Published in final edited form as:

Mol Cell. 2017 July 06; 67(1): 148–161.e5. doi:10.1016/j.molcel.2017.06.003.

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

Yan Song[1,§], **Olga B. Botvinnik**[1,2,§], **Michael T. Lovci**[1,3], **Boyko Kakaradov**[1,2,6], **Patrick Liu**[1], **Jia L. Xu**[1], and **Gene W. Yeo**[1,2,3,4,5,7]

[1]Department of Cellular and Molecular Medicine, Stem Cell Program and Institute for Genomic Medicine; University of California, San Diego; La Jolla, California, 92093; USA

[2]Bioinformatics and Systems Biology Graduate Program; University of California, San Diego; La Jolla, California, 92093; USA

[3]Biomedical Sciences Graduate Program; University of California, San Diego; La Jolla, California, 92093; USA

[4]Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[5]Molecular Engineering Laboratory, A*STAR, Singapore

Abstract

Alternative splicing (AS) generates isoform diversity for cellular identity and homeostasis in multicellular life. Although AS variation has been observed among single cells, little is known about the biological or evolutionary 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

[7]**Correspondence to** be addressed to geneyeo@ucsd.edu.
[6]Present address: Human Longevity Institute
[§]These authors contributed equally.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

DATA AND SOFTWARE AVAILABILITY

All Python code in the form of Jupyter notebooks is available at https://github.com/YeoLab/singlecell_pnm, and the Expedition suite is available here: https://github.com/YeoLab/Expedition, with individual outrigger, (https://github.com/YeoLab/outrigger), anchor (https://github.com/YeoLab/anchor), and bonvoyage (https://github.com/YeoLab/bonvoyage) packages available separately.

ADDITIONAL RESOURCES

Detailed Protocols

Methods S1 in the Supplemental Information details three protocols. Protocol 1 describes the procedure of single cell capture and RNA-sequencing library preparation. Protocol 2 describes the procedure of single cell capture for qPCR. Protocol 3 describes single molecule RNA-FISH.

Author Contributions

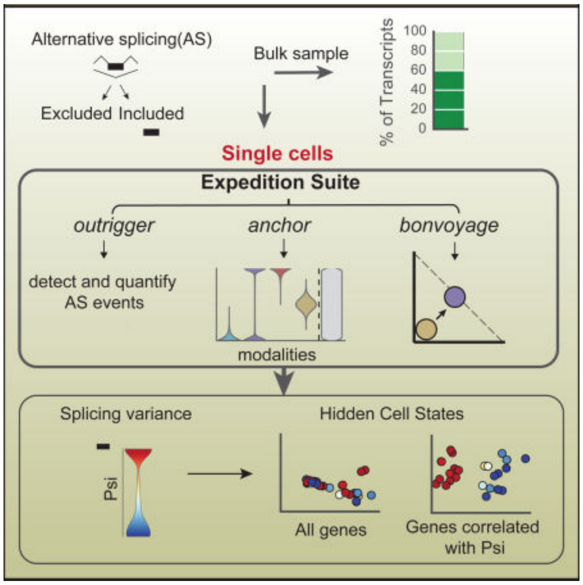
Y.S., O.B.B. and G.W.Y. conceived and designed experiments; Y.S. and J.L.X. performed the experiments; O.B.B. wrote the *Expedition* suite and performed computational analysis; P.L., M.T.L. and B.K. assisted with computational analysis; Y.S. performed and analyzed the sc-qPCR and RNA-FISH data; Y.S., O.B.B. and G.W.Y. wrote the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

single pluripotent stem cells undergoing neuronal differentiation, we discover that up to 20% of AS exons exhibit bimodality. Bimodal exons are flanked by more conserved intronic sequences harboring distinct *cis*-regulatory motifs, constitute much of cell-type specific splicing, are highly dynamic during cellular transitions, preserve reading frame and reveal intricacy of cell states invisible to conventional gene expression analysis. Systematic AS characterization in single cells redefines our understanding of AS complexity in cell biology.

Graphical abstract



Introduction

Over 90% of multi-exon human genes undergo alternative splicing (AS) (Johnson et al., 2003; Pan et al., 2008; Takeda et al., 2010; Wang et al., 2008). Transcriptome profiling by sequencing (RNA-seq) is a powerful means to detect and quantify AS in tissue or cell populations (Barbosa-Morais et al., 2012; Merkin et al., 2012; Wang et al., 2008). Advances in single-cell RNA-seq (scRNA-seq) now enables the detection of AS at the single cell level. Previous studies that investigated AS in single cells were limited to a few exons (Shalek et al., 2013; Waks et al., 2011) or focused on discovering novel splice junctions (Marinov et al., 2014). However, the complexity of AS in single cells remains unappreciated. There is an urgent need to develop robust computational tools to detect, measure and interpret variation in percent-spliced-in (Psi/Ψ) values as a measure of the inclusion rates of alternative events from scRNA-seq datasets.

Many computational tools for AS analysis, such as DEXSeq (Anders et al., 2012) and rMATs (Shen et al., 2014) were developed for bulk RNA-seq datasets. These algorithms focused on determining the change in Ψ of events when comparing two groups (or samples). Algorithms such as MISO (Katz et al., 2010) utilize probabilistic priors, which can result in incorrect assignment of Ψ values (Supplementary Software Figure 1). This is otherwise innocuous when performing pairwise comparisons, however, for hundreds of single cells, calculating all pairwise comparisons is impractical. Other available methods that reconstruct isoforms or estimate read dispersion (Cufflinks, TIGAR2, WemIQ) (Nariai et al., 2013; Trapnell et al., 2012; Zhang et al., 2015) are inappropriate due to the current low molecular capture rate and uneven transcript coverage in scRNA-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 approaches to understand AS complexity and dynamics on a global scale.

Designs

Three key design concepts are important 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. To address these concerns, 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 (*bonvovage*). 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) (Figure 1A). AS events were quantitated by *outrigger* and classified into five distinct modalities by *anchor*. Approximately 75% of AS events exhibit unimodality, where exons are primarily included or excluded with low variance in each cell population. Up to ~20% of AS events are highly varying, composed mostly of bimodal AS events. Interestingly, these bimodal AS events account for essentially all AS events that change modalities during neuronal differentiation, reflecting cell-type specific splicing. We further validated these events by single molecule RNA-FISH (fluorescent *in situ* hybridization) and single cell qPCR. Moreover, we demonstrate that individual bimodal and multimodal events reveal the subpopulations of cells that were homogeneous by conventional global gene expression analysis. Finally, our study revealed that high variance AS events exhibit evolutionary and sequence characteristics distinct from unimodal events, emphasizing the importance of single-cell analysis of RNA processing.

Results

Identification of alternative splicing events in single cells with *outrigger*

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 1A, Figure S1A). Using the Fluidigm C1 system, scRNA-seq libraries were prepared (Ramskold et al., 2012) and sequenced to an average depth of 15–25 million, 100 bp paired-end (PE) reads per cell (Figure S1B). Bulk sequencing libraries were generated from ~1,000 cells. Reads were mapped to the hg19 genome using RNA-STAR (Dobin et al., 2013) and gene expression was estimated as transcripts per million (TPM) using sailfish (Patro et al., 2014). Genes detected in at least 10 cells were retained and ~4,000–11,000 genes were identified per cell in each population (Figure S1C–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 S1E). Lineage-specific transcription factors (POU5F1, PAX6 and ISL1) and RNA binding proteins (LIN28A, MSI1 and RBFOX1) that distinguished each cell-type were observed (Figure S1F). Principal and independent component analysis (PCA and ICA) confirmed distinct iPSC, NPC and MN populations that were each relatively homogenous (Figure S1G–H).

To analyze alternative splicing (AS) events in scRNA-seq, we developed *outrigger*, an algorithm that uses junction-spanning reads to detect and quantify AS. Outrigger builds a *de novo* index based on the aligned reads to identify known and novel AS events (Figure S1I, Supplementary Software Figures 2–4). Strict rules were applied to report only events with sufficient read coverage, valid splice sites, and definitions compatible with skipped exon (SE) and mutually exclusive exon (MXE) annotations (Figure S1J). 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 S1K–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 S1M–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 (Christofk et al., 2008; Takenaka et al., 1989) (Figure 1B). *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) of MNs were observed to possess both isoforms in each cell (Figure 1C–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 1E). Furthermore, ICA based on the Ψ value for each AS event within non-differentially expressed genes generalized our findings with PKM splicing. Single-cell alternative splicing profiles identified by *outrigger* distinguish the three cell-types (Figure 1F, G), demonstrating that AS discerns cell identities independent of gene expression.

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

To categorize the distribution of single cell Ψ values, we developed a Bayesian framework, *anchor*, to designate each AS exon's distribution into one of five modalities: (1) *excluded*, where most cells contain the excluded isoform (Ψ ~ 0); (2) *bimodal*, where two subpopulations with either the excluded (Ψ ~ 0) or included isoform (Ψ ~ 1) can be observed; (3) *included*, where most cells contain the inclusion isoform (Ψ ~ 1); (4) *middle*, where most individual cells have both the inclusion and exclusion isoforms (Ψ ~ 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 Ψ distribution for each AS event was modeled using a Beta distribution (Barash et al., 2010). A two-step process was used to assign modality (Figure 2C). 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 were then assessed for their fit to the two-parameter models, namely middle and bimodal. Remaining events were assigned to the multimodal modality. Using *anchor*, detection of unimodality was robust up to the addition of ~50% uniform random noise (Figure S2A–G) and bimodality was detected up to a 9:1 ratio of inclusion to exclusion, and was robust with up to 70% uniform random noise (Figure S2H–R).

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 2D, Figure S2S), with events largely shared across cell-types (Figure S2T). 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 AS estimates from bulk RNA-seq with a $\Delta\Psi$>0.1 for 40–80% of the events Figure S2U). Simulations indicate that the observed percentages of unimodal and bimodal AS events are statistically unexpected (random permutations expect 99% bimodality and ~0% unimodality; Figure 2E). 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 S2V–Y), suggesting that sampling biases cannot account for the observation of bimodality. Therefore, *anchor* estimated that most AS events are included or excluded in single cells, with up to a fifth of events exhibiting bimodality or multimodality, which are undetected in bulk splicing analyses.

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, exons 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 3A). 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 3B and Figure S3A, 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 levels for excluded modality decrease in the same region. We also examined the evolutionary history of genes containing bimodal and multimodal exons. Interestingly, 98 genes harboring multimodal and 1,832 genes containing bimodal AS events are found in more recently evolved genes, as evidenced by their phylostrata classification (Domazet-Loso and Tautz, 2008), in comparison to genes containing excluded, included AS events or all genes containing any AS exon (Figure 3C). Additionally, orthologous exons of 28 bimodal and 3 multimodal AS are more frequently alternatively spliced across mammals (Figure 3D). The lengths of exon and flanking introns of bimodal AS events are significantly longer than those of the included modality and constitutive exons (Figure 3E, Figure S3C). Repetitive elements such as *Alu* are known to be stochastically exonized (Stower, 2013), and we found *Alu* elements are more enriched within excluded exons, fewer within bimodal exons, and almost absent from AS events in the included modality (Figure S3D). Other features analyzed, including splice site strengths, GC content, showed that bimodal and multimodal exons as intermediate between excluded and included modalities (Figure S3E–I). We conclude that bimodal and multimodal events are enriched for longer flanking introns with higher conservation, present in recently evolved genes, and have orthologs in mammals that are also subject to AS.

Next, we asked whether there are c/s-regulatory elements within flanking intron sequences. We performed PCA on RBP motif (Ray et al., 2013) enrichment scores for conserved flanking introns of AS exons in each modality (Figure S3J–O). We found that bimodal and included modalities are enriched for U-rich and G-rich motifs, respectively, regardless of the cell-types. Moreover, upstream intronic sequences of exons within the included modality are enriched for GC and the downstream counterparts are enriched for GA motifs (Figure 3F). 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.

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 MNs or NPCs (Figure 4A, Figure S4A). To our surprise, we found 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 fifth (~18%) of the AS events detected in two cell-types (iPSCs and NPCs or iPSCs and MNs) exhibited a change in modality (Figure 4B), At least 98% of these switching events are comprised of bimodal AS events (Figure 4C). 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 Gene Ontology categories, such as ‘protein localization or transportation,’ and ‘RNA processing’ (Figure S4B).

Since bimodal and multimodal events are more likely to switch modality during differentiation, we asked whether they are more likely to preserve protein-coding capacity. We required that either the excluded or included isoform (Figure 4D) is part of an annotated coding transcript and utilized *hmmscan* (Eddy, 1998; Finn et al., 2015) to search Pfam (Bateman et al., 2004; Finn et al., 2016) for protein domain clades (Figure 4E). 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 with 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 4F).

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

As highly variant bimodal and multimodal AS events appear to be most sensitive to differentiation, we surmised that they provide an opportunity to identify subpopulations that would otherwise be difficult to discern when analyzing gene expression in scRNA-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 high variance multimodal events in MNs (Figure 5A–C, Figure S5A). Exon 5b is more included in adult brain (Johansson et al., 2008) which may facilitate faster exocytosis (Nagy et al., 2008). We identified genes that correlated with the Ψ values of exon 5a (Spearman correlation $|R| > 0.5$; Figure S5B), which separated the MNs into two clusters (Figure 5D–G). Excitingly, MNs which included exon 5a ($\Psi > 0.5$) express genes essential in cytoskeletal reorganization required for axon guidance and dendritic spine formation and maturation (KATNAL1, ZMYND10, WASF2 and STX16). They also express genes associated with repression of cell proliferation (Figure 5D, 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 genes associated with synapse organization and synaptic vesicle trafficking (SYNGR3, DCTN1, COPA and PCLO) genes associated with intracellular vesicle trafficking, as well as plasma membrane receptors and cell-cell contact genes (Figure 5D, blue labels). Thus, MNs utilizing exon 5b reflect mature neurons with active protein transport and vesicle trafficking. To summarize, genes that correlate with these Ψ values distinguish the two subgroups by PCA, whereas a complete list of expressed genes from MNs fail to do so (Figure 5F, G).

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 5H–M, Figure S5C). DYNC112 encodes a non-catalytic component of the cytoplasmic dynein 1 complex, which acts as a retrograde microtubule motor to transport organelles and vesicles (Crackower et al., 1999). NPCs were clustered into two groups by genes that correlate with Ψ scores of this SE exon (Figure 5J, K). The subgroup with Ψ ~1 are enriched for genes associated with various neuronal genes, such as ONECUT2, a generic transcription factor of motor neurons and genes related with axon guidance and cytoskeleton reorganization (Figure 5J). This subgroup is also enriched for multiple neuron-specific RNA binding proteins (RBPs), including ELAVL2-4 and SRRM4. The subgroup of NPCs with Ψ ~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 that correlate with Ψ scores reveal the substructures of NPC population by PCA (Figure 5L, M). Thus, the bimodality of this SE event is a sufficient statistic to delineate NPCs into a more proliferative subgroup (Ψ ~1) consistent with their progenitor fate and a subgroup ~0) that appears farther on the neuronal trajectory. Many additional examples were found including AS exons in PKM, SUGT1, BRD8, MDM4, MEAF6, and RPN2 (Figure S5D–O), demonstrating 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 analysis.

Transformation of splicing distributions to “waypoints” reveals dynamic of AS events

To visualize changes in modalities, we developed *bonvoyage*, where the distribution of Ψ values of each AS event across single cells from a cell-type is first discretized, then reduced *via* non-negative matrix factorization (NMF) (Figure 6A, left and middle), an algorithm that decomposes data into its constituent parts (Lee and Seung, 1999). The Ψ values are factorized into two components, excluded (x-axis) and included (y-axis), which depict the “waypoint” space (Figure 6A, right). Usage of the waypoint space is illustrated using simulated modality data (Figure S6A–D). Each AS event is depicted as a point in waypoint space, which represents the distribution of Ψ scores in single cells (Figure 6B). 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 6C, D). In such a representation, each modality occupies a discrete region in waypoint space. Also, AS events that change their Ψ 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 6E). 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 S6E–F).

We next sought to establish a global view of AS changes between cell-types. Focusing on exons with large voyages (Figure S6G), we visualized the voyaging exons using vectors between iPSCs and MNs. Consistent with our modality-based analysis (Figure 4A), majority of cell-type specific exons changed from or to the bimodal modality (Figure 6F–G, Figure S6H). To evaluate the consequences of voyages on the protein properties of resulting isoforms, we transformed each protein property into a waypoint-weighted score, enabling an evaluation of protein property based on both isoforms and their distribution in single cells. Among properties investigated, we found that MNs favor splicing that generates more disordered and basic proteins, such as the AS events in RPS24 (ribosomal subunit protein S24), and ZNF207/BuGZ (Figure 7A, B).

To validate the Ψ distributions of bimodal and high-magnitude voyaging AS events during MN 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 like our findings using scRNA-seq (Figure 7C–F, Figure S7A–N). For example, the SE event in RPS24 that introduces a stop codon and removes 3 amino acids from its C-terminal, was partially included in individual iPSCs (middle modality), and became completely included in almost all NPCs and MNs (Figure 7C), which was confirmed by sc-qPCR (Figure 7D). 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 7E). The modalities and voyages of these and many other exons were validated by sc-qPCR (Figure 7F, Figure S7A–N). In conclusion, *bonvoyage* is an effective method to visualize and identify AS events that change across populations.

Discussion

We have developed the *Expedition* software suite, integrating *outrigger, anchor* and *bonvoyage*, to address key issues of AS analysis from scRNA-seq data. Many studies have performed RNA sequencing from bulk samples to measure AS, where the "relative" inclusion (ΔΨ) of alternative exons in a comparison (e.g. treatment versus control or between tissues) is the primary metric used. However, ΔΨ comparisons across all single cells are impractical. Thus, robust estimation of Ψ is required to assess the distribution of Ψ amongst a population of single cells. It is also important that Ψ values reflect the actual biological phenomenon, such that a Ψ value of 0.5 indicates that 50% of transcripts include the alternative exon while the other 50% exclude it. Thus, using Ψ of 0.5 as a prior in probabilistic models and assessing the confidence of estimates by resampling data (Katz et al., 2010) may not be 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 Figure 1). In contrast, *outrigger* identifies splicing events by constructing *de novo* splicing annotation based on only junction-spanning reads, and reconstructs the exon trio (quartet) for SE (MXE) events using graph traversal. Outrigger then applies user-defined rules to ensure compatibility and sufficient read coverage of AS events.

Anchor enables robust classification of AS exons into five modalities (included, middle, excluded, bimodal and multimodal). *Anchor* characterizes distribution and variation at the population level using a Bayesian approach, instead of estimating the noise or cell-to-cell variation of AS events (Marinov et al., 2014). 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 in single cells. 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 7G). 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 introns containing c/s-motifs enriched for U or UA nucleotides, in contrast to the G-rich sequences in included modality. G-rich sequences have been shown to create G-quadruplexes that increase the efficiency of splicing (Marcel et al., 2011; Ribeiro et al., 2015; Zizza et al., 2016), and thus the lack of G-rich sequences proximal to bimodal events may promote their regulatory flexibility. 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 (Domazet-Loso and Tautz, 2008).

Orthologous exons of the human bimodal AS events detected in our cells are also more frequently regulated as AS across other mammalian lineages (Merkin et al., 2012).

A distinct property of bimodal and multimodal AS exons is their preference to maintain protein translatability, possibly with a different function between the two isoforms. Bimodal and multimodal exons in the same cell provide cells the flexibility to increase protein diversity without severely compromising protein-coding capacity. This is in contrast to the exons within the included or excluded modalities, which 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 likely beneficial for the development of a flexible repertoire of protein variants to efficiently cope with evolutionary or environmental changes.

Lastly, we illustrate that high variance AS events reveal cellular states invisible to conventional gene expression analysis alone, emphasizing the utility of analyzing AS at the single cell level. The findings that high variance AS events are primary determinants of cell-type-specific splicing is reminiscent of the findings that the cell-type-or state-specific master regulators are more likely to be variable in either gene expression (Shalek et al., 2013; Shalek et al., 2014) or epigenetic control (Buenrostro et al., 2015).

In summary, our study provides a computational 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 pseudo-uridine and N^6-methyl adenosine, alternative polyadenylation sites, and polyA tail lengths), providing advanced analysis to categorize, and describe these molecular features at single-cell resolution.

Limitations

Currently, the accuracy of scRNA-seq is confounded by the low molecular capture rate and uneven coverage of transcripts. Thus, we have captured AS profiles for moderate to highly expressed genes but not for genes with the lowest abundance. Additionally, we are unlikely to capture AS events that occur closer to the 5′-ends of transcripts. Although we have found that the vast majority of genes use one dominant isoform per cell, it is possible that minor isoforms are not sampled adequately. With more efficient molecular capture rates, the middle and multimodal modalities may comprise larger proportions than we currently estimate. In the future, a comprehensive comparison of outrigger with all available AS algorithms will be useful for scRNA-seq applications. Lastly, while we expect the main conclusions to be robust, applying *Expedition* to greater numbers of cells in diverse cell populations will be informative.

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Request should be directed and will be fulfilled by Lead Contact G.W.Y. (geneyeo@ucsd.edu)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines and Culture—iPSCs (male)(Gore et al., 2011) were cultured on matrigel (Corning) coated plates using mTeSR(Stem Cell Technologies) media with mTeSR supplement (Stem Cell Technologies) at 37 °C incubator with 5% CO2.

Differentiation—Neuron progenitor cells 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 placed 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 2ug/ml FGF) to allow neuron progenitor cell differentiation. NPCs were maintained in NPC media.

Motor neurons were directly differentiated from iPSCs as previous described (Chambers et al., 2009). Briefly, iPSCs were cultured on matrigel coated plates until fully confluent in mTeSR then switch to knock-out serum replacement media (KSR) containing Dorsomorphin(1uM) and SB431542(10uM). Upon day 4 of differentiation, increasing amounts of N2 media (25%, 50%) was added to the KSR. From day 7 of differentiation, 1.5uM retinoic acid and 200nM Smoothened Agonist (SAG, EMD Millipore) were added to induce patterning. Cells were dissociated on day 18 of differentiation and replated in poly-D-lysine and laminin coated plates. Maturation was performed using BDGF(2ng/ml), GDNF(2ng/ml), CNTF(2ng/ml), ascorbid acid, sonic hedgehog and retinoic acid in N2 and B27 media up until 35 days of differentiation.

METHOD DETAILS

Single cell capture and library construction—iPSCs, NPCs and MNs were dissociated using accutase(Stem cell Biotech) and filtered through 40um cell strainers to obtain single cell suspension. Single cells were captured on C1 auto prep platform (Fluidigm) according to manufacturer’s instructions. C1 auto prep chips were visually inspected with a light microscopy at 20× 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 cylcle s 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 1ng to generate libraries using the Nextera XT DNA kit(Illumina). Libraries were multiplexed and sequenced on Illumina HiSeq2000 to generate 100bp PE reads.

Single cell qPCR and primer designs—Single iPSCs, NPCs and 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 (Table S1). 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. 3 housekeeping genes (RPL22, RPL27, PGK) and lineage genes (POU5F1, LIN28A, DPPA2, PAX6, NES, ISL1, MNX1, STMN2) were included.

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, Table S2) 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 Quasar570 and another probe set against exon16 was designed and labeled with Quasar 670.

iPSCs and MNs grown on matrigel coated coverslip were fixed with 3.7% formaldehyde PFA for 10 minutes at room temperature. The probes for constitutive (1.25uM) and alternative exons (1.25uM) 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.

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 (UC San Diego). 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 0.12503BCm for the depth of cell diameter, usually around 10-1203BCm. The resulting stacks of images were deconvoluted using manufacturer software. 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 percentage of the maximal percentage.

QUANTIFICATION AND STATISTICAL ANALYSIS

Primary RNA-Sequencing data processing and outlier cell detection—RNA-sequencing reads were trimmed using *cutadapt* (v1.8.1) of adapter sequences TCGTATGCCGTCTTCTGCTTG, ATCTCGTATGCCGTCTTCTGCTTG, CGACAGGTTCAGAGTTCTACAGTCCGACGATC, GATCGGAAGAGCACACGTCTGAACTCCAGTCAC, $[A]_{50}$, $[T]_{50}$, and mapped to repetitive elements (RepBase v18.05) using the STAR (v2.4.01) (Dobin et al., 2013). Reads did not map to repetitive elements were then mapped to the human genome (hg19), using GENCODE (v19) gene annotations to create the splice junction database. SJ.out.tab files from STAR were used to create alternative splicing annotations and calculate percent spliced-in.

Gene expression was quantified with sailfish using GENCODE v19 protein-coding and long noncoding RNA annotation. Transcript-level expression was then aggregated to genes. Genes with TPM >1 in at least 10 cells were identified (18,594 genes). Cells with < 4,000 expressed genes were filtered out. 63 iPSCs, 73 NPCs, and 70 MNs pass gene expression level quality control. K-means clustering was performed with k=3 on gene expression matrix, with 1000 random initializations. Cells did not clustered into their designated populations were identified as outliers and discarded from splicing analysis. For iPSC: 71 were captured, 63 passed expression QC, 1 was assigned as outlier and 62 were retained. For NPC: 98 were captured, 73 passed QC, 4 were assigned as outliers and 69 were retained; for MN: 93 were captured, 70 passed QC, 10 were assigned as outliers and 60 were retained.

Estimation of alternative splicing—Outrigger (see Supplementary Software) created a custom alternative splicing index on the splice junction (SJ.out.tab) files created by STAR, and GENCODE v19 was used to define possible exons. A total of 40,534 skipped exon (SE) and 13,217 mutually exclusive exon (MXE) possible alternative events were created. Percent spliced-in (Psi/Ψ) score is used to measure the degree of alternative exon inclusion and calculated as inclusion reads/(inclusion reads + 2* exclusion reads). Psi/Ψ scores were calculated for events with a minimum of 10 junction reads. Alternative events were defined by, $0 < \Psi < 1$, $\Psi \neq 0,1$ in at least one cell. AS events were further filtered to be detected in at least 10 cells of a given cell-type, resulting in 13,910 events. Constitutive exons were defined as not appear as the alternative exon in any of the splice types (MXE and SE), with at least 10 reads on both upstream and downstream junctions, in at least 10 cells per cell type.

ICA, hierarchical clustering and GO analysis—To perform ICA (Independent Component analysis) on non-differentially expressed genes (non-DE genes), non-DE genes (12,685) were identified across the three populations using a nonparametric Kruskal-Wallis test with Bonferroni-corrected p-value, called q, with q > 1 as the cutoff. AS events were extracted from non-DE genes and their Psi scores were subjected to ICA. The NAs in splicing matrix were replaced with an arbitrary number (100) out of the range of Psi values. Choice of the arbitrary number does not affected the ICA results.

Hierarchical clustering was performed using the *fastcluster* and the *polo* package (optimal leaf ordering) in Python with Euclidean distance metric and Ward's method.

Gene Ontology (GO) enrichment was performed using *mygene* package in Python with only the "biological process" category. The significance was corrected for multiple hypothesis testing using Bonferroni correction as performed in the Python package *goatools* (https://github.com/tanghaibao/goatools).

Assignment of modalities to AS Ψ distributions—Ψs are continuous value between (0,1), thus distribution of Ψ can be modeled as a Beta distribution. The probability density function for the Beta distribution, Pr(α, β) is defined between (0,1), with parameters α >0 and β >0. The Beta distributions can be described by four parameterizations, which correspond to the four modalities: excluded (1=< α < β), middle (α = β >1), included (α > β >=1) and bimodal (α = β <1). Multimodal modality corresponds to α = β = 1, and was used as null model. The excluded and included modalities vary only one parameter at a time, whereas middle and bimodal modalities vary both α and β simultaneously. Models with more parameters are more likely to fit, thus we fit AS distributions to one-parameter models first, assessing whether $K > K_{cutoff}$ for either excluded or included. If so, it is assigned to the modality with highest K. The distributions don't fit the one-parameter model are then fitted to the two-parameter bimodal and middle models, to assess whether $K > K_{cutoff}$. If the distributions cannot fit to any of the four modalities, they are assigned to multimodal. Modalities are estimated of *anchor* software (see Supplementary Software) using the default parameters. Only the AS events observed in at least 10 cells per cell-type are considered. The performance of anchor was tested extensively using simulated data in comparison to existing bimodality detecting methods (see Supplementary Software).

Molecular features of alternative exons and isoforms—Placental Mammal PhastCons scores were used to represent evolutionary conservation. For average conservation of exons, *bigWigAverageOverBed* (Kent et al., 2010) was used to calculate the mean conservation across each exon. Bases with no annotated conservation were considered as NAs. For base-wise conservation, a memory-mapped GenomicArray was created by HTSeq Python package, which was then queried with the intronic intervals.

To identify repetitive elements in AS exons, Repeat Masker track was downloaded from UCSC Genome Browser and intersected with AS exons by *bedtools intersect*. Repeats were grouped into families defined by the Dfam database of repetitive DNA elements.

Phylostratum scores were used to describe gene age, as previously reported(Domazet-Loso and Tautz, 2008). Since different AS exons in a given gene could be assigned with different AS modalities, this gene was considered in multiple modalities.

To calculate k-mer enrichment, placental mammal conserved elements was downloaded from UCSC and filtered for regions upstream and downstream of AS exons. *Kvector* (https://github.com/olgabot/kvector) was used to count k-mers in these conserved elements. Z-scores of k-mer enrichment were calculated for each intron group defined by cell-type, intron context, and modality against total k-mer counts in the same intron context and celltype, but for all modalities (Figure S3K–L). PCA was performed with Z-scores using the Python package scikit-learn (Figure S3L). k-mers were labeled with the color for the most common nucleotide in the motif (if there was a tie between nucleotides, the k-mer was assigned grey) and for which the squared PCA distance were greater than two squared standard deviations from the center, i.e. an ellipse around the origin of the plot. Python package *adjustText* was used to adjust the text labels for readability. To calculate motif enrichment, the CISBP-RNA binding database (version 0.6) was used. Each position-weight matrix (PWM) was transformed into a Boolean vector of k-mers with no mismatches (Figure S3M). All values ≥

0.1 were set to zero. The resulting motif k-mers matrix was used to calculate motif k-mers enrichment using a t-test, by comparing each motif k-mer to all k-mers of that intron group. PCA was performed on the resulting motif t-statistics (Figure S3O, Figure 3F). Motifs were labeled for those with greater than two squared standard deviations from the center.

To compare Ψ between the bulk sample and single cells, we computed the mean of each pairwise difference of the pooled sample Ψ and every single-cell Ψ.

To evaluate splice site strength, 5′ of exon-intron boundary (−20nt into intron and +3nt into exon) and 3′ of exon-intron boundary (−3nt into exon and +6nt into intron), together with the transcript sequences for these regions were obtained by *bedtools* and *pybedtools*. *MaxEntScan*(Yeo and Burge, 2004) was used to calculate the strength of the splice sites for the AS exons (Figure S3E–F).

To address whether inclusion of AS exons would change coding capacity, we curated translatable transcripts for the ones that have at least one isoform annotated to contain a CDS based on GENCODE v19. A total of 22,152 SE and MXE events reside in such transcripts. If the AS exons participated in transcripts with multiple reading frames, all the reading frames were included. To identify protein domains for the translatable transcripts, *hmmscan* command from the HMMER software suite (v3.1b1)(Finn et al., 2011) was used against Pfam-A database, with a domainindependent E-value cutoff of 10^{-5}. Domains were further aggregated into clades based on Pfam's annotations. Finally, we annotated whether inclusion of the AS exons leads to an annotated translation, with or without a clade or with the same or different clades (Figure 4D–F).

Identification of genes that correlate with AS events—To identify the genes correlating with bimodal and multimodal AS events, we first identified variant genes for which the variances are more than two standard deviations away from the mean variance of all genes. Then, genes with Spearman correlation $|r| > 0.5$ between genes and Ψ scores of each tested AS event were retained as correlated genes. The correlated genes were subsequently used for hierarchical clustering and PCA (Figure 5, Figure S5).

Transformation of splicing distribution into 2-dimensional space—To facilitate visualization and quantitation of splicing distribution changes, we have developed *bonvoyage* to transform Ψ distribution into 2-dimensional space (see Supplementary Software,). First, Ψ distribution was discretization into 10 bins, each of size 0.1. The binned splicing matrix is $B_\psi[k, j]$, where the value of feature (AS event) j are contained in b_k. After transformation, $B_\psi[j,k]$ was reduced *via* non-negative matrix factorization (NMF), to generate a W[j,2] matrix, where each feature(AS event) j can be summarized by two prominent values as exclusion and inclusion. The resulting 2-dimensional space is called 'waypoint space' and the distance between two points in waypoint space is named as 'voyage'. Python package scikit-learn was used for NMF implementation.

Waypoint-weighted protein properties—To obtain protein properties, we used IUPRED(Dosztanyi et al., 2005) to calculate protein disorder and the ProtParam module in BioPython 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-integrated protein properties by using the waypoint space coordinates as weight indexes. p_{included} and p_{excluded} were used to represent the protein property value (e.g. molecular weight or disordered protein score) of each isoform, and w_{included} and w_{excluded} were used to represent the splicing event's waypoint space coordinates for the included (y) and excluded (x) axes. The weighted protein property, p_w, for each cell population was calculated as

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

For properties that have a relative center, e.g. isolectric point for which 7 is the neutral point, the center value, p_{center}, was subtracted for each protein property:

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

To identify protein properties that changed significantly between cell types, Mahalonobis distance (d_m), a non-parametric method to identify outliers from distributions was used. We used $3d_m$ as the threshold for highly changed protein properties.

qPCR data processing—The log expression of each primer set ‘g’ was computed as $\log E_{g,c} = 25 - Ct_{g,c}$, where c is the cell and $Ct_{g,c}$ is the Ct value for corresponding primer set. iPSCs were filtered by (RPL22 > 5, LIN28A > 8 and POU5F1 > 8), NPCs were filtered by (RPL27 > 9, PAX6 > 1, NES > 1) and MNs were filtered by (RPL27 > 9, ISL1 > 2 and STMN2 > 5). A total of 216 single iPSCs, 77 single NPCs and 146 single MNs were retained for further analysis. If $Ct_{exp,c}$ is > 25 (Ct value for the expression primer), the corresponding $Ct_{inc,c}$ (Ct value for the inclusion primer) and Ct_{excc} (Ct value for the exclusion primer) were excluded from analysis. Percentage of inclusion is calculated by $2\hat{}Ct_{inc}/(2\hat{}Ct_{inc} + 2\hat{}Ct_{exc})$. Distribution of percentage of inclusion is plot by violin plot or decomposed into 2-dimension space (nmf(dataset, 2, ‘lee’)) and projected into waypoint space in R.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Yeo lab for critical reading of the manuscript. This work was supported by grants from the NIH (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).

References

Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome research. 2012; 22:2008–2017. [PubMed: 22722343]

Barash Y, Blencowe BJ, Frey BJ. Model-based detection of alternative splicing signals. Bioinformatics. 2010; 26:i325–333. [PubMed: 20529924]

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012; 338:1587–1593. [PubMed: 23258890]

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al. The Pfam protein families database. Nucleic acids research. 2004; 32:D138–141. [PubMed: 14681378]

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015; 523:486–490. [PubMed: 26083756]

Chambers SM, Fasano CA, Papapetrou EP, Tomishima M, Sadelain M, Studer L. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nature biotechnology. 2009; 27:275–280.

Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, Fleming MD, Schreiber SL, Cantley LC. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. Nature. 2008; 452:230–233. [PubMed: 18337823]

Crackower MA, Sinasac DS, Xia J, Motoyama J, Prochazka M, Rommens JM, Scherer SW, Tsui LC. Cloning and characterization of two cytoplasmic dynein intermediate chain genes in mouse and human. Genomics. 1999; 55:257–267. [PubMed: 10049579]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]

Domazet-Loso T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. Molecular biology and evolution. 2008; 25:2699–2707. [PubMed: 18820252]

Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005; 21:3433–3434. [PubMed: 15955779]

Eddy SR. Profile hidden Markov models. Bioinformatics. 1998; 14:755–763. [PubMed: 9918945]

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. HMMER web server: 2015 update. Nucleic acids research. 2015; 43:W30–38. [PubMed: 25943547]

Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011; 39:W29–37. [PubMed: 21593126]

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The Pfam protein families database: towards a more sustainable future. Nucleic acids research. 2016; 44:D279–285. [PubMed: 26673716]

Gore A, Li Z, Fung HL, Young JE, Agarwal S, Antosiewicz-Bourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, et al. Somatic coding mutations in human induced pluripotent stem cells. Nature. 2011; 471:63–67. [PubMed: 21368825]

Johansson JU, Ericsson J, Janson J, Beraki S, Stanic D, Mandic SA, Wikstrom MA, Hokfelt T, Ogren SO, Rozell B, et al. An ancient duplication of exon 5 in the Snap25 gene is required for complex neuronal development/function. PLoS Genet. 2008; 4:e1000278. [PubMed: 19043548]

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science. 2003; 302:2141–2144. [PubMed: 14684825]

Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods. 2010; 7:10091–015.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010; 26:2204–2207. [PubMed: 20639541]

Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401:788–791. [PubMed: 10548103]

Marcel V, Tran PL, Sagne C, Martel-Planche G, Vaslin L, Teulade-Fichou MP, Hall J, Mergny JL, Hainaut P, Van Dyck E. G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. Carcinogenesis. 2011; 32:271–278. [PubMed: 21112961]

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome research. 2014; 24:496–510. [PubMed: 24299736]

Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012; 338:1593–1599. [PubMed: 23258891]

Nagy G, Milosevic I, Mohrmann R, Wiederhold K, Walter AM, Sorensen JB. The SNAP-25 linker as an adaptation toward fast exocytosis. Mol Biol Cell. 2008; 19:3769–3781. [PubMed: 18579690]

Nariai N, Hirose O, Kojima K, Nagasaki M. TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. Bioinformatics. 2013; 29:2292–2299. [PubMed: 23821651]

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics. 2008; 40:1413–1415. [PubMed: 18978789]

Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature biotechnology. 2014; 32:462–464.

Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. Full-length mRNA-Seq from singlecell levels of RNA and individual circulating tumor cells. Nature biotechnology. 2012; 30:777–782.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499:172–177. [PubMed: 23846655]

Ribeiro MM, Teixeira GS, Martins L, Marques MR, de Souza AP, Line SR. G-quadruplex formation enhances splicing efficiency of PAX9 intron 1. Hum Genet. 2015; 134:37–44. [PubMed: 25204874]

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013; 498:236–240. [PubMed: 23685454]

Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014; 510:363–369. [PubMed: 24919153]

Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111:E5593–5601. [PubMed: 25480548]

Stower H. Alternative splicing: Regulating Alu element ‘exonization’. *Nature reviews*. Genetics. 2013; 14:152–153.

Takeda J, Suzuki Y, Sakate R, Sato Y, Gojobori T, Imanishi T, Sugano S. H-DBAS: human-transcriptome database for alternative splicing: update 2010. Nucleic acids research. 2010; 38:D86–90. [PubMed: 19969536]

Takenaka M, Noguchi T, Inoue H, Yamada K, Matsuda T, Tanaka T. Rat pyruvate kinase M gene. Its complete structure and characterization of the 5′-flanking region. The Journal of biological chemistry. 1989; 264:2363–2367. [PubMed: 2914912]

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012; 7:562–578. [PubMed: 22383036]

Waks Z, Klein AM, Silver PA. Cell-to-cell variability of alternative RNA splicing. Mol Syst Biol. 2011; 7:506. [PubMed: 21734645]

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–476. [PubMed: 18978772]

Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology : a journal of computational molecular cell biology. 2004; 11:377–394. [PubMed: 15285897]

Zhang J, Kuo CC, Chen L. WemIQ: an accurate and robust isoform quantification method for RNA-seq data. Bioinformatics. 2015; 31:878–885. [PubMed: 25406327]

Zizza P, Cingolani C, Artuso S, Salvati E, Rizzo A, D'Angelo C, Porru M, Pagano B, Amato J, Randazzo A, et al. Intragenic G-quadruplex structure formed in the human CD133 and its biological and translational relevance. Nucleic acids research. 2016; 44:1579–1590. [PubMed: 26511095]

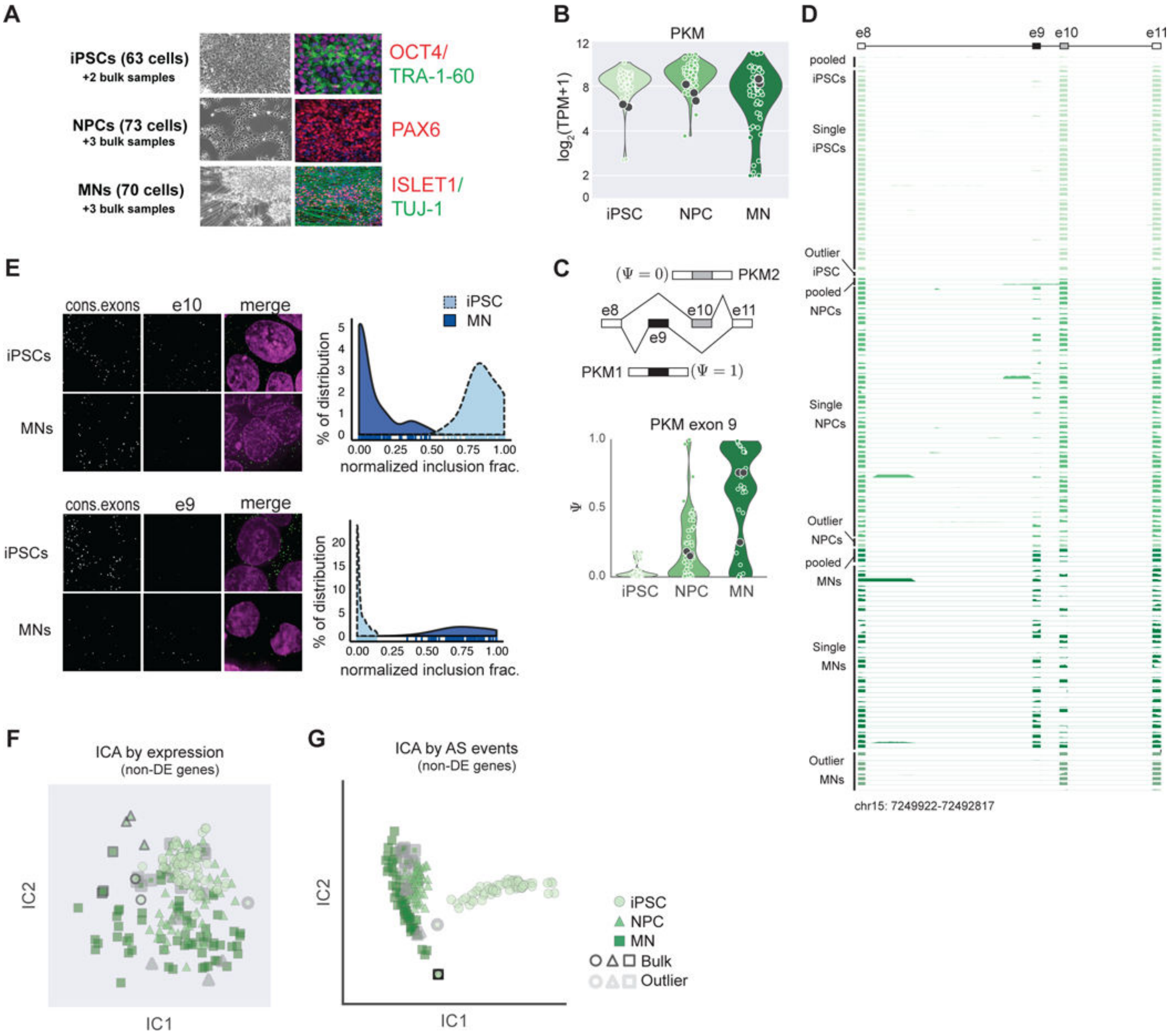


Figure 1. Cell-type specific alternative splicing is an independent feature of cell identity

(A) Human iPSCs were directly differentiated into neuron progenitor cells (NPC) or motor neurons (MN) *in vitro.* Cell identity was verified by immunofluorescence staining. 63 iPSCs, 73 NPCs and 70 MNs 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.

(C) Differential inclusion of a mutually exclusive exon (MXE) alternative splicing (AS) event in PKM is observed in the three cell-types from scRNA-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. Ψ score is estimated by *outrigger* (see STAR 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 are shown 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.

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 Bonferroni-corrected q-values > 1.

(F) ICA on gene expression values of non-DE genes fails to distinguish the three cell-types.

(G) ICA on Ψ scores of the AS events residing in non-DE genes groups iPSCs, NPCs and MNs independent of gene expression. See also Figure S1.

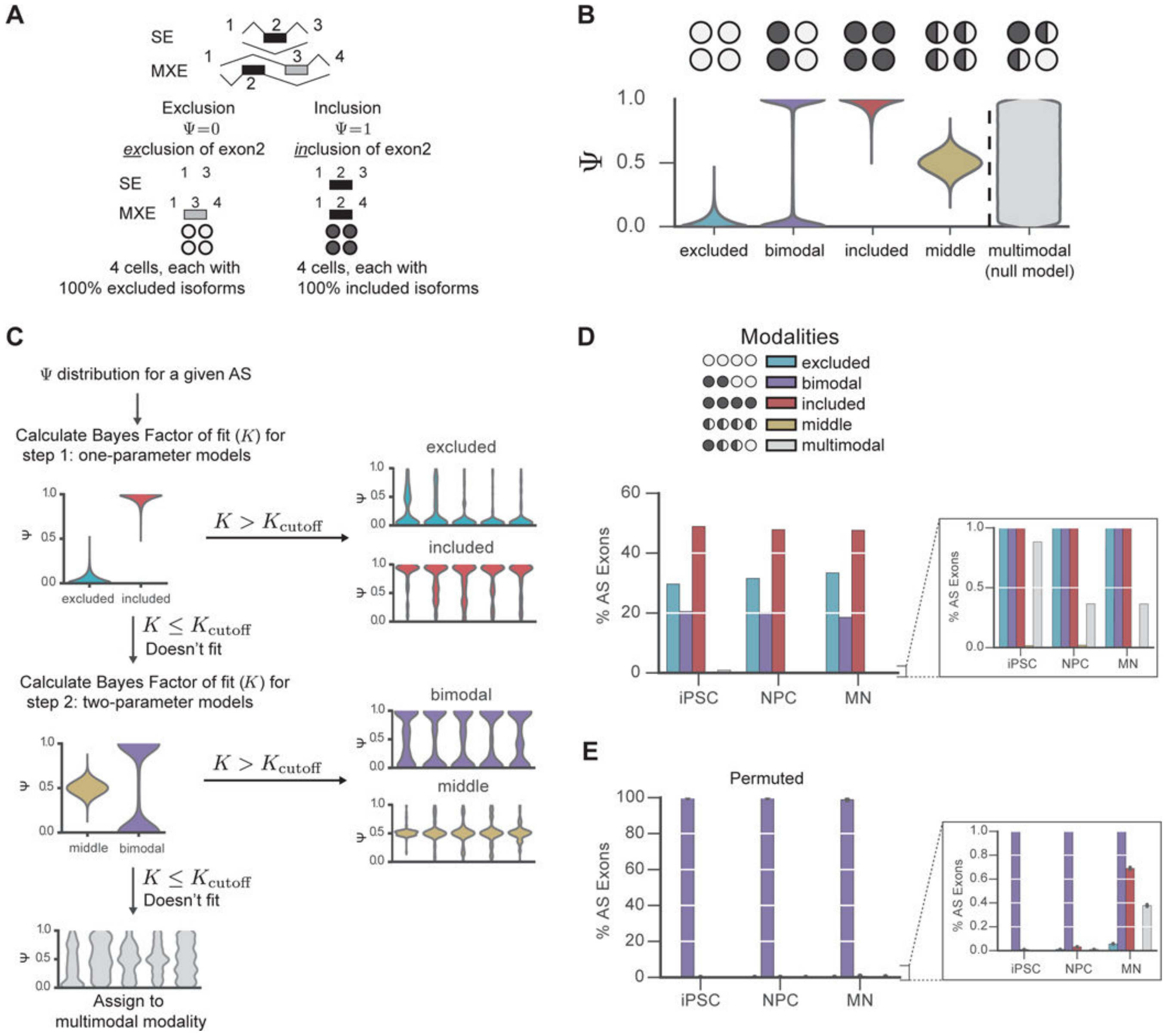


Figure 2. Assignment of single cell alternative splicing distributions to modalities using *anchor* algorithm

(A) Schematic of SE and MXE alternative splicing events. “Exclusion isoform” 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 “Inclusion isoform” 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 inclusion isoform and dark grey represents exclusion isoform.

(B) A schematic of the proposed five modalities tested by *anchor*. Distribution of Ψ for each AS event can be modeled as a Beta probability distribution parameterized by a and p. 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 is 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 α or β is parameterized), including included and excluded modalities. If $K > K_{\text{cutoff}}$, modality is assigned to the modality with highest K. When K_{cutoff} 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 α and β are parameterized), indicating bimodal and middle modalities. If an event cannot fit at either step, it will be assigned to multimodal modality. $K_{\text{cutoff}} = 2^5 = 32$ 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. Ψ 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 represents 95% confidence interval from 1,000 bootstrapped intervals. *right*, zoomed-in panel shows low percentage of unimodal events in permuted data. See also Figure S2.

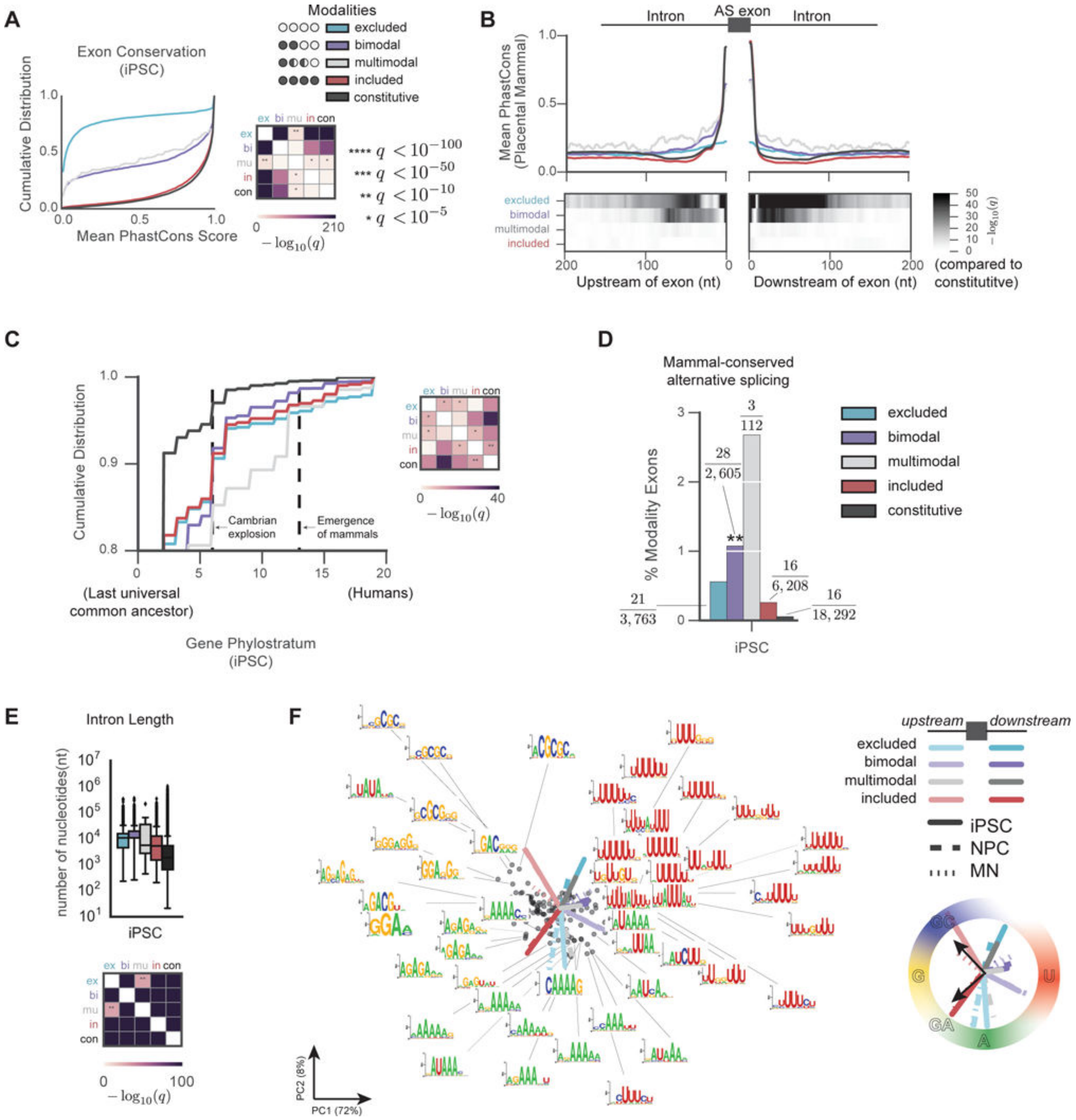


Figure 3. Bimodal AS events exhibit distinct sequence and evolutionary features

All results are shown for iPSCs with highest number of AS events (12,690). All *q*-values of significance were derived from multiple hypothesis corrected (Bonferroni) non-parametric Mann-Whitney U test, unless otherwise indicated.

(A) *Left*, Cumulative distributions of the mean Placental Mammal PhastCons score in each modality, together 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.

(B) Mean Placental Mammal PhastCons scores of flanking introns of AS exons in excluded (blue), bimodal (purple), multimodal (grey), included (red) modalities, and constitutive (black) exons in all cell-types. *Bottom*, nucleotide-level significance of PhastCons scores is presented $0 < -\log_{10}(q) \leq 50$ for clarity.

(C) Phylostratum scores are summarized for genes harboring AS events in each modality together with genes containing constitutive exons. *Right*, pairwise significance scores.

(D) Mammal-conserved AS exons and their percentage in each modality. Hypergeometric test (multiple hypothesis corrected with Bonferroni) indicated $q < 10^{-5}$ statistical significance. Fraction indicates number of conserved AS exons divided by the total AS exons in modality.

(E) Intron lengths in excluded, bimodal, multimodal and included modalities, with constitutive exons as comparison. *Bottom*, pairwise significance scores.

(F) Conserved intronic sequences in each modality are enriched with distinct nucleotides. Motifs enriched for each modality are presented by PCA, with each circle as a motif and the vectors as component loadings of intronic groups. *Left*, motifs are annotated with motif sequences. *Right*, A simplified illustration of nucleotide enrichment in each modality. See also Figure S3.

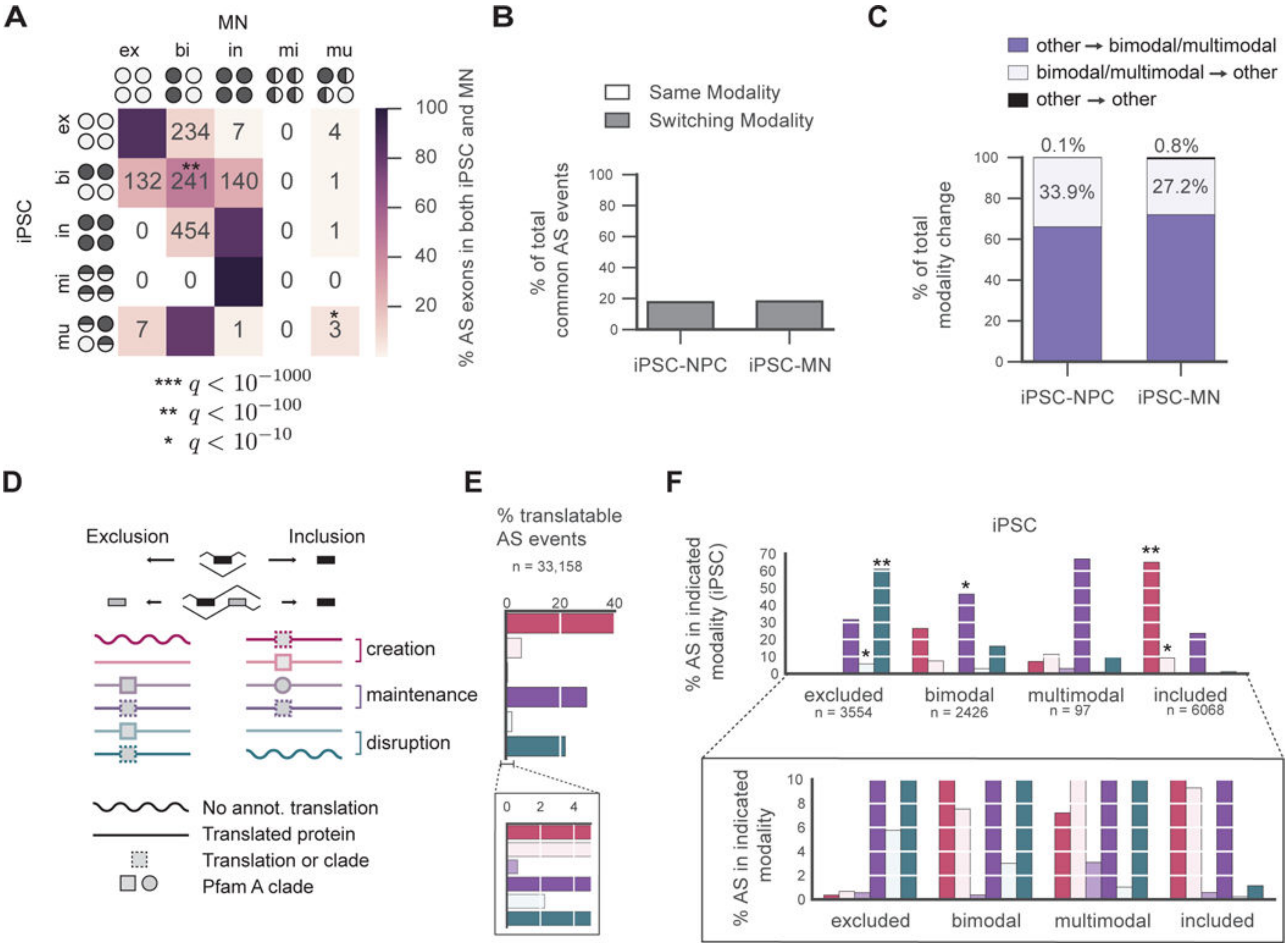


Figure 4. Dynamic AS events are highly variant bimodal and multimodal events

(A) AS events change modalities during the transition from iPSCs to MNs, presented as events in iPSCs (y-axis) against their corresponding modalities in MNs (x-axis). Heat map represents the % of overlapping events in the iPSCs and MNs, annotated with the exact number of events. Notably, 88% of excluded events in iPSCs remained in the 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 (Bonferroni) hypergeometric tests were used to determine significance.

(B) During the differentiation from iPSCs to MNs or from iPSCs to NPCs, we found that 1,586 (17.6%) or 1,029 (18.0%) AS events switched modality, respectively.

(C) Within the switching events, 99% of AS events either switched from a bimodal/ multimodal state or switched towards a bimodal/multimodal state. Less than 1% of switching events were among other types of modality changes.

D–F AS events in bimodal modality exhibit flexibility in protein coding.

(D) Schematic of predicted protein coding changes associated with AS exon inclusion. 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 exons are included. The square and circle illustrate different Pfam domain clades. The square with dashed outlines represents translated protein, which may contain a Pfam domain.

(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 are enriched for maintaining reading frame and presence of domain. The dominant isoforms in included and excluded modalities favor protein or domain creation and switching to the other isoform results in disruption of reading frame. Enrichment is calculated against population average as shown in (E) in each category using multiple hypothesis corrected hypergeometric tests. *: $q< 10^{-10}$ **: $q< 10^{-10}$ See also Figure S4.

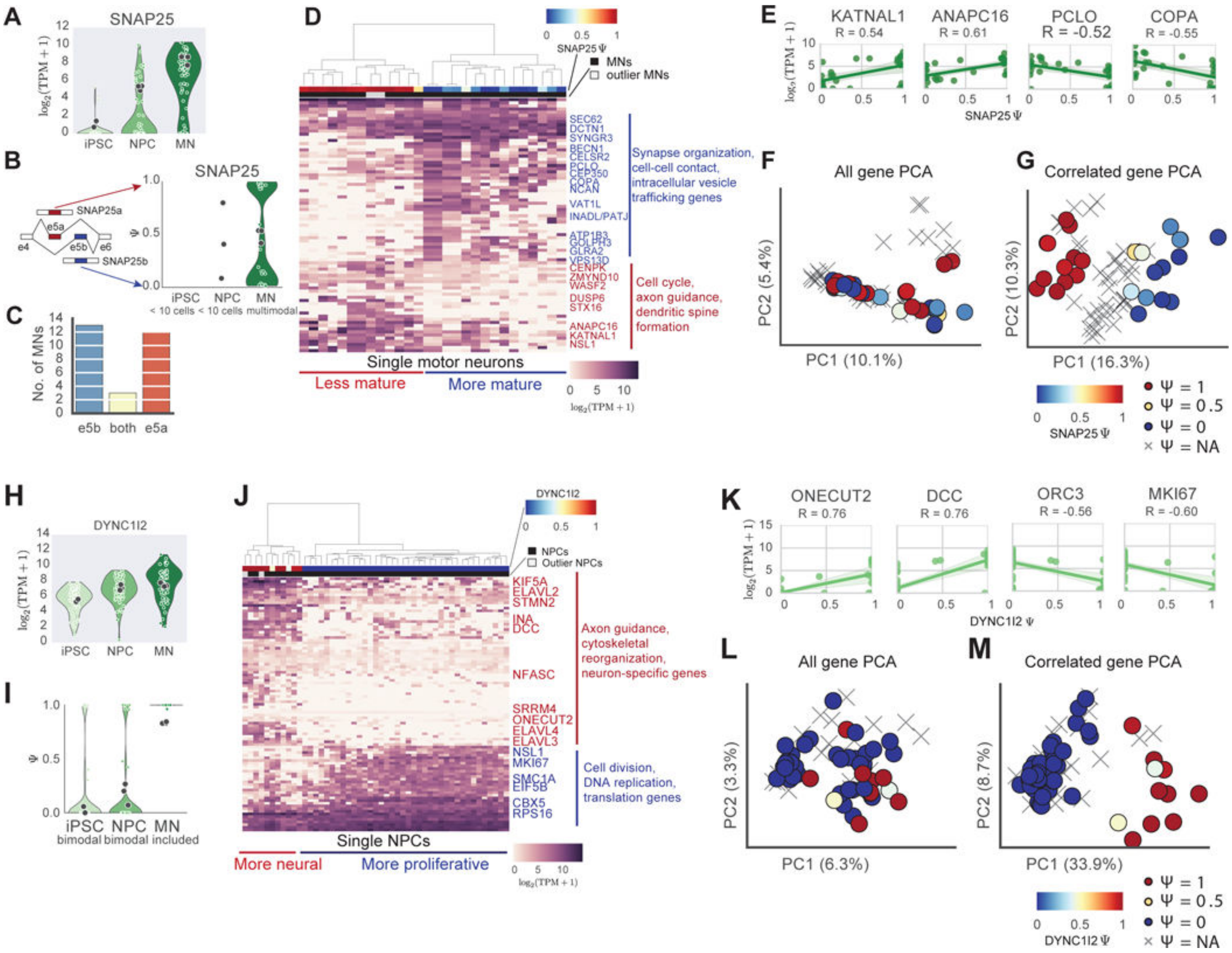


Figure 5. Bimodal and multimodal AS events reveal subpopulations invisible by conventional gene expression analysis

A–G SNAP25 AS reveals a more mature subpopulation in motor neuron population.

(A) SNAP25 is primarily expressed in MNs.

(B) Inclusion of exon 5a in SNAP25 in the three populations.

(C) Number of cells that contain primarily exon 5a or 5b (or both) in motor neurons.

(D) Preferential usage of exon 5a or exon 5b of SNAP25 in MNs reveals intricate cell states. Genes correlated with the Ψ score of this MXE clustered MNs into two main subgroups, Ψ~1 (*red* in the legend bar) and Ψ~0 (*blue* in the legend bar). Rows represent the genes and columns represent single cells. Cells with Ψ around 0.5 are illustrated by yellow in the legend bar. *Black* and *light grey* indicate qualified and outlier MNs based on *k*-means clustering, respectively. Gradient of purple indicates gene expression in $\log_2$(TPM+1), with darker being highly expressed. A few representative genes from the two subgroups are highlighted in red or blue.

(E) Examples of representative genes that correlate with Ψ of exon 5a of SNAP25. KATNAL1 and ANAPC16 are more enriched in the cells with Ψ ~1. DCTN1 and PCLO are more enriched in the cells with Ψ ~0. X-axis represents the Ψ score, and y-axis represent gene expression in $\log_2$(TPM+1). Each MN is depicted as a green circle. Solid green line represents linear regression between Ψ and the expression of indicated genes. Shaded green represents 95% confidence interval of the regression.

F–G Genes that correlate with exon 5a of SNAP25 distinguish MNs into two subgroups. Each MN is depicted as a dot in PCA. *Red:* cells with Ψ ~1; *blue*: Ψ ~0; *yellow*: Ψ ~0.5; X: cells with a Ψ assigned as NA.

(F) PCA of all expressed genes in MNs failed to separate the two subgroups.

(G) Using only the genes correlated with Ψ of exon 5a in SNAP25, two subgroups are readily separated. Percent of variance explained are indicated at each PC.

H–M A bimodal SE event in DYNC1I2 separates NPCs into a more proliferative subgroup and a subgroup on the trajectory of neuronal differentiation.

(H) Gene expression of DYNC112.

(I) Ψ distribution of a SE event in DYNC1I2. This event is bimodal in both iPSCs, NPCs and becomes included in MNs.

(J) Genes that correlate with Ψ of the SE event in DYNC1I2 cluster the NPCs into two subgroups. Green: NPC. Blue: cells with Ψ around 0. Red: cells with Ψ around 1. Light blue to yellow: cells with Ψ 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.

(K) Example genes enriched in the two subgroups of NPCs. Ψ scores of the SE in DYNC1I2 is on x-axis and expression of indicated genes is on y-axis.

L–M Only genes that correlate with Ψ separate two subgroups in NPCs, with each NPC depicted as a dot in the PCA. *Blue:* cells with Ψ ~0; *Red:* cells with Ψ ~1; *yellow:* Ψ ~0.5; X: cells with a Ψ assigned as NA.

(L) PCA of all genes expressed in NPCs failed to separate the two subgroups.

(M) Genes that correlate with Ψ separate the two subgroups by PCA. See also Figure S5.

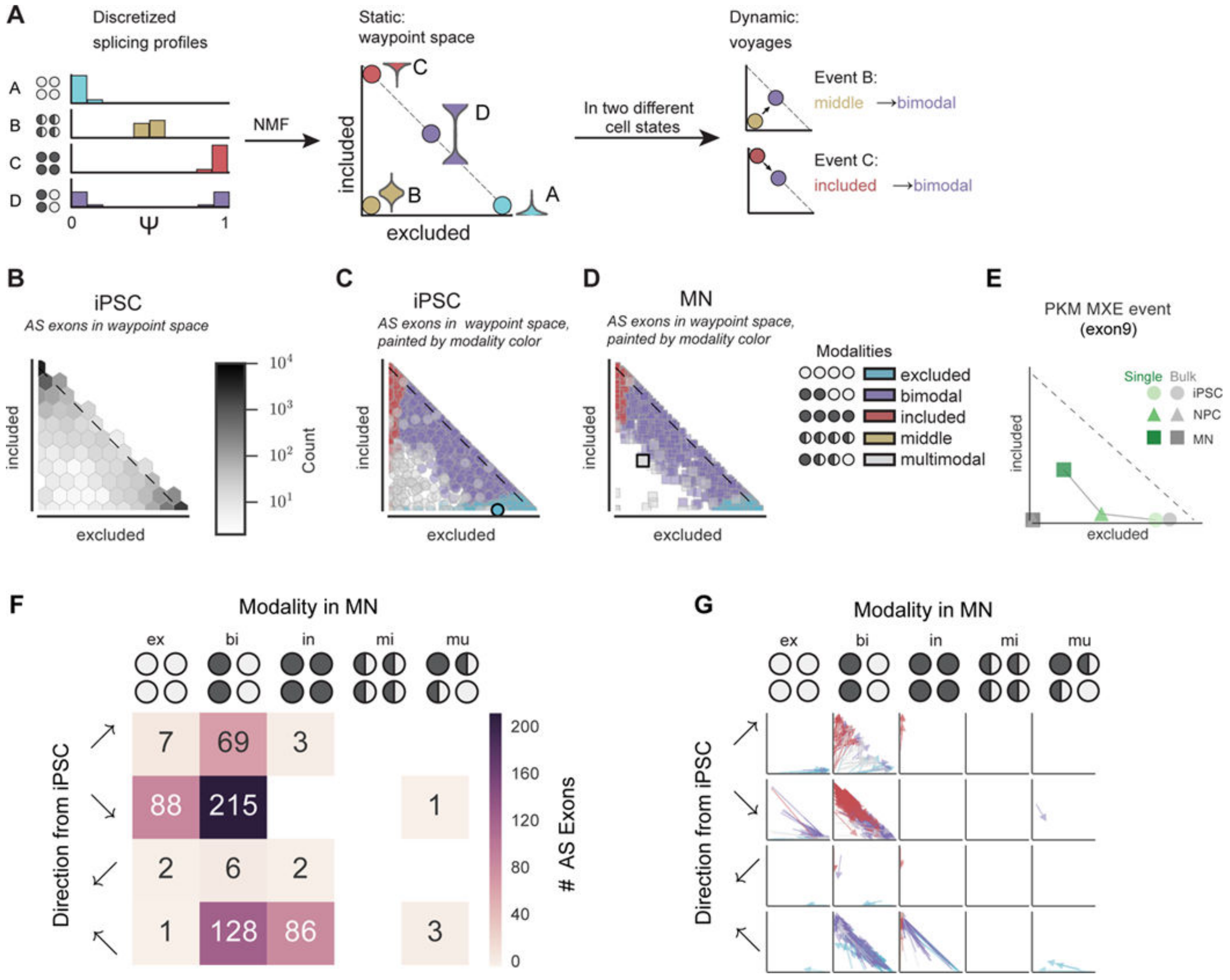


Figure 6. *Bonvoyage* visualizes dynamic AS changes

(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 NMF and projected onto a 2-dimensional space (*middle*), such that each data point summarizes a distribution of AS. The origin represents a distribution that all cells contain 50% inclusion and 50% exclusion isoforms. When the distributions of the same event (either event B or C) are visualized in two different cell-types or states, the change in 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 in B), colored by the modality estimated by *anchor.* Each dot represents the 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 PKM, which is included in both iPSCs and NPCs and becomes bimodal in MNs. Greys represent Ψ measurements in bulk samples.

F–G Global splicing dynamics between iPSCs and MNs are shown and categorized by voyage direction instead of modalities. Only the events with voyage distance ≥ 0.2 are shown for clarity (Figure S6G).

(F) Number of AS events in iPSCs that transitioned to (as indicated by the directionality of the arrows) excluded, bimodal, included, middle, or multimodal modality in MNs.

(G) Same data as in (F), visualized by vectors representing the iPSC (tail) and MN (tip) position of the alternative exon. Colors of arrows reflect the event modalities in iPSCs. See also Figure S6.

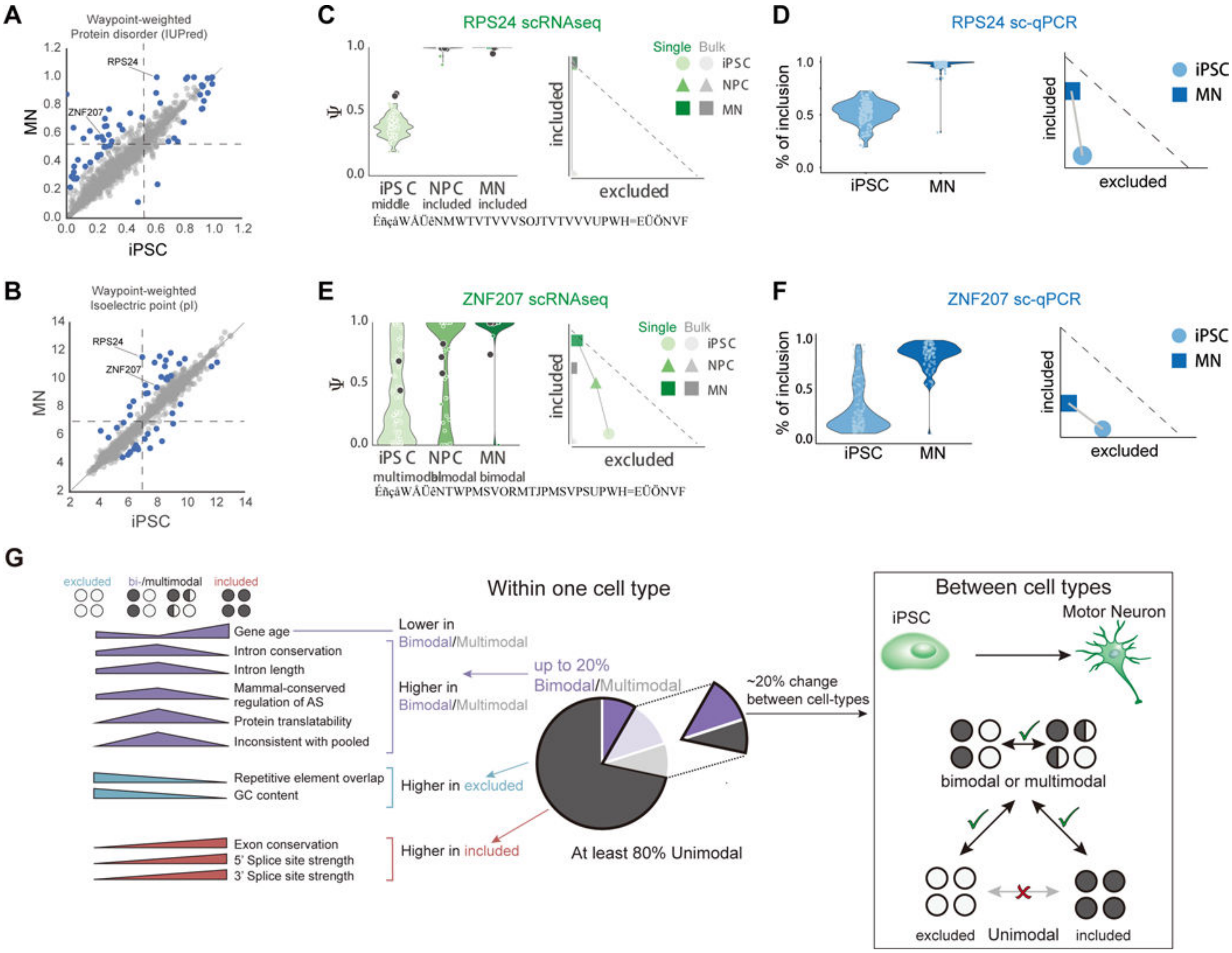


Figure 7. Single-cell qRT-PCR validation and summary of biological findings

A–B. Waypoint-weighted protein properties that change between iPSCs and MNs. Significant changes (in blue) are identified by a factor of three on Mahalanobis distance relative to all iPSC-MN comparisons. X- and y-axis labels refer to, weighted protein property in iPSC and in MN, respectively.

(A) Protein disorder, where a score above 0.5 by IUPred (black dashed line) indicates disorder.

(B) Isoelectric point (pl), where the black dashed line indicates pI=7.

C–F. Distribution of AS inclusion is verified by single cell qRT-PCR (sc-qPCR). See also Figure S7.

(C) Percent spliced-in (Ψ) distributions for RPS24 exon 5 measured by scRNA-seq.

(D) Percent exon inclusion distributions for RPS24 exon 5 measured by sc-qPCR.

(E) Percent spliced-in (Ψ) distributions for ZNF207 exon 9 measured by scRNA-seq.

(F) Percent exon inclusion distributions for ZNF207 exon 9 measured by sc-qPCR.

(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 transcriptome.