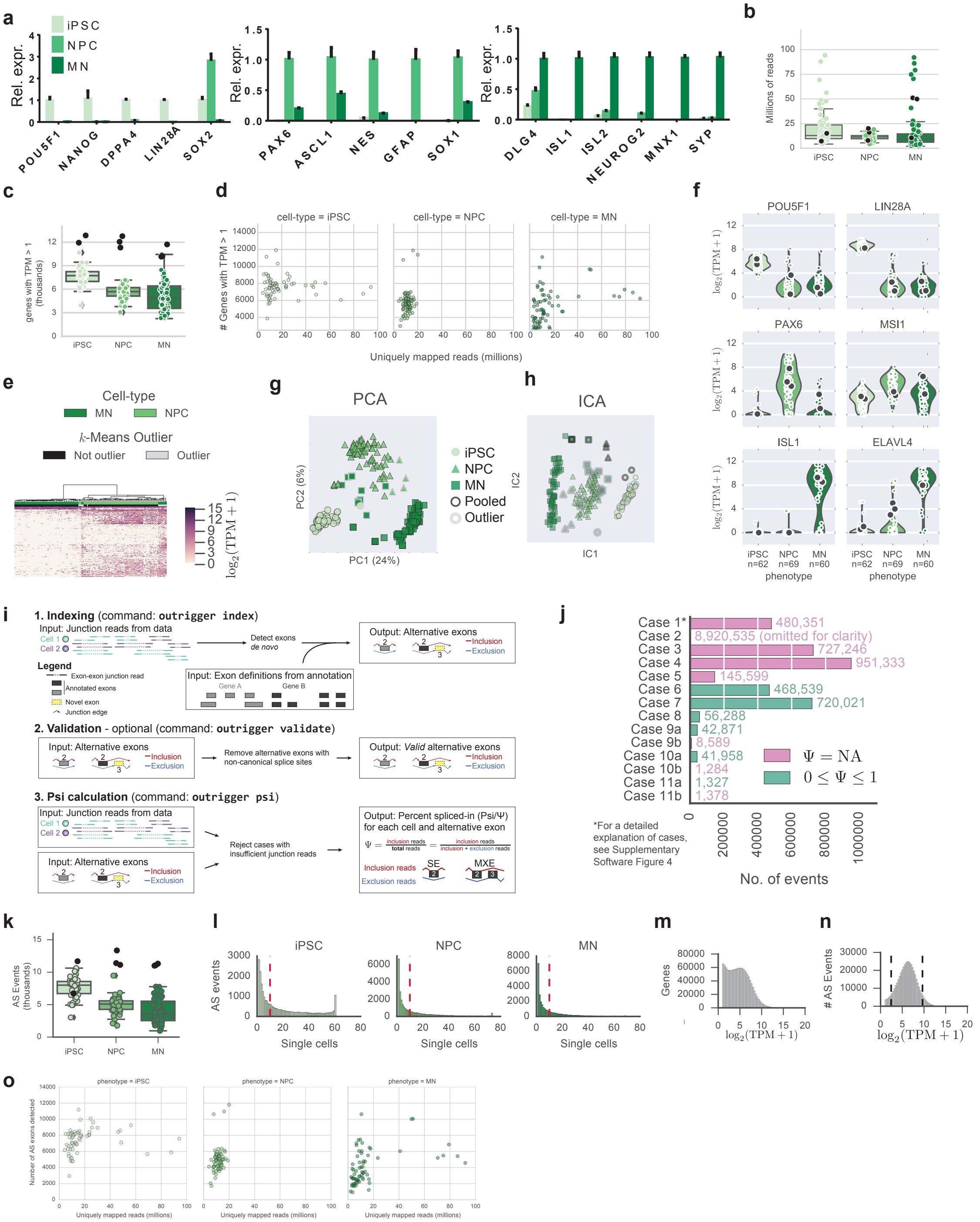
f. ICA on gene expression values of non-DE genes failed to distinguish the three cell-types.

g. ICA on Ψ scores of the AS events residing in non-DE genes, showing AS events are able to group iPSCs, NPCs and MNs, independent of gene expression.


a
iPSCs (63 cells)
+2 bulk samples
OCT4/
TRA-1-60
NPCs (73 cells)
+3 bulk samples
PAX6
MNs (70 cells)
+3 bulk samples
ISLET1/
TUJ-1
b
PKM
12
8
4
0
$\log_2(\mathrm{TPM}+1)$
iPSC
NPC
MN
c
$(\Psi = 0)$
PKM2
e8
e9
e10
e11
PKM1
$(\Psi = 1)$
PKM exon 9
1.0
0.5
0.0
Ψ
iPSC
NPC
MN
d
e8
e9
e10
e11
pooled
iPSCs
Single
iPSCs
Outlier
iPSC
pooled
NPCs
Single
NPCs
Outlier
NPCs
pooled
MNs
Single
MNs
Outlier
MNs
chr15: 7249922-72492817
e
cons.exons
e10
merge
iPSCs
MNs
iPSC
MN
5
4
3
2
1
0
% of distribution
0.00
0.25
0.50
0.75
1.00
normalized inclusion frac.
cons.exons
e9
merge
iPSCs
MNs
20
15
10
5
0
% of distribution
0.00
0.25
0.50
0.75
1.00
normalized inclusion frac.
f
ICA by expression
(non-DE genes)
IC2
IC1
g
ICA by AS events
(non-DE genes)
IC2
IC1
iPSC
NPC
MN
Bulk
Outlier

To identify and quantify alternative splicing (AS) events in scRNA-seq, we developed `outrigger`, an algorithm that uses only junction-spanning scRNA-seq reads to detect and quantify AS. `outrigger` then builds a *de novo* index based on the aligned reads to identify known and novel AS events (**Figures 2.1**, **2.2**, **2.5** and **2.6**). Strict rules were applied to ensure only events with sufficient read coverage, contained valid splice sites, and were compatible with skipped exon (SE) and mutually exclusive exon (MXE) definitions were reported (**Figure 3.3**j). Requiring at least 10 reads per junction, `outrigger` detected 2,000-10,000 SE and MXE events in each cell. Single iPSCs contained a higher number of AS events (5,000-10,000) compared to NPCs or MNs (2,000-6,000) (**Figure 3.3**k,l), likely due to higher RNA content in iPSCs. The bulk samples consistently comprised of 10,000 events, more than most single cells. When an AS event is detected in only a few cells, it may be due to biological variation, aberrant splicing or technical noise. Thus, we retained 13,910 AS events that were detected in at least 10 non-outlier cells in each population within genes that satisfy an expression threshold of TPM > 1 (**Figure 3.3**m-o). An example of an AS event detected by `outrigger` is a MXE event of exons 9 (e9) and 10 (e10) in the PKM gene, encoding pyruvate kinase, which is known to be differentially spliced between committed and proliferative tissues[220;221] (**Figure 3.2**b). PKM is highly expressed across the three cell-types, yet individual iPSCs almost exclusively utilizes e10 whereas e9 is the major AS event in MNs, although 20% (14 out of 60) MNs were observed to possess both isoforms (**Figure 3.2**c,d). To verify the differential inclusion of e10 and e9 in iPSCs and MNs, we designed RNA-FISH probes that target constitutive exons of PKM and two probe sets targeting e9 or e10, exclusively. Our RNA-FISH results agreed with `outrigger` predictions (**Figure 3.2**e). Furthermore, ICA based on the Psi value for each AS event within non-differentially expressed genes generalized our findings with PKM splicing. Indeed, single-cell alternative splicing profiles identified by `outrigger` distinguish the three cell-types (**Figure 3.2**f,g), revealing that AS discerns single cell identities, independent of gene expression.

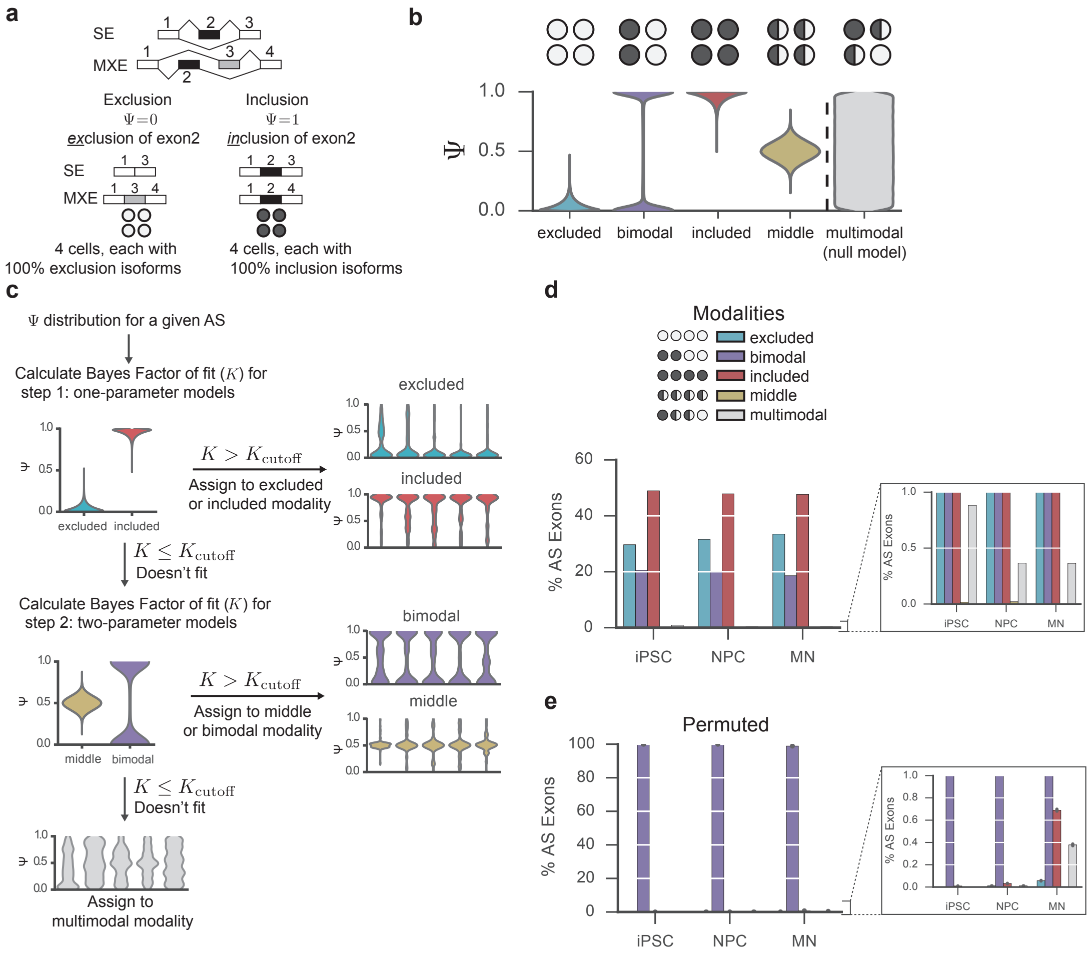
Figure 3.3 *(next page)*: Quality control of single cell expression and splicing data.
a. RT-qPCR validation of biomarker expression in the bulk populations of iPSCs (light green), NPCs (medium green), MNs (dark green). Relative expression of the indicated genes were normalized to housekeeping genes RPL27 and PGK.
b. Sequencing depth for single cell libraries were depicted in box plots. On average, 10-20 million reads of 100bp length was obtained.
c. Number of detected genes for single cell libraries shown as boxplots. Approximately 4,000-6,000 genes were detected at TPM > 1 in single cells.
d. Number of detected genes compared to the sequencing depth for each sample. x-axis, number of reads that mapped uniquely to the genome (fewer than 10 locations), y-axis, number of genes with TPM > 1 detected in each sample. Bulk samples are indicated with a black outline and outlier samples are indicated with a grey outline. Left, iPSC samples, middle, NPC samples, right, MN samples.
e. Outlier MN cells identified by K-means clustering exhibited a transcriptome resembling NPCs. Unsupervised hierarchical clustering demonstrated that MN outliers are clustered together with NPCs.
f. Expression of lineage-specific transcription factors (left) and RNA binding proteins (right). Specifically, POU5F1/OCT4 and LIN28A are specific to iPSCs, PAX6 and MSI1 are more highly expressed in NPCs, and ISL1 and ELAVL4 are only expressed in MNs.
g. PCA of highly variant gene expression. Highly variant is defined as two standard deviations away from mean gene-level variance across all samples.
h. ICA on highly variant gene expression. Highly variant is defined as two standard deviations away from mean gene-level variance across all samples.
i. Barplot showing the
texttttoutrigger cases found across all splicing events and all samples.
j. The number of AS exons (both SE and MXE event types) detected per single cell library.
k. Histograms of number of cells per detected AS exon, in each cell type. Many AS exons were found in only one cell. A minimum of 10 cells per phenotype used, indicated by a dashed red line.
l. Histogram of gene expression across all single cells in iPSC, NPC and MN populations.
m. Expression of genes containing AS exons. 90% of the detected splicing events reside in transcripts expressed between 2.5-10 of $\log_2(\mathrm{TPM}+1)$, as indicated by a dashed black line.
n. Number of detected AS events compared to the sequencing depth for each sample. x-axis, number of reads that mapped uniquely to the genome (fewer than 10 locations), y-axis, number of non-NA AS events detected in each sample. Bulk samples are indicated with a black outline and outlier samples are indicated with a grey outline. Left, iPSC samples, middle, NPC samples, right, MN samples.


a
iPSC
NPC
MN
Rel. expr.
POU5F1
NANOG
DPPA4
LIN28A
SOX2
PAX6
ASCL1
NES
GFAP
SOX1
DLG4
ISL1
ISL2
NEUROG2
MNX1
SYP
b
Millions of reads
iPSC
NPC
MN
c
genes with TPM > 1 (thousands)
iPSC
NPC
MN
d
cell-type = iPSC
cell-type = NPC
cell-type = MN
# Genes with TPM > 1
Uniquely mapped reads (millions)
e
Cell-type
MN
NPC
k-Means Outlier
Not outlier
Outlier
$\log_2(\mathrm{TPM}+1)$
f
POU5F1
LIN28A
PAX6
MSI1
ISL1
ELAVL4
$\log_2(\mathrm{TPM}+1)$
iPSC n=62
NPC n=69
MN n=60
phenotype
g
PCA
PC2 (6%)
PC1 (24%)
iPSC
NPC
MN
Pooled
Outlier
h
ICA
IC2
IC1
i
1. Indexing (command: outrigger index)
Input: Junction reads from data
Cell 1
Cell 2
Detect exons de novo
Output: Alternative exons
Inclusion
Exclusion
Legend
Exon-exon junction read
Annotated exons
Novel exon
Junction edge
Input: Exon definitions from annotation
Gene A
Gene B
2. Validation - optional (command: outrigger validate)
Input: Alternative exons
Remove alternative exons with non-canonical splice sites
Output: Valid alternative exons
3. Psi calculation (command: outrigger psi)
Input: Junction reads from data
Input: Alternative exons
Reject cases with insufficient junction reads
Output: Percent spliced-in (Psi/Ψ) for each cell and alternative exon
$\Psi = \frac{\text{inclusion reads}}{\text{total reads}} = \frac{\text{inclusion reads}}{\text{inclusion + exclusion reads}}$
Inclusion reads
Exclusion reads
SE
MXE
j
Case 1* 480,351
Case 2 8,920,535 (omitted for clarity)
Case 3 727,246
Case 4 951,333
Case 5 145,599
Case 6 468,539
Case 7 720,021
Case 8 56,288
Case 9a 42,871
Case 9b 8,589
Case 10a 41,958
Case 10b 1,284
Case 11a 1,327
Case 11b 1,378
$\Psi = \mathrm{NA}$
$0 \le \Psi \le 1$
*For a detailed explanation of cases, see Supplementary Software Figure 4
No. of events
k
AS Events (thousands)
iPSC
NPC
MN
l
iPSC
NPC
MN
AS events
Single cells
m
Genes
$\log_2(\mathrm{TPM}+1)$
n
# AS Events
$\log_2(\mathrm{TPM}+1)$
o
phenotype = iPSC
phenotype = NPC
phenotype = MN
Number of AS exons detected
Uniquely mapped reads (millions)

3.2.2 Assignment of single cell alternative splicing events to modalities using anchor

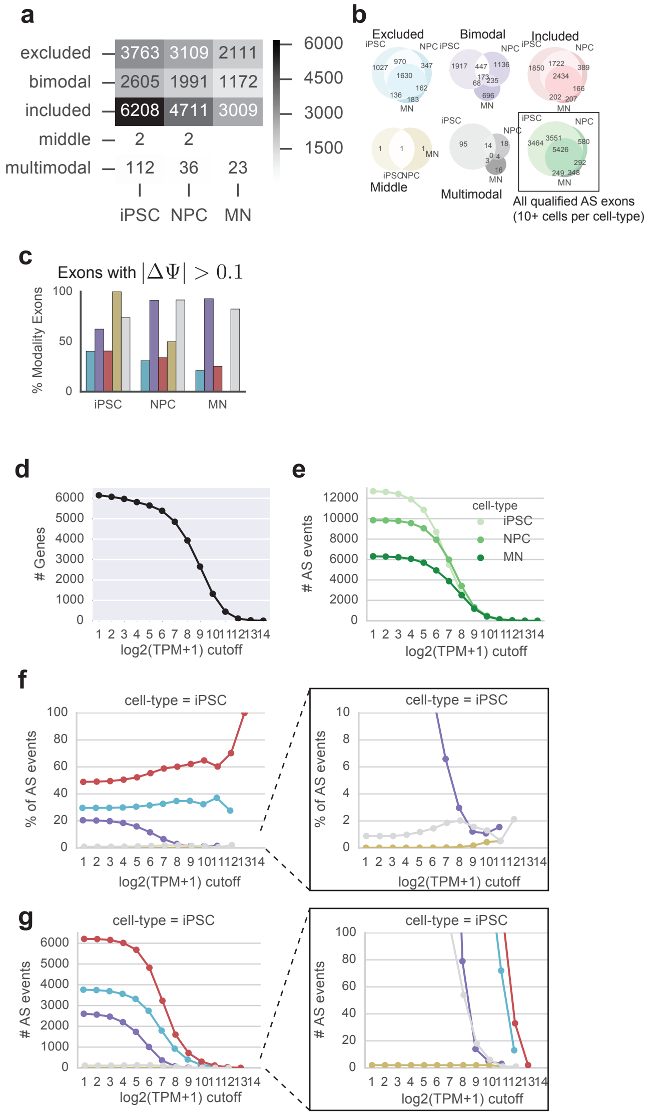
To categorize the distribution of single cell Psi values, we developed a Bayesian framework, anchor, to designate each AS exon's distribution into one of five modalities (**Figure 3.4**b): (1) *excluded*, where most cells contain the excluded isoform and Psi is close to 0; (2) *bimodal*, where two subpopulations with either the excluded (Psi near 0) or included isoform (Psi close to 1) can be observed; (3) *included*, where most cells contain the inclusion isoform (Psi close to 1); (4) *middle*, where most individual cells have both the inclusion and exclusion isoforms (Psi distribution is centered around 0.5); and (5) *multimodal*, where the distribution of inclusion and exclusion isoforms does not fit any of the previous categories (Figures 2a,b). Within each cell-type, the Psi distribution for each AS event was modeled using a Beta distribution[222]. We use a two-step process to assign modality (**Figure 3.4**c), a Bayes Factor (K) of fit was first calculated for the one-parameter models, namely included and excluded. If K did not meet the cutoff ($\log_2 K < 5$), these events are then assessed for their fit to the two-parameter models, namely middle and bimodal. Remaining events were assigned to the multimodal modality. Detection of unimodality was robust up to the addition of 50% uniform random noise (**Figure 2.10**) and bimodality was detected up to a 9:1 ratio of inclusion to exclusion, and is robust with up to 70% uniform random noise (**Figure 2.11**). Thus, we conclude that anchor is a robust classifier of alternative splicing modalities.

Figure 3.4 *(next page)*: Assignment of single cell alternative splicing events to modalities using `anchor` algorithm.
a. Schematic of SE and MXE alternative splicing events. Isoform A refers to exclusion of alternative exon (exon 2 in SE and exclusion of exon 2 (black) but inclusion of exon 3 (grey) in MXE), and isoform B refers to inclusion of alternative exon (exon 2 in SE and MXE) of alternative exon. Circles illustrate a single cell containing RNA molecules of a given AS event. Light grey represents isoform A and dark grey represents isoform B.
b. A schematic of the proposed five modalities tested by `anchor`. Distribution of Ψ for each AS event can be modeled as beta probability distribution parameterized by and . Modality of excluded (Ψ density concentrated around 0), bimodal (Ψ density concentrated towards 0 and 1), included (Ψ density around 1), middle (Ψ density around 0.5) or multimodal (Ψ density spread out uniformly across 0 to 1). The first four modalities are tested by `anchor`, and the final multimodal modality represents the null model.
c. Two-step modality assignment process is utilized by `anchor`. For the Ψ distribution of a given AS event, the Bayes Factor (K) of fit is first calculated for one-parameter models (only one of or is parameterized), including included and excluded modalities. If , modality is assigned to the modality with highest . When is not satisfied, an event will be tested in the 2nd step, in which the Bayes Factor (K) of fit is calculated for two-parameter models (where both and are parameterized), indicating bimodal and middle modalities. If an event cannot fit at either step, it will be assigned to multimodal modality. for both steps. Five events from each modality assigned by `anchor` were randomly selected as examples. **d.** Composition of AS modalities is similar in iPSCs, NPCs, and MNs. right, zoomed-in panel shows middle and multimodal modality are less than 1% in the three populations.
e. Composition of modalities of permuted splicing data. Psi scores from all identified AS events in all cells were randomly permuted 1,000 times, then `anchor` was applied to estimate modalities. Almost 100% of permuted events are assigned as bimodal. Error bars represent 95% confidence interval from 1,000 bootstrapped intervals. Right, zoomed-in panel shows low percentage of unimodal events in permuted data.


a
SE
1
2
3
MXE
1
2
3
4
Exclusion
Ψ=0
exclusion of exon2
Inclusion
Ψ=1
inclusion of exon2
SE
1 3
MXE
1 3 4
1 2 3
1 2 4
4 cells, each with
100% exclusion isoforms
4 cells, each with
100% inclusion isoforms
b
Ψ
1.0
0.5
0.0
excluded
bimodal
included
middle
multimodal
(null model)
c
Ψ distribution for a given AS
Calculate Bayes Factor of fit (K) for
step 1: one-parameter models
excluded
included
K > K_cutoff
Assign to excluded
or included modality
excluded
included
K ≤ K_cutoff
Doesn't fit
Calculate Bayes Factor of fit (K) for
step 2: two-parameter models
middle
bimodal
K > K_cutoff
Assign to middle
or bimodal modality
bimodal
middle
K ≤ K_cutoff
Doesn't fit
Assign to
multimodal modality
d
Modalities
excluded
bimodal
included
middle
multimodal
% AS Exons
60
40
20
0
iPSC
NPC
MN
1.0
0.5
0.0
e
Permuted
% AS Exons
100
80
60
40
20
0
iPSC
NPC
MN
1.0
0.8
0.6
0.4
0.2
0.0

In all three cell-types, exons within the excluded and included modalities account for 25-30% and 45-50% of all AS exons analyzed, respectively, indicating that up to 70-80% of AS events in a given cell-type exhibit unimodality (**Figure 3.4**d, **Figure 3.5**a), with events largely shared across cell-types **Figure 3.5**b). In comparison, AS events that exhibit bimodality account for up to 20% of detected AS events, whereas the middle and multimodal modalities account for less than 1% of AS events. The high-variance bimodal and multimodal events differ the most from bulk samples' AS estimates with a $\Delta\Psi$>0.1 for 40-80% of the events (**Figure 3.5**c). Simulations indicate that the observed percentages of unimodal and bimodal AS events are statistically unexpected (random permutations expect 99% bimodality and 0% unimodality; **Figure 3.4**e). As we increased the gene expression thresholds, the total number of reliably detected AS events decrease for all modalities. Yet, bimodal events continue to be observed even in the genes with the highest expression ($\log_2$ TPM > 9, **Figure 3.5**d-g), suggesting that sampling biases cannot account for the observation of bimodality. Therefore, our algorithm `anchor` estimated that most AS events are either included or excluded in single cells, with up to a fifth of events exhibiting bimodality or multimodality, which are undetected in bulk splicing analyses.

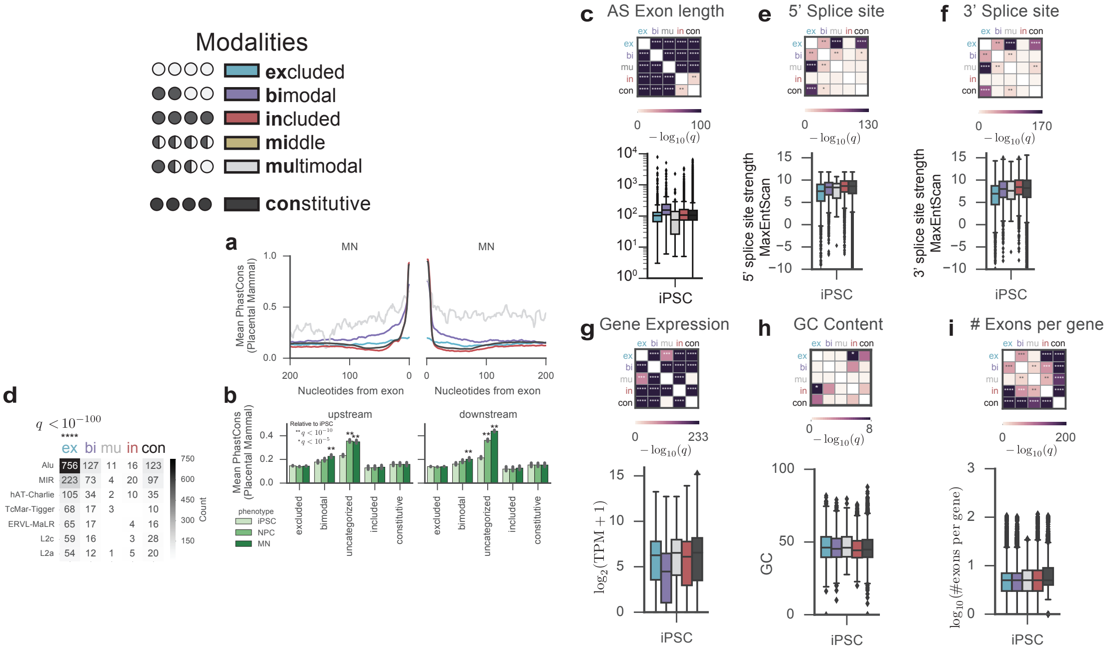
Figure 3.5 *(next page)*: Modality estimation at increasing gene expression cutoffs.
a. Summary of total number of AS events identifed by
textttoutrigger and their modality identified by
texttttanchor for each cell type.
b. Venn diagrams of events shared in modalities between cell types. AS events in included and excluded modality are largely shared across the three cell types, but fewer bimodal events are shared across three cell types. Boxed, all AS events, regardless of modality.
c. Percentage of modality AS events inconsistent with pooled estimates, where the mean difference of psi between singles and pooled ($|\Delta\bar{\Psi}|$) is greater than 0.2.
d-g. Effect of the expression level per AS event on modality estimation.
d. Number of genes remaining at the expression cutoffs.
e. Number of AS exons at varying expression cutoffs.
f. Percentage of modality estimated at different expression cutoffs (right, zoomed in panel).
g. Number of modality events estimated at different expression cutoffs (right, zoomed in panel).


a
excluded
bimodal
included
middle
multimodal
3763 3109 2111
2605 1991 1172
6208 4711 3009
2 2
112 36 23
iPSC NPC MN
6000
4500
3000
1500
b
Excluded
iPSC
NPC
MN
1027 970 347 1630 162 136 183
Bimodal
1917 447 1136 173 68 235 696
Included
1850 1722 389 2434 166 202 207
Middle
1 1 1
Multimodal
95 14 18 0 3 4 16
3464 3551 580 5426 292 249 348
All qualified AS exons
(10+ cells per cell-type)
c
Exons with $|\Delta\Psi| > 0.1$
% Modality Exons
100
50
0
iPSC NPC MN
d
# Genes
6000 5000 4000 3000 2000 1000 0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
log2(TPM+1) cutoff
e
# AS events
12000 10000 8000 6000 4000 2000 0
cell-type
iPSC
NPC
MN
log2(TPM+1) cutoff
f
cell-type = iPSC
% of AS events
100 80 60 40 20 0
10 8 6 4 2 0
log2(TPM+1) cutoff
g
cell-type = iPSC
# AS events
6000 5000 4000 3000 2000 1000 0
100 80 60 40 20 0
log2(TPM+1) cutoff

3.2.3 Splicing modalities exhibit distinct sequence and evolutionary characteristics.

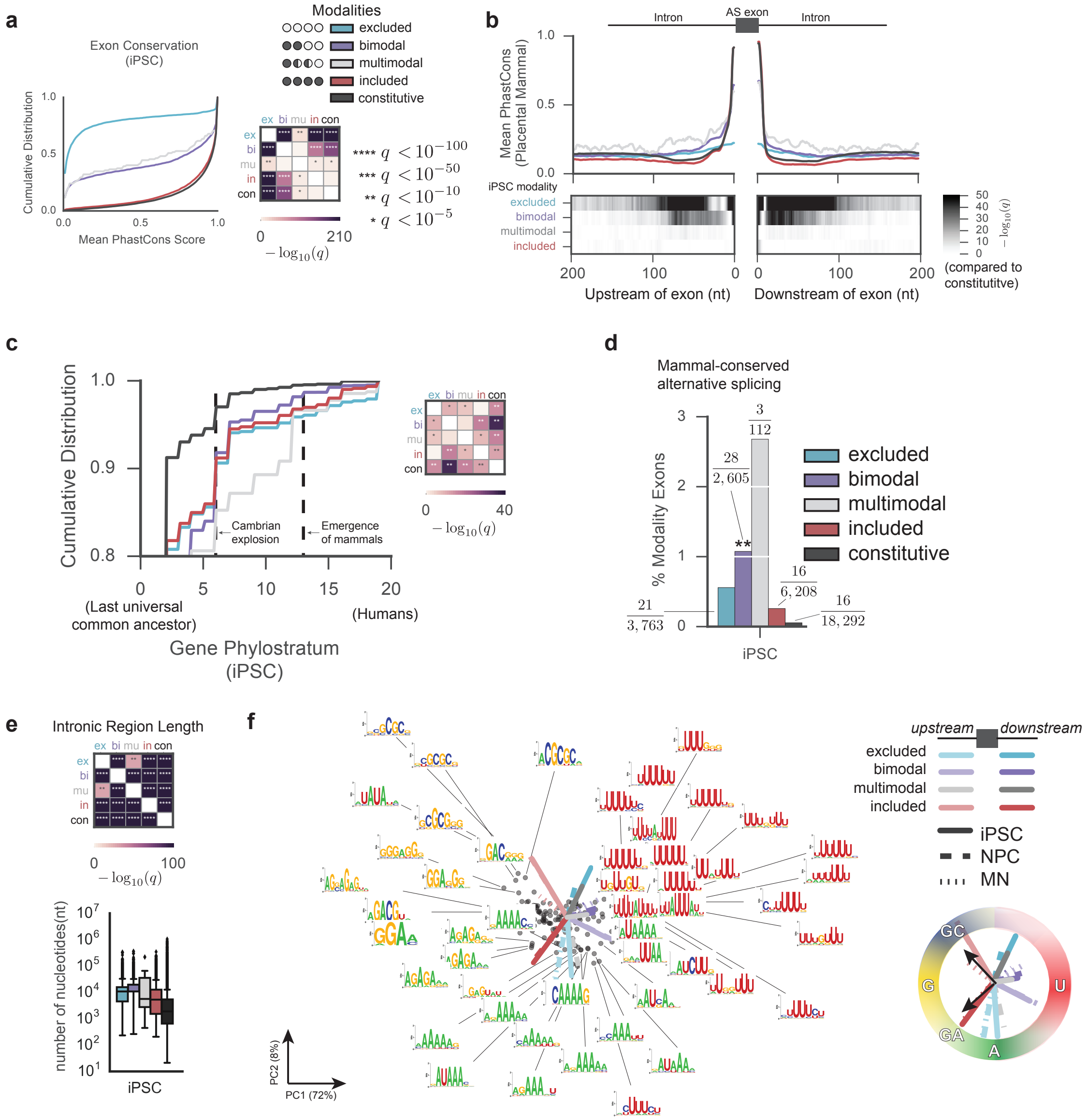
To investigate whether events in different modalities had distinct properties, we first measured the degree of evolutionary conservation of exon sequences across placental mammals. Expectedly, exon sequences within AS events in the included modality show the highest degree of sequence conservation equivalent to that of constitutive exons, whereas exons in the excluded modality are least conserved (**Figure 3.7**a). Bimodal exons exhibit an intermediate level of evolutionary conservation, which is statistically significantly different from excluded and included modalities ($q < 10^{-50}$, $q < 10^{-100}$, respectively). However, intronic sequences flanking excluded and bimodal AS are both significantly more conserved than introns flanking included or constitutive exons, a trend that increased along neural differentiation (**Figure 3.7**b and **Figure 3.6**a,b). While both excluded and bimodal introns are highly conserved, bimodal introns are more conserved in the 5-20bp window adjacent to the exon-intron junction, whereas conservation for excluded modality decreases in the same region. We also examined the evolutionary history of genes containing bimodal and multimodal exons. Human protein-coding genes have been categorized into 20 phylostrata, with archea as phylostratum 1 (ps1) and human as ps20[223]. Interestingly, 98 genes harboring multimodal and 1832 genes containing bimodal AS events are more likely found in recent phylostrata in comparison to genes containing excluded, included AS events or all genes containing any AS exon (**Figure 3.7**c). Additionally, orthologous exons of 28 bimodal and 3 multimodal AS are more frequently alternatively spliced across mammals (**Figure 3.7**d). The exon lengths and the flanking introns of bimodal AS events are significantly longer than those of the included modality and constitutive exons (**Figure 3.7**e, **Figure 3.6**c). Repetitive elements such as *Alu* are known to be stochastically exonized[224], and we find *Alu* elements more enriched in excluded exons, fewer within bimodal exons, and almost absent from AS events in the included modality (**Figure 3.6**d). Other features analyzed, including splice site strengths, GC content, showed that bimodal and multimodal exons as intermediate between excluded and included modalities (**Figure 3.6**e-i). We conclude that bimodal and multimodal events are enriched for longer flanking introns with higher conservation, present in recently evolved genes, have orthologs in mammals that are also AS events, in agreement we previous findings[225].

Figure 3.6 *(next page)*: Supplementary molecular features of each splicing modality.
a. Flanking intron sequence is more conserved in bimodal modality. Shown in motor neurons, intron conservation of bimodal events is slightly higher than excluded AS events. **b.** Barplot of mean placental mammal PhastCons score in introns flanking modality exons, across cell types. Bimodal exons in motor neurons and NPCs are statistically enriched for higher conservation as compared to iPSCs (Kolmogorov-Smirnov test, Bonferroni-corrected).
c. Significance (top) and boxplots (bottom) of the length of the alternative exons of different modalities. Constitutive exons are statistically enriched for longer exons, compared to excluded modality (Kolmogorov-Smirnov test, Bonferonni-corrected).
d. Heatmap of the number of AS events in each modality overlapping with repetitive elements with AS exons, shown in iPSC. Excluded modality is statistically enriched for overlap ($q < 10^{-50}$, Hypergeometric test).
e. Significance (top) and boxplots (bottom) of the 5′ splice site scores of the exon, specifically the splice donor site as measured by MaxEntScan. Bimodal and excluded exons have statistically significantly lower splice site scores than included exons (Kolmogorov-Smirnov test, Bonferonni-corrected).
f. Significance (top) and boxplots (bottom) of the 3′ splice site scores of the exon, specifically the splice acceptor site as measured by MaxEntScan. Bimodal and excluded exons have statistically significantly lower splice site scores than included exons (Kolmogorov-Smirnov test, Bonferonni-corrected).
g. Significance (top) and boxplots (bottom) of the mean expression level of genes ($\log_2(\mathrm{TPM}+1)$, x axis) harboring corresponding AS events in each modality. While events from all five modalities are detected across entire range of gene expression, genes containing bimodal exons are statistically enriched for lower expression (Kolmogorov-Smirnov test, Bonferonni-corrected).
h. Significance (top) and boxplots (bottom) of the GC content of the alternative exons of different modalities. Excluded exons are statistically enriched for higher GC content, compared to included exons (Kolmogorov-Smirnov test, Bonferonni-corrected).
i. Significance (top) and boxplots (bottom) of the number of exons per gene harboring corresponding modalities, measured by the maximum number of genes in any transcript of a gene. Genes containing excluded exons are statistically enriched for fewer exons per gene (Kolmogorov-Smirnov test, Bonferonni-corrected).



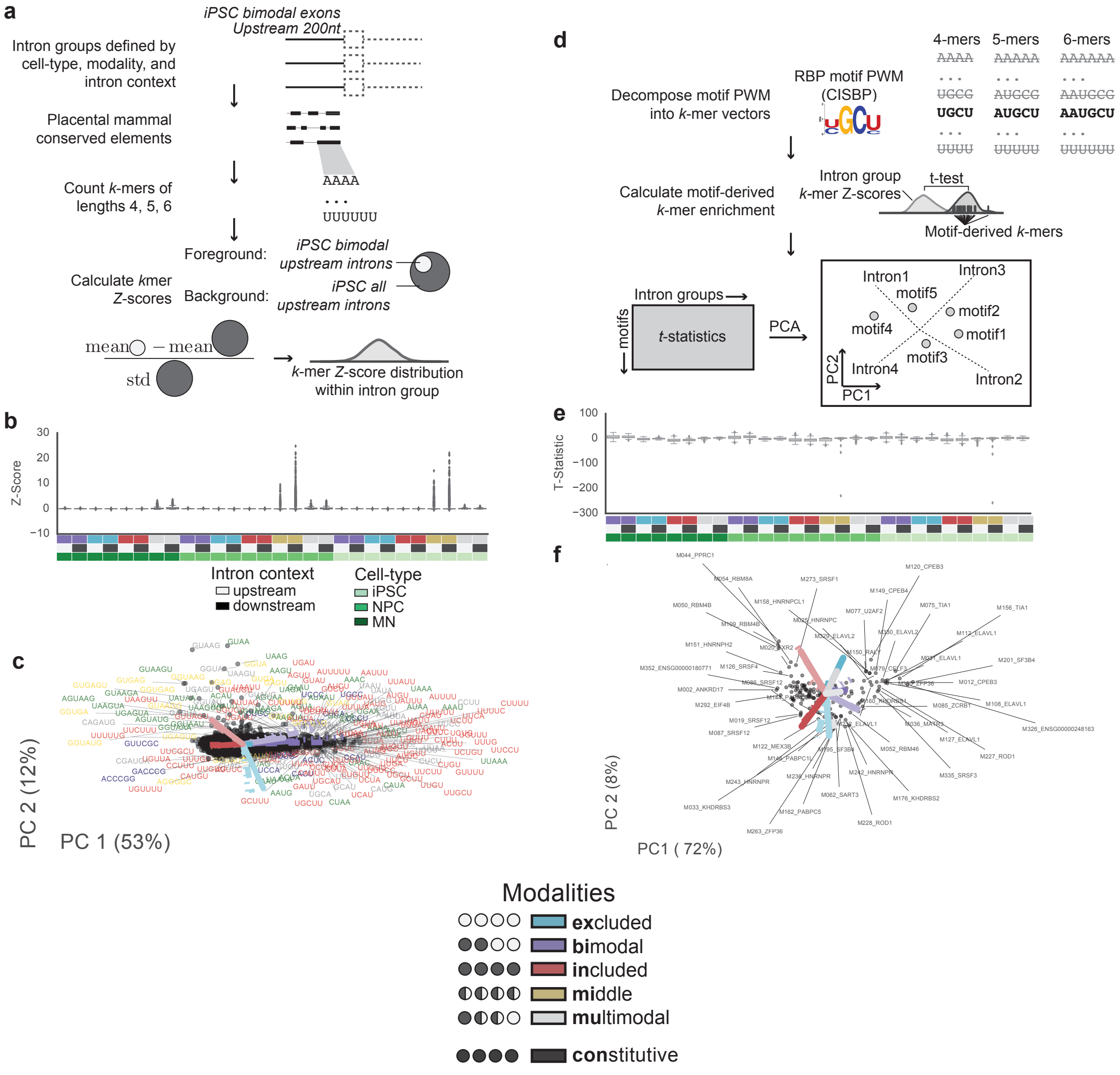
Modalities
excluded
bimodal
included
middle
multimodal
constitutive
a
MN
MN
Mean PhastCons (Placental Mammal)
1.0
0.5
0.0
200 100 0
0 100 200
Nucleotides from exon
Nucleotides from exon
b
upstream
downstream
Relative to iPSC
$^{**}q < 10^{-10}$
$^{*}q < 10^{-5}$
Mean PhastCons (Placental Mammal)
0.4
0.2
0.0
excluded
bimodal
uncategorized
included
constitutive
phenotype
iPSC
NPC
MN
d
$q < 10^{-100}$
****
ex bi mu in con
Alu 756 127 11 16 123
MIR 223 73 4 20 97
hAT-Charlie 105 34 2 10 35
TcMar-Tigger 68 17 3 10
ERVL-MaLR 65 17 4 16
L2c 59 16 3 28
L2a 54 12 1 5 20
Count
750
600
450
300
150
c AS Exon length
e 5' Splice site
f 3' Splice site
g Gene Expression
h GC Content
i # Exons per gene
ex bi mu in con
$-\log_{10}(q)$
0 100
0 130
0 170
0 233
0 8
0 200
10^0 10^1 10^2 10^3 10^4
5' splice site strength MaxEntScan
3' splice site strength MaxEntScan
15 10 5 0 −5 −10
$\log_2(\text{TPM}+1)$
GC
100 50 0
$\log_{10}(\text{\#exons per gene})$
3 2 1 0
iPSC

Figure 3.7 *(next page)*: Bimodal AS events exhibit distinct sequence and evolutionary features.
All results are shown for iPSCs that have highest number of AS events (12,690). Results are similar in three cell-types, except where indicated. All q-values of significance were derived from multiple hypothesis corrected (Bonferonni) non-parametric Mann-Whitney U test, unless otherwise indicated.
a. Cumulative distributions of the mean Placental Mammal PhastCons score in each modality are shown, with constitutive exons as comparison. AS exons from included modality (red) are as conserved as constitutive exons (black), while excluded exons (blue) are least conserved, followed by bimodal (purple) and multimodal (grey) exons. right, heatmap of pairwise significance scores between each modality or constitutive exons (right panel).
b. Mean Placental Mammal PhastCons scores of flanking intronic regions of exons in excluded (blue) bimodal (purple), multimodal (grey), included (red) modalities, and constitutive (black) exons in all cell-types. bottom, heatmap of base-wise significance of PhastCons scores is presented 0 < for clarity.
c. Phylostratum scores are summarized for genes harboring AS events in each modality together with genes containing constitutive exons. right, heatmap of pairwise significance scores between each modality or constitutive exons.
d. Alternative splicing events conserved in mammals were extracted from Merkin et al, 2012 (Merkin et al., 2012) and their percentage among each modality is calculated. Hypergeometric test (multiple hypothesis corrected with Bonferonni) indicated q < 10-5 statistical significance. Fraction indicates No. of conserved events in each modality(nominator)/total events in the modality(denominator)
e. Intron lengths summarized in excluded, bimodal, multimodal, included modality together with constitutive exons. top, heatmap of pairwise significance scores between each modality or constitutive exons.
f. Conserved intronic sequences in each modality are enriched with distinct nucleotides. Motifs enriched for each modality are presented by PCA, shown with each circle as a motif and the vectors as component loadings of intron groups. left, Representative motifs are annotated with logos from the CISBP database. right, A simplified illustration of distinct nucleotide enrichment in each intron group. An interactive version of this plot is available at https://plot.ly/~OlgaBotvinnik/32/cisbp-motif-t-test-enrichments-background-phenotype/


a
Exon Conservation
(iPSC)
Cumulative Distribution
Mean PhastCons Score
Modalities
excluded
bimodal
multimodal
included
constitutive
ex bi mu in con
**** $q < 10^{-100}$
*** $q < 10^{-50}$
** $q < 10^{-10}$
* $q < 10^{-5}$
0
210
$-\log_{10}(q)$
b
Intron
AS exon
Intron
Mean PhastCons
(Placental Mammal)
iPSC modality
excluded
bimodal
multimodal
included
Upstream of exon (nt)
Downstream of exon (nt)
(compared to
constitutitve)
$-\log_{10}(q)$
c
Cumulative Distribution
Cambrian
explosion
Emergence
of mammals
(Last universal
common ancestor)
(Humans)
Gene Phylostratum
(iPSC)
ex bi mu in con
0
40
$-\log_{10}(q)$
d
Mammal-conserved
alternative splicing
% Modality Exons
$\frac{3}{112}$
$\frac{28}{2,605}$
$\frac{16}{6,208}$
$\frac{16}{18,292}$
$\frac{21}{3,763}$
**
iPSC
excluded
bimodal
multimodal
included
constitutive
e
Intronic Region Length
ex bi mu in con
0
100
$-\log_{10}(q)$
number of nucleotides(nt)
iPSC
f
upstream
downstream
excluded
bimodal
multimodal
included
iPSC
NPC
MN
PC2 (8%)
PC1 (72%)
GC
G
GA
A
U

Next, we asked whether there are *cis*-regulatory elements within flanking intronic sequences. Position weight matrices (PWMs) for motifs recognized by RBPs were obtained from the CISBP motif database[226] and transformed into k-mers[227]. We defined an intron group as 200 intronic bases upstream or downstream of alternative exons of a specific modality and cell-type. Within each intron group, we calculated Z-scores of k-mer enrichment (**Figure 3.8**a,b). By PCA analysis, we found bimodal and included modalities are separated on the first principal component (PC1) and enriched for U-rich and G-rich sequences, respectively (**Figure 3.8**c). Curious whether such U-G division is present at the motif level, enriched motifs were identified by calculating a t-statistic between the motif-derived k-mer Z-scores against the Z-scores of all identified k-mers in the same intron group (**Figure 3.8**d,e). We then subjected the t-statistics of motif-derived k-mer enrichments in each intron group to PCA (**Figure 3.7**f, **Figure 3.8**f). Principal component 1 (PC1) explains 72% of the variance of k-mer enrichment and readily separates the included modality from bimodal modality. Meanwhile, principal component 2 (PC2) distinguishes motifs located upstream or downstream of the alternative exons and account for 8% of total variance. Consistent with k-mer results, bimodal and included modalities are enriched for U-rich and G-rich motifs, respectively, regardless of the cell-types. Moreover, upstream intronic sequences of included modality are enriched for GC and the downstream counterpart are enriched for GA motifs (**Figure 3.7**f, right). This finding suggests that the sequence properties of the introns, together with the trans-factors associated with these motifs distinguish each AS modality, independent of cell-type. Together, our results reveal that exons with highly variant AS events have sequence and evolutionary attributes distinct from other modalities.

Figure 3.8 *(next page)*: Sequence enrichment of modality introns.
a. Overview of defining “Intron groups” defined by cell-type, modality, and intron context, and process for obtaining their conserved k-mer Z-scores.
b. Boxplots of the Z-scores of k-mer enrichment in the different intron groups, labeled with a colorbar of modality, intron context, and cell-type.
c. PCA on k-mer Z-scores, with each point as a k-mer and the vector components as the introns. k-mers with principal comoponent greater than 2.5 standard deviations away from zero were labeled with the sequence, colored by the majority nucleotide. If there was a tie for the majority nucleotide, it was assigned the color grey. An interactive version of this plot can be viewed here: https://plot.ly/~OlgaBotvinnik/20/modality-k-mer-Z-scores-background-phenotype/. Multimodal is not shown because its k-mer enrichment has a much larger range than the other modalities and overwhelms the plot.
d. Overview of motif enrichments calculated from intron groups using a t-test and their transformation into PCA for visualization.
e. Boxplots of the t-statistics of motif enrichment in different intron groups, labeled with colorbars of modality, intron context, and cell-type.
f. PCA on the t-statistics of the Motif enrichment, labeled with the motif ID and RPB name from CISBP v0.6. An interactive version of this plot is available at https://plot.ly/~OlgaBotvinnik/32/cisbp-motif-t-test-enrichments-\
background-phenotype/


a
iPSC bimodal exons
Upstream 200nt
Intron groups defined by cell-type, modality, and intron context
Placental mammal conserved elements
Count k-mers of lengths 4, 5, 6
AAAA
. . .
UUUUUU
Calculate kmer Z-scores
Foreground:
iPSC bimodal upstream introns
Background:
iPSC all upstream introns
mean ○ − mean ●
std ●
k-mer Z-score distribution within intron group
b
Z-Score
30
20
10
0
−10
Intron context
upstream
downstream
Cell-type
iPSC
NPC
MN
c
PC 2 (12%)
PC 1 (53%)
GUAAG
GUAA
UAAG
GUAAGU
GUGAGU
GUGAG
AGUAAG
GGUGA
CAGAUG
UUUUUG
GGUAGG
GUUCGC
ACCCGG
UGUUUU
GACCCG
GCUUU
UGCUU
CUAA
AUUUUU
AAUUU
UUAUU
CUUUU
UUAAA
UUGCU
d
Decompose motif PWM into k-mer vectors
RBP motif PWM (CISBP)
4-mers
5-mers
6-mers
~~AAAA~~
~~AAAAA~~
~~AAAAAA~~
~~UGCG~~
~~AUGCG~~
~~AAUGCG~~
UGCU
AUGCU
AAUGCU
~~UUUU~~
~~UUUUU~~
~~UUUUUU~~
Calculate motif-derived k-mer enrichment
Intron group k-mer Z-scores
t-test
Motif-derived k-mers
Intron groups
motifs
t-statistics
PCA
Intron1
Intron3
motif5
motif2
motif4
motif1
motif3
Intron4
Intron2
PC2
PC1
e
T-Statistic
100
0
−100
−200
−300
f
PC 2 (8%)
PC1 (72%)
M044_PPRC1
M054_RBM8A
M273_SRSF1
M120_CPEB3
M149_CPEB4
M050_RBM4B
M158_HNRNPCL1
M075_TIA1
M156_TIA1
M077_U2AF2
M025_HNRNPC
M109_RBM4B
M030_ELAVL2
M112_ELAVL1
M151_HNRNPH2
M201_SF3B4
M352_ENSG00000180771
M126_SRSF4
M012_CPEB3
M002_ANKRD17
M292_EIF4B
M085_ZCRB1
M108_ELAVL1
M019_SRSF12
M326_ENSG00000248163
M087_SRSF12
M127_ELAVL1
M122_MEX3B
M052_RBM46
M227_ROD1
M335_SRSF3
M242_HNRNPR
M243_HNRNPR
M062_SART3
M176_KHDRBS2
M033_KHDRBS3
M162_PABPC5
M228_ROD1
M263_ZFP36
Modalities
○○○○ excluded
●●○○ bimodal
●●●● included
◐◐◐◐ middle
●◐◐○ multimodal
●●●● constitutive

3.2.4 Cell-type specific AS are largely comprised of high variance events.

We next asked whether there are AS events that change modalities during the differentiation of iPSCs to NPCs or MNs (**Figure 3.9**a, **Figure 3.10**a). To our surprise, we find that only 20% of AS events shared between pluripotent stem cells and the neuronal derivatives exhibit a change in modality ($q < 10^{-100}$, hypergeometric test, corrected for multiple hypothesis testing). As these events have a unique modality in each cell-type, they are cell-type specific. Less than a quarter (18%) of the AS events detected in two cell-types (iPSCs and NPCs or iPSCs and MNs) exhibited a change in modality (**Figure 3.9**b), At least 98% of these switching events are comprised of bimodal AS events (**Figure 3.9**c). As cells transition from iPSCs to NPCs or to MNs, 66% and 72% of the unimodal events became bimodal or multimodal, and conversely, 34% and 27% of bimodal events switched to a unimodal modality. These “switching” AS events are enriched for GO functional categories, such as ′protein localization or transportation,′ and ′RNA processing′ (**Figure 3.10**b). Thus, we conclude that bimodal and multimodal AS events likely play an important role in cell-type specificity and are more malleable during differentiation, in contrast to included and excluded events.

Figure 3.9 *(next page)*: Dynamic AS events are primarily contributed by highly variant bimodal and multimodal events.

a. AS events change modalities during iPSC to MN transition. A total of 5,675 AS events was identified as common ones in both iPSCs and MNs. The compartmentalization of these common events in five modalities is presented in iPSCs (y-axis) against their corresponding modalities in MNs (x-axis). Gradient of heat map represents the percent of events in the iPSC modality row, annotated with the exact number of events. The diagonal indicates events remained in the same modality. Notably, 88% of excluded events in iPSCs remained in excluded modality, and 86% of included events in iPSCs remained as included in MNs. In contrast 52% of bimodal events in iPSCs switch to either included or excluded modalities in MNs. Multiple hypothesis corrected (Bonferonni) hypergeometric tests were used to calculate significance.

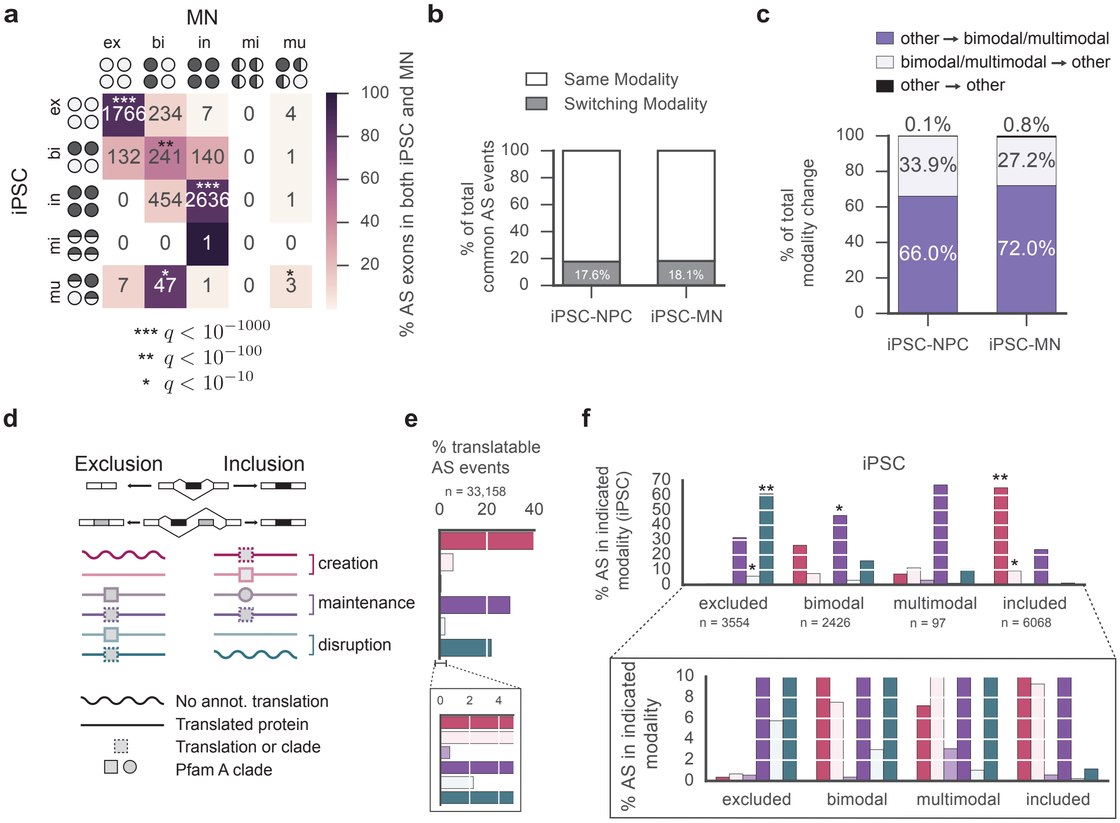
b. During the differentiation from iPSCs to MNs or from iPSCs to NPCs, we found 1,586 (17.6%) or 1,029 (18.0%) AS events switched modality, respectively. **c.** Within the switching events, 99% events either switched from a bimodal/multimodal state or switched towards a bimodal/multimodal state. Around 1% of switching events were observed among other types of modality changes.

d-f. AS events in bimodal modality exhibits flexibility in protein coding. **d.** Schematic of predicted translation changes associated with AS exon inclusion.

Exclusion and inclusion of AS exon is termed as Isoform A and Isoform B, respectively. Six categories of coding outcomes are depicted when the isoform switch occurs. Pink, highlights creation of translated proteins or protein domain clades when AS exon is included. Purple, represents maintenance of protein clades with or without change of domain clades. Blue, represents loss of domain clades or disruption of translation when AS exon become included. The square and circle illustrate different Pfam domain clades. The square with dashed outline represents translated protein, possibly containing a Pfam domain clade.

e. The coding outcomes are summarized in the six categories based on all AS events. The percentage of each translation configuration is used as the background distribution for significance calculations in **f**.

f. AS events in bimodal modality favor protein and domain maintenance. The dominant isoforms in included and excluded modalities favor protein or domain creation and switching to the other isoform results in overwhelming disruption of protein coding. Enrichment is calculated against population average (shown in e) in each category using multiple hypothesis test corrected hypergeometric tests. *: $q < 10^{-10}$ **: $q < 10^{-100}$


a
MN
ex
bi
in
mi
mu
iPSC
ex
bi
in
mi
mu
1766
234
7
0
4
132
241
140
0
1
0
454
2636
0
1
0
0
1
0
0
7
47
1
0
3
100
80
60
40
20
% AS exons in both iPSC and MN
*** $q < 10^{-1000}$
** $q < 10^{-100}$
* $q < 10^{-10}$
b
Same Modality
Switching Modality
% of total common AS events
100
80
60
40
20
0
17.6%
18.1%
iPSC-NPC
iPSC-MN
c
other → bimodal/multimodal
bimodal/multimodal → other
other → other
% of total modality change
100
80
60
40
20
0
0.1%
0.8%
33.9%
27.2%
66.0%
72.0%
iPSC-NPC
iPSC-MN
d
Exclusion
Inclusion
creation
maintenance
disruption
No annot. translation
Translated protein
Translation or clade
Pfam A clade
e
% translatable AS events
n = 33,158
0
20
40
0
2
4
f
iPSC
% AS in indicated modality (iPSC)
70
60
50
40
30
20
10
0
excluded
n = 3554
bimodal
n = 2426
multimodal
n = 97
included
n = 6068
% AS in indicated modality
10
8
6
4
2
0
excluded
bimodal
multimodal
included

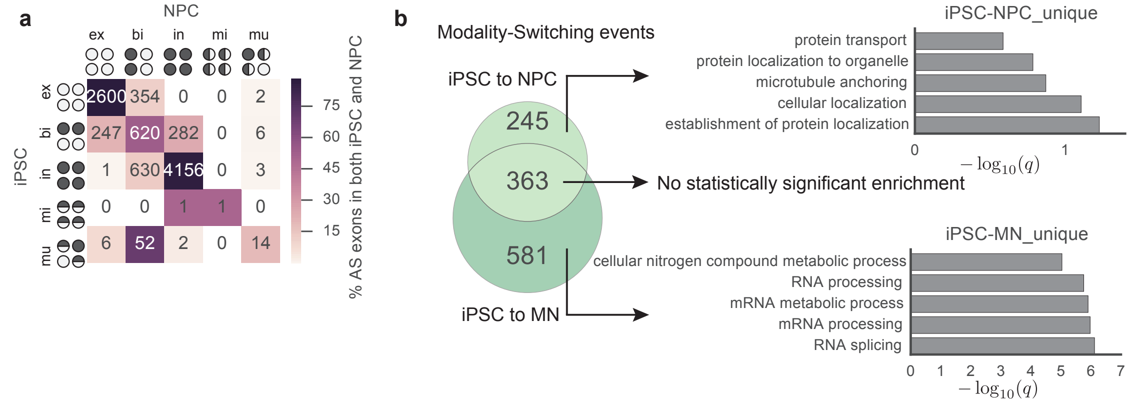


Figure 3.10: Switching AS events are enriched for transcriptome and post-transcriptional regulation GO terms.
a. AS events change modalities during iPSC to NPC transition. A total of 7,962 AS events was identified as common events in both iPSCs and MNs. Notably, ≈ 82% of excluded events in iPSCs remained in excluded modality, and ≈ 84% of included events in iPSCs remained as included in NPCs. In contrast 42% of bimodal events in iPSCs switch to either included or excluded modalities in NPCs.
b. Of the common events shared by all three populations, the events changing between iPSCs to NPCs (light green) and iPSCs to MNs (dark green). Venn diagram show the overlap between the two sets of switching AS events and GO function terms for each section of switching events.

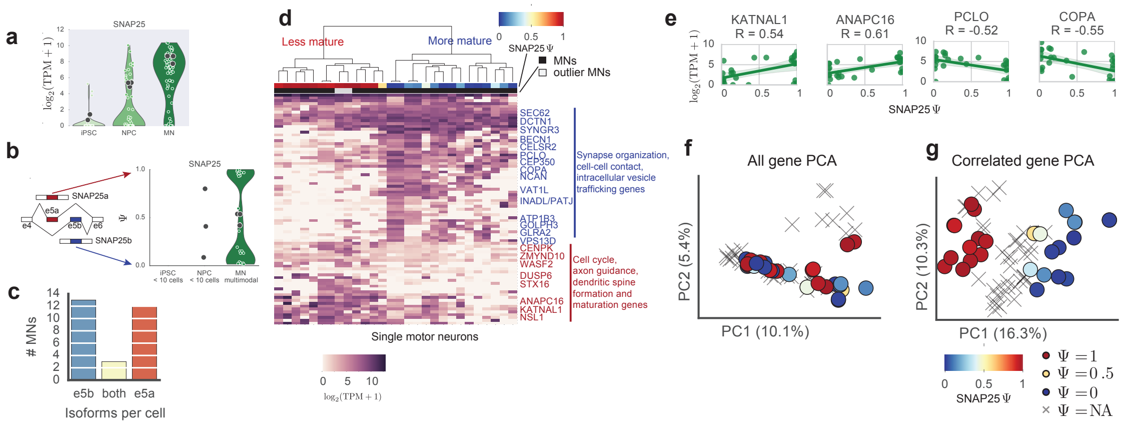
Since bimodal and multimodal events are more dynamic, we asked whether they are more likely to preserve protein-coding capacity. For simplicity, the transcripts with excluded and included AS exons are designated as isoform A and isoform B, respectively (**Figure 3.9**d). We required that at least one isoform is a GENCODE-annotated coding transcript and utilized hmmscan[228;229] to search Pfam[230;231] for protein domain clades (**Figure 3.9**e). Both included and excluded modality exons were enriched for the presence of known protein domain clades in their dominant isoform ($q < 10^{-10}$, hypergeometric test corrected for multiple hypothesis testing). Switching to the other isoform either disrupted the reading frame or the functional protein domain, underscoring the importance of maintaining their dominant isoform. Surprisingly, the bimodal and multimodal

AS events appear to balance domain creation, maintenance and disruption between isoforms. In particular, 65% of multimodal and 50% of bimodal events result in domain maintenance where a functional domain has been exchanged or preserved, in contrast to 15-30% of excluded and included modalities (**Figure 3.9**f). Thus, the highly variant AS events adapt their coding capacity during differentiation.

3.2.5 Highly variant AS events can reveal subpopulations invisible to gene expression analysis

As highly variant bimodal and multimodal AS events appear to be most sensitive to differentiation, we surmised that they can provide an opportunity to identify subpopulations that were otherwise invisible when analyzing gross expression differences in single cell RNA-seq data. To illustrate, SNAP25 (synaptosomal-associated protein 25) is a presynaptic plasma membrane protein of the trans-SNARE complex that mediates synaptic vesicle membrane docking and fusion. Mutually exclusive exons 5a and 5b are characterized as a high variance multimodal event in MNs (**Figure 3.11**a-c, **Figure 3.13**a). Exon 5b is more included in adult brain[232] which may facilitate faster exocytosis[233]. We identified genes that correlated with the Psi values of this event (Spearman correlation $|R| > 0.5$; **Figure 3.13**b). The correlated genes separated the MNs into two clusters that correspond to Psi values of greater than 0.5 or less than 0.5 (**Figure 3.11**d-g). Excitingly, MNs which included exon 5a ($\Psi > 0.5$) are enriched for genes essential in cytoskeletal reorganization required for axon guidance and dendritic spine formation and maturation, such as KATNAL1, ZMYND10, WASF2 and STX16. They also express genes associated with repression of cell proliferation (**Figure 3.11**d, red labels). Thus, MNs utilizing exon 5a are less 'mature', may have recently exited cell proliferation and are forming synapses. In contrast, MNs that included exon 5b ($\Psi < 0.5$) are enriched with many genes associated with synapse organization and synaptic vesicle trafficking, such as SYNGR3, DCTN1, COPA and PCLO, as well as plasma membrane receptors and cell-cell contact genes such as CELSR2, INADL/PATJ, ATP1B3, and GLRA2. At the same time, these MNs expressed multiple genes associated with intracellular vesicle trafficking (**Figure 3.11**d, blue labels), reflecting a more mature neuronal state with active protein transport and vesicle trafficking (**Figure 3.11**d). Finally, genes correlating with Psi scores are able to separate the two subgroups in PCA, whereas a complete list of expressed genes from MNs fail to do so (**Figure 3.11**f, g). Thus, the variation of the MXE event in SNAP25 reveals substructure in MN populations.

Figure 3.11 *(next page)*: Mulitmodal AS event in SNAP25 reveals subpopulations invisible by gene expression alone.
a-g. SNAP25 alternative splicing reveals a more mature subpopulation in motor neuron population.
a. SNAP25 is primarily expressed in MNs.
b. Usage of alternative exon 5 (a MXE containing exon 5a and exon 5b) in the three populations. Shown is the usage of alternative exon 5a of SNAP25.
c. Summary of exon 5 usage in motor neurons.
d. Preferential usage of exon 5a or exon 5b of SNAP25 in MNs reveals intricate cell states. Genes correlated with the Psi score of this MXE in SNAP25 (above an empirical threshold) were used to cluster all MNs containing this event. Two main subgroups are observed, one with Psi close to 1 (red in the legend bar), the other with Psi close to 0 (blue in the legend bar). Cells with Psi around 0.5 are illustrated with yellow. Black and light grey indicate qualified and outlier MNs based on k-means clustering, respectively. Gradient of purple indicates gene expression in $\log_2(\mathrm{TPM}+1)$, with darker being highly expressed. A few representative genes from the two subgroups are highlighted.
e. Examples of representative genes correlating with Psi of this MXE in SNAP25. KATNAL1 and ANAPC16 are more enriched in the cells with $\Psi \approx 1$. DCTN1 and PCLO are more enriched in the cells with $\Psi \approx 0$. X-axis represents the Psi score, and y-axis represent gene expression in $\log_2(\mathrm{TPM}+1)$. Each MN is depicted as a green circle. Solid green line represents simple linear regression line between Psi and the expression of indicated genes. Shaded green represents 95% confidence interval of the regression.
f-g. Genes correlating with this MXE event distinguish the two subgroups of MNs. Each MN is depicted as a dot in PCA. Red: cells with $\Psi \approx 1$; blue: $\Psi \approx 0$; yellow: $\Psi \approx 0.5$; X: cells with a Psi assigned as NA.
f. PCA of all expressed genes in MNs failed to separate the two subgroups.
g. Using only the genes correlated with Psi of the MXE in SNAP25, two subgroups are readily separated. Percentage of variance explained are labeled at each PC.


a
SNAP25
$\log_2(\mathrm{TPM}+1)$
0
2
4
6
8
10
12
iPSC
NPC
MN
b
SNAP25a
e5a
e4
e5b
e6
SNAP25b
SNAP25
Ψ
0.0
0.5
1.0
iPSC
< 10 cells
NPC
< 10 cells
MN
multimodal
c
# MNs
0
2
4
6
8
10
12
14
e5b
both
e5a
Isoforms per cell
d
Less mature
More mature
0
0.5
1
SNAP25 Ψ
■ MNs
□ outlier MNs
SEC62
DCTN1
SYNGR3
BECN1
CELSR2
PCLO
CEP350
COPA
NCAN
VAT1L
INADL/PATJ
ATP1B3
GOLPH3
GLRA2
VPS13D
CENPK
ZMYND10
WASF2
DUSP6
STX16
ANAPC16
KATNAL1
NSL1
Synapse organization, cell-cell contact, intracellular vesicle trafficking genes
Cell cycle, axon guidance, dendritic spine formation and maturation genes
Single motor neurons
0
5
10
$\log_2(\mathrm{TPM}+1)$
e
KATNAL1
R = 0.54
ANAPC16
R = 0.61
PCLO
R = -0.52
COPA
R = -0.55
$\log_2(\mathrm{TPM}+1)$
0
5
10
0
0.5
1
SNAP25 Ψ
f
All gene PCA
PC2 (5.4%)
PC1 (10.1%)
g
Correlated gene PCA
PC2 (10.3%)
PC1 (16.3%)
0
0.5
1
SNAP25 Ψ
● $\Psi = 1$
● $\Psi = 0.5$
● $\Psi = 0$
× $\Psi = \mathrm{NA}$

As another example, we observed a SE event from DYNC1I2 (Dynein Cytoplasmic 1 Intermediate Chain 2), which is bimodal in both iPSCs and NPCs (**Figure 3.12**a-f, **Figure 3.13**c). DYNC1I2 encodes a non-catalytic component of the cytoplasmic dynein 1 complex, which acts as a retrograde microtubule motor to transport organelles and vesicles[234]. NPCs were clustered into two groups by genes that correlated with Psi scores of the SE exon (**Figure 3.12**c,d). The subgroup with $\Psi \approx 1$ are enriched for genes associated with a variety of mature neuronal genes, such as ONECUT2, a generic transcription factor of motor neurons and numerous genes related with axon guidance and cytoskeleton reorganization (**Figure 3.12**c). This subgroup is also enriched for multiple neuron-specific RNA binding proteins (RBPs), including ELAVL2-4 and SRRM4. On the other hand, the subgroup of NPCs with $\Psi \approx 0$ is strongly enriched with genes associated with cell division, DNA replication and translation. Again, in contrast to all genes detected in NPCs, only genes correlating with Psi scores reveal the substructures of NPC population in PCA (**Figure 3.12**e,f). Thus, the bimodality of this SE event is a sufficient statistic to delineate NPCs into a more proliferative subgroup ($\Psi \approx 1$) consistent with their progenitor fate and a subgroup ($\Psi \approx 0$) that appears farther on the trajectory of neuronal fate.

Figure 3.12 *(next page)*: Bimodal AS event in DYNC1I2 reveals subpopulations invisible by gene expression alone.

a-m. A bimodal SE event in DYNC1I2 as an example to dissect NPCs into a more proliferating subgroup and a subgroup on the trajectory of neuronal differentiation.

b. Expression of DYNC1I2 in the three populations.

c. Psi distribution of a SE event in DYNC1I2 in the three populations. This event is bimodal in both iPSCs, NPCs and becomes included in MNs.

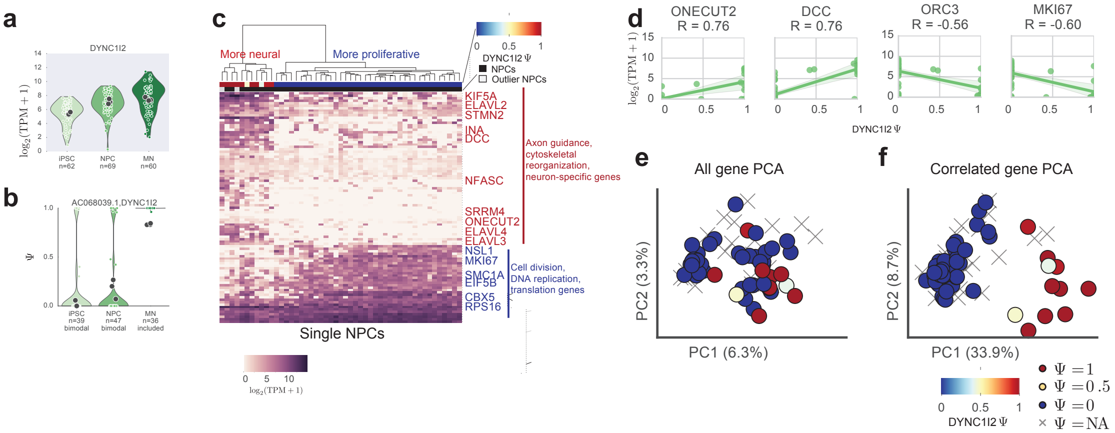
d. Genes correlating with Psi of this SE event is able to cluster the NPCs into two subgroups. Rows represent the genes and columns represent single cells in NPCs. Genes detected in NPC and correlated with Psi (Spearman $R > 0.5$). Green: NPC. Blue: cells with Psi around 0. Red: cells with Psi around 1. Light Blue to yellow: cells with Psi around 0.5. Black and grey: cells designated as qualified cells versus outlier-cells based on k-means clustering. Representative genes enriched in the two subgroups are highlighted in blue or red.

e. Example genes enriched in the two subgroups of NPCs. ONECUT2 and DCC are more highly expressed in cells with $\Psi \approx 1$; ORC3 and MKI67 are more highly expressed in cells with $\Psi \approx 0$. Psi scores of the SE in DYNC1I2 is plot on x-axis and expression of indicated genes is plotted on y-axis.

f-g. Only genes correlating with Psi are able to separate two subgroups in NPCs, with each NPC depicted as a dot in the PCA. Red: cells with $\Psi \approx 1$; blue: $\Psi \approx 0$; yellow: $\Psi \approx 0.5$; X: cells with a Psi assigned as NA.

f. PCA of all expressed genes in NPCs failed to separate the two subgroups.

g. Genes correlating with Psi are able to segregate the two subgroups by PCA.


a
DYNC1I2
$\log_2(\mathrm{TPM}+1)$
14
12
10
8
6
4
2
0
iPSC n=62
NPC n=69
MN n=60
b
AC068039.1,DYNC1I2
Ψ
1.0
0.5
0.0
iPSC n=39 bimodal
NPC n=47 bimodal
MN n=36 included
c
More neural
More proliferative
0
0.5
1
DYNC1I2 Ψ
■ NPCs
□ Outlier NPCs
KIF5A
ELAVL2
STMN2
INA
DCC
NFASC
SRRM4
ONECUT2
ELAVL4
ELAVL3
NSL1
MKI67
SMC1A
EIF5B
CBX5
RPS16
Axon guidance, cytoskeletal reorganization, neuron-specific genes
Cell division, DNA replication, translation genes
Single NPCs
0
5
10
$\log_2(\mathrm{TPM}+1)$
d
ONECUT2 R = 0.76
DCC R = 0.76
ORC3 R = -0.56
MKI67 R = -0.60
$\log_2(\mathrm{TPM}+1)$
15
10
5
0
0
0.5
1
DYNC1I2 Ψ
e
All gene PCA
PC2 (3.3%)
PC1 (6.3%)
f
Correlated gene PCA
PC2 (8.7%)
PC1 (33.9%)
0
0.5
1
DYNC1I2 Ψ
● Ψ = 1
● Ψ = 0.5
● Ψ = 0
× Ψ = NA

Lastly, we examined how the multimodal MXE event containing e9 and e10 in PKM distinguishes cell states in MNs. Notably, MNs were partitioned into three subgroups by genes that correlated with the Psi score of this event (**Figure 3.13**d-f). The first subgroup is primarily composed of outlier MNs previously characterized by k-means clustering and PCA, which prefers inclusion of exon 9 ($\Psi < 0.5$) and is enriched with genes related to cell proliferation or signaling in progenitor cells (**Figure 3.13**d, labeled in light blue). The second subgroup represents MNs also preferring exon 9 ($\Psi < 0.5$), but have lower expression of progenitor genes, and have not expressed neuron-specific genes (**Figure 3.13**d, labeled in dark blue). The third subgroup MNs using exon 10 ($\Psi > 0.5$) is highly enriched with neuron-specific genes (**Figure 3.13**d, labeled in red) confirming their motor neuron fate. Therefore, a single variance event in PKM provides a sufficient information that unravels distinct cell states (**Figure 3.13**f). Many additional examples were found including AS exons in SUGT1, BRD8, MDM4, MEAF6, and RPN2 that demonstrate that high variance AS events extracted from single cells offer an additional layer of information to demarcate cell states that are otherwise hidden in overall gene expression (**Figure 3.14**a-i).

Figure 3.13 *(next page)*: Highly variant AS events reveal intricacies of cell states.
a. Read coverage tracks for SNAP25 in MNs. Numbers indicate observed junction reads.
b. Spearman correlation values of a gene's alternative splicing score (Ψ) to gene expression values, with a dotted line at the threshold of $R > 0.5$.
c. Tracks from NPCs were shown to illustrate the bimodal inclusion of exon 5. Numbers indicate observed junction reads covering this SE in DYNC1I2.
d-f. A multimodal MXE event in PKM as an example to dissect MNs into three subgroups.
d. Genes correlating with Psi of the MXE event containing exon 9 and exon 10 (**Figure 3.2**) is able to cluster the MNs into three subgroups. Subgroup 1, mostly composed of outliers identified by *k*-means clustering (Supplementary **Figure 3.4**), contain characteristic genes for progenitors. Subgroup 2 and 3 are enriched for neuronal genes. Rows represent the genes and columns represent single cells in MNs. Genes detected in MNs and correlated with the Psi, using an emipircally-defined threshold of Spearman's *R* greater than two standard deviations away from the mean permuted correlation values. Psi/Ψ ranged from 0 (blue) to 0.5 (yellow) to 1 (red). Black and grey: cells designated as qualified cells versus outlier-cells based on *k*-means clustering. Representative genes enriched in two of the subgroups are highlighted in blue (high with exon 10 inclusion) or red (high with exon 9 inclusion).
e. Example genes enriched in two of the subgroups of MNs. MAP2 and NRXN1 are more highly expressed in cells with $\Psi \approx 1$; ETV5 and MASTL are more highly expressed in cells with $\Psi \approx 0$. Psi scores of the MXE in PKM is plot on x-axis and $\log_2(\text{TPM} + 1)$ of indicated genes is plot on y-axis.
f. Genes correlating with Psi is able to separate the three subgroups in MNs. Left, PCA using all detected genes in MNs. Right, PCA using genes correlating with Psi.

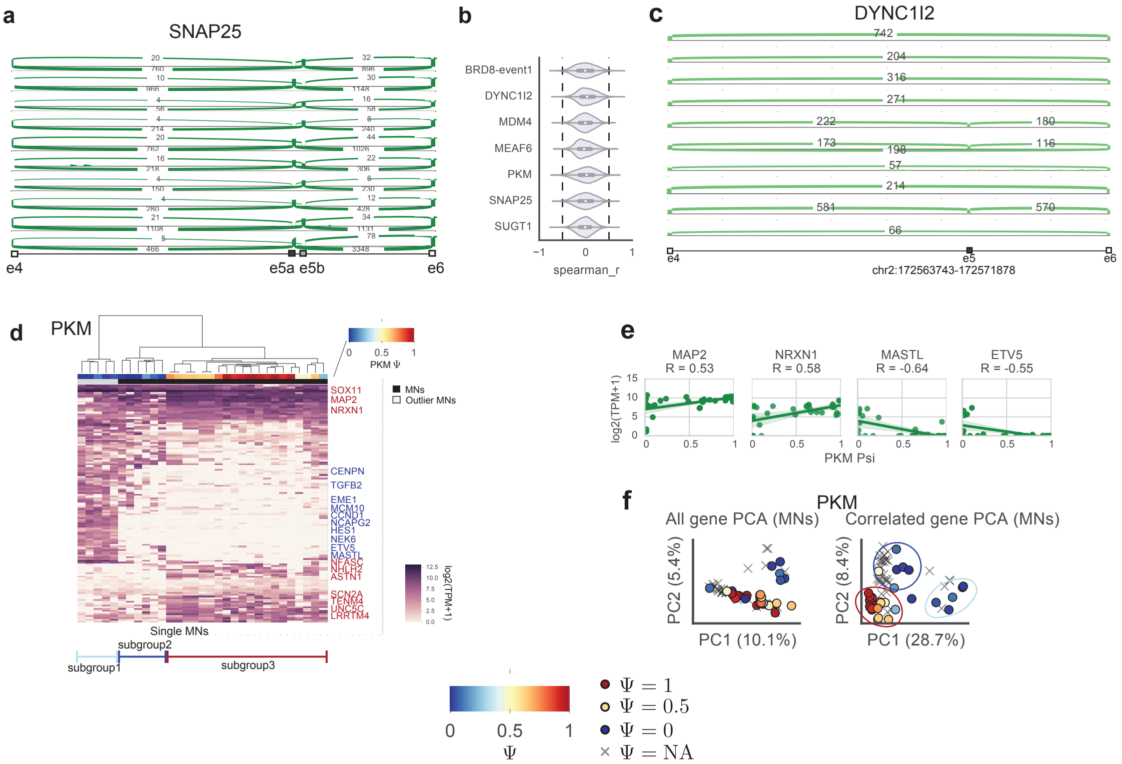
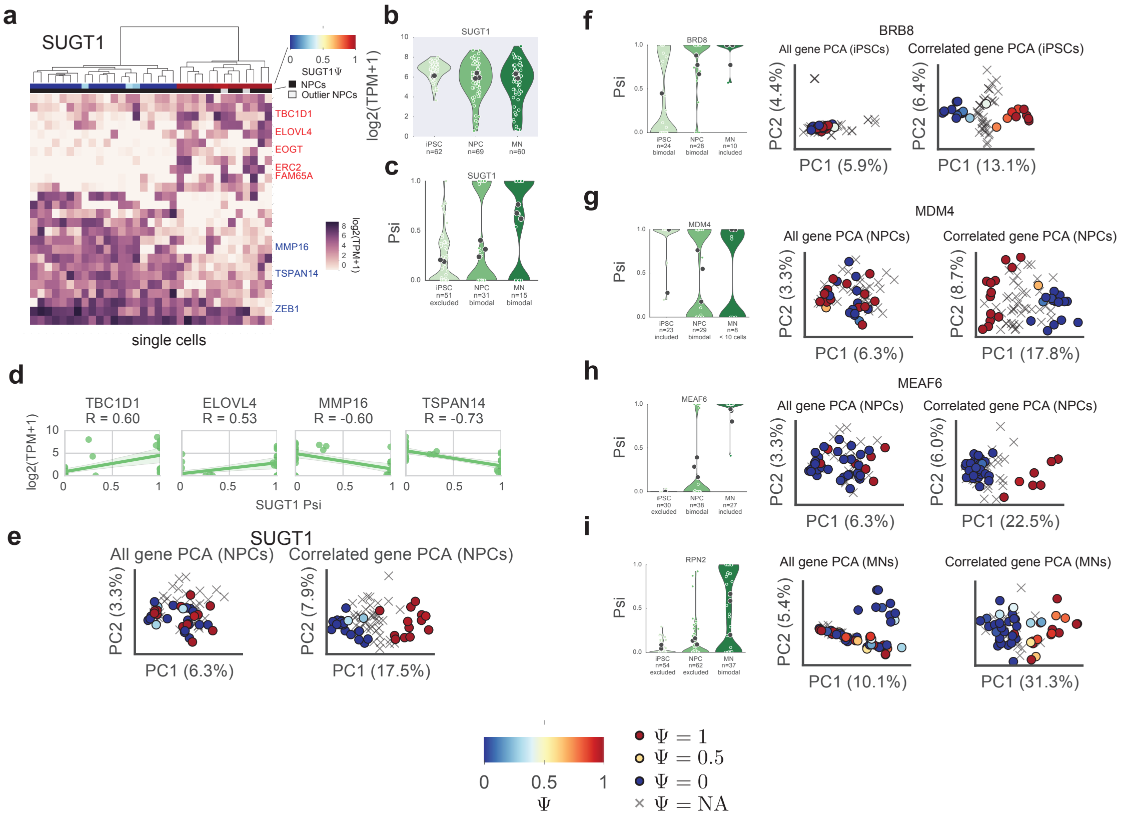

a
SNAP25
20
760
32
896
10
966
30
1148
4
56
16
58
4
214
6
240
20
762
44
1026
16
218
22
306
4
150
6
230
4
280
12
428
21
1108
34
1131
5
466
78
3348
e4
e5a e5b
e6
b
BRD8-event1
DYNC1I2
MDM4
MEAF6
PKM
SNAP25
SUGT1
−1
0
1
spearman_r
c
DYNC1I2
742
204
316
271
222
180
173
198
116
57
214
581
570
66
e4
e5
e6
chr2:172563743-172571878
d
PKM
0
0.5
1
PKM Ψ
MNs
Outlier MNs
SOX11
MAP2
NRXN1
CENPN
TGFB2
EME1
MCM10
CCND1
NCAPG2
HES1
NEK6
ETV5
MASTL
NFASC
NHLH2
ASTN1
SCN2A
TENM4
UNC5C
LRRTM4
12.5
10.0
7.5
5.0
2.5
0.0
log2(TPM+1)
Single MNs
subgroup1
subgroup2
subgroup3
e
MAP2
R = 0.53
NRXN1
R = 0.58
MASTL
R = -0.64
ETV5
R = -0.55
log2(TPM+1)
15
10
5
0
0
0.5
1
PKM Psi
f
PKM
All gene PCA (MNs)
Correlated gene PCA (MNs)
PC2 (5.4%)
PC1 (10.1%)
PC2 (8.4%)
PC1 (28.7%)
0
0.5
1
Ψ
Ψ = 1
Ψ = 0.5
Ψ = 0
Ψ = NA

Figure 3.14 *(next page)*: Highly variant AS events in SUGT1, BRD8, MDM4, MEAF6, and RPN2 reveal intricacies of cell states.
a-f. A bimodal SE event in SUGT1 as an example to dissect NPCs into two subgroups.
b. Genes correlating with Psi of the SE event cluster the NPCs into two subgroups. Genes detected in NPCs and correlated with the Psi. Blue: cells with Psi around 0. Red: cells with Psi around 1. Light Blue to yellow: cells with Psi around 0.5. Black and grey: cells designated as qualified cells versus outlier cells based on *k*-means clustering. Representative genes enriched in two of the subgroups are highlighted in blue (high upon exon exclusion) or red (high upon exon inclusion).
c. Expression of SUGT1 in the three populations.
d. Psi distribution of a SE event (lower) in SUGT1 in the three populations. This event is excluded in iPSCs, and bimodal in both NPCs and MNs.
e. Example genes enriched in the two subgroups of NPCs. TBC1D1 and ELOVL4 are more highly expressed in cells with Psi ≈ 1; MMP16 and TSPAN14 are more highly expressed in cells with Psi 0. Psi scores of the SE event in SUGT1 is plot on x-axis and $\log_2(\text{TPM}+1)$ of indicated genes is plotted on y-axis.
f. Only genes correlating with Psi is able to separate the two subgroups in NPCs. Left: PCA using all detected genes in NPCs. Right: PCA using genes correlating with Psi.
g-j. PCA using all detected genes in perspective population fail to identify substructures of seemingly homogenous cells (left panel). PCA using gene correlating with each AS events (right panel) is able to identify the delicate substructures of cells.
g. Bimodal SE event in BRD8 distinguishes iPSC substructure.
h. Bimodal SE event in MDM4 distinguishes NPC substructure.
i. Bimodal SE event in MEAF6 distinguishes NPC substructure.
j. Bimodal SE event in RPN2 distinguishes MN substructure.


a
SUGT1
0 0.5 1
SUGT1 Ψ
NPCs
Outlier NPCs
TBC1D1
ELOVL4
EOGT
ERC2
FAM65A
MMP16
TSPAN14
ZEB1
8 6 4 2 0
log2(TPM+1)
single cells
b
SUGT1
log2(TPM+1)
10 8 6 4 2 0
iPSC n=62
NPC n=69
MN n=60
c
SUGT1
Psi
1.0 0.5 0.0
iPSC n=51 excluded
NPC n=31 bimodal
MN n=15 bimodal
d
TBC1D1 R = 0.60
ELOVL4 R = 0.53
MMP16 R = -0.60
TSPAN14 R = -0.73
log2(TPM+1)
10 5 0
0 0.5 1
SUGT1 Psi
e
SUGT1
All gene PCA (NPCs)
Correlated gene PCA (NPCs)
PC2 (3.3%)
PC1 (6.3%)
PC2 (7.9%)
PC1 (17.5%)
f
BRD8
Psi
1.0 0.5 0.0
iPSC n=24 bimodal
NPC n=28 bimodal
MN n=10 included
BRB8
All gene PCA (iPSCs)
Correlated gene PCA (iPSCs)
PC2 (4.4%)
PC1 (5.9%)
PC2 (6.4%)
PC1 (13.1%)
g
MDM4
Psi
1.0 0.5 0.0
iPSC n=23 included
NPC n=29 bimodal
MN n=8 < 10 cells
MDM4
All gene PCA (NPCs)
Correlated gene PCA (NPCs)
PC2 (3.3%)
PC1 (6.3%)
PC2 (8.7%)
PC1 (17.8%)
h
MEAF6
Psi
1.0 0.5 0.0
iPSC n=30 excluded
NPC n=38 bimodal
MN n=27 included
MEAF6
All gene PCA (NPCs)
Correlated gene PCA (NPCs)
PC2 (3.3%)
PC1 (6.3%)
PC2 (6.0%)
PC1 (22.5%)
i
RPN2
Psi
1.0 0.5 0.0
iPSC n=54 excluded
NPC n=62 excluded
MN n=37 bimodal
All gene PCA (MNs)
Correlated gene PCA (MNs)
PC2 (5.4%)
PC1 (10.1%)
PC1 (31.3%)
0 0.5 1
Ψ
● Ψ = 1
● Ψ = 0.5
● Ψ = 0
× Ψ = NA

3.2.6 Transformation of splicing distributions to "waypoints" reveals dynamic of AS events

To visualize changes in modalities, we developed bonvoyage, where the distribution of Psi values of each AS event across single cells from a cell-type is first discretized, then reduced via non-negative matrix factorization (NMF) (**Figure 3.15**a, left and middle). NMF is a dimensionality reduction algorithm which factorizes data into its components using a parts-based approach[209]. The Psi values are factorized into two components, excluded (x-axis) and included (y-axis), which depict the "waypoint" space (**Figure 3.15**a, right). Usage of the waypoint space is illustrated using simulated modality data (**Figure 2.12**a-d). Each AS event is depicted as a point in waypoint space, which represents the distribution of Psi scores in single cells (**Figure 3.15**b). All the AS events measured in a cell-type were projected into waypoint space, and colored by their corresponding modalities identified previously by anchor (**Figure 3.15**c, d). In such a representation, each modality occupies a discrete region in waypoint space. Also, AS events that change their Psi distributions during differentiation undergo "voyages". To illustrate, exon 9 of PKM is excluded in iPSCs, becomes more included in NPC and is a bimodal exon in MNs. Such a change of modality creates a voyage in waypoint space (**Figure 3.15**e). In contrast, projection of this event measured in bulk MNs failed to capture the bimodality. Additionally, MAP4K4 encodes a member of the serine/threonine protein kinase family and inclusion of exon 16 extends MAP4K4's protein kinase-like domain. This event became progressively more included along MN differentiation, readily observed in a voyage plot, which we independently confirmed by RNA-FISH (**Figure 3.16**a-b). Thus, bonvoyage is an effective method to visualize and identify AS events that change across populations.

Figure 3.15 *(next page)*: `bonvoyage` visualizes dynamic AS changes.

a. A schematic to illustrate the transformation of splicing profiles into the two-dimensional waypoint space by `bonvoyage`. Splicing distribution of each event (A, B, C and D represent 4 different AS events) was discretized into bins (left), factorized by non-negative matrix factorization (NMF) and projected onto 2-dimensional space (middle), such that each data point represents a distribution of alternative splicing. The origin point represents a distribution that all cells have 50% of inclusion and 50% exclusion reads observed in scRNA-seq. When the distributions of the same event (either event B or C) are visualized in two different cell-types or states, the dynamic of the event is illustrated by its voyage in the waypoint space (right panel).

b. AS events in iPSCs projected in the waypoint space. The shade of hexagon indicates the number of events.

c. AS events in iPSCs (same as **b**), colored by the modality estimated by `anchor`. Each dot represents distribution of one AS event. Note, each modality occupies a distinct region of the waypoint space. Black-outlined circle highlights PKM MXE event.

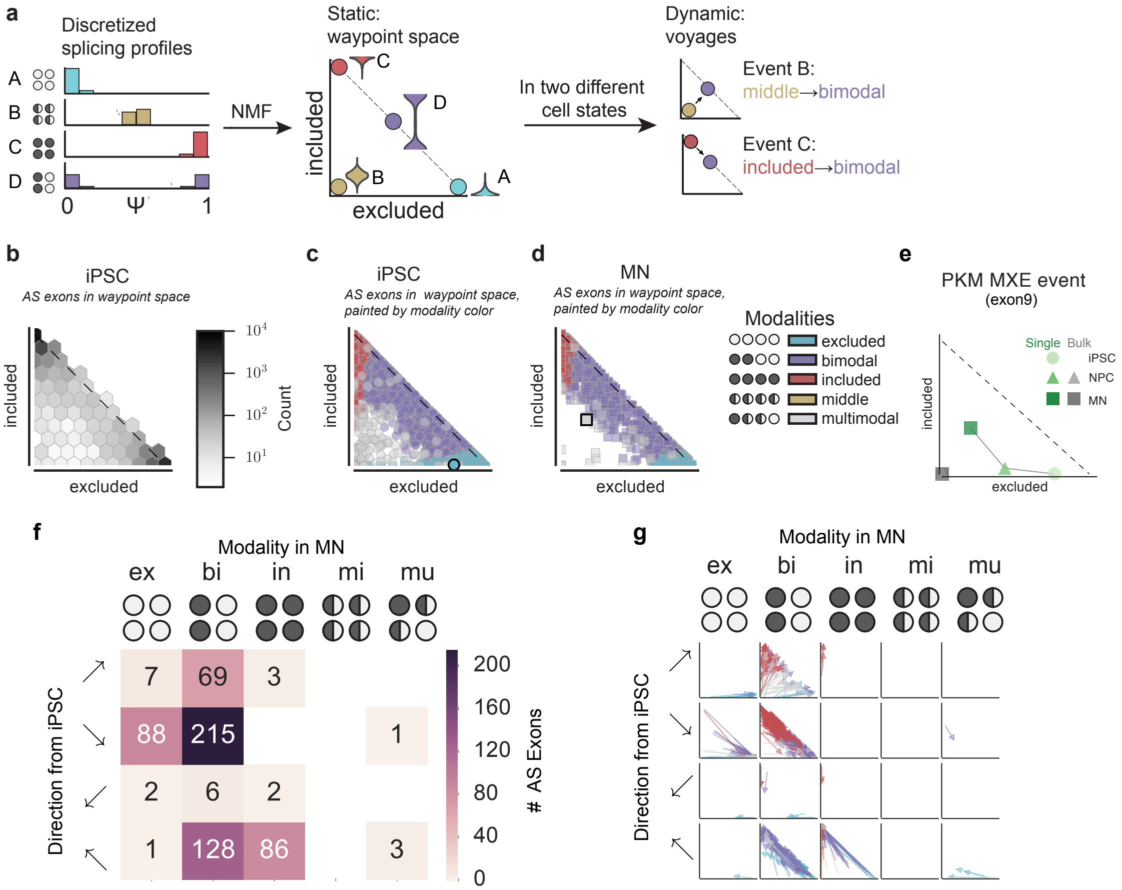
d. AS events in MNs are colored by their modalities and presented in waypoint space. Black-outlined square highlights PKM MXE event.

e. Dynamics of the MXE event in PKM is illustrated in the waypoint space. Shown is the inclusion of exon 9 of the MXE, which is included in both iPSCs and NPCs and becomes bimodal in MNs.

f-g. Global splicing dynamics between iPSC and MN, aggregated by voyage direction instead of modalities.

f. Number of events originated in iPSC and travel in the indicated directions to land in excluded, bimodal, included, middle, or multimodal modality in MN.

g. Same data as (**f**), visualized by vectors representing the iPSC (tail) and MN (tip) position of the alternative exon. Color of arrows are coded based on event modalities in iPSCs.


a
Discretized splicing profiles
A
B
C
D
0
Ψ
1
NMF
Static: waypoint space
included
excluded
A
B
C
D
In two different cell states
Dynamic: voyages
Event B: middle→bimodal
Event C: included→bimodal
b
iPSC
AS exons in waypoint space
included
excluded
10^4
10^3
10^2
10^1
Count
c
iPSC
AS exons in waypoint space, painted by modality color
included
excluded
d
MN
AS exons in waypoint space, painted by modality color
included
excluded
Modalities
excluded
bimodal
included
middle
multimodal
e
PKM MXE event
(exon9)
Single
Bulk
iPSC
NPC
MN
included
excluded
f
Modality in MN
ex
bi
in
mi
mu
Direction from iPSC
↗
↘
↙
↖
7
69
3
88
215
1
2
6
2
1
128
86
3
200
160
120
80
40
0
# AS Exons
g
Modality in MN
ex
bi
in
mi
mu
Direction from iPSC
↗
↘
↙
↖

We next sought to establish a global view of AS changes between cell-types. Focusing on exons with large voyages (**Figure 3.16**c), we visualized the voyaging exons using vectors between iPSC and MNs. We regard voyages as complementary to delta Psi ($\Delta\Psi$) used in two-sample AS comparisons of bulk RNA-seq data. Consistent with our modality-based analysis (**Figure 3.9**a), the majority of the dynamic exons changed from or to the bimodal modality (**Figure 3.15**f-g, **Figure 3.16**d). To evaluate the consequences of voyages on the protein properties of resulting isoforms, we transformed each property into a waypoint-weighted score by multiplying the property of each isoform with its corresponding coordinate in the waypoint space, enabling a more integrated evaluation of protein property based on both isoforms and their distribution in single cells. Among many properties investigated, we found that MNs favor splicing that generates more disordered and basic proteins such as the events in RPS24, and ZNF207/BuGZ (**Figure 3.18**a, b). Thus, AS voyages allow for population-based investigation of the protein outcomes of isoform preferences.

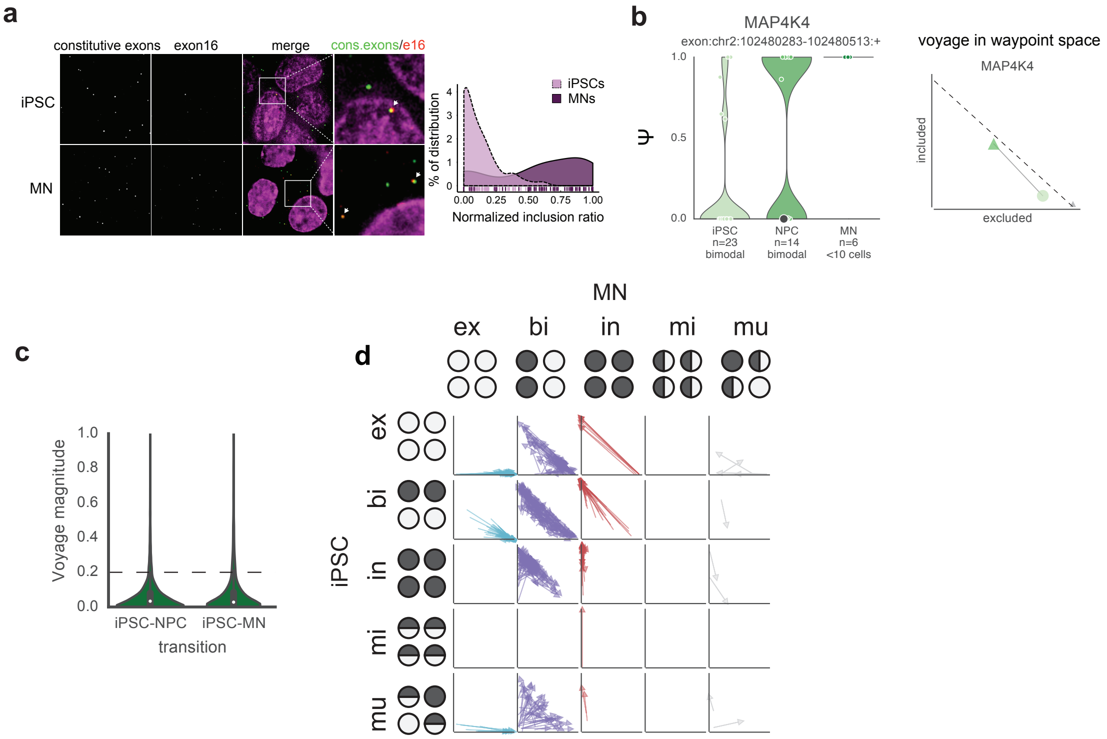
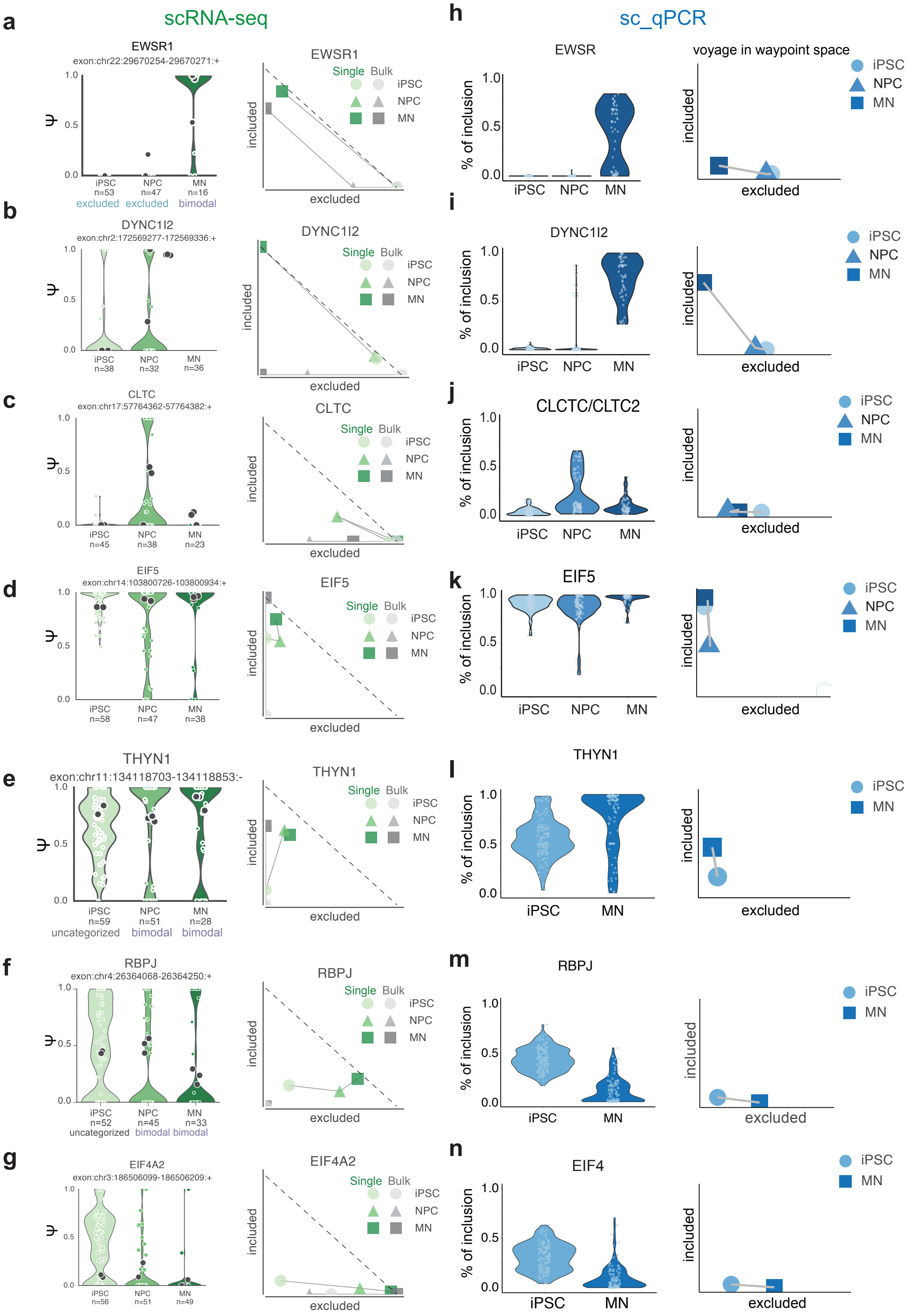


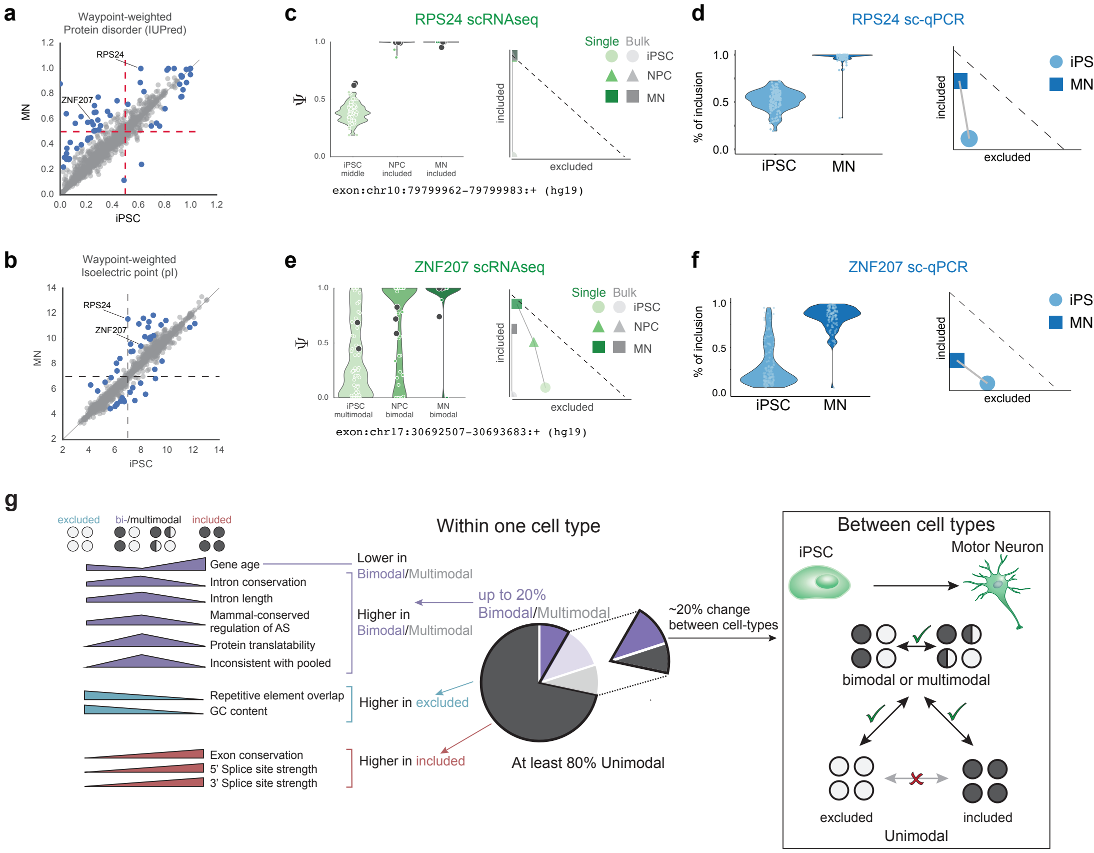
Figure 3.16: Validation of voyaging events between iPSC and MN.
a. Validation of a SE event in MAP4K4 by smRNA-FISH.
b. MAP4K4 smRNA-FISH. Left, probe sets are designed for constitutive exons and alternative exon 16. Exon 16 is excluded in iPSCs (n = 113, light purple with dashed line) and become more included in MNs (n = 68, dark purple with solid outline. Middle, quantitation of normalized inclusion of exon 16. Arrows point out foci overlapped for both constitutive and exon 16 probes. Normalized inclusion ratio is calculated by percentage of e16 probes co-localized with constitutive probes/constitutive probes, and resulting percentage is normalized by 95 percentage of the maximal percentage.
c. MAP4K4 single-cell RNA-Seq. Left, violinplots percent spliced-in inclusion values, and right, waypoint space of exon 16.
d. Magnitude of change in waypoint space (voyages) from iPSC to NPC, and iPSC to MN, with a cutoff shown as a black dashed line at 0.2.
e. Global splicing dynamics between iPSC and MN modalities, visualized as vectors from iPSC to MN in waypoint space. Underlying data is the same as **Figure Figure 3.9a**. Color of arrows are coded based on event modalities in MNs.

Figure 3.17 *(next page)*: Validation of alternative splicing events by sc-qPCR.
a-g. Distribution of alternative exon inclusion by single-cell RNA-Seq for indicated events in EWSR1 (**a**), DYNC1I2 (**b**), CLTC/CLCT2 (**c**), EIF5 (**d**), THYN1 (**e**), RBPJ (**f**), and EIF4A2 (**g**), shown in violin plots (left) and in waypoint plots (right). Percent spliced-in (Psi/Ψ) is calculated based on single cell RNA-seq data, illustrated in green. Black dots indicate bulk samples (1,000 cells) for each cell type.
h-n. Distribution of percentage of inclusion by single-cell qPCR of indicated events EWSR1 (**h**), DYNC1I2 (**i**), CLTC/CLCT2 (**j**), EIF5 (**k**), THYN1 (**l**), RBPJ (**m**), and EIF4A2 (**n**), based on single cell qPCR shown in violin plot (left) and waypoint plot (right), illustrated in blue.


scRNA-seq
sc_qPCR
a
EWSR1
exon:chr22:29670254-29670271:+
Ψ
1.0
0.5
0.0
iPSC n=53 excluded
NPC n=47 excluded
MN n=16 bimodal
EWSR1
Single Bulk
iPSC
NPC
MN
included
excluded
b
DYNC1I2
exon:chr2:172569277-172569336:+
iPSC n=38
NPC n=32
MN n=36
DYNC1I2
c
CLTC
exon:chr17:57764362-57764382:+
iPSC n=45
NPC n=38
MN n=23
CLTC
d
EIF5
exon:chr14:103800726-103800934:+
iPSC n=58
NPC n=47
MN n=38
EIF5
e
THYN1
exon:chr11:134118703-134118853:-
iPSC n=59 uncategorized
NPC n=51 bimodal
MN n=28 bimodal
THYN1
f
RBPJ
exon:chr4:26364068-26364250:+
iPSC n=52 uncategorized
NPC n=45 bimodal
MN n=33 bimodal
RBPJ
g
EIF4A2
exon:chr3:186506099-186506209:+
iPSC n=56
NPC n=51
MN n=49
EIF4A2
h
EWSR
% of inclusion
voyage in waypoint space
iPSC
NPC
MN
i
DYNC1I2
j
CLCTC/CLTC2
k
EIF5
l
THYN1
m
RBPJ
n
EIF4

To validate the Psi distributions of bimodal and high-magnitude voyaging AS events during motor neuron differentiation, we designed splicing-sensitive primers to assess exon usage by qPCR at single cell resolution in iPSCs, NPCs and MNs. We observed that 60% AS events recapitulated an exon inclusion distribution similar to our findings using scRNA-seq (**Figure 3.18**c-f, **Figure 3.17**a-n). For example, a SE event that introduces a stop codon and removes three amino acids from C-terminal in RPS24, encoding a ribosomal subunit protein S24, previously reported in different human tissues[235]. In single cells, this event was partially included in individual iPSCs (middle modality), and became completely included in almost all NPCs and MNs (**Figure 3.18**c). These dynamics were confirmed by sc-qPCR (**Figure 3.18**d). Also, exon 9 in ZNF207 encoding serine-rich sequences that may affect post-translational modifications, starts as multimodal in iPSCs and becomes more included in MNs (**Figure 3.18**e). The dynamics and voyages of these and many other exons were validated by sc-qPCR (**Figure 3.18**f, **Figure 3.17**a-n). Thus, by enabling comparison of splicing profiles and protein properties, the `bonvoyage` resource enables visualization of AS dynamics across cell populations.

Figure 3.18 *(next page)*: qPCR validation and summary of biological findings.
a-b. Waypoint-weighted protein properties changing between iPSC and MN. Significant changes(blue) are identified by a factor of three on Mahalanobis distance relative to all iPSC-MN comparisons.
a. Protein disorder by IUPred, where a score above 0.5 (red dashed line) indicates disorder.
b. Isoelectric point (pI), where the black dashed line indicates pI = 7. X-axis, weighted protein property in iPSC and y-axis, weighted protein property in MN.
c-f. Distribution of AS inclusion is verified by single cell qRT-PCR (sc-qPCR). Primer sets for inclusion, exclusion and gene expression were designed for each event tested. Percent inclusion measured in sc-qPCR is calculated by $\frac{2^{\text{inclusion Ct}}}{2^{\text{inclusion Ct}}+2^{\text{exclusion Ct}}}$ (See Methods for more details) in both iPSCs (n = 134) and MNs (n = 95).
c. Percent spliced-in (Psi/Ψ) distributions for RPS24 exon 5 measured by single-cell RNA-Seq shown as violinplots (left) and voyages (right).
d. Percent exon inclusion distributions for RPS24 exon 5 measured by single-cell qPCR shown as violinplots (left) and voyages (right).
e. Percent spliced-in (Psi/Ψ) distributions for ZNF207 exon 9 measured by single-cell RNA-seq shown as violinplots (left) and voyages (right).
f. Percent exon inclusion distributions for ZNF207 exon 9 measured by single-cell qPCR shown as violinplots (left) and voyages (right).
g. Summary: At single cell resolution, three main categories of modalities can be identified: included, excluded and bimodal. Each modality has unique sequence, coding and evolutionary features. During cell differentiation, majority of unimodal events are static, whereas the highly variance events are dynamic, playing a key role in shaping the transcriptome.


a
Waypoint-weighted
Protein disorder (IUPred)
RPS24
ZNF207
MN
iPSC
0.0
0.2
0.4
0.6
0.8
1.0
1.2
b
Waypoint-weighted
Isoelectric point (pI)
RPS24
ZNF207
MN
iPSC
2
4
6
8
10
12
14
c
RPS24 scRNAseq
Ψ
0.0
0.5
1.0
iPSC
middle
NPC
included
MN
included
included
excluded
Single Bulk
iPSC
NPC
MN
exon:chr10:79799962-79799983:+ (hg19)
d
RPS24 sc-qPCR
% of inclusion
0.0
0.5
1.0
iPSC
MN
included
excluded
iPSC
MN
e
ZNF207 scRNAseq
Ψ
0.0
0.5
1.0
iPSC
multimodal
NPC
bimodal
MN
bimodal
included
excluded
Single Bulk
iPSC
NPC
MN
exon:chr17:30692507-30693683:+ (hg19)
f
ZNF207 sc-qPCR
% of inclusion
0.0
0.5
1.0
iPSC
MN
included
excluded
iPSC
MN
g
excluded
bi-/multimodal
included
Gene age
Intron conservation
Intron length
Mammal-conserved
regulation of AS
Protein translatability
Inconsistent with pooled
Repetitive element overlap
GC content
Exon conservation
5' Splice site strength
3' Splice site strength
Within one cell type
Lower in
Bimodal/Multimodal
Higher in
Bimodal/Multimodal
up to 20%
Bimodal/Multimodal
Higher in excluded
Higher in included
At least 80% Unimodal
~20% change
between cell-types
Between cell types
iPSC
Motor Neuron
bimodal or multimodal
excluded
included
Unimodal

3.3 Methods

3.3.1 Cell culture and differentiation

iPSCs were cultured on matrigel coated plated using mTeSR (Stem Cell Technologies) media with mTeSR supplement at 37° C incubator with 5% CO_2.

Neuronal progenitor cells (NPCs) were differentiated from iPSCs. Briefly, iPSCs were cultured in Matrigel coated plates and dislodged by dispase. To form embryonic bodies, the dislodged colonies were cultured in DMEM/F12 (Invitrogen) with GlutaMax and N2 supplement in non-adhere petri dish. Media were replaced every other day for 7 days. EBs were then plated onto matrigel coated plate to allow rosette formation. Clean rosette were picked manually and maintained in EB media for 7 days and subsequently dissociated with accutase and cultured in NPC media (DMEM/F12, GlutaMax, N2 and B27 with 2 µg/µL FGF) to allow neuron progenitor cell differentiation. NPCs were maintained in NPC media.

Motor neurons were directly differentiated from iPSCs as previous described[236]. Briefly, iPSCs were cultured on matrigel coated plates until fully confluent in mTeSR then switch to knock-out serum replacement media (KSR) containing Dorsomorphin(1 µM) and SB431542 (10 µM). Upon day 4 of differentiation, increasing amounts of N2 media (25%, 50%) was added to the KSR. From day 7 of differentiation, 1.5 µM retinoic acid and 200 nM Smoothened Agonist (SAG, EMD Millipore) were added to induce patterning. Cells were dissociated on day 17 of differentiation and replated in poly-D-lysine and laminin coated plates. Maturation was performed using BDGF (2 ng/µL), GDNF (2 ng/µL), CNTF (2 ng/µL), ascorbid acid, sonic hedgehog and retinoic acid in N2 and B27 media up until 35 days of differentiation.

3.3.2 Single-cell capture and library preparation

iPSCs, NPCs and MNs were dissociated using Accutase (Stem Cell Biotechnologies) and filtered through 40 μm cell strainers to obtain single cell suspension. Single cells were captured on C1 auto prep platform (Fluidigm, CA) according to manufacturer's instructions. C1 auto prep chips were visually inspected with a light microscopy at 20X to ensure singularity of captured cells. All non-single cells were discarded from analysis. SMARTer Ultra Low RNA cDNA Synthesis Kit (Clontech) was used to reverse transcribe polyA-tailed RNA. cDNA was amplified using Advantage 2 Polymerase Mix by PCR at 95 °C for 1 minutes, followed by 21 cycles of 15 seconds at 95 °C, 30 seconds at 65 °C and 6 minutes at 68 °C, followed by another 10 minutes at 72 °C as a final extension. cDNAs were inspected using Agilent Bioanalyzer High Sensitivity DNA chips and quantitated by PicoGreen dsDNA Assay kit (ThermoFisher). cDNAs were diluted to 1 ng to generate libraries using the Nextera XT DNA kit (Illumina, La Jolla, CA). Libraries were multiplexed and sequenced on Illumina HiSeq 2000 to generate 100bp PE reads.

3.3.3 RNA-Seq processing

RNA-seq reads were trimmed using `cutadapt` v1.8.1 of adapter sequences `TCGTATGCCGTCTTCTGCTTG`, `ATCTCGTATGCCGTCTTCTGCTTG`, `CGACAGGTTCAGAGTTCTACAGTCCGACGATC`, `GATCGGAAGAGCACACGTCTGAACTCCAGTCAC`, [A]$_{50}$, [T]$_{50}$, mapped to repetitive elements (RepBase v18.05[237]) using the STAR[199] splicing-aware aligner (v2.4.01). Reads that did not map to repetitive elements were then mapped to the human genome (hg19), using GENCODE[238] (v19) gene annotations to create the splice junction database. We used the `SJ.out.tab` files from STAR to create alternative splicing annotations and calculate percent spliced-in (see **Section 2.1**). Gene expression was quantified with sailfish[219] using GENCODE v19 protein-coding and long non-coding RNA annotation, and we then aggregated transcript-level expression to genes.

3.3.4 Single-cell expression-level quality control and outlier detection

We retained genes expressed with TPM > 1 in at least 10 cells for a total of 18,594 genes, and filtered out cells which had $< 4,000$ expressed genes, which was a natural cutoff in the data. For the three cell types, $n = 63$ iPSCs, $n = 73$ NPCs, and $n = 70$ MNs had enough expressed genes to pass gene expression level quality control.

We performed K-means clustering with $k = 3$ on the gene expression matrix, with 1000 different random initializations. For each cell that clustered into a group that consisted of a majority of a different cell type (e.g. a motor neuron that was clustered in the group with majority NPCs), we called these cells outliers and discarded them from analysis. Overall, for iPSC: 71 were captured, 63 passed QC, 1 outlier for 62 total; for NPC: 98 were captured, 73 passed QC, 4 outliers for 69 total; for MN: 93 were captured, 70 passed QC, 10 outliers for 60 total.

3.3.5 Estimation of alternative splicing

We used `outrigger` to create a custom alternative splicing index on the splice junction (`SJ.out.tab`) files created by STAR, and used GENCODE v19 to define possible exons. This created $40,534$ skipped exon (SE) and $13,217$ mutually exclusive exon (MXE) possible alternative events, and we calculated percent spliced-in (Psi/Ψ) with a minimum of 10 junction reads. We then filtered for events that were alternative, not constitutively included or excluded across all cells. Alternative events were defined by, $0 < \Psi < 1, \Psi \neq 0, 1$ in at least one cell. Events were then filtered for events that were detected in at least 10 cells of any celltype, resulting in $13,910$ events.

3.3.6 Constitutive exons

We defined constitutive exons as those that did not appear as the alternative exon in any of the splice types (MXE and SE), and had at least 10 reads on both upstream and downstream junctions, in at leat 10 cells per cell type.

3.3.7 ICA on constitutively expressed genes and their splicing events

First, $12,685$ genes were identified as non-DE genes across the three populations using a non-parametric Kruskal-Wallis test with Bonferroni-corrected p-value, called q, with $q > 10$ as the cutoff.

Second, AS events were extracted from non-DE genes and their Psi scores are subjected to Independent Component Analysis (ICA). To impute the null values widespread in splicing data, we replaced NAs with an arbitrary number

(100) out the of range of Psi values. We did not find that the choice of the arbitrary number affected the ICA results. We then calculated ICA on the imputed matrix.

3.3.8 Hierarchical clustering

We performed hierarchical clustering on samples in Python, using the `fastcluster`[239] package and performing optimal leaf ordering[240] using the `polo`[241] package. All clustering was performed using the Euclidean distance metric with Ward's method[242]. We visualized using the matplotlib[243] and seaborn[244] visualization libraries in Python.

3.3.9 Gene Ontology Enrichment

We calculated Gene Ontology (GO) enrichment by using the Gene Ontology mapping queried to the Entrez gene database using the Python package `mygene`[245;246]. We calculated GO enrichment using only the "biological process" category, and corrected for multiple hypothesis testing using Bonferroni correction as performed in the Python package `goatools`[247].

3.3.10 Categorization of alternative splicing "modes"

We calculated modality using the default parameters of the texttanchor software (see **Section 2.2**) only on splicing events observed in at least 10 cells per cell-type. The performance of `anchor` was tested extensively using simulated data in comparison to existing bimodality detecting methods.

3.3.11 Sequence annotation of alternative isoforms

We annotated alternative events and their biological features at different levels of the Central Dogma.

DNA-level

Evolutionary conservation. We used units of evolutionary conservation as measured by Placental Mammal PhastCons[248] scores calculated previously[102] (**Figure 3a-b**, **Supplementary Figure 3.7**).

For average conservation of exons, we used `bigWigAverageOverBed`[249] to calculate the mean conservation (treating bases without annotated conservation as NA) across each exon. For base-wise conservation, we used the HTSeq[250] Python package to create a memory-mapped `GenomicArray`, and queried this object with the intronic intervals.

Repetitive element overlap. We used the Repeat Masker track[251] from UCSC's Genome Browser[252] and used `bedtools intersect`[253] to overlap with our exon definitions. We grouped repeats into families defined by the Dfam[254] database of repetitive DNA elements (**Figure 3.6**e). For simplicity of interpretation, we used only repetitive elements that appeared at least 10 times in the excluded modality, as it was the modality with the most repetitive elements.

Gene age (Phylostratum) We used the Phylostratum classification of genes as found previously[223] (**Figure 3.7**e). For each splicing event, we found all overlapping genes in the same genomic locus, and aggregated all genes with at least one event in each modality. Meaning, a gene could appear in multiple modality categories if it had one exon in the included modality and another in the bimodal category.

***k*-mer counting and motif (PWM) enrichment** We used placental mammal conserved elements as downloaded from UCSC[251], taking only conserved elements upstream and downstream of alternative exons. We used `kvector`[255] to count *k*-mers in these conserved elements, and calculated a *Z*-score of *k*-mer enrichment for each intron group defined by cell-type, intron context, and modality (**Figure 3.8**a-b, **Figure 3.7**d). Interested in which *k*-mers were enriched in each modality, we used the total *k*-mer counts in the intron context and celltype, for all modalities, as the background. We then performed principal component analysis using the Python package `scikit-learn`[207] on the modality introns (**Figure 3.7**m, **Figure 3.8**c). We labeled *k*-mers by the standard color of the majority nucleotide (if there was a tie for the winner, the *k*-mer was assigned grey) whose squared PCA distance was greater than two squared standard deviations from the center, i.e. an ellipse around the origin of the plot. We used the Python package `adjustText`[256] to move the text labels away from each other and make them readable.

To find which RNA binding protein motifs were enriched for different modalities, we used version 0.6 of the CISBP-RNA binding database[226] and transformed each position-weight matrix (PWM) into a Boolean vector of *k*-mers that could exactly fit into the PWM, with no mis-matches (**Figure 3.8**d). We ignored psuedocounts by setting all values ≤ 0.1 to zero. We then used this Boolean matrix to obtain motif *k*-mers and calculate enrichment using a *t*-test, as compared to all *k*-mers of that intron group. We then performed PCA on the motif *t*-statistics, using the intron groups as features (**Figure 3.6**p, **Figure 3.7**e-f). We labeled motifs whose squared PCA distance was greater than two squared standard deviations from the center, i.e. an ellipse around the origin of the plot. We used the Python package `adjustText`[256] to move the text labels away from each other and make them readable.

RNA-level

Consistency of splicing between bulk and single-cells To calculate the total difference between the bulk Ψ and single-cell Ψ estimates, for each event, we calculated the average difference between the pooled sample Ψ and every single-cell Ψ, much like a sample mean calculation (**Figure 3.6**a).

Splice site strength We used `bedtools`[253] and `pybedtools`[257] to obtain the 5′ (relative to exon-intron boundary: -20nt into intron and +3nt into exon) and 3′ (relative to exon-intron boundary: -3 into exon and +6 into intron), and obtained the transcript sequences for these regions. We used MaxEntScan[258] to calculate the strength of the alternative exon (exon 2 in both the SE and MXE cases) splice sites (**Figure 3.6**f-g).

Expression of splicing events For finding the gene expression per splicing event, for each event, we used all genes that could map to it. Sometimes multiple genes could map to a single event, as a result of poor annotation, or multiple read-through transcripts. To mitigate this, for each event, we summed all gene expression by the $\log_2(\mathrm{TPM}+1)$ values, and plotted the distribution of expression per modality (**Figure 3.6**h).

Intron and exon length As we used `outrigger` to calculate splicing, it also output the lengths of the introns and exons for each alternative event, which is what we used (**Figure 3.7**c and **Figure 3.6**d).

Protein-level

We are in the process of packaging the splicing event isoform translation and domain scanning code into a package called `poshsplice`[259].

Protein translation Using events which had at least one isoform annotated with a CDS in GENCODE v19 (22,152 SE and MXE events), we translated the exon trio and duo (SE, included isoform has three exons and and excluded has two) or exon trios (MXE, both included and excluded isoforms contain three exons) to its transcript-annotated reading frame. If these exons participated in transcripts with multiple reading frames, we used all translations.

Domain search We used the `hmmscan` command from the HMMer[228;229] software suite (v3.1b1) to search for protein domains matching those in the manually curated Pfam-A database[230]. We used a domain-independent E-value cutoff of 10^{-5}. With this raw data, we observed "domain switching" between isoforms in instances such as "Kinase" to "Tyrosine Kinase", when indeed the exact characters of domain name changed, but the overall function didn't. To alleviate this problem, we aggregated domains into clades using Pfam's annotations. We then annotated each individual event with whether only the exclusion or inclusion isoforms had an annotated translation, only one isoform, contained a clade, both contained the same clade, or the clades switched (**Figure 3.9**d).

3.3.12 Correlation of splicing to expression

We correlated bimodal and multimodal splicing events to genes with variant expression, defined as two standard deviations away from the mean variance of all genes. We used Spearman correlation to compare splicing profiles to gene expression, and used a threshold of absolute correlation values $|R| > 0.5$ across all samples.

3.3.13 Transformation of splicing profiles to 2d space

We used `bonvoyage` (see **Section 2.3**) to transform one-dimensional splicing profiles into two-dimensional space (**Figure 3.15**a-c), using the default parameters. We performed the transformation within cell-type, and required at least 10 cells per splicing event to transform.

3.3.14 Waypoint-weighted protein properties

To obtain protein properties, we used IUPRED[260] to calculate protein disorder and the `ProtParam` module in BioPython[261] to calculate aromaticity, instability index, molecular weight, secondary structure properties (alpha-helix, beta-sheet, and turns), flexibility, grand average of hydropathy (GRAVY) and isoelectric point.

We summarized isoform protein properties for each phenotype by using the NMF-transformed waypoint space into a weighted average. Using p_{included} and p_{excluded} to represent the protein property value (e.g. molecular weight or disordered protein score) of each isoform, and w_{included} and w_{excluded} to represent the splicing event's waypoint space position for the included (y) and excluded (x) axes. We calculated the weighted protein property, p_w, within each phenotype, as we did for the modality and waypoint calculation.

$$p_w = p_{\text{included}} w_{\text{included}} + p_{\text{excluded}} w_{\text{excluded}} \tag{3.1}$$

For properties that had a relative center, e.g. isolectric point which has a neutral value of 7, we subtracted the center value for each protein property,

p_{center} so the multiplication by the waypoint space would amplify the distance from center.

$$p_w = p_{\text{center}} + (p_{\text{included}} - p_{\text{center}})w_{\text{included}} + (p_{\text{excluded}} - p_{\text{center}})w_{\text{excluded}} \qquad (3.2)$$

Voyaging protein properties

Interested in which protein properties which changed significantly between cell types, we used Mahalonobis distance[262] (d_m), a non-parametric method of finding outliers from distributions. In the two-dimensional case, this means values that are significantly "off-diagonal" when comparing two cell types, e.g. iPSC to MN. We used a multiplier of $3d_m$ as the threshold for highly changing protein properties.

3.3.15 Single-cell qPCR and primer design

Single iPSCs and differentiated MNs were captured on C1 auto prep platform (Fluidigm, CA). All non-single cells were discarded from analysis. cDNA from single cells were prepared using the Single-Cell-to-Ct kit (ThermoFisher, USA) and pre-amplified with a pool of primers designed for the splicing events and the expression of corresponding genes. Inclusion and exclusion primers were specifically designed to quantitate inclusion and exclusion of AS exons and expression primers were designed from constitutive exons. All primers were tested for amplification efficiency. High-throughout quantitative PCR was performed on 96.96 Dynamic Arrays on BioMark system (Fluidigm) according to manufacturer's instructions. Each pre-amplified STA sample was diluted 1:15 for iPSCs and 1:10 for MNs. 3 housekeeping genes (RPL22, RPL27, PGK) and lineage genes (POU5F1, LIN28A, DPPA2, ISL1, MNX1, STMN2, NFEL, DCX) were included.

3.3.16 qPCR data processing

The log expression of each primer set g was computed as $\log(E_{g,c}) = 25 - \text{Ct}_{(g,c)}$ where c is the cell and $\text{Ct}_{(g,c)}$ is the Ct value for corresponding primer set. iPSCs were filtered by (RPL22 > 5, LIN28A > 8 and POU5F1 > 8) and MNs were filtered by (RPL27 > 9, ISL1 > 2 and STMN2 > 5). A total of 134 single iPSCs and 95 single MNs were retained for further analysis. If $\text{Ct}_{\text{xp},c} > 25$ (Ct value for the expression primer), the corresponding $\text{Ct}_{(\text{inc},c)}$ (Ct value for the inclusion primer) and $\text{Ct}_{(\text{exc},c)}$ (Ct value for the exclusion primer) were excluded from analysis. Percentage of inclusion is calculated by $\frac{2^{\text{Ct}_{\text{inc}}}}{2^{\text{Ct}_{\text{inc}}}+2^{\text{Ct}_{\text{exc}}}}$. Distribution of percentage of inclusion is plotted by violinplot or decomposed into 2-dimension space (`nmf(dataset, 2, “lee”)`) and projected into waypoint space in R.

3.3.17 RNA fluorescence in situ hybridization (FISH)

To verify alternative splicing of MXE event composed of exon 9 and 10 in PKM, we designed 3 probe sets (Custom Stellaris® FISH Probes, Biosearch Technologies, Inc., CA) using the Stellaris® RNA FISH Probe Designer available online. One set against constitutive exons of PKM labeled with Quasar 570, two probe sets specifically against exon9 or exon 10, respectively, labeled with Quasar 670. For Exon16 SE event in MAP4K4, one probe set against constitutive exons was designed and labeled with Quasar 570 and another probe set against exon16 was designed and labeled with Quasar 670.

iPSCs and MNs grown on coverslip were fixed with 3.7% formaldehyde

PFA for 10 minutes at room temperature. The probes for constitutive (1.25 μм) and alternative exons (1.25 μм) were mixed and hybridized to the cells in 10% deionized formamide for overnight at 37 °C, according to manufacturer's instructions. For MNs, a probe set against ISL1 is designed and labeled with fluorescein to allow the counting of only motor neurons.

3.3.18 RNA-FISH image acquisition and data processing

Images were acquired on Applied Precision OMX Super Resolution System at the Microscopy Core in the School of Medicine. Specifically, transmission and acquisition time were set at 100% and 2 minutes for both FISH probes (constitutive and alternative exons). DAPI was acquired at 10% transmission and 20 second to localize the cells. Sections were taken at 0.125 μm for the diameter of the cells, usually around 10–12μm. The resulting stacks of images were deconvoluted on Applied Precision OMX workstation. Foci of RNA molecules were quantified using Volocity 6.3 (PerkinElmer). The raw count files were then processed in R to compute ratio of exon inclusion. To limit non-specific foci, only the foci identified by both inclusion probe and constitutive probe were counted for included exons. Normalized inclusion ratio is calculated by percentage of included probes co-localized with constitutive probes/constitutive probes, and resulting percentage is normalized by 95 percent of the maximal percentage.

3.4 Supplementary Notes

3.4.1 Bimodal AS events that partition cell populations

Another example is a bimodal SE event in SUGT1 gene (MIS12 Kinetochore Complex Assembly Cochaperone), encoding a protein involved in kinetochore function and required for the G1/S and G2/M transition. Though alternative variants have been observed, their functions are largely unknown. By clustering global expression with Psi of this event, we identified two distinct subgroups of cells clustered by their Psi score. Noticeably, the subgroup with <0.5 Psi score, indicating exclusion of the alternative exon, demonstrates consistently high expression of ZEB1 (Zinc Finger E-Box Binding Homeobox 1), a master transcription factor regulating epithelial polarity, and was recently reported to be highly expressed in neuron progenitor cells to control neuronal differentiation by repressing polarity genes. Progenitor cells losing ZEB1 expression are likely to exit proliferation and become polarized[263]. Additionally, this subgroup is enriched with MMP16, reported to be expressed in less differentiated cells[264] and a few genes associated with signaling (TSPAN14, involved in presentation of ADAM10, and YES1, a src family tyrosin kinase). In contrast, the other subgroup utilizing the alternative exon highly expresses ERC2 (ELKS/RAB6-Interacting/CAST Family Member 2), encoding a protein actively involved in presynaptic organization of cytomatrix at the active zone (CAZ) complex and function as regulators of neurotransmitter release[265], suggesting this subgroup may be on the path to become nascent neurons. Supporting such a possibility, this subgroup is enriched with genes associated with different aspects of neuronal differentiation, such as TBC1D1 (acts as a GTPase-activating protein for Rab family protein(s) involving in vesicle trafficking), ELOVL4 (Very Long Chain 3-Ketoacyl-CoA Synthase 4), EOGT (EGF Domain Specific O-Linked N-Acetylglucosamine Transferase, modifying Notch receptor), FAM60A (Subunit of the Sin3 deacetylase complex (Sin3/HDAC), repressing components of the TGF-beta signaling pathway). Lastly, the two outlier NPCs (demonstrated sufficient coverage of this event and highlighted in grey) presenting higher inclusion of this alternative exon, are projected more towards MNs on PCA (Supplementary Fig 1g) in comparison to the rest of NPCs. Thus, among the NPCs demonstrating bimodality of this SE event in SUGT1, the subgroup with exclusion Psi appears to be more 'progenitor-cell' like, whereas the subgroup with inclusion Psi is likely to be geared toward nascent neurons.

3.5 Discussion

We developed the Expedition software suite to address key aspects of AS analysis from single-cell RNA-seq data. The Expedition suite consists of three packages that integrate the detection and quantification of AS events (`outrigger`) with the assignment of modalities (`anchor`), and a method for visualization of changes in modality (`bonvoyage`). As an application, Expedition was used to analyze AS in single cells from three homogenous cell-types, specifically human pluripotent stem cells, neural progenitors and motor neurons.

Many studies have performed RNA sequencing from bulk samples to measure AS, where the "relative" inclusion ($\Delta\Psi$) of alternative exons in a comparison (e.g. treatment versus control or between tissues) is the primary metric used. However, $\Delta\Psi$ comparison across all single cells are impractical. Thus, robust estimation of Psi is required to assess the distribution of Psi amongst a population of single cells. It is also important that Psi values reflect the actual biological phenomenon, such that a Psi value of 0.5 indicates that 50% of transcripts include the alternative exon while the other 50% exclude it. Thus, using Psi of 0.5 as a prior in probabilistic models and assessing the confidence of estimates by resampling data[194] is not appropriate in single cell splicing analysis as it does not eliminate cases where the observed data and annotation are incompatible (examples shown in Supplementary Software Fig. 1). In contrast, `outrigger` identifies splicing events by constructing *de novo* splicing annotation based on only junction-spanning reads, reconstructs the exon trio (quartet) for SE (MXE) events using graph traversal, and quantifies Psi. `outrigger` also applies user-defined rules that ensure compatibility and sufficient read coverage of the AS events.

`anchor` enables the robust classification of AS exons into five modalities (included, middle, excluded, bimodal and multimodal). `anchor` characterizes AS events by their distribution and variation at the population level using a Bayesian approach, instead of estimating the noise or cell-to-cell variation of AS events[25]. The representation of modalities in all three cell-types is remarkably consistent: 30% excluded, 50% included and 20% bimodal modalities, with small contributions from middle and multimodal modalities, indicating that AS is largely unimodal at the single-cell level. The ability to categorize AS distribution and variation into modalities allowed us to identify distinct sequence and evolutionary features for the three major modalities (summarized in **Figure 3.18**g). While high variance bimodal and multimodal AS events exhibit some features intermediate between included and excluded modalities, other features suggest that these AS events reflect an evolutionarily important class of exons distinct from included and excluded. High variance events contain more highly conserved and longer flanking intronic sequences. The conserved flanking intronic sequences contain cis-motifs enriched for U or UA nucleotides, in contrast to the G rich sequence in included modality. G-rich sequences have been shown to create G-quadruplexes that increase efficiency of splicing[266–268], and thus the lack of G-rich sequences in bimodal may promote their flexibility to be regulated by trans-factors. Interestingly, high variance AS events are also enriched for genes present in more recently evolved phylostrata. This enrichment is concomitant with a peak of gene emergence associated with the evolution of multicellularity, shortly before the Cambrian explosion[223]. At the same time, orthologous exons of the human bimodal AS events detected in our cells are also more frequently regulated as AS across other mammalian lineages[213;214].

Lastly, a distinct property of multimodal AS exons is their preference to maintain protein translatability, possibly with a different function, between the two isoforms. It appears that multimodal exons provide cells flexibility to increase protein diversity without severely compromising protein-coding capacity. This is in contrast to the exons within the included or excluded modalities that tend to create or disrupt reading frames. While it is currently unknown whether these multimodal AS events are a consequence of selective allelic expression or splicing, our evidence suggests that the creation and preservation of bimodal AS exons is required to build a flexible repertoire of protein variants to efficiently cope with evolutionary or environmental changes. Moreover, we illustrate that high variance AS events reveals cellular states invisible to gene expression analysis alone, emphasizing the need to analyze AS at the single cell level. Our findings in single cells that high variance AS events are primary determinants of cell-type-specific splicing is reminiscent of findings that the cell-type- or state-specific master regulators are more likely to be variable in either gene expression[27;269] or epigenetic control[270].

In summary, our study provides a technological framework to deconvolute the complexity of AS at a single cell level. Prospectively, Expedition can be applied to other increasingly popular data types represented by distributions of continuous variables (including but not limited to RNA-editing, nucleotide modifications such as psuedo-uridine and N6-methyl adenosine, alternative polyadenylation sites, and polyA tail lengths), providing advanced analysis to categorize, and describe these molecular features at single-cell resolution.

3.6 Acknowledgements

The authors would like to thank members of the Yeo lab for critical reading of the manuscript. This work was supported by grants from the National Institute of Health (HG004659, NS075449, AI095277, MH107369 and HD085902), the California Institute of Regenerative Medicine (RB3-05009, RB4-06045) and the Brain Research Foundation (BRFSG-2014-14) to G.W.Y. G.W.Y. is an Alfred P. Sloan Research Fellow. O.B.B is a National Defense Science and Engineering Graduate Fellow and a NumFOCUS John Hunter Technology Fellow. We thank Jennifer Santini for technical support from Microscopy Core in the School of Medicine at UC San Diego (NS047101).

Chapter 3, in full, has been accepted for publication as it may appear in Molecular Cell, 2017, Yan Song*, Olga B Botvinnik*, Michael T Lovci, Boyko Kakaradov, Patrick Liu, Jia L. Xu and Gene W Yeo (* These authors contributed equally to this work). The dissertation author was one of the primary investigators and authors of this paper.

Bibliography

[1] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, Soledad Perez-Amodio, Pierluigi Strippoli, and Silvia Canaider. An estimation of the number of cells in the human body. *Ann. Hum. Biol.*, 40(6):463–471, November 2013.

[2] Hooke, Allestry, and Martyn. *Micrographia, or, Some physiological descriptions of minute bodies made by magnifying glasses :with observations and inquiries thereupon /by R. Hooke*. London :Printed by Jo. Martyn and Ja. Allestry, printers to the Royal Society ... ,, 1665.

[3] A Van Leeuwenhoek. Microscopical observations about animals in the scurf of the teeth. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 14:568–574.

[4] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell RNA sequencing. *Mol. Cell*, 58(4):610–620, 21 May 2015.

[5] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Martha Smets, Heinrich Leonhardt, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell RNA sequencing methods. 29 June 2016.

[6] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, 17:63, 7 April 2016.

[7] Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.*, 46(11):2496–2506, November 2016.

[8] Serena Liu and Cole Trapnell. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res.*, 5, 17 February 2016.

[9] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Res.*, 25(10):1491–1498, October 2015.

[10] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, 16 (3):133–145, March 2015.

[11] Valerie Blanc and Nicholas O Davidson. C-to-U RNA editing: mechanisms leading to genetic diversity. *J. Biol. Chem.*, 278(3):1395–1398, 17 January 2003.

[12] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nat. Rev. Genet.*, 15(12):829–845, December 2014.

[13] Gene W Yeo. *RNA Processing: Disease and Genome-wide Probing*. Springer, 2 June 2016.

[14] Julia K Nussbacher, Ranjan Batra, Clotilde Lagier-Tourenne, and Gene W Yeo. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends Neurosci.*, 38(4):226–236, April 2015.

[15] Guramrit Singh, Gabriel Pratt, Gene W Yeo, and Melissa J Moore. The clothes make the mRNA: Past and present trends in mRNP fashion. *Annu. Rev. Biochem.*, 84:325–354, 11 March 2015.

[16] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res.*, 22(10):2008–2017, October 2012.

[17] C L Kleinman, Adoue, V, and J Majewski. RNA editing of protein sequences: A rare event in human transcriptomes. *RNA*, 18(9):1586–1596, 2012.

[18] S Lianoglou, Garg, V, J L Yang, C S Leslie, and C Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, 2013.

[19] Kazuko Nishikura. Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.*, 79:321–349, 2010.

[20] E Park, B Williams, B J Wold, and A Mortazavi. RNA editing in the human ENCODE RNA-seq data. *Genome Res.*, 22(9):1626–1633, 2012.

[21] Zhiyu Peng, Yanbing Cheng, Bertrand Chin-Ming Tan, Lin Kang, Zhijian Tian, Yuankun Zhu, Wenwei Zhang, Yu Liang, Xueda Hu, Xuemei Tan, Jing Guo, Zirui Dong, Yan Liang, Li Bao, and Jun Wang. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.*, 30(3):253–260, 12 February 2012.

[22] 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. *Proc. Natl. Acad. Sci. U. S. A.*, 111(51):E5593–601, 23 December 2014.

[23] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[24] Kasper Karlsson and Sten Linnarsson. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics*, 18(1):126, 3 February 2017.

[25] 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 Res.*, 24(3):496–510, 2014.

[26] Ernesto Picardi, David Stephen Horner, and Graziano Pesole. Single cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA*, 3 March 2017.

[27] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schragi Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John T 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, 2013.

[28] Lars Velten, Simon Anders, Aleksandra Pekowska, Aino I Järvelin, Wolfgang Huber, Vicent Pelechano, and Lars M Steinmetz. SingleâĂŘcell polyadenylation site mapping reveals 3âĂš isoform choice variability. *Mol. Syst. Biol.*, 11(6):812, 1 June 2015.

[29] Joshua D Welch, Yin Hu, and Jan F Prins. Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Res.*, 44(8):e73, 5 May 2016.

[30] A M Femino, F S Fay, K Fogarty, and R H Singer. Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590, 24 April 1998.

[31] Arjun Raj and Alexander van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Annu. Rev. Biophys.*, 38:255–270, 2009.

[32] Todd M Gierahn, Marc H Wadsworth, 2nd, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love,

and Alex K Shalek. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods*, 13 February 2017.

[33] Iain C Macaulay, Chris P Ponting, and Thierry Voet. Single-Cell multiomics: Multiple measurements from single cells. *Trends Genet.*, 33(2):155–168, February 2017.

[34] 1m_neurons - datasets - single cell - official 10x genomics support. https://support.10xgenomics.com/single-cell/datasets/1M_neurons. Accessed: 2017-4-20.

[35] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 21 May 2015.

[36] Lanier Jaron. You are not a gadget: A manifesto. *New York: Knopf*, 2010.

[37] Adina R Buxbaum, Gal Haimovich, and Robert H Singer. In the right place at the right time: visualizing and understanding mRNA localization. *Nat. Rev. Mol. Cell Biol.*, 16(2):95–109, February 2015.

[38] Sheel Shah, Eric Lubeck, Maayan Schwarzkopf, Ting-Fang He, Alon Greenbaum, Chang Ho Sohn, Antti Lignell, Harry M T Choi, Viviana Gradinaru, Niles A Pierce, and Long Cai. Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing. *Development*, 143(15):2862–2867, 1 August 2016.

[39] Jennifer B Treweek, Ken Y Chan, Nicholas C Flytzanis, Bin Yang, Benjamin E Deverman, Alon Greenbaum, Antti Lignell, Cheng Xiao, Long Cai, Mark S Ladinsky, Pamela J Bjorkman, Charless C Fowlkes, and Viviana Gradinaru. Whole-body tissue stabilization and selective extractions via tissue-hydrogel hybrids for high-resolution intact circuit mapping and phenotyping. *Nat. Protoc.*, 10(11):1860–1896, November 2015.

[40] Bin Yang, Jennifer B Treweek, Rajan P Kulkarni, Benjamin E Deverman, Chun-Kan Chen, Eric Lubeck, Sheel Shah, Long Cai, and Viviana Gradinaru. Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell*, 158(4):945–958, 14 August 2014.

[41] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. RNA imaging. spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 24 April 2015.

[42] Nicola Crosetto, Magda Bienko, and Alexander van Oudenaarden. Spatially resolved transcriptomics and beyond. *Nat. Rev. Genet.*, 16(1):57–66, January 2015.

[43] Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods*, 10(9):857–860, September 2013.

[44] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.*, 10(3):442–458, March 2015.

[45] Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*, 11(4):360–361, April 2014.

[46] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357, 19 October 2016.

[47] Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A Smallwood, Chris P Ponting, Thierry Voet, Gavin Kelsey, Oliver Stegle, and Wolf Reik. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, 13(3):229–232, March 2016.

[48] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, and Jirun Peng. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.*, 26(3):304–319, March 2016.

[49] Jason A Reuter, Damek V Spacek, Reetesh K Pai, and Michael P Snyder. Simul-seq: combined DNA and RNA sequencing for whole-genome and transcriptome profiling. *Nat. Methods*, 13(11):953–958, November 2016.

[50] Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, 6(6):451–464, June 2005.

[51] Benjamin B Kaufmann and Alexander van Oudenaarden. Stochastic gene expression: from single molecules to the proteome. *Curr. Opin. Genet. Dev.*, 17(2):107–112, April 2007.

[52] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 17 October 2008.

[53] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.*, 4(10): e309, October 2006.

[54] Diana Y Vargas, Khyati Shah, Mona Batish, Michael Levandoski, Sourav Sinha, Salvatore A E Marras, Paul Schedl, and Sanjay Tyagi. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell*, 147(5): 1054–1065, 23 November 2011.

[55] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, 33(2):155–160, February 2015.

[56] Kenneth J Livak, Alex J Tipping, Tariq Enver, Andrew J Goldson, Darren W Sexton, Chris Holmes, and Quin F Wills. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.*, 31(8):748–752, 2013.

[57] Keren Bahar Halpern, Inbal Caspi, Doron Lemze, Maayan Levy, Shanie Landen, Eran Elinav, Igor Ulitsky, and Shalev Itzkovitz. Nuclear retention of mRNA in mammalian tissues. *Cell Rep.*, 13(12):2653–2662, 29 December 2015.

[58] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of transcript variability in single mammalian cells. *Cell*, 163(7):1596–1610, 17 December 2015.

[59] Adam Ameur, Ammar Zaghlool, Jonatan Halvardson, Anna Wetterbom, Ulf Gyllensten, Lucia Cavelier, and Lars Feuk. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural & Molecular Biology*, 18(12):1435–1440, 2011.

[60] Douglas L Black. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, 72:291–336, 27 February 2003.

[61] Javier F Cáceres and Alberto R Kornblihtt. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.*, 18 (4):186–193, 2002.

[62] Christopher R Day, Huimin Chen, Antoine Coulon, Jordan L Meier, and Daniel R Larson. High-throughput single-molecule screen for small-molecule perturbation of splicing and transcription kinetics. *Methods*, 96: 59–68, 1 March 2016.

[63] Alberto R Kornblihtt, Ignacio E Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J Muñoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.*, 14(3):153–165, 2013.

[64] Yeon Lee and Donald C Rio. Mechanisms and regulation of alternative Pre-mRNA splicing. *Annu. Rev. Biochem.*, 84:291–323, 12 March 2015.

[65] Zeev Waks, Allon M Klein, and Pamela A Silver. Cell-to-cell variability of alternative RNA splicing. *Mol. Syst. Biol.*, 7:1–12, 2011.

[66] Lior Faigenbloom, Nimrod D Rubinstein, Yoel Kloog, Itay Mayrose, Tal Pupko, and Reuven Stein. Regulation of alternative splicing at the single-cell level. *Mol. Syst. Biol.*, 11(12):845, 28 December 2015.

[67] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, Darren Bertagnolli, Jeff Goldy, Nadiya Shapovalova, Sheana Parry, Changkyu Lee, Kimberly Smith, Amy Bernard, Linda Madisen, Susan M Sunkin, Michael Hawrylycz, Christof Koch, and Hongkui Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.*, 19(2):335–346, February 2016.

[68] Alan Derr, Chaoxing Yang, Rapolas Zilionis, Alexey Sergushichev, David M Blodgett, Sambra Redick, Rita Bortell, Jeremy Luban, David M Harlan, Sebastian Kadener, Dale L Greiner, Allon Klein, Maxim N Artyomov, and Manuel Garber. End sequence analysis toolkit (ESAT) expands the extractable information from single-cell RNA-seq data. *Genome Res.*, 26(10): 1397–1410, October 2016.

[69] Rinat Arbel-Goren, Asaf Tal, and Joel Stavans. Phenotypic noise: effects of post-transcriptional regulatory processes affecting mRNA. *Wiley Interdiscip. Rev. RNA*, 5(2):197–207, March 2014.

[70] Hannah Dueck, James Eberwine, and Junhyong Kim. Variation is function: Are single cell differences functionally important?: Testing the hypothesis that single cell variation is required for aggregate function. *Bioessays*, 38(2): 172–180, February 2016.

[71] Orsolya Symmons and Arjun Raj. What's luck got to do with it: Single cells, multiple fates, and biological nondeterminism. *Mol. Cell*, 62(5):788–802, 2 June 2016.

[72] K Yap and E V Makeyev. Functional impact of splice isoform diversity in individual cells. *Biochemical Society Transactions*, 44(4):1079–1085, August 2016. doi: 10.1042/BST20160103. URL http://biochemsoctrans.org/cgi/doi/10.1042/BST20160103.

[73] Antoine Coulon, Matthew L Ferguson, Valeria de Turris, Murali Palangat, Carson C Chow, and Daniel R Larson. Kinetic competition during the transcription cycle results in stochastic RNA processing. *Elife*, 3, 1 October 2014.

[74] Karen Yap, Yixin Xiao, Brad A Friedman, H Shawn Je, and Eugene V Makeyev. Polarizing the neuron through sustained co-expression of alternatively spliced isoforms. *Cell Rep.*, 15(6):1316–1328, 10 May 2016.

[75] J M Gott and R B Emeson. Functions and mechanisms of RNA editing. *Annu. Rev. Genet.*, 34:499–531, 2000.

[76] Ian A Mellis, Rohit K Gupte, Arjun Raj, and Sara H Rouhanifard. Visualizing adenosine to inosine RNA editing in single mammalian cells. 16 November 2016.

[77] Ahmadreza Niavarani, Erin Currie, Yasmin Reyal, Fernando Anjos-Afonso, Stuart Horswell, Emmanuel Griessinger, Jose Luis Sardina, and Dominique Bonnet. APOBEC3A is implicated in a novel class of G-to-A mRNA editing in WT1 transcripts. *PLoS One*, 10(3):e0120089, 25 March 2015.

[78] Nils Knie, Felix Grewe, Simon Fischer, and Volker Knoop. Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns – a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. *BMC Evol. Biol.*, 16(1):134, 2016.

[79] Shalev Itzkovitz and Alexander van Oudenaarden. Validating transcripts with probes and imaging technology. *Nat. Methods*, 8(4 Suppl):S12–9, April 2011.

[80] Arjun Raj. Single-Molecule RNA FISH. In Gordon C K Roberts, editor, *Encyclopedia of Biophysics*, pages 2340–2343. Springer Berlin Heidelberg, 2013.

[81] Tatjana Trcek, Hanae Sato, Robert H Singer, and Lynne E Maquat. Temporal and spatial characterization of nonsense-mediated mRNA decay. *Genes Dev.*, 27(5):541–551, 1 March 2013.

[82] Zakary S Singer, John Yong, Julia Tischler, Jamie A Hackett, Alphan Altinok, M Azim Surani, Long Cai, and Michael B Elowitz. Dynamic heterogeneity

and DNA methylation in embryonic stem cells. *Mol. Cell*, 55(2):319–331, 17 July 2014.

[83] Kirsten L Frieda, James M Linton, Sahand Hormoz, Joonhyuk Choi, Ke-Huan K Chow, Zakary S Singer, Mark W Budde, Michael B Elowitz, and Long Cai. Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111, 5 January 2017.

[84] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 1 July 2016.

[85] Jing Liu, Yangqing Xu, Dan Stoleru, and Adrian Salic. Imaging protein synthesis in cells and tissues with an alkyne analog of puromycin. *Proc. Natl. Acad. Sci. U. S. A.*, 109(2):413–418, 10 January 2012.

[86] Robert A J Signer, Jeffrey A Magee, Adrian Salic, and Sean J Morrison. Haematopoietic stem cells require a highly regulated protein synthesis rate. *Nature*, 509(7498):49–54, 1 May 2014.

[87] James M Halstead, Timothée Lionnet, Johannes H Wilbertz, Frank Wippich, Anne Ephrussi, Robert H Singer, and Jeffrey A Chao. Translation. an RNA biosensor for imaging the first round of translation from single cells to living animals. *Science*, 347(6228):1367–1671, 20 March 2015.

[88] Bin Wu, Carolina Eliscovich, Young J Yoon, and Robert H Singer. Translation dynamics of single mRNAs in live cells and neurons. *Science*, 352(6292): 1430–1435, 17 June 2016.

[89] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9(1):72–74, 2011.

[90] S Nakielny, U Fischer, W M Michael, and G Dreyfuss. RNA transport. *Annu. Rev. Neurosci.*, 20:269–301, 1997.

[91] Mutsuhito Ohno. Size matters in RNA export. *RNA Biol.*, 9(12):1413–1417, December 2012.

[92] Nadeem Siddiqui and Katherine L B Borden. mRNA export and cancer. *Wiley Interdiscip. Rev. RNA*, 3(1):13–25, January 2012.

[93] Katja Strässer, Seiji Masuda, Paul Mason, Jens Pfannstiel, Marisa Oppizzi, Susana Rodriguez-Navarro, Ana G Rondón, Andres Aguilera, Kevin Struhl, Robin Reed, and Ed Hurt. TREX is a conserved complex coupling transcription with messenger RNA export. *Nature*, 417(6886):304–308, 16 May 2002.

[94] Paul Anderson and Nancy Kedersha. RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nat. Rev. Mol. Cell Biol.*, 10 (6):430–436, June 2009.

[95] Paul Anderson and Nancy Kedersha. RNA granules. *J. Cell Biol.*, 172(6): 803–808, 13 March 2006.

[96] Michael D Blower. Molecular insights into intracellular RNA localization. *Int. Rev. Cell Mol. Biol.*, 302:1–39, 2013.

[97] Michael A Kiebler and Gary J Bassell. Neuronal RNA granules: movers and makers. *Neuron*, 51(6):685–690, 21 September 2006.

[98] María Gabriela Thomas, Mariela Loschi, María Andrea Desbats, and Graciela Lidia Boccaccio. RNA granules: the good, the bad and the ugly. *Cell. Signal.*, 23(2):324–334, February 2011.

[99] Andrey Damianov, Yi Ying, Chia-Ho Lin, Ji-Ann Lee, Diana Tran, Ajay A Vashisht, Emad Bahrami-Samani, Yi Xing, Kelsey C Martin, James A Wohlschlegel, and Douglas L Black. Rbfox proteins regulate splicing as part of a large multiprotein complex LASR. *Cell*, 165(3):606–619, 21 April 2016.

[100] B Kate Dredge and Kirk B Jensen. NeuN/Rbfox3 nuclear and cytoplasmic isoforms differentially regulate alternative splicing and nonsense-mediated decay of rbfox2. *PLoS One*, 6(6):e21585, 29 June 2011.

[101] 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. *CellReports*, pages 1–39, 2014.

[102] Michael T Lovci, Dana Ghanem, Henry Marr, Justin Arnold, Sherry Gee, Marilyn Parra, Tiffany Y Liang, Thomas J Stark, Lauren T Gehman, Shawn Hoon, Katlin B Massirer, Gabriel A Pratt, Douglas L Black, Joe W Gray, John G Conboy, and Gene W Yeo. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.*, 20(12):1434–1442, December 2013.

[103] Curtis A Nutter, Elizabeth A Jaworski, Sunil K Verma, Vaibhav Deshmukh, Qiongling Wang, Olga B Botvinnik, Mario J Lozano, Ismail J Abass, Talha Ijaz, Allan R Brasier, Nisha J Garg, Xander H T Wehrens, Gene W Yeo, and Muge N Kuyumcu-Martinez. Dysregulation of RBFOX2 is an early event in cardiac pathogenesis of diabetes. *Cell Rep.*, 15(10):2200–2213, 7 June 2016.

[104] Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and Gideon Rechavi. Topology of the human and mouse m6a RNA methylomes revealed by m6a-seq. *Nature*, 485(7397):201–206, 29 April 2012.

[105] Dan Dominissini, Sharon Moshitch-Moshkovitz, Mali Salmon-Divon, Ninette Amariglio, and Gideon Rechavi. Transcriptome-wide mapping of n6-methyladenosine by m6a-seq based on immunocapturing and massively parallel sequencing. *Nat. Protoc.*, 8(1):176–189, 3 January 2013.

[106] Bastian Linder, Anya V Grozhik, Anthony O Olarerin-George, Cem Meydan, Christopher E Mason, and Samie R Jaffrey. Single-nucleotide-resolution mapping of m6a and m6am throughout the transcriptome. *Nat. Methods*, 12(8):767–772, August 2015.

[107] Nejc Haberman, Ina Huppertz, Jan Attig, Julian König, Zhen Wang, Christian Hauer, Matthias W Hentze, Andreas E Kulozik, Hervé Le Hir, Tomaž Curk, Christopher R Sibley, Kathi Zarnack, and Jernej Ule. Insights into the design and interpretation of iCLIP experiments. *Genome Biol.*, 18(1):7, 16 January 2017.

[108] Ina Huppertz, Jan Attig, Andrea DâĂŹAmbrogio, Laura E Easton, Christopher R Sibley, Yoichiro Sugimoto, Mojca Tajnik, Julian König, and Jernej Ule. iCLIP: Protein–RNA interactions at nucleotide resolution. *Methods*, 65 (3):274–287, 2014.

[109] Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, 13(6):508–514, June 2016.

[110] Brian J Zarnegar, Ryan A Flynn, Ying Shen, Brian T Do, Howard Y Chang, and Paul A Khavari. irCLIP platform for efficient characterization of protein-RNA interactions. *Nat. Methods*, 13(6):489–492, June 2016.

[111] David A Nelles, Mark Y Fang, Mitchell R O'Connell, Jia L Xu, Sebastian J Markmiller, Jennifer A Doudna, and Gene W Yeo. Programmable RNA tracking in live cells with CRISPR/Cas9. *Cell*, 165(2):488–496, 7 April 2016.

[112] U Deppe, E Schierenberg, T Cole, C Krieg, D Schmitt, B Yoder, and G von Ehrenstein. Cell lineages of the embryo of the nematode caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.*, 75(1):376–380, January 1978.

[113] H M Ellis and H R Horvitz. Genetic control of programmed cell death in the nematode c. elegans. *Cell*, 44(6):817–829, 28 March 1986.

[114] Kevin Da Silva. Neurodegeneration: Mechanistic overlap in ALS. *Nat. Med.*, 20(7):714–714, 7 July 2014.

[115] Magdalini Polymenidou and Don W Cleveland. The seeds of neurodegeneration: prion-like spreading in ALS. *Cell*, 147(3):498–508, 28 October 2011.

[116] Anaïs Aulas, Stéphanie Stabile, and Christine Vande Velde. Endogenous TDP-43, but not FUS, contributes to stress granule assembly via G3BP. *Mol. Neurodegener.*, 7:54, 24 October 2012.

[117] Jozsef Gal, Jiayu Zhang, David M Kwinter, Jianjun Zhai, Hongge Jia, Jianhang Jia, and Haining Zhu. Nuclear localization sequence of FUS and induction of stress granules by ALS mutants. *Neurobiol. Aging*, 32(12): 2323.e27–40, December 2011.

[118] Fernando J Martinez, Gabriel A Pratt, Eric L Van Nostrand, Ranjan Batra, Stephanie C Huelga, Katannya Kapeli, Peter Freese, Seung J Chun, Karen Ling, Chelsea Gelboin-Burkhart, Layla Fijany, Harrison C Wang, Julia K Nussbacher, Sara M Broski, Hong Joo Kim, Rea Lardelli, Balaji Sundararaman, John P Donohue, Ashkan Javaherian, Jens Lykke-Andersen, Steven Finkbeiner, C Frank Bennett, Manuel Ares, Jr, Christopher B Burge, J Paul Taylor, Frank Rigo, and Gene W Yeo. Protein-RNA networks regulated by normal and ALS-Associated mutant HNRNPA2B1 in the nervous system. *Neuron*, 92(4):780–795, 23 November 2016.

[119] Kaia Achim, Jean-Baptiste Pettit, Luis R Saraiva, Daria Gavriouchkina, Tomas Larsson, Detlev Arendt, and John C Marioni. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.*, 33(5):503–509, May 2015.

[120] Tomoya Mori, Junko Yamane, Kenta Kobayashi, Nobuko Taniyama, Takanori Tano, and Wataru Fujibuchi. Development of 3D tissue reconstruction method from single-cell RNA-seq data. *Genomics and Computational Biology*, 3(1):53, 31 January 2017.

[121] Jean-Baptiste Pettit, Raju Tomer, Kaia Achim, Sylvia Richardson, Lamiae Azizi, and John Marioni. Identifying cell types from spatially referenced single-cell expression datasets. *PLoS Comput. Biol.*, 10(9):e1003824, September 2014.

[122] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, 33(5):495–502, May 2015.

[123] Sten Linnarsson and Sarah A Teichmann. Single-cell genomics: coming of age. *Genome Biol.*, 17:97, 10 May 2016.

[124] Aviv Regev and Others. The human cell atlas, 2016.

[125] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11):1145–1160, 8 November 2016.

[126] Adrian R Ferré-DâĂŹAmaré. RNA-modifying enzymes. *Curr. Opin. Struct. Biol.*, 13(1):49–55, 2003.

[127] Wendy V Gilbert, Tristan A Bell, and Cassandra Schaening. Messenger RNA modifications: Form, distribution, and function. *Science*, 352(6292): 1408–1412, 17 June 2016.

[128] Sheng Li and Christopher E Mason. The pivotal regulatory landscape of RNA modifications. *Annu. Rev. Genomics Hum. Genet.*, 15:127–150, 2 June 2014.

[129] Yogesh Saletore, Kate Meyer, Jonas Korlach, Igor D Vilfan, Samie Jaffrey, and Christopher E Mason. The birth of the epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.*, 13(10):175, 31 October 2012.

[130] Wen-Ju Sun, Jun-Hao Li, Shun Liu, Jie Wu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res.*, 44(D1):D259–65, 4 January 2016.

[131] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie. Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.*, 25(11):1750–1756, November 2015.

[132] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, 17(1):239, 25 November 2016.

[133] Nicholas J Loman and Aaron R Quinlan. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, 30(23):3399–3401, 1 December 2014.

[134] Alexander S Mikheyev and Mandy M Y Tin. A first look at the oxford nanopore MinION sequencer. *Mol. Ecol. Resour.*, 14(6):1097–1102, November 2014.

[135] Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, 7(6):461–465, June 2010.

[136] Christopher Carr. Detection of inosine via strand sequencing, 2016.

[137] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Andrew J Heron, Mark Bruce, Joseph Lloyd, Anthony Warland, Nadia Pantic, Tigist Admassu, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, Botond Sipos, Stephen Young, Sissel Juul, James Clarke, and Daniel J Turner. Highly parallel direct RNA sequencing on an array of nanopores. 12 August 2016.

[138] Julia Salzman, Raymond E Chen, Mari N Olsen, Peter L Wang, and Patrick O Brown. Cell-type specific features of circular RNA expression. *PLoS Genet.*, 9(9):e1003777, 5 September 2013.

[139] Helen A King and André P Gerber. Translatome profiling: methods for genome-scale analysis of mRNA translation. *Brief. Funct. Genomics*, 15(1): 22–31, January 2016.

[140] Adam J Lesiak and John F Neumaier. RiboTag: Not lost in translation. *Neuropsychopharmacology*, 41(1):374–376, 1 January 2016.

[141] G A Brar, M Yassour, N Friedman, A Regev, N T Ingolia, and J S Weissman. High-Resolution view of the yeast meiotic program revealed by ribosome profiling. *Science*, 335(6068):552–557, 2012.

[142] N T Ingolia, S Ghaemmaghami, J R S Newman, and J S Weissman. Genome-Wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, 2009.

[143] Nicholas T Ingolia, Liana F Lareau, and Jonathan S Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, 2011.

[144] Philip C Bevilacqua, Laura E Ritchey, Zhao Su, and Sarah M Assmann. Genome-Wide analysis of RNA secondary structure. *Annu. Rev. Genet.*, 50: 235–266, 23 November 2016.

[145] Ryan A Flynn, Qiangfeng Cliff Zhang, Robert C Spitale, Byron Lee, Maxwell R Mumbach, and Howard Y Chang. Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. *Nat. Protoc.*, 11(2):273–290, February 2016.

[146] Miles Kubota, Catherine Tran, and Robert C Spitale. Progress and challenges for chemical probing of RNA structure inside living cells. *Nat. Chem. Biol.*, 11(12):933–941, December 2015.

[147] Silvi Rouskin, Meghan Zubradt, Stefan Washietl, Manolis Kellis, and Jonathan S Weissman. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505(7485):701–705, 30 January 2014.

[148] Robert C Spitale, Pete Crisalli, Ryan A Flynn, Eduardo A Torre, Eric T Kool, and Howard Y Chang. RNA SHAPE analysis in living cells. *Nat. Chem. Biol.*, 9(1):18–20, January 2013.

[149] Yue Wan, Michael Kertesz, Robert C Spitale, Eran Segal, and Howard Y Chang. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.*, 12(9):641–655, 18 August 2011.

[150] Yue Wan, Kun Qu, Qiangfeng Cliff Zhang, Ryan A Flynn, Ohad Manor, Zhengqing Ouyang, Jiajing Zhang, Robert C Spitale, Michael P Snyder, Eran Segal, and Howard Y Chang. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485): 706–709, 30 January 2014.

[151] Jesse M Engreitz, Noah Ollikainen, and Mitchell Guttman. Long noncoding RNAs: spatial amplifiers that control nuclear structure and gene expression. 17(12):756–770, October 2016. doi: 10.1038/nrm.2016.126. URL http://dx.doi.org/10.1038/nrm.2016.126.

[152] Guobing Chen and Nan-Ping Weng. Analyzing the phenotypic and functional complexity of lymphocytes using CyTOF (cytometry by time-of-flight). *Cell. Mol. Immunol.*, 9(4):322–323, July 2012.

[153] Charlotte Giesen, Hao A O Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, 11(4):417–422, April 2014.

[154] Jonathan M Irish, Nikesh Kotecha, and Garry P Nolan. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. *Nat. Rev. Cancer*, 6(2):146–155, February 2006.

[155] Matthew H Spitzer and Garry P Nolan. Mass cytometry: Single cells, many features. *Cell*, 165(4):780–791, 5 May 2016.

[156] Meiye Wu and Anup K Singh. Single-cell protein analysis. *Curr. Opin. Biotechnol.*, 23(1):83–88, February 2012.

[157] Daniel N Wilson and Jamie H Doudna Cate. The structure and function of the eukaryotic ribosome. *Cold Spring Harb. Perspect. Biol.*, 4(5), 1 May 2012.

[158] Alessandro Brombin, Jean-Stéphane Joly, and Françoise Jamen. New tricks for an old dog: ribosome biogenesis contributes to stem cell homeostasis. *Curr. Opin. Genet. Dev.*, 34:61–70, October 2015.

[159] Michael Buszczak, Robert A J Signer, and Sean J Morrison. Cellular differences in protein synthesis regulate tissue homeostasis. *Cell*, 159(2): 242–251, 9 October 2014.

[160] Zhen Shi and Maria Barna. Translating the genome in time and space: specialized ribosomes, RNA regulons, and RNA-binding proteins. *Annu. Rev. Cell Dev. Biol.*, 31:31–54, 5 October 2015.

[161] Wendy V Gilbert. Functional specialization of ribosomes? *Trends Biochem. Sci.*, 36(3):127–132, March 2011.

[162] Shifeng Xue and Maria Barna. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat. Rev. Mol. Cell Biol.*, 13(6): 355–369, 23 May 2012.

[163] Qunxing Ding, William R Markesbery, Qinghua Chen, Feng Li, and Jeffrey N Keller. Ribosome dysfunction is an early event in alzheimer's disease. *J. Neurosci.*, 25(40):9171–9175, 5 October 2005.

[164] Emily F Freed, Franziska Bleichert, Laura M Dutca, and Susan J Baserga. When ribosomes go bad: diseases of ribosome biogenesis. *Mol. Biosyst.*, 6 (3):481–493, March 2010.

[165] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 13 March 2003.

[166] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer,

Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber, and Bernhard Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, 29 May 2014.

[167] Anurada D Arya, David I Wilson, Diana Baralle, and Michaela Raponi. RBFOX2 protein domains and cellular activities. *Biochem. Soc. Trans.*, 42(4): 1180–1183, August 2014.

[168] 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, 15 March 2014.

[169] Long Cai, Chiraj K Dalal, and Michael B Elowitz. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*, 455(7212): 485–490, 25 September 2008.

[170] Cellina Cohen-Saidon, Ariel A Cohen, Alex Sigal, Yuvalal Liron, and Uri Alon. Dynamics and variability of ERK2 response to EGF in individual living cells. *Mol. Cell*, 36(5):885–893, 11 December 2009.

[171] Nan Hao and Erin K O'Shea. Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nat. Struct. Mol. Biol.*, 19(1): 31–39, 18 December 2011.

[172] Galit Lahav, Nitzan Rosenfeld, Alex Sigal, Naama Geva-Zatorsky, Arnold J Levine, Michael B Elowitz, and Uri Alon. Dynamics of the p53-mdm2 feedback loop in individual cells. *Nat. Genet.*, 36(2):147–150, February 2004.

[173] Joe H Levine, Yihan Lin, and Michael B Elowitz. Functional roles of pulsing in genetic circuits. *Science*, 342(6163):1193–1200, 6 December 2013.

[174] Yihan Lin, Chang Ho Sohn, Chiraj K Dalal, Long Cai, and Michael B Elowitz. Combinatorial gene regulation by modulation of relative pulse timing. *Nature*, 527(7576):54–58, 5 November 2015.

[175] James C W Locke, Jonathan W Young, Michelle Fontes, María Jesús Hernández Jiménez, and Michael B Elowitz. Stochastic pulse regulation in bacterial stress response. *Science*, 334(6054):366–369, 21 October 2011.

[176] D E Nelson, A E C Ihekwaba, M Elliott, J R Johnson, C A Gibney, B E Foreman, G Nelson, V See, C A Horton, D G Spiller, S W Edwards, H P McDowell, J F Unitt, E Sullivan, R Grimley, N Benson, D Broomhead, D B Kell, and M R H White. Oscillations in NF-kappaB signaling control the dynamics of gene expression. *Science*, 306(5696):704–708, 22 October 2004.

[177] Jeremy E Purvis and Galit Lahav. Encoding and decoding cellular information through signaling dynamics. *Cell*, 152(5):945–956, 28 February 2013.

[178] Nir Yosef and Aviv Regev. Impulse control: Temporal dynamics in gene transcription. *Cell*, 144(6):886–896, 2011.

[179] Sami Hocine, Pascal Raymond, Daniel Zenklusen, Jeffrey A Chao, and Robert H Singer. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nat. Methods*, 10(2):119–121, February 2013.

[180] Sanjay Tyagi. Imaging intracellular RNA distribution and dynamics in living cells. *Nat. Methods*, 6(5):331–338, May 2009.

[181] Thomas C Custer and Nils G Walter. In vitro labeling strategies for in cellulo fluorescence microscopy of single ribonucleoprotein machines. *Protein Sci.*, 28 December 2016.

[182] Philip J Santangelo. Molecular beacons and related probes for intracellular RNA imaging. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.*, 2(1):11–19, January 2010.

[183] Pierluigi Strippoli, Silvia Canaider, Francesco Noferini, Pietro D'Addabbo, Lorenza Vitale, Federica Facchin, Luca Lenzi, Raffaella Casadei, Paolo Carinci, Maria Zannotti, and Flavia Frabetti. Uncertainty principle of genetic information in a living cell. *Theor. Biol. Med. Model.*, 2:40, 30 September 2005.

[184] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, 14(9):618–630, 2013.

[185] Iain C Macaulay, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, Mabel J Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M Shirley, Miriam Smith, Niels Van der Aa, Ruby Banerjee, Peter D Ellis, Michael A Quail, Harold P Swerdlow, Magdalena Zernicka-Goetz, Frederick J Livesey, Chris P Ponting, and Thierry Voet. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods*, 12(6): 519–522, June 2015.

[186] Ana Fiszbein and Alberto R Kornblihtt. Alternative splicing switches: Important players in cell differentiation. *Bioessays*, 27 April 2017.

[187] Hagen Tilgner and Roderic Guigó. From chromatin to splicing: RNA-processing as a total artwork. *Epigenetics*, 5(3), 2010.

[188] Eliezer Calo and Joanna Wysocka. Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, 49(5):825–837, 2013.

[189] H H Ng and A Bird. DNA methylation and chromatin modification. *Curr. Opin. Genet. Dev.*, 9(2):158–163, April 1999.

[190] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'Shea, Peter J Park, Bing Ren, Joan C Ritland, Jay Shendure, Sheng Zhong, and The 4D Nucleome Network. The 4D nucleome project. 26 January 2017.

[191] Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, Ryan A Pak, Andrew N Gray, Carol A Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan S Weissman. A multiplexed Single-Cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, 15 December 2016.

[192] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable Single-Cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 15 December 2016.

[193] Aaron McKenna, Gregory M Findlay, James A Gagnon, Marshall S Horwitz, Alexander F Schier, and Jay Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, 29 July 2016.

[194] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, November 2010. doi: 10.1038/nmeth.1528. URL http://www.nature.com/doifinder/10.1038/nmeth.1528.

[195] Jing Wang, Sijin Wen, W Fraser Symmans, Lajos Pusztai, and Kevin R Coombes. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer informatics*, 7:199–216, 2009. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19718451&retmode=ref&cmd=prlinks.

[196] J A Hartigan and P M Hartigan. The dip test of unimodality. *The Annals of Statistics*, 1985. doi: 10.2307/2241144. URL http://www.jstor.org/stable/2241144.

[197] Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, January 1991. doi: 10.1234/12345678. URL http://dl.acm.org/citation.cfm?id=129837&coll=DL&dl=GUIDE&CFID=555373975&CFTOKEN=99027636.

[198] SAM BAM Format Specification Working Group. Sequence alignment/map format specification, 2014. URL http://scholar.google.com/scholar?q=related:vDsyBhxBicMJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5.

[199] 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, January 2013. doi: 10.1093/bioinformatics/bts635. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23104886&retmode=ref&cmd=prlinks.

[200] Ryan K Dale. daler/gffutils. URL https://github.com/daler/gffutils.

[201] eugene-eeo/graphlite. URL https://github.com/eugene-eeo/graphlite.

[202] C Joel McManus and Brenton R Graveley. RNA structure and the mechanisms of alternative splicing. *Current Opinion in Genetics & Development*, 21(4):373–379, August 2011. doi: 10.1016/j.gde.2011.04.001. URL http://dx.doi.org/10.1016/j.gde.2011.04.001.

[203] Mariano A Garcia-Blanco, Andrew P Baraniak, and Erika L Lasda. Alternative splicing in disease and therapy. *Nature Biotechnology*, 22(5), May 2004. doi: 10.1038/nbt964. URL http://www.nature.com/doifinder/10.1038/nbt964.

[204] Z Ye, Z Chen, X Lan, S Hara, B Sunkel, T H M Huang, L Elnitski, Q Wang, and V X Jin. Computational analysis reveals a correlation of exon-skipping events with splicing, transcription and epigenetic factors. *Nucleic Acids Research*, 42(5):2856–2869, March 2014. doi: 10.1093/nar/gkt1338. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1338.

[205] T E Oliphant. Python for Scientific Computing. *Computing in Science &amp; Engineering*, 9(3):10–20, 2007. doi: 10.1109/MCSE.2007.58. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4160250.

[206] K J Millman and M Aivazis. Python for Scientists and Engineers. *Computing in Science &amp; Engineering*, 13(2):9–12, 2011. doi: 10.1109/MCSE.2011.36. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5725235.

[207] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12, February 2011. URL http://portal.acm.org/citation.cfm?id=1953048.2078195&coll=DL&dl=ACM&CFID=422733224&CFTOKEN=93183407.

[208] Alistair Muldal. alimuldal/diptest. URL https://github.com/alimuldal/diptest.

[209] D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999. doi: 10.1038/44565. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=10548103&retmode=ref&cmd=prlinks.

[210] Jason M Johnson, John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M Loerch, Christopher D Armour, Ralph Santos, Eric E Schadt, Roland Stoughton, and Daniel D Shoemaker. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302(5653):2141–2144, December 2003. doi: 10.1126/science.1090100. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1090100.

[211] 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. *Nature Genetics*, 40(12):1413–1415, December 2008. doi: 10.1038/ng.259. URL http://www.nature.com/doifinder/10.1038/ng.259.

[212] J i Takeda, Y Suzuki, R Sakate, Y Sato, T Gojobori, T Imanishi, and S Sugano. H-DBAS: human-transcriptome database for alternative splicing: update 2010. *Nucleic Acids Research*, 38(Database):D86–D90, December 2009. doi: 10.1093/nar/gkp984. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkp984.

[213] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B Burge. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, 338(6114):1593–1599, December 2012. doi: 10.1126/science.1228186. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1228186.

[214] N L Barbosa-Morais, M Irimia, Q Pan, H Y Xiong, S Gueroussov, L J Lee, V Slobodeniuc, C Kutter, S Watt, R Colak, T Kim, C M Misquitta-Ali, M D Wilson, P M Kim, D T Odom, B J Frey, and B J Blencowe. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, 338 (6114):1587–1593, December 2012. doi: 10.1126/science.1230612. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1230612.

[215] Naoki Nariai, Kaname Kojima, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, and Masao Nagasaki. TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads. *BMC Genomics*, 15 Suppl 10:S5–S5, January 2014. doi: 10.1186/1471-2164-15-S10-S5. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-S10-S5.

[216] Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, Lior Pachter, and Cole Trapnell. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc.*, 7(3):562–578, 2012.

[217] Jing Zhang, C-C Jay Kuo, and Liang Chen. WemIQ: an accurate and robust isoform quantification method for RNA-seq data. *Bioinformatics*, 31(6):878–885, March 2015. doi: 10.1093/bioinformatics/btu757. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu757.

[218] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, 30(8):777–782, 2012.

[219] Rob Patro, Stephen M Mount, and Carl Kingsford. sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature Biotechnology*, pages 1–6, April 2014. doi: 10.1038/nbt.2862. URL http://dx.doi.org/10.1038/nbt.2862.

[220] Heather R Christofk, Matthew G Vander Heiden, Marian H Harris, Arvind Ramanathan, Robert E Gerszten, Ru Wei, Mark D Fleming, Stuart L Schreiber, and Lewis C Cantley. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature*, 452(7184):230–233, March 2008. doi: 10.1038/nature06734. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=18337823&retmode=ref&cmd=prlinks.

[221] M Takenaka, T Noguchi, H Inoue, K Yamada, T Matsuda, and T Tanaka. Rat pyruvate kinase M gene. its complete structure and characterization of the 5'-flanking region. *J. Biol. Chem.*, 264(4):2363–2367, 5 February 1989.

[222] Yoseph Barash, John A Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J Blencowe, and Brendan J Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, 2010. doi: 10.1038/nature09000. URL http://www.nature.com/doifinder/10.1038/nature09000.

[223] Tomislav Domazet-Lošo and Diethard Tautz. An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular biology and evolution*, 25(12):2699–2707, December 2008. doi: 10.1093/molbev/msn214. URL http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msn214.

[224] Hannah Stower. Alternative splicing: Regulating Alu element 'exonization'. *Nature Reviews Genetics*, pages 1–1, February 2013. doi: 10.1038/nrg3428. URL http://dx.doi.org/10.1038/nrg3428.

[225] Gene Wei-Ming Yeo, Eric Lyman Van Nostrand, and Tiffany Y Liang. Discovery and Analysis of Evolutionarily Conserved Intronic Splicing Regulatory Elements in Mammalian Genomes. *PLoS Genetics*, preprint (2007):e85, 2005. doi: 10.1371/journal.pgen.0030085.eor. URL http://dx.plos.org/10.1371/journal.pgen.0030085.

[226] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H Matzat, Ryan K Dale, Sarah A Smith, Christopher A Yarosh, Seth M Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S Laishram, Mei Qiao, Howard D Lipshitz, Fabio Piano, Anita H Corbett, Russ P Carstens, Brendan J Frey, Richard A Anderson, Kristen W Lynch, Luiz O F Penalva, Elissa P Lei, Andrew G Fraser, Benjamin J Blencowe, Quaid D Morris, and Timothy R Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, July 2013. doi: 10.1038/nature12311. URL http://www.nature.com/doifinder/10.1038/nature12311.

[227] M Xu and Z Su. A novel alignment-free method for comparing transcription factor binding site motifs. *PLoS ONE*, 2010. doi: 10.1371/journal.pone.0008797.t001. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0008797.

[228] S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=9918945&retmode=ref&cmd=prlinks.

[229] R D Finn, J Clements, and S R Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl):W29–W37, June 2011. doi: 10.1093/nar/gkr367. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr367.

[230] Robert D Finn, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A Salazar, John Tate, and Alex Bateman. The Pfam protein families database: towards a more sustainable future.

Nucleic Acids Research, 44(D1):D279–85, January 2016. doi: 10.1093/nar/gkv1344. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1344.

[231] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L L Sonnhammer, David J Studholme, Corin Yeats, and Sean R Eddy. The pfam protein families database. *Nucleic Acids Res.*, 32(Database issue): D138–41, 1 January 2004.

[232] Jenny U Johansson, Jesper Ericsson, Juliette Janson, Simret Beraki, Davor Stanić, Slavena A Mandic, Martin A Wikström, Tomas Hökfelt, Sven Ove Ögren, Björn Rozell, Per-Olof Berggren, and Christina Bark. An Ancient Duplication of Exon 5 in the Snap25 Gene Is Required for Complex Neuronal Development/Function. *PLoS Genetics*, 4(11):e1000278, November 2008. doi: 10.1371/journal.pgen.1000278.s004. URL http://dx.plos.org/10.1371/journal.pgen.1000278.s004.

[233] Gábor Nagy, Ira Milosevic, Ralf Mohrmann, Katrin Wiederhold, Alexander M Walter, and Jakob B Sørensen. The SNAP-25 linker as an adaptation toward fast exocytosis. *Mol. Biol. Cell*, 19(9):3769–3781, September 2008.

[234] M A Crackower, D S Sinasac, J Xia, J Motoyama, M Prochazka, J M Rommens, S W Scherer, and L C Tsui. Cloning and characterization of two cytoplasmic dynein intermediate chain genes in mouse and human. *Genomics*, 55(3): 257–267, 1 February 1999.

[235] W B Xu and D J Roufa. The gene encoding human ribosomal protein S24 and tissue-specific expression of differentially spliced mRNAs. *Gene*, 169 (2):257–262, 9 March 1996.

[236] Stuart M Chambers, Christopher A Fasano, Eirini P Papapetrou, Mark Tomishima, Michel Sadelain, and Lorenz Studer. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat Biotech*, 27(3):275–280, March 2009. doi: 10.1038/nbt.1529. URL http://www.nature.com/doifinder/10.1038/nbt.1529.

[237] J Jurka, V V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4):462–467, 2005. doi: 10.1159/000084979. URL http://www.karger.com/Article/FullText/84979.

[238] J Harrow, A Frankish, J M Gonzalez, E Tapanari, M Diekhans, F Kokocinski, B L Aken, D Barrell, A Zadissa, S Searle, I Barnes, A Bignell, V Boychenko, T Hunt, M Kay, G Mukherjee, J Rajan, G Despacio-Reyes, G Saunders, C Steward, R Harte, M Lin, C Howald, A Tanzer, T Derrien, J Chrast,

N Walters, S Balasubramanian, B Pei, M Tress, J M Rodriguez, I Ezkurdia, J van Baren, M Brent, D Haussler, M Kellis, A Valencia, A Reymond, M Gerstein, R Guigo, and T J Hubbard. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, September 2012. doi: 10.1101/gr.135350.111. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.135350.111.

[239] Daniel Müllner. fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software*, 53:1–18, 2011. URL http://www.jstatsoft.org/v53/i09/paper.

[240] Z Bar-Joseph, D K Gifford, and T S Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17 Suppl 1:S22–9, 2001. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11472989&retmode=ref&cmd=prlinks.

[241] adrianveres/polo: Python Optimal Leaf Ordering for Hierarchical Clustering, . URL https://github.com/adrianveres/polo.

[242] Joe H Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, April 2012. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

[243] matplotlib: python plotting — Matplotlib 1.5.3 documentation, . URL http://matplotlib.org/.

[244] Seaborn: statistical data visualization — seaborn 0.7.1 documentation, . URL http://seaborn.pydata.org/.

[245] C Wu, I MacLeod, and A I Su. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Research*, 41(D1):D561–D565, December 2012. doi: 10.1093/nar/gks1114. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1114.

[246] Jiwen Xin, Adam Mark, Cyrus Afrasiabi, Ginger Tsueng, Moritz Juchler, Nikhil Gopal, Gregory S Stupp, Timothy E Putman, Benjamin J Ainscough, Obi L Griffith, Ali Torkamani, Patricia L Whetzel, Christopher J Mungall, Sean D Mooney, Andrew I Su, and Chunlei Wu. High-performance web services for querying gene and variant annotation. *Genome Biology*, pages 1–7, June 2016. doi: 10.1186/s13059-016-0953-9. URL http://dx.doi.org/10.1186/s13059-016-0953-9.

[247] H Tang, D Klopfenstein, B Pedersen, P Flick, and K Sato. *GOATOOLS: tools for gene ontology*. Zenodo., 2015. URL http://scholar.google.com/scholar?q=related:BVSb8_z8_3YJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5.

[248] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, August 2005. doi: 10.1101/gr.3715005. URL http://www.genome.org/cgi/doi/10.1101/gr.3715005.

[249] W J Kent, A S Zweig, G Barber, A S Hinrichs, and D Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, August 2010. doi: 10.1093/bioinformatics/btq351. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq351.

[250] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, January 2015. doi: 10.1093/bioinformatics/btu638. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu638.

[251] K R Rosenbloom, J Armstrong, G P Barber, J Casper, H Clawson, M Diekhans, T R Dreszer, P A Fujita, L Guruvadoo, M Haeussler, R A Harte, S Heitner, G Hickey, A S Hinrichs, R Hubley, D Karolchik, K Learned, B T Lee, C H Li, K H Miga, N Nguyen, B Paten, B J Raney, A F A Smit, M L Speir, A S Zweig, D Haussler, R M Kuhn, and W J Kent. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681, January 2015. doi: 10.1093/nar/gku1177. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1177.

[252] W J Kent. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4): 656–664, March 2002. doi: 10.1101/gr.229202. URL http://www.genome.org/cgi/doi/10.1101/gr.229202.

[253] A R Quinlan and I M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010. doi: 10.1093/bioinformatics/btq033. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq033.

[254] Robert Hubley, Robert D Finn, Jody Clements, Sean R Eddy, Thomas A Jones, Weidong Bao, Arian F A Smit, and Travis J Wheeler. The Dfam database of repetitive DNA families. *Nucleic Acids Research*, 44(D1):D81–9, January 2016. doi: 10.1093/nar/gkv1272. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1272.

[255] Olga B Botvinnik. olgabot/kvector. URL https://github.com/olgabot/kvector.

[256] Phlya/adjustText. . URL https://github.com/Phlya/adjustText.

[257] Pybedtools a flexible Python library for manipulating genomic datasets and annotations, March 2012. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21949271&retmode=ref&cmd=prlinks.

[258] Gene Yeo, Dirk Holste, Gabriel Kreiman, and Christopher B Burge. Variation in alternative splicing across human tissues. *Genome Biology*, 5(10):R74–R74, January 2004. doi: 10.1186/gb-2004-5-10-r74. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=15461793&retmode=ref&cmd=prlinks.

[259] olgabot/poshsplice. . URL https://github.com/olgabot/poshsplice.

[260] Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, and István Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, August 2005. doi: 10.1093/bioinformatics/bti541. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti541.

[261] P J A Cock, T Antao, J T Chang, B A Chapman, C J Cox, A Dalke, I Friedberg, T Hamelryck, F Kauff, B Wilczynski, and M J L de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, May 2009. doi: 10.1093/bioinformatics/btp163. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp163.

[262] R De Maesschalck, D Jouan-Rimbaud, and D L Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, January 2000. doi: 10.1016/S0169-7439(99)00047-7. URL http://linkinghub.elsevier.com/retrieve/pii/S0169743999000477.

[263] Shalini Singh, Danielle Howell, Niraj Trivedi, Ketty Kessler, Taren Ong, Pedro Rosmaninho, Alexandre ASF Raposo, Giles Robinson, Martine F Roussel, Diogo S Castro, David J Solecki, and Jonathan A Cooper. Zeb1 controls neuron differentiation and germinal zone exit by a mesenchymal-epithelial-like transition. *eLife*, 5:e12717, May 2016. doi: 10.7554/eLife.12717. URL http://elifesciences.org/lookup/doi/10.7554/eLife.12717.

[264] Erhan Astarci, Ayşe E Erson Bensan, and Sreeparna Banerjee. Matrix metalloprotease 16 expression is downregulated by microRNA-146a in spontaneously differentiating Caco-2 cells. *Development, Growth &amp; Differentiation*, 54(2):216–226, February 2012. doi: 10.1111/j.1440-169X.2011.

01324.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1440-169X.2011.01324.x/full.

[265] Jaewon Ko, Chan Yoon, Giovanni Piccoli, Hye Sun Chung, Karam Kim, Jae-Ran Lee, Hyun Woo Lee, Hyun Kim, Carlo Sala, and Eunjoon Kim. Organization of the Presynaptic Active Zone by ERC2/CAST1-Dependent Clustering of the Tandem PDZ Protein Syntenin-1. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(3):963–970, January 2006. doi: 10.1523/JNEUROSCI.4475-05.2006. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4475-05.2006.

[266] Virginie Marcel, Phong L T Tran, Charlotte Sagne, Ghyslaine Martel-Planche, Laurence Vaslin, Marie-Paule Teulade-Fichou, Janet Hall, Jean-Louis Mergny, Pierre Hainaut, and Eric Van Dyck. G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. *Carcinogenesis*, 32(3):271–278, March 2011.

[267] Mariana Martins Ribeiro, Gleidson Silva Teixeira, Luciane Martins, Marcelo Rocha Marques, Ana Paula de Souza, and Sergio Roberto Peres Line. G-quadruplex formation enhances splicing efficiency of PAX9 intron 1. *Hum. Genet.*, 134(1):37–44, January 2015.

[268] Pasquale Zizza, Chiara Cingolani, Simona Artuso, Erica Salvati, Angela Rizzo, Carmen D'Angelo, Manuela Porru, Bruno Pagano, Jussara Amato, Antonio Randazzo, Ettore Novellino, Antonella Stoppacciaro, Eric Gilson, Giorgio Stassi, Carlo Leonetti, and Annamaria Biroccio. Intragenic g-quadruplex structure formed in the human CD133 and its biological and translational relevance. *Nucleic Acids Res.*, 44(4):1579–1590, 29 February 2016.

[269] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P May, and Aviv Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505): 363–369, 19 June 2014.

[270] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, June 2015. doi: 10.1038/nature14590. URL http://www.nature.com/doifinder/10.1038/nature14590.