Published online 12 January 2023

Nucleic Acids Research, 2023, Vol. 51, No. 5 ***e29***
https://doi.org/10.1093/nar/gkac1260

MARVEL: an integrated alternative splicing analysis platform for single-cell RNA sequencing data

Wei Xiong Wen[1,2], Adam J. Mead[1,3,*] and Supat Thongjuea[1,2,3,*]

[1]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK, [2]MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK and [3]NIHR Oxford Biomedical Research Centre, University of Oxford, Oxford OX4 2PG, UK

Received June 28, 2022; Revised December 13, 2022; Editorial Decision December 14, 2022; Accepted January 11, 2023

ABSTRACT

Alternative splicing is an important source of heterogeneity underlying gene expression between individual cells but remains an understudied area due to the paucity of computational tools to analyze splicing dynamics at single-cell resolution. Here, we present MARVEL, a comprehensive R package for single-cell splicing analysis applicable to RNA sequencing generated from the plate- and droplet-based methods. We performed extensive benchmarking of MARVEL against available tools and demonstrated its utility by analyzing multiple publicly available datasets in diverse cell types, including in disease. MARVEL enables systematic and integrated splicing and gene expression analysis of single cells to characterize the splicing landscape and reveal biological insights.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a powerful tool for studying transcriptional heterogeneity in normal tissues (1–5) and pathological conditions (6–11). The vast majority of scRNA-seq analyses focus on gene-level expression, however, alternative splicing represents an important additional layer of transcriptional complexity underlying gene expression (12). Alternative splicing has not been widely investigated at single-cell resolution and thus remains an untapped source of knowledge in both health and disease states. This is potentially due to the lack of available computational tools to address the challenges of alternative splicing analysis at single-cell resolution, such as high dropout rates, large cell numbers, and PCR amplification biases that may distort isoform expression (13,14). Although existing analysis pipelines such as Seurat (15), Monocle (16) and Scanpy (17) enabled integrative analysis workflows for single-cell gene expression, they do not support comprehensive analyses to combine gene-level and alternative splicing information.

Recently, analysis tools, such as BRIE (versions 1 and 2) (18,19), Expedition (20), SCATS, (21), DESJ-detection (22) and VALERIE (23) were developed to analyze alternative splicing in scRNA-seq datasets generated from the plate-based platforms, e.g. Smart-seq2 (24) or microfluidic-based platforms, e.g. Fluidigm C1 instrument. BRIE uses a Bayesian approach to learn informative sequence features for percent spliced-in (PSI) estimation, leading to the improvement of PSI estimation for splicing events that have low-to-no coverage in scRNA-seq data (18,19). Expedition introduces the concept of 'modalities' to stratify PSI distributions into discrete categories (20). SCATS aggregates spliced reads from a group of exons generated from the same isoform(s), allowing the detection of splicing events with low sequencing depth, and it also supports analysis of scRNA-seq data with or without unique molecular identifiers (UMIs) (21). DESJ-detection performs splicing analysis at the splice junction level to detect differential splicing between groups of cells (22). Lastly, VALERIE enables visual-based validation of candidate splicing events across groups of a large number of single cells to identify true positive events for downstream studies (23).

However, a number of functionalities required to comprehensively characterize alternative splicing dynamics at the single-cell level are not yet available. For instance, current analysis tools focus on PSI quantification for skipped-exons (SE) and mutually exclusive exons (MXE) splicing events (18,20,21) but did not include retained-introns (RI), alternative 5′ and 3′ splice sites (A5SS and A3SS), and alternative first and last exons (AFE and ALE). While SE are the major splicing event type (25), other types of splicing events are also important sources of gene expression heterogeneity and have been shown to contribute to the cellular phenotype. For example, RI are a source of neoantigens in melanoma (26), whereas A5SS, A3SS, AFE and ALE

*To whom correspondence should be addressed. Tel: +49 015201091154; Email: supat.thongjuea@kitz-heidelberg.de
Correspondence may also be addressed to Adam J. Mead. Email: adam.mead@imm.ox.ac.uk
Present address: Supat Thongjuea, Hopp Children's Cancer Center Heidelberg (KiTZ) & Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Heidelberg, Germany.

are often dysregulated in myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML) patients carrying mutations in genes encoding for splicing factors (25,27,28).

Modality classification enables the changes in splicing patterns across different cell populations (20). Biases from PCR amplification and library preparation prevalent in scRNA-seq have been shown to lead to a high proportion of false positives, in particular for the bimodal classification (14). Therefore, modality assignment should incorporate these technical biases to enable better classification of splicing patterns.

Taken together, current computational tools may not comprehensively facilitate the characterization of alternative splicing dynamics at single-cell resolution. Moreover, existing analysis workflows do not integrate gene expression and alternative splicing information into a single framework. Here, we introduce MARVEL, an R package for integrative single-cell alternative splicing and gene expression analysis. We benchmarked MARVEL against existing computational tools for single-cell alternative splicing analysis and demonstrated its utility by analyzing publicly available datasets generated from the plate- and droplet-based library preparation methods derived from induced pluripotent stem cells (iPSCs) differentiated into endoderm and cardiomyocytes, respectively (29,30).

MATERIALS AND METHODS

Plate-based scRNA-seq datasets

Processing of publicly available datasets. To assess and validate the performance of MARVEL on scRNA-seq data generated from plate-based library preparation protocols, we retrieved five datasets from previous studies (16,20,29,31,32). Raw sequencing reads (FASTQ) were downloaded from the Sequence Reads Archive (SRA). Adapters and 3′ bases with Phred quality scores <20 were trimmed using Trim Galore 0.6.5 (33). Trimmed reads were mapped to the GRCh38 reference genome using STAR 2.6.1d in 2-pass mode (34). STAR was also used to detect and quantify splice junction counts, while RSEM v1.2.31 was used to quantify gene expression in transcripts per million (TPM). Binary Alignment Map (BAM) file statistics including total mapped reads and mitochondrial reads were computed using Samtools 1.9 (35).

The first dataset consisted of human-induced pluripotent stem cells (iPSCs), neural progenitor cells (NPCs), and motor neurons (MNs) (20). Cells with >100 000 mapped reads, >70% alignment rate and <15% mitochondrial reads were retained for data with paired-end reads. For the dataset with single-end reads, cells with >5 000 000 mapped reads, >90% alignment rate and <10% mitochondrial reads were retained (Supplementary Figure S1A–F). Single cells that were annotated as outliers by the original study were excluded. In total, 62 iPSCs, 68 NPCs and 60 MN cells were included for analysis. In addition, 2 iPSC, 3 NPC and 3 MN matched-bulk samples were included for analysis.

The second dataset consisted of human myoblasts cultured and sequenced at 0-, 24-, 48- and 72-h (16). Cells with >100 000 mapped reads, >75% alignment rate and <20% mitochondrial reads were retained (Supplementary Figure S1G–I). Single cells that were annotated as control wells by the original study were excluded. In total, 82, 85, 88 and 72 myoblasts at 0-, 24-, 48- and 72-h time points, respectively, were included for analysis. In addition, three matched-bulk samples for each time point were included for analysis.

The third dataset consisted of iPSC and endoderm cells (29). Cells with >100 000 mapped reads, >75% alignment rate and <20% mitochondrial reads were retained (Supplementary Figure S1J–L). Five cells that were annotated as the unknown cell type by the original study were excluded. In total, 83 iPSC and 53 endoderm cells were included for analysis.

The fourth dataset consisted of single cells derived from the spinal cord of mice induced with experimental autoimmune encephalomyelitis (EAE) and control mice (31). Cells that passed sequencing QC were defined as having read alignment >50%, >40 000 mapped reads, and mitochondrial reads <55% (Supplementary Figure S1M–O). Eight cells annotated as doublets by the original publication were removed. In total, 1078 EAE and 978 control mice cells were included for analysis.

The fifth dataset consisted of single cells derived from mouse endothelial-to-hematopoietic stem cell (HSC) transition (32,36). In total, 18 aortic endothelial cells (AECs), 24 hemogenic endothelial cells (HECs), 28 *CD201*^{high} T1 pre-HSCs, 44 *CD201*^{high} T2 pre-HSCs, 21 E12 HSCs, 32 E14 HSCs and 47 adult HSCs were included for analysis.

The first four datasets were used for benchmarking MARVEL. The third and fifth datasets were used to demonstrate the analyses provided by MARVEL.

Isoform detection. We analyzed isoform usage at the exon level for scRNA-seq data. For each cell type, the bulk samples were used to create a cell-type-specific gene transfer file (GTF) using StringTie2 (37). When the bulk samples were not available, pseudo-bulk samples were generated by merging the single-cell BAM files. The GENCODE GTF v31 file was used as a guide to generate the cell-type-specific GTF files (38). The GTF represents the transcriptome assembly and hence the catalog for all genes, transcripts, and exons detected for a particular cell type. The cell-type-specific GTF files were then merged to obtain the final GTF file for isoform detection in BRIE (19) and rMATS (39). Expedition performed *de novo* detection of splice junctions, exons, and alternative splicing events directly on the single-cell and bulk BAM files (20).

Next, rMATS was used to identify SE, MXE, RI, A5SS and A3SS splicing events using the aforementioned merged GTF file as previously described (14,39). Splice junction counts were generated using STAR 2.6.1d in a two-pass mode (34). The gene expression matrix, splicing junction count matrix, coordinates of rMATS-detected exon-level alternative splicing events, and GENCODE GTF v31 file, were used as inputs for MARVEL. MARVEL created an R object from these inputs using the *CreateMarvelObject* function for downstream data processing and analyses (Figure 1A). After creating the MARVEL object, additional splicing event types, AFE and ALE, were detected by MARVEL from the GENCODE GTF v31 provided by using the *DetectEvents* function.

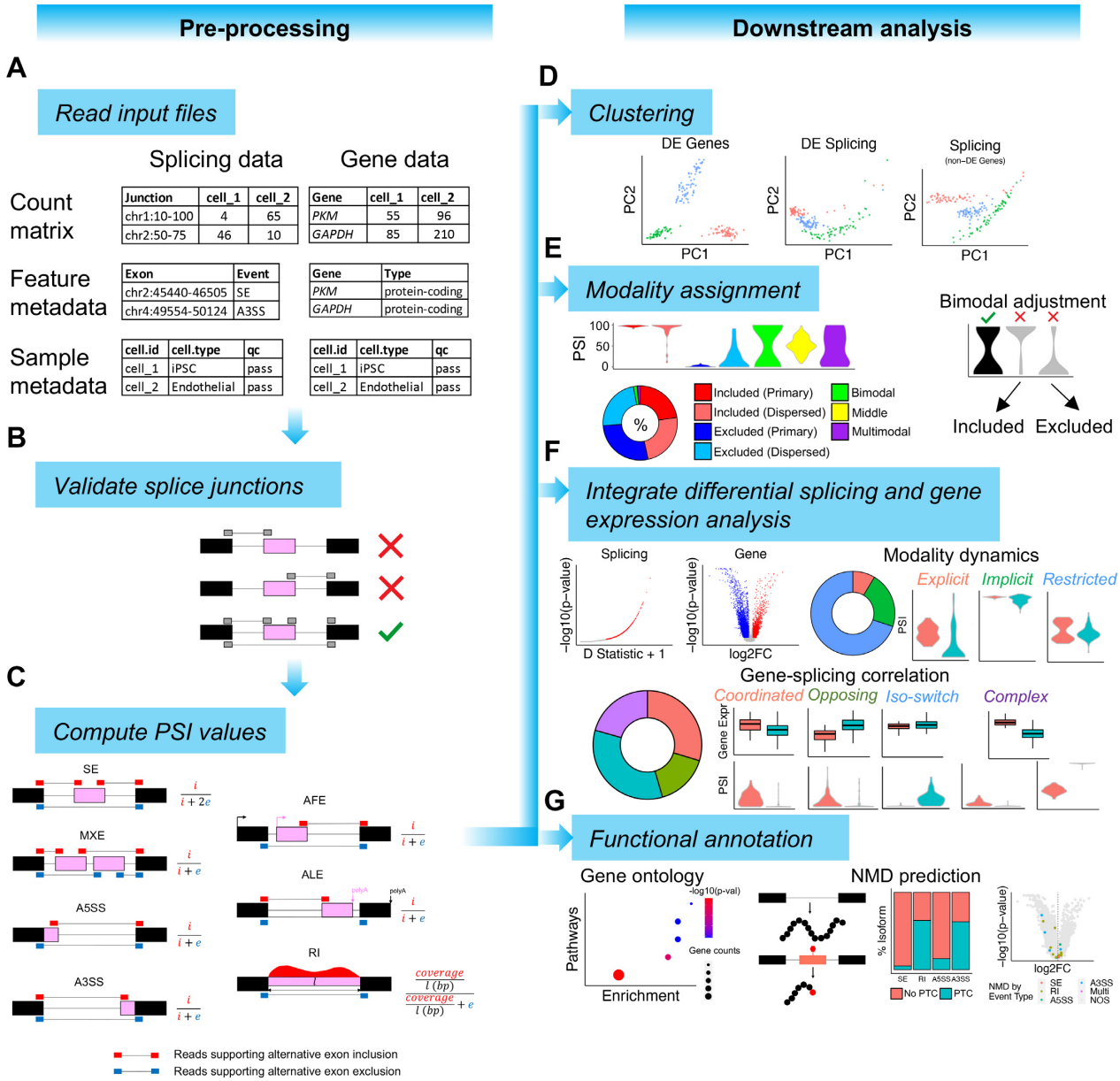

Figure 1. MARVEL workflow for single-cell alternative splicing analysis in RNA-sequencing dataset generated from plate-based methods. (A–C) Workflow for pre-processing of splicing and gene expression data by MARVEL. (**A**) Input files required by MARVEL include splice junction count and normalized gene expression matrix, alternative splicing events, and gene and sample metadata. (**B**) Only alternative splicing events supported by at least 10 splice junction reads are retained. (**C**) The PSI values of the confident alternative splicing events identified in (B) are computed for main exon-level alternative splicing event types. PSI values are calculated as the total number of reads supporting the alternative exons (pink) divided by the total number of reads supporting both alternative exons and constitutive exons (black). (D–G) Downstream analyses using computed PSI and gene expression values. (**D**) Dimension reduction analysis using differentially expressed genes (left), PSI values (middle), and PSI values of non-differentially expressed genes (right). (**E**) The assignment of PSI distributions into seven modalities (as indicated by colors) and the bimodal classification adjustment to reduce false bimodal classification. (**F**) Differential splicing and gene expression analysis and characterization of the alternative splicing in different modality changes or relative to gene expression changes across different cell populations. (**G**) Pathway enrichment analysis of differentially spliced genes and NMD prediction of alternative splicing events to understand the functional consequences of differential alternative splicing events. Genes subjected to NMD are visualized on the volcano plot generated from differential gene expression analysis. A3SS: alternative 3′ splice site; A5SS: alternative 5′ splice site; AFE: alternative first exon; ALE: alternative last exon; DE: differentially expressed; FC: fold change; iPSC: induced pluripotent stem cell; MN: motor neuron; MXE: mutually exclusive exons; NMD: nonsense-mediated decay; NPC: neural progenitor cell; PC: principal component; PSI: percent spliced-in; PTC: premature terminal codon; RI: retained-intron; SE: skipped-exon

Isoform validation and quantification. The percent spliced-in (PSI) values were used to measure the degree of alternative exon inclusion for scRNA-seq data generated from the plate-based protocol. The *briekit-event* function in BRIE was used to detect alternative splicing events from the GTF provided (see 'Isoform detection' section). Next, the *briekit-event-filter* function was used together with the options [*–add_chrom chrX –as_exon_min 10 –as_exon_max 100000000 –as_exon_tss 10 –as_exon_tts 10 –no_splice_site*] to filter for high-quality alternative splicing events. The *briekit-factor* function was then used to calculate the set of sequence features to infer PSI values for each alternative splicing event. Only skipped-exon (SE) splicing event was analyzed by BRIE here. The PSI values of these detected splicing events were subsequently quantified in three different modes using the *brie-quant* function with the options *–interceptMode None*, *–interceptMode cell* and *–interceptMode gene*. The first mode (mode 0) uses a prior distribution centered at 0.5 to impute PSI values for alternative splicing events. The second mode (mode 1) combines a prior distribution centered at 0.5 with an informative prior inferred from genomic sequence-based features to impute PSI values. The third mode (mode 2) uses a prior distribution centered on the mean PSI values across the cell population to impute PSI values.

For Expedition analysis, we performed *de novo* detection of alternative splicing events using the *outrigger index* function and computed the PSI values for each alternative splicing event using the *outrigger psi* function. Each PSI value represents the fraction of splice junction reads supporting the alternative exons over the total splice junction reads supporting or skipping the alternative exons. For each cell, alternative splicing events supported by <10 splice junction reads were annotated as missing values. The types of alternative splicing events analyzed by expedition included SE and MXE.

The types of exon-level alternative splicing events analyzed by MARVEL included seven main exon-level alternative splicing events comprising SE, MXE, RI, A5SS, A3SS, AFE and ALE. To ensure high-quality alternative splicing events for downstream analyses, both alternative and constitutive exons of these alternative splicing events needed to be supported by the splice junction reads. Only alternative splicing events whose exons were supported by splice junction reads were retained (Figure 1B and Supplementary Figure S2A). Furthermore, for the RI event, introns that overlap with any alternative or constitutive exons were filtered away (40). This resulted in high-quality introns and was termed 'independent' intron because they do not overlap with any annotated exons.

Similar to Expedition, MARVEL uses a splice junction-based approach to compute PSI values. For SE, MXE, A5SS, A3SS, AFE and ALE alternative splicing events, PSI values are computed as a fraction of splice junction reads supporting the alternative exons over the total splice junction reads supporting or skipping the alternative exons (Figure 1C and Supplementary Figure S2B–G).

$$\psi_{exon,cell} = \frac{Counts_{sj(included),cell}}{Counts_{sj(included),cell} + Counts_{sj(excluded),cell}}$$

Two approaches are introduced to calculate PSI values for RI alternative splicing events (40). The first approach computes values using the number of intron read counts divided by the total transcript read counts. The second approach computes values using the normalized intron coverage divided by the sum of the normalized intron coverage and splice junction reads of skipping intron. The normalized intron coverage is calculated as the average per-base coverage over the intron interval.

The first approach requires full-length transcript quantification as its denominator, which is not suitable for inferring full-length transcript expression from short-read RNA-sequencing data (41). Therefore, MARVEL implements the second approach to calculate the PSI values for RI. In addition, MARVEL filters away introns that overlap with annotated exonic regions because sequencing reads mapping to exonic regions may bias RI quantification (40).

To this end, the PSI value of a given intron is computed as the total intron coverage normalized by the intronic length and then divided by the sum of the length-normalized intron coverage and total splice junction counts skipping the intron (Supplementary Figure S2H). The total intron coverage is computed as the sum of coverage across each intronic base. The unit of intronic length is in base-pair (bp).

$$\psi_{intron,cell} = \frac{\frac{Coverage_{intron,cell}}{Length_{intron(bp)}}}{\frac{Coverage_{intron,cell}}{Length_{intron(bp)}} + Counts_{sj(excluded),cell}}$$

The *ComputePSI* function was used to validate the alternative splicing events and calculate the corresponding PSI values. SE, MXE, A5SS, A3SS, AFE and ALE alternative splicing events supported by <10 of splice junction reads supporting or skipping the alternative exons in a given cell were annotated as missing values. RI alternative splicing events supported by <10 of length-normalized intronic coverage or <10 of splice junction skipping introns in a given cell were annotated as missing values.

Benchmarking processing time for isoform quantification. To compare processing time between BRIE and MARVEL, we measured the time taken to compute the PSI values for the same set of 1000 SE splicing events. To compare processing time between Expedition and MARVEL, we measured the time taken to compute the PSI values for the same set of 500 SE and 500 MXE splicing events. We additionally measured the time taken to compute the PSI values for 1000 splicing events per each of RI, A5SS, A3SS, AFE and ALE. We measured the processing time and random-access memory (RAM) using the Slurm Workload Manager (v20.02.0) on CentOS Linux 7 (Core) on the Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz CPU.

Sequence conservation analysis. The sequence conservation scores for the 5′ and 3′ constitutive exons, and alternative exons were computed using the *phastCons100way.UCSC.hg38* R package (42). The Pearson correlation between alternative exon conservation scores and mean PSI values for each cell line was analyzed (29).

Linear dimension reduction analysis. Principal component analysis (PCA) was used for linear dimension reduction

analysis (Figure 1D). Only genes and alternative splicing events expressed in at least 3 and 25 cells, respectively, were included for analysis. For alternative splicing events whose PSI values were NA, i.e. coverage <10 in a given cell, they were re-coded randomly with values ranging from 0–100 prior to dimension reduction analysis. PCA was performed and visualized by MARVEL using the *RunPCA* function.

Modality assignment. Song *et al.* proposed that the PSI values for a given alternative splicing event can be categorized into five modalities comprising included, excluded, bimodal, middle and multimodal (20). Here, MARVEL models each alternative splicing event as a beta distribution and estimates the alpha and beta parameters using the maximum likelihood approach. Based on the parameters' values, each alternative splicing event was categorized sequentially into their respective modality as follows. First, PSI distributions with $\alpha < 0.5$ or $\beta < 0.5$ will be classified as bimodal (PSI $\approx$ 0, 100) (Supplementary Figure S3A). PSI distributions not meeting the bimodal criteria will be classified as included (PSI $\approx$ 100) when $\alpha > 2$ and $\beta < 1$ or α:β ratio > 2 (Supplementary Figure S3B), or classified as excluded (PSI $\approx$ 0) when $\beta > 2$ and $\alpha < 1$ or β: α ratio > 2 (Supplementary Figure S3C). Next, PSI distributions not meeting included or excluded classification criteria will be classified as middle (PSI $\approx$ 50) when $\alpha > 1$ and $\beta > 1$ and $\alpha = \beta$ (Supplementary Figure S3D). Finally, the remaining PSI distributions will be classified as multimodal (uniform distribution) (20) (Supplementary Figure S3E).

MARVEL further expands the current repertoire of modalities by stratifying the included and excluded modalities into primary and dispersed (Figure 1E). In included and excluded primary modalities, the PSI values cluster tightly around 100 and 0. In included and excluded dispersed modalities, PSI values cluster towards 100 and 0 with the addition of some values that trended towards opposite ends. Therefore, the dispersed modality has a higher variance among PSI values than the primary modality. Here, we applied a heuristic threshold of variance at 0.001 to categorize the included and excluded modalities into primary (<0.001) and dispersed (≥0.001; Supplementary Figure S3B and C).

Modality assignment for alternative splicing events was performed using the *AssignModality* function. In this study, alternative splicing events supported by at least 10 reads in at least 25 cells were included for modality assignment by Expedition and MARVEL.

Bimodality adjustment. A significant proportion of bimodal splicing patterns detected were previously reported as artifacts of single-cell RNA-sequencing (14). To distinguish between true and false (spurious) bimodal splicing patterns, we generated a set of false and true positive bimodal splicing patterns from a previous study with experimental validation of alternative splicing events detected from RNA-sequencing using quantitative polymerase chain reaction (qPCR) and small molecular fluorescent *in situ* hybridization (smFISH) (20). We further expanded our search for true bimodal splicing patterns in two additional studies (16,29), whereby bimodal splicing patterns constituted of single cells in which the corresponding gene had 10 or more mRNA molecules as previously described (14). The mRNA count for each gene was computed using the *monocle* R package (16). In total, 45 true and 7 false bimodal splicing patterns were included for analysis. We assessed the ability of three features to distinguish true from false bimodal patterns: 1) fold difference between the proportion of cells with PSI > 75 and PSI < 25 (and vice versa), 2) difference between the proportion of cells with PSI > 75 and PSI < 25 (and vice versa), and 3) average PSI value. Heuristic thresholds of 75 and 25 were chosen because they distinguished true from false bimodal patterns (see 'Modality classification and correction' of the Results section). The argument *bimodal.adjust = TRUE* can be used in the *AssignModality* function to detect true and false and subsequently reassign false bimodality into either included or excluded modality.

We then assessed the ability of Expedition and MARVEL to distinguish bimodality and non-bimodality. From the datasets (16,20,29), we generated a set of alternative splicing events consisting of 45 bimodal and 17 259 non-bimodal (included, excluded, middle, and multimodal) modalities as previously described in (14). We cross-tabulated the bimodal and non-bimodal assignment of Expedition and MARVEL against this set of ground truths to create a confusion matrix. This allowed us to compute and compare several evaluation metrics consisting of sensitivity, specificity, negative predictive value, and precision for Expedition and MARVEL.

Differential splicing analysis. To detect differences in splicing patterns between groups of single cells, we need to take into account both mean and variance. For example, it would not be possible to distinguish between bimodal, middle, and multimodal splicing patterns based on mean alone. To this end, MARVEL implements three nonparametric statistical tests for assessing the differences in splicing patterns between groups of single cells. These are the Kolmogorov–Smirnov, Anderson–Darling (AD) (43) and D Test Statistic (DTS) approaches (44). These tests take into account the overall PSI distribution and assess the differences in PSI distribution for each alternative splicing event between groups of single cells. MARVEL combines differential splicing analysis using AD and DTS, followed by the outlier removal. MARVEL also includes Wilcoxon rank-sum test, *t*-test and permutation test for differential alternative splicing analysis, used for comparing PSI values in bulk samples. Differential splicing analysis can be performed using the *CompareValues* function and subsequently visualized using the *PlotDEValue* function. In this study, alternative splicing events supported by at least 10 reads in at least 25 cells were included for differential splicing analysis. Alternative splicing events with FDR < 0.10 were considered to be differentially spliced.

Leveraging on the bimodality assignment, the modality dynamics of differentially spliced events between two cell populations may be classified into explicit, implicit, or restricted (Figure 1F). Explicit modality change involves five original modality types (included, excluded, bimodal, middle, and multimodal) (20). Implicit modality change involves the change between primary and dispersed, for example, included primary to included dispersed. Restricted modality change indicates no difference in modality

between the two cell populations, notwithstanding significantly different PSI distributions between the two cell populations.

For BRIE, differential splicing analysis was performed using the *brie-quant* function with default parameters. Alternative splicing events with evidence of lower bound (ELBO) gain >4 were considered to be differentially spliced as previously described (19).

Differential gene expression analysis. Differential gene expression analysis of differentially spliced genes was performed using the *CompareValues.Exp.Spliced* function and subsequently visualized using the *PlotDEValues.Exp.Spliced* function in MARVEL. Wilcoxon rank-sum test was used to assess the differences in normalized and log2-transformed gene expression values between two cell populations. Genes with FDR <0.10 and $\log_2$ fold change of >0.5 or <−0.5 were considered to be differentially expressed.

Gene–splicing relationships. MARVEL incorporates differential splicing and gene expression analyses to investigate, for a given differentially spliced gene, the change in its gene expression relative to the change in its corresponding splicing event(s) between two cell populations (Figure 1F). MARVEL classifies gene-splicing relationships into coordinated, opposing, isoform-switching, and complex using the *IsoSwitch* function. Coordinated relationships indicate that the change in mean gene expression values is in the same direction as the change in the mean PSI values from one cell population to the next. Opposing relationships indicate that the change in mean gene expression values is in the opposite direction to the change in the mean PSI values. Isoform-switching indicates that the PSI distributions are significantly different but mean gene expression values are not significantly different between the two cell populations. Complex relationships refer to genes with a combination of coordinated, opposing, and/or isoform-switching with their corresponding splicing events.

Gene ontology analysis. MARVEL implements the gene ontology analysis provided by the *clusterProfiler* R package (45,46) (Figure 1G). Gene ontology analysis to detect enriched pathways among differentially spliced genes can be performed using the *BioPathways* function.

Nonsense-mediated decay (NMD) prediction. For a given alternative exon with >5 PSI difference between iPSCs and endoderm cells and FDR <0.10, MARVEL will retrieve the gene identifier from which the alternative exon is related. All protein-coding isoforms from this gene that encode the alternative exon are retrieved. MARVEL inserts the alternative exon sequence into these isoforms and predicts the resulting amino acid sequences using the *translate* function implemented by the *Biostrings* R package (Figure 1G). The position(s) of any stop codon and its relative position in base-pair to the final exon-exon junction is noted. Consequently, there are four categories of isoforms: 1) alternative exons belonging to novel isoforms (no matching record in GTF), 2) non-protein-coding isoforms (isoforms with no open reading frame), 3) protein-coding isoforms with a premature terminal codon (PTC) introduced by the alternative exons and 4) protein-coding isoforms whose open reading frame are not disrupted by the alternative exons.

Protein-coding isoforms are further stratified into isoforms that are subjected to nonsense-mediate decay (NMD) or not (25). For the former, PTC(s) are located > 50 bp upstream of the final exon-exon junction. For the latter, the isoforms either have PTC(s) located within 50 bp upstream of the final exon-exon junction or the isoforms do not have any PTC(s) introduced by the alternative exons.

10× genomics dataset

Processing of publicly available datasets. To demonstrate the utility of MARVEL for single-cell alternative splicing on a dataset from a droplet-based platform, we retrieved scRNA-seq data from two previous studies (30,47).

The first dataset consists of iPSC and iPSC-derived cardiomyocytes on days 2, 4 and 10 generated using 10× Genomics Chromium Single Cell 3′ Reagent Kit (version 2) (30). Raw sequencing reads (FASTQ) were downloaded from the Sequence Reads Archive (SRA) and were aligned to the GRCh38 reference genome using Cell Ranger v2.1.1. The resulting BAM files for each sample were used as inputs for STARsolo (available in STAR v2.7.8a) to generate the gene expression count matrices (48). SingCellaR (1,49) was subsequently used to identify and retain good-quality cells based on per-cell UMI counts and the number of detected genes (Supplementary Figure S4A–D) (1). Additionally, only cells with < 15% mitochondrial counts and genes expressed in at least 10 cells were retained. The gene expression values of the good-quality cells were normalized using SingCellaR by scaling UMI counts per library size to 10 000. Good-quality cells from iPSC and day-10 cardiomyocytes were subsequently integrated, and the t-Distributed Stochastic Neighbor Embedding (tSNE) coordinates were generated using SingCellaR. In total, 11 244 iPSCs and 6240, 8635 and 5937 of cardiomyocytes at day-2, -4 and -10, were included for analysis.

The second dataset consists of brain tissues from 15 Autism Spectrum Disorder (ASD) patients and 16 controls (47). This dataset was generated using single-nucleus RNA sequencing based on the 10x Genomics platform. Raw sequencing reads (FASTQ) were downloaded from the Sequence Reads Archive (SRA) and were aligned to the GRCh38 reference genome using Cell Ranger v7.0.0 with the '*include-introns true*' option because nuclei mRNAs contain a higher proportion of unspliced intronic reads compared to cytoplasmic mRNAs (47). STARsolo was subsequently used to generate the gene and splice junction count matrices (50), and the normalized gene expression matrix was generated by SingCellaR. We analyzed 104 559 cells and used tSNE coordinates from Supplementary Table S2 of the original study for analysis.

Isoform detection. We analyzed isoform usage at the splice junction level for scRNA-seq data generated from a droplet-based platform. Splice junction counts were generated using STAR v2.7.8a (STARsolo). The filtered gene count matrix and tSNE coordinates from SingleCellaR, raw splice junction count from STARsolo, and reference gene transfer file (GTF) were used as inputs for MARVEL. MARVEL cre-

ated an R object from inputs using the *CreateMarvelObject.10x* function for downstream data processing and analyses (Supplementary Figure S5A).

Isoform validation and quantification. To ensure the inclusion of high-quality splicing junctions for downstream analyses, the exons of each splice junction were cross-checked with the GTF file and were categorized as annotated, multi-mapped, and unannotated. An annotated exon is an exon that maps to a single gene. A multi-mapped exon is an annotated exon that maps to multiple genes. An unannotated exon is an exon with no matching record in the GTF file. Only splice junctions consisting of both annotated exons were retained for downstream analyses (Supplementary Figure S5B). The *AnnotateSJ.10x* function was used to annotate each splice junction, while the *FilterSJ.10x* function was used to filter splicing junctions consisting of unannotated and multi-mapped exons.

Splice junction usage was used to measure the degree of splice junction inclusion. For a given cell type, the splice junction usage was computed as the fraction of the sum of splice junction counts across all cells over the sum of gene counts for the corresponding splice junction across all cells (50) (Supplementary Figure S5C).

$$U_{sj,cell\ type} = \frac{\sum_{Cell \in cell\ type} Counts_{sj,cell}}{\sum_{cell \in cell\ type} Counts_{gene[sj],cell}}$$

The *ComputeSJusage.10x* function was used to compute the cell type-specific usage of validated splice junctions.

Differential splicing analysis. Due to the sparsity of genes and splice junctions detected in scRNA-seq data generated from a droplet-based method, we performed differentially splicing analysis at the cell type level (pseudo-bulk) as opposed to at the single-cell level (50) (Supplementary Figure S5D). First, for two cell populations identified for differential splicing analysis, we retained genes expressed in at least 10% in both cell populations (Supplementary Figure S6A). We further retained splice junctions expressed in at least 10% of either cell population (Supplementary Figure S6B). The *PlotPctExprCells.Genes.10x* and *PlotPctExprCells.SJ.10x* functions were used to explore gene and splice junction expression rates. Next, MARVEL utilizes a permutation approach for assessing differentially spliced junctions between two cell populations (Supplementary Figure S6C) (51). For a given splice junction, the cell-type labels of single cells in the two cell populations are shuffled (permutated). PSI values per cell population are computed. Differences in the PSI values between populations are noted ($\Delta PSI_{permutated}$). The differential process is iterated 100 times, and these differences in the PSI values will form the null distribution. Then, the observed differences in the PSI values between the cell populations ($\Delta PSI_{observed}$) are compared against the null distribution to obtain *P*-values. Differentially spliced junctions were defined as mean $\log_2$-transformed normalized gene expression > 1.0, $|\Delta\ PSI_{observed}| > 5$, and *P*-value < 0.05. Differential splicing analysis can be performed using the *CompareValues.PSI.10x* function and subsequently visualized on the PCA/tSNE/UMAP plot using the *PlotDEValues.PCA.10x* function.

To assess the ability of MARVEL to identify biologically relevant genes that are differentially spliced, we performed differential splicing analysis on a dataset generated from iPSC and iPSC-derived cardiomyocytes on day 10 (30).

Differential gene expression analysis. For a droplet-based scRNA-seq dataset, differential gene expression analysis of differentially spliced genes was performed using the *CompareValues.Genes.10x* function and subsequently visualized using the *PlotDEValues.Genes.10x* function in MARVEL. Wilcoxon rank-sum test was used to assess the differences in normalized and log2-transformed gene expression values between two cell populations. Genes with FDR <0.10 and $\log_2$ fold change of >1 or <−1 were considered to be differentially expressed.

Wilcoxon rank-sum test is the default statistical test for differential gene expression analysis in MARVEL for both plate- and droplet-based scRNA-seq data (15). MAST is also available as an option for differential gene expression analysis (52). For MAST, MARVEL computes the number of genes detected per cell (gene detection rate) and includes this variable as a covariate in the zero-inflated regression model. The gene detection rate is recommended as a covariate by MAST (52). To identify differentially expressed genes, a likelihood ratio test (LRT) is performed by comparing the model with and without the user-specified cell groups, e.g. iPSCs and endoderms for plate-based analysis or iPSCs and day-10 cardiomyocytes for droplet-based analysis. The overlap of differentially expressed genes returned by Wilcoxon rank-sum test and MAST was compared using the hypergeometric test provided by the *GeneOverlap* R package (http://shenlab-sinai.github.io/shenlab-sinai/).

Gene-splicing relationship. Similar to plate-based analysis, MARVEL classifies gene-splicing relationships into coordinated, opposing, isoform-switching, and complex (Supplementary Figure S5E) using the *IsoSwitch.10x* function.

Gene ontology analysis. Similar to plate-based analysis, MARVEL implements the *BioPathways.10x function* to identify enriched gene sets that are coordinatedly spliced (Supplementary Figure S5F).

RESULTS

MARVEL is benchmarked against established packages

Percent spliced-in estimation. To estimate PSI values from scRNA-Seq data, Bayesian regression prediction- and sequencing read-based approaches have been used (13,18–20,29). The former incorporates genomic features such as nucleotide context and cell-specific features such as cell-type, together with sequencing reads into the PSI value prediction, whereas the latter uses only sequencing reads, specifically splice junction reads, to compute PSI values.

The Bayesian approach based on genomic features for PSI estimation has been applied to SE splicing events only (18). Here, we assessed the predictive value of a genomic feature in inferring PSI values for other types of splicing events using Pearson correlation (29). We observed relatively low-level correlations between PSI values from MXE, RI, A5SS,

A3SS, AFE and ALE splicing events and the phastCons conservation scores compared to SE (Figure 2A and B). The phastCons score was identified as the most predictive feature for PSI value estimation using BRIE, as previously described (29). This would suggest reliable estimation of PSI values if the Bayesian regression based on genomic features was to be applied to SE, but other methods of PSI estimation, such as those based on sequencing reads alone, may be more suitable for non-SE splicing events. Therefore, MARVEL implements the splice junction read-based for PSI estimation of SE, MXE, RI, A5SS, A3SS, AFE, and ALE.

To assess the precision of estimated PSI values computed using a splice junction read-based approach, we evaluated the reproducibility of PSI quantification across homogenous cell populations in different cell types (53). Compared to all three modes of computing PSI values by BRIE, the median cell-to-cell correlation in PSI values for SE splicing event was higher for Expedition and MARVEL (Figure 2C). Expedition and MARVEL showed similar median cell-to-cell correlation in PSI values for SE and MXE splicing events. This is because Expedition and MARVEL both used splice junction counts to compute PSI values. In addition, MARVEL computes PSI values for RI, A5SS, A3SS, AFE and ALE that are not provided by BRIE and Expedition. Furthermore, we assessed the median cell-to-bulk correlation for SE splicing events and observed a significantly higher correlation for MARVEL and Expedition compared to BRIE modes 0 and 1 (Figure 2D). There was no significant difference in median cell-to-bulk correlation in PSI values for SE and MXE splicing events between MARVEL and Expedition. The overall median cell-to-cell and cell-to-bulk correlation for SE, MXE, RI, A5SS, A3SS, AFE and ALE splicing events computed by MARVEL were generally higher than 0.82 per category, suggesting robust PSI values generated using MARVEL.

We further compared the computational efficiency in processing time and Random-Access Memory (RAM) usage for computing the PSI values by BRIE, Expedition, and MARVEL. MARVEL required less time to compute the PSI values than BRIE and Expedition for the same dataset of SE and MXE splicing events (Figure 2E). Except for RI, the processing time across all datasets for A5SS, A3SS, AFE, and ALE was less than one minute. Lastly, the RAM usage was slightly higher for MARVEL than BRIE but lower than Expedition for computing the PSI values for the same dataset (Supplementary Figure S7A).

Taken together, MARVEL enables computationally efficient PSI quantification of all exon-level splicing event types and demonstrates reproducible PSI values across different cell populations that are comparable with existing single-cell splicing software.

Modality classification and correction. Modality classification concept was previously introduced to stratify the PSI distribution in a cell population into discrete categories, comprising excluded (PSI $\sim$ 0), included (PSI $\sim$ 100), middle (PSI $\sim$ 50), bimodal (PSI $\sim$ 0, 100), and multimodal (uniform distribution) (20). However, the current modality assignment does not identify and correct false classification caused by the technical noise of scRNA-Seq experiments. Specifically for bimodal distributions, recent simulated and empirical data showed that a significant proportion of bimodality was spurious due to PCR amplification bias during the single-cell library preparation (14). Analyzing highly expressed alternative splicing events defined as genes with high mRNA count, and hence genes with low possibility of dropouts, e.g. genes with at least 10 molecules, has been shown to mitigate the false bimodal classification (14). However, this approach would preclude the majority of genes from downstream alternative splicing analysis. For example, we observed that $>$ 90% of genes were excluded when at least 10 molecules were required (Supplementary Figure S7B and C).

To retain alternative splicing events for analysis irrespective of gene abundance and at the same time mitigate false bimodal classification, we sought to identify distinguishable features between true and false bimodal distributions. We tabulated a set of true and false bimodal distributions encompassing alternative splicing events previously validated using qPCR or smFISH (20). Additionally, we tabulated a set of true bimodal distributions consisting of highly expressed alternative splicing events as previously described (14) (Figure 2F). We observed that the fold difference or difference in the proportion of cells at both ends of PSI distribution could delineate true from false bimodal distributions with thresholds of $<$3 and $<$50%, respectively (Figure 2G and H). Therefore, MARVEL incorporated these heuristic thresholds to identify true bimodality. Moreover, true bimodality revealed an average PSI (from both ends) of about 50, whereas the PSI values of false bimodality trended towards 100 or 0 (Supplementary Figure S7D). Therefore, MARVEL reclassifies the false bimodality into included or excluded modalities when the average PSI value is above or below 50.

Next, using our bimodal-adjusted modality approach, we compared Expedition's and MARVEL's ability to distinguish bimodal distributions from other modalities (e.g. included, excluded, middle, and multimodal). We tabulated a set of presumed true bimodal and non-bimodal distributions based on qPCR and smFISH validation and mRNA counts comprising 17 304 splicing events (14) (Supplementary Figure S7E). Expedition and MARVEL showed similar sensitivity, specificity, and negative predictive values in classifying bimodal and non-bimodal distributions (Figure 2I). However, Expedition showed a higher number of non-bimodality, leading to higher false-positive rates than MARVEL ($P < 0.01$; Fisher's exact test; Supplementary Figure S7F). MARVEL, therefore, is more precise in classifying bimodality than expedition.

Finally, we compared the percentage of splicing events classified as bimodal distribution by MARVEL and Expedition for highly expressed alternative splicing events and alternative splicing events that do not meet the criteria for high mRNA counts. Expedition classified a median of 7.8% of all splicing events as bimodal distribution compared to 1.4% by MARVEL, whereas only 0.2% of highly expressed splicing events were classified as bimodal distribution (Supplementary Figure S7G). However, MARVEL identified a bimodal distribution of the lowly expressed *PKM* gene, previously validated using smFISH (20). This gene would be missed if only highly expressed alternative splicing events were included for analysis (Supplementary

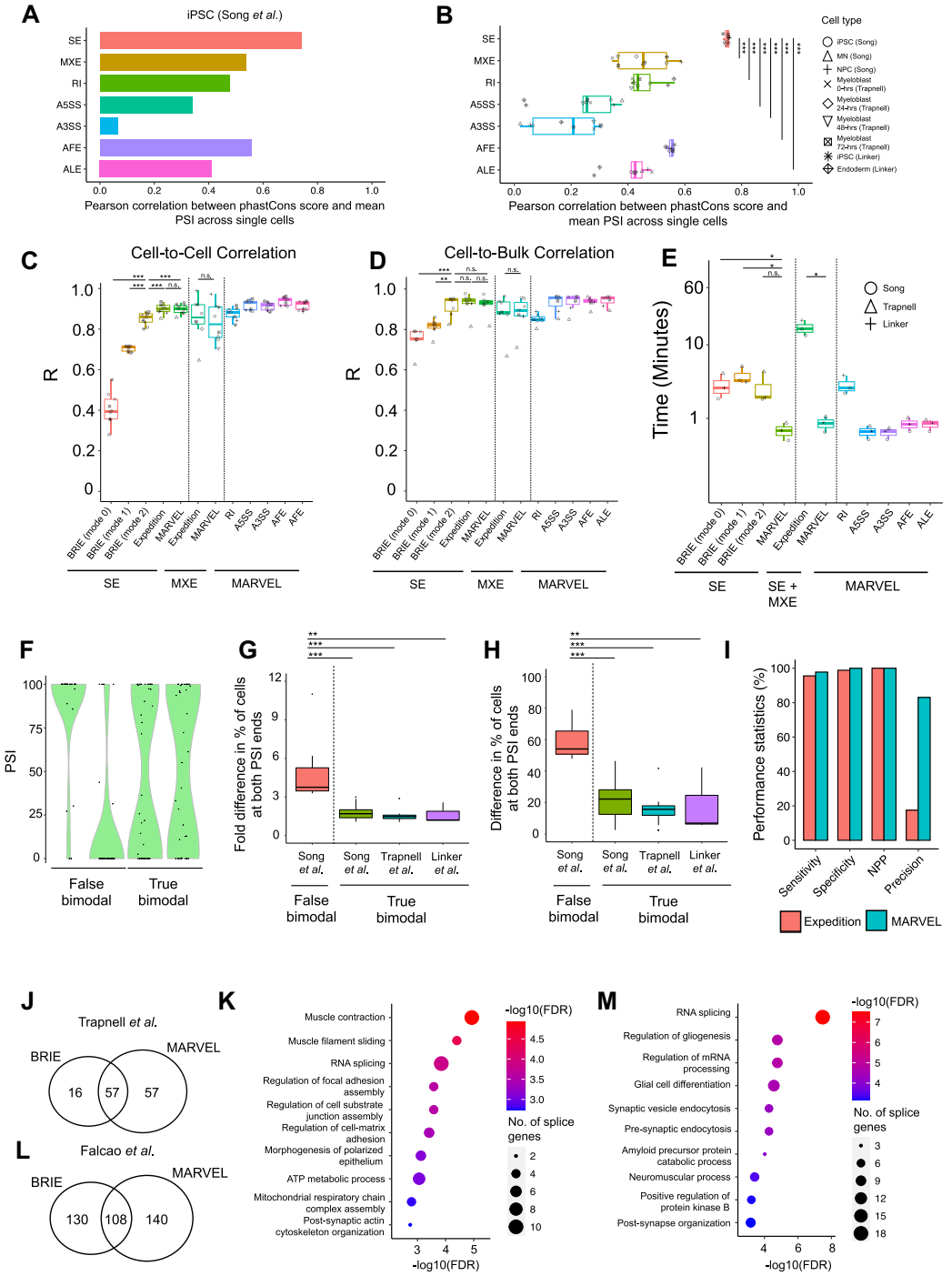

Figure 2. Benchmarking MARVEL against existing computational tools for single-cell alternative splicing analysis. (A, B) Pearson correlation between sequence conservation (phastCons) scores of alternative exons and their corresponding average **PSI** values across all single cells for each splicing event type in (**A**) iPSCs and (**B**) across nine cell lines. (**C**) Pearson correlation of PSI values between all possible single-cell pairs in nine cell lines compared across available splicing analysis tools. (**D**) Pearson correlation of PSI values between single cells and matched bulk sample in seven cell lines with both single cell and matched bulk sample available. (**E**) Comparison of processing time used to compute PSI values for 1000 splicing events for each splicing event type in three datasets. (**F**) Representative false and true bimodal distributions from qPCR, smFISH, and mRNA count-based approach [15, 19, 26]. (G, H) Comparison of the (**G**) fold change (ratio) and (**H**) fold difference in the percentage of cells with PSI > 75 and PSI < 25 (and vice versa) between false and true bimodal distributions. (**I**) Evaluation metrics compared MARVEL versus Expedition for sensitivity, specificity, negative predictive value, and precision. The classification of false and true bimodal distributions predicted by MARVEL and Expedition was compared against the catalog of ground truths comprising 17 304 false and true bimodal distributions. (**J**) A Venn diagram showing the number of differentially spliced SE events detected between 72- versus 0-h myoblast by BRIE and MARVEL. (**K**) MARVEL identified muscle-related pathways among differentially spliced genes between 72- versus 0-h myoblast. (**L**) A Venn diagram showing the number of differentially spliced SE events detected between EAE and control mice using BRIE and MARVEL. (**M**) MARVEL identified neuron-related pathways among differentially spliced genes between EAE and control mice. h: hours; iPSC: induced pluripotent stem cell; MN: Motor neuron; NPC: Neural progenitor cell. *** FDR < 0.01 ** FDR < 0.05 * FDR < 0.1.

Figure S7H and I). It is also noteworthy that almost four times more splicing events were eligible for modality assignment by MARVEL than when only highly expressed alternative splicing events were included for analysis.

Taken together, MARVEL leverages the concept of modality assignment introduced previously (20) while adjusting for technical biases from scRNA-seq library preparation (14) to enable robust classification of splicing patterns.

Differential splicing analysis. Statistical approaches for differential splicing analysis for scRNA-seq, such as BRIE (18), BRIE2 (19), and Expedition (20) were recently developed. However, BRIE requires a pairwise comparison between all possible pairs of cells that might use high computational resources and processing time to detect differential splicing events in a large number of cells (18). BRIE's approach is suitable for detecting differential splicing events within a cell population but not across two different populations. Expedition detects differential splicing events based on modality change, excluding splicing events that show no modality change across cell populations (20). Detection of differential splicing events based on modality change is further limited by assigning modality without considering biases in PCR amplification, leading to inaccurate modality assignment, such as false bimodal classification (14). Lastly, BRIE2 differential splicing analysis is recommended for comparing PSI values across homogenous, but not, heterogenous cell populations (19). This is because, for a given cell type, BRIE2 imputes alternative splicing events for missing PSI values, using the mean PSI values across the cell population. Moreover, imputed PSI values may not represent actual biological phenomenon (20) and underappreciate the cell-to-cell heterogeneity present in the cell population. Therefore, we sought to apply alternative approaches to compare PSI values between two cell populations and distinguish splicing distributions with similar average PSI values but different PSI distributions, such as bimodal, middle, and multimodal distributions.

To this end, we applied nonparametric tests, namely Kolmogorov–Smirnov (KS), Anderson–Darling (AD), D Test Statistics (DTS) (44), and Wilcoxon rank-sum test, to identify the number of differential splicing events during myoblast differentiation cultured and sequenced at 0- and 72-h (16). These tests were implemented into MARVEL and identified 39 (KS), 69 (AD), 175 (DTS) and 65 (Wilcoxon rank-sum test) differentially spliced events. Using a comparable cut-off, DTS detected a much higher number of differentially spliced events than other tests, suggesting higher detection sensitivity. However, we observed that differentially spliced events detected by DTS were driven by outlier cells with extremely large or small PSI values relative to most of the cell population (Supplementary Figure S8A–D).

To mitigate differentially spliced events driven by these outlier cells, we applied MARVEL's bimodal-adjusted modality assignment to identify the events that demonstrated only included to included or excluded to excluded modalities. We set a heuristic threshold of at least 10 cells with PSI values >0 or <100 in at least one of two of the cell populations for the differentially spliced events to be retained for downstream analysis (Supplementary Figure S8E–H). We implemented this outlier removal technique into MARVEL. MARVEL's method removed outliers and retained a higher number of differentially spliced events detected by DTS than other nonparametric tests and BRIE2, which uses the Bayesian model selection method (19) (Supplementary Figure S8I). Because AD and DTS captured most differential splicing events (Supplementary Figure S8J), we combined AD and DTS tests followed by the outlier removal as the default method for MARVEL's differential splicing analysis. Using a comparable cut-off, MARVEL identified 114 differentially spliced events compared to 73 by BRIE2 (Figure 2J). Next, we investigated whether differentially spliced events detected by MARVEL were biologically relevant to myoblast differentiation. As expected, muscle-related genes were differentially spliced (Supplementary Figure S8K–N). Gene ontology analysis of all differentially spliced genes showed the enrichment of muscle-related pathways when immature myoblasts were differentiated into mature cells (Figure 2K). Moreover, gene ontology analysis of differentially spliced genes detected exclusively by MARVEL additionally identified pathways related to protein translation and localization, and cell cycle pathways (Supplementary Figure S8O).

To assess the generalizability of our method, we performed differential splicing analysis on single-cell neurons derived from mice induced with multiple sclerosis compared to healthy mice (31). We compared differentially spliced events detected by MARVEL against BRIE2 (in mode 2-diff), which effectively identified differential alternative splicing events for this dataset (19). Using a comparable cut-off, MARVEL identified 248 differentially spliced events compared to 238 by BRIE2 (Figure 2L). Both MARVEL and BRIE2 identified a splicing event that was previously validated using qPCR (31) (Supplementary Figure S8P). Gene ontology analysis of differentially spliced genes detected by MARVEL identified RNA splicing and neuron-related pathways to be enriched, as expected when mice were induced to manifest an autoimmune disease that attacks the nervous system (Figure 2M). Moreover, gene ontology analysis of differentially spliced genes detected exclusively by MARVEL identified lysosome and neurotransmission pathways (Supplementary Figure S8Q).

Taken together, MARVEL identified biologically relevant pathways during muscle cell maturation and in a mouse model with multiple sclerosis. MARVEL complements existing single-cell alternative splicing tools by detecting additional differentially spliced genes and biological pathways.

MARVEL application for analyzing a plate-based scRNA-seq dataset

After benchmarking MARVEL, we showed the functional features provided by MARVEL to characterize the single-cell alternative splicing landscape in induced pluripotent stem cells (iPSCs) and iPSC-derived endoderm cells (29).

MARVEL detected 13 125 and 5308 splicing events in iPSCs and endoderm cells. The most prevalent splicing event in iPSCs and endoderm cells was SE, followed by RI, AFE, A3SS, A5SS, ALE and MXE (Supplementary Figure S9A and B). We investigated whether alternative splicing repre-

sented an underappreciated layer of complexity underlying gene expression profile by performing linear dimension reduction analysis using gene expression and PSI values of alternative splicing events. Differentially expressed genes and spliced events robustly distinguished the cell types (Figure 3A and B), whereas non-differentially expressed genes could not separate the cell types (Figure 3C). Interestingly, differential splicing events from non-differentially expressed genes could clearly delineate the two cell types (Figure 3D; Supplementary Figure S9C–I), except for MXE, which was due to the low number of MXE events detected in this analysis (Supplementary Figure S9D). Lastly, all splicing events expressed in both cell types, regardless of whether the events were differentially spliced or not, were also able to separate the cell types (Supplementary Figure S9J).

Similar to previous studies (20,29), we next explored PSI distributions (modalities) of individual alternative splicing events identified in iPSCs and endoderm cells. Modality assignment can inform whether a predominant isoform (included and excluded) or both isoforms (bimodal, middle and multimodal) are expressed and contribute to cellular identity in a cell population. MARVEL assigned distinct modalities to alternative splicing events for each cell type (Figure 3E and Supplementary Figure S9K). Both cell types showed a high proportion of included and excluded modalities (∼97% of all modality types), whereas other modalities (bimodal, middle, and multimodal) showed only ∼3%. This is consistent with previous empirical and simulated studies (22,54). In addition to the original modalities previously proposed (20), MARVEL could stratify the included and excluded modalities into primary and dispersed. In iPSCs (Figure 3E), we observed primary and dispersed modalities constituted 40.9% and 59.1% of the included modality, whereas primary and dispersed modalities constituted 44.5% and 55.5% of the excluded modality. Similar proportions were observed in endoderm cells (Supplementary Figure S9K). Further stratification of modality types by alternative splicing event types showed different proportions of modality types by alternative splicing events (Figure 3F and Supplementary Figure S9L). For example, excluded modality was most prevalent in RI events, constituting 73.6% and 75.2% of all modality types in this splicing event type in iPSCs and endoderm cells. This is consistent with the reported role of intron retention in gene regulation (55).

We performed differential splicing analysis to detect differentially spliced events to understand the splicing dynamics when iPSCs were differentiated into endoderm cells. MARVEL identified 1,614 differential alternative splicing events comprising 816 genes (Figure 3G; Supplementary Tables S2 and S3). Top differentially spliced genes included *DNAJC15*, *SNRPN*, *RPL26*, *RPS24* and *RPS10*. *DNAJC15* regulates cellular metabolism (56), whereas *SNRPN*, *RPL26*, *RPS24* and *RPS10* are ribonuclear proteins involved in transcription and translation (57). We selected representative differential alternative splicing events from SE, MXE, RI, A5SS and A3SS for visual validation using VALERIE (23) (Supplementary Figure S10A–E) (23).

MARVEL further categorized differential alternative splicing events based on modality changes between iPSCs and endoderm cells. We defined modality changes as explicit, implicit, and restricted. Explicit is defined as clear changes within the five main modalities (included, excluded, bimodal, middle and multimodal). Implicit is defined as changes involving the sub-modalities primary and dispersed. Restricted is defined as no modality changes across the two cell populations. During iPSCs to endoderm cell differentiation, we observed 160, 300 and 1154 explicit, implicit, and restricted modality changes, respectively, among the differential alternative splicing events (Figure 3H). Notably, most differential alternative splicing events, 1454 (90%) events, would have been missed if differences in splicing patterns were detected based on explicit modality change alone as used in the previous study (20). Examples of genes that underwent explicit, implicit, and restricted modality change from iPSCs and endoderm cells were *CNBP*, *SOX4* and *DPPA4* (Figure 3I–K). CNBP is dysregulated in iPSC derived from patients with myotonic dystrophy (58), whereas SOX4 and DPPA4 are transcription factors shown to be dynamically regulated during endoderm induction from iPSCs (59).

General scRNA-seq analysis pipelines perform either differential gene expression or alternative splicing analysis alone but do not integrate both analyses into a single framework. MARVEL allows the integration of differential alternative splicing and gene expression analysis. This enabled us to investigate the changes in alternative splicing relative to changes in gene expression when iPSCs were differentiated into endoderm cells. We identified 816 differentially spliced genes among the 1,614 differentially splicing events. 479 (58%) genes were concurrently differentially expressed based on Wilcoxon rank-sum test (Figure 3L). Differential gene expression analysis using MAST identified 234 (27%) differentially spliced genes to be concurrently differentially expressed, and 219 of them overlapped with genes identified from Wilcoxon rank-sum test ($P < 2.2e{-}16$; Supplementary Figure S9M). To explore the relationship between differential genes and alternative splicing changes, MARVEL categorized the changes in gene expression relative to changes in PSI values into coordinated, opposing, isoform switching, and complex (Figure 3M). Coordinated and opposing relationships are defined as changes in gene expression between two cell populations in the same or opposite direction to the change in average PSI values (Figure 3N–Q). For example, DHX9, involved in chromatin remodeling during stem cell differentiation (60), showed coordinated gene-splicing changes, whereby the gene expression and PSI values were decreased from iPSCs to endoderm cells. On the other hand, *BCLAF1* encodes for an anti-apoptotic protein that promotes maintenance and self-renewal of stem cells (61), showed opposing gene-splicing changes, whereby there was a decrease in gene expression from iPSCs to endoderm cells, but PSI values of an RI event were increased. Isoform switching is defined as genes showing differential splicing but not differentially expressed (Figure 3R–S). CELF1, involved in regulating the stability and translation of mRNA during the differentiation process (62), showed no significant difference in gene expression, and mean PSI values in both cell populations were similar, but the overall PSI distribution for an A3SS event was changed from excluded dispersed in iPSCs to bimodal in endoderm cells (explicit modality change). Lastly, a complex relationship involves a combination of coordinated,

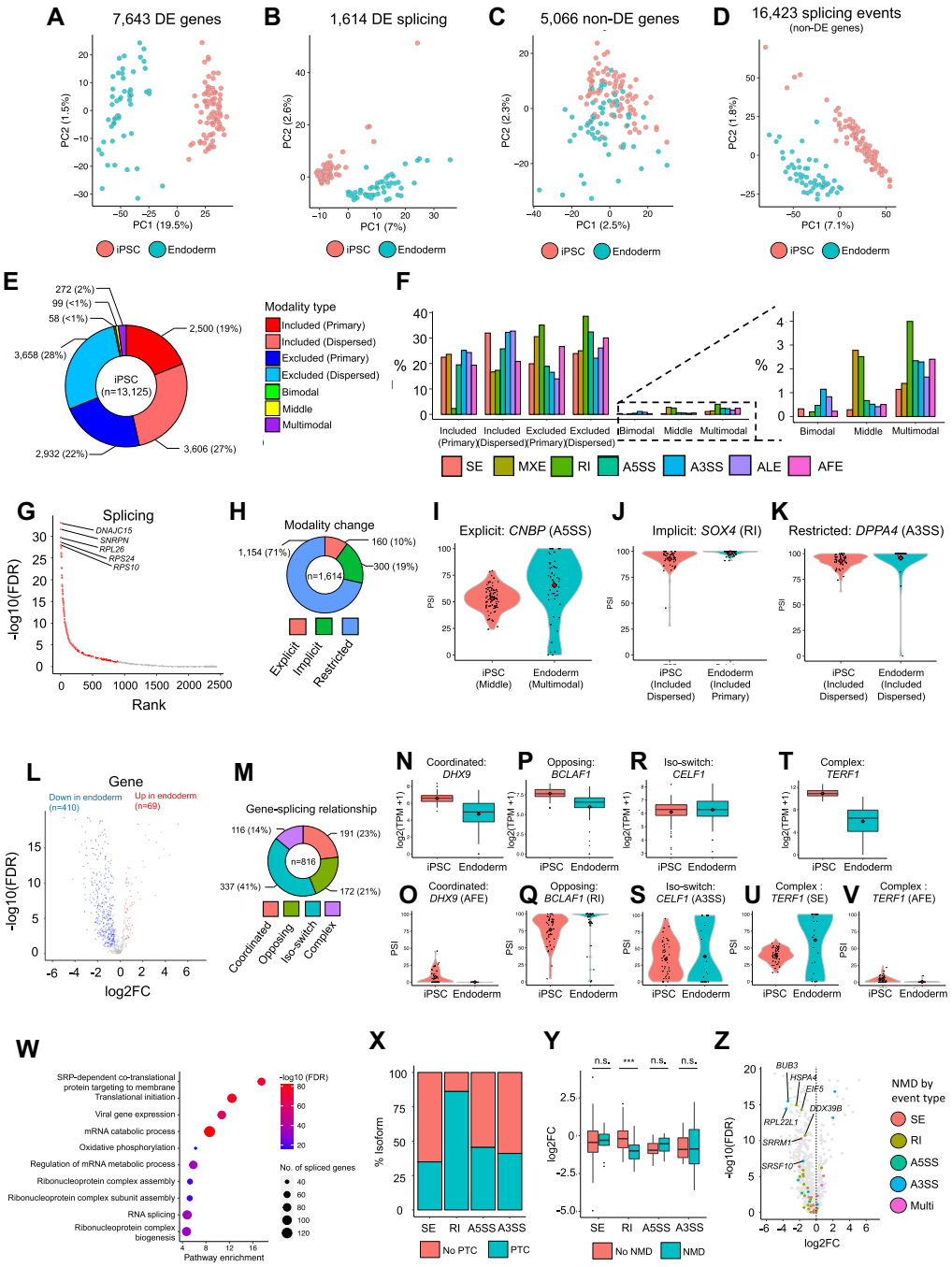

Figure 3. Application of MARVEL on plate-based scRNA-seq data from iPSCs differentiated to endoderm cells. (A–D) PCA plots using (**A**) differentially expressed genes, (**B**) differential alternative splicing events, (**C**) non-differentially expressed genes, and (**D**) all alternative splicing events from non-differentially expressed genes. (**E**) The proportion of each modality class in iPSCs. (**F**) The proportion of each modality class by splicing event type in iPSCs. (**G**) Ranked list of differentially alternative splicing events identified using MARVEL, comparing PSI distributions between iPSCs and endoderm cells. (**H**) The proportion of each modality dynamic class, i.e. the type of changes in modality of alternative splicing events from iPSCs to endoderm cells. (I–K) Representative alternative splicing events of each modality dynamic class (**I**) explicit, (**J**) implicit and (**K**) restricted. (**L**) Differential expression of genes that were differentially spliced. Blue denotes down-regulated genes ($log_2FC < -0.5$ and FDR < 0.10), red denotes up-regulated genes ($log_2FC > 0.5$ and FDR < 0.10), and grey denotes non-differentially expressed genes ($-0.5 < log_2FC < 0.5$ or FDR ≥ 0.10). (**M**) The proportion of each gene-splicing relationship class, i.e. the type of changes in average gene expression value relative to change in average PSI value for the corresponding alternative splicing event from iPSCs to endoderm cells. (N–V) A representative gene and corresponding alternative splicing event for each gene-splicing relationship class, (**N, O**) coordinated, (**P**, **Q**) opposing, (**R**, **S**) isoform switching and (**T**–**V**) complex. (**W**) Enrichment scores, FDR values, and gene set sizes of selected biological pathways enriched among differentially spliced genes. (**X**) The proportion of isoforms with PTC introduced by SE, RI, A5SS and A3SS. (**Y**) Boxplots showing the comparison of log2FC between genes that were not predicted or predicted to be subjected to alternative splicing-mediated NMD. (**Z**) A volcano plot showing differential gene expression analysis between iPSCs and endoderm cells annotated with genes predicted to be subjected to alternative splicing-mediated NMD. FC: fold change; NMD: nonsense-mediated decay; PTC: premature stop codon. *** FDR < 0.01, ** FDR < 0.05, * FDR < 0.1.

opposing, and/or isoform switching relationships (Figure 3T–V). For instance, *TERF1*, involved in telomere elongation and maintenance of pluripotency in the iPSCs (63), showed a significant decrease in gene expression in endoderm cells, whereas a SE event of this gene showed a clear modality change from middle to bimodal with a slight increase in PSI values, while a separate AFE splicing event of this gene showed a decrease in PSI values in endoderm cells. Opposing, isoform switching, and complex gene-splicing relationship constituted the majority of the relationships, 632/816 (77%). Therefore, most PSI changes may not be inferred directly from gene expression changes alone. This highlights the value of differential splicing analysis in revealing additional differentially regulated genes.

To assess whether functionally related genes or genes that belong to the same biological pathways are coordinatedly and differentially spliced, we performed gene ontology analysis using 816 differentially spliced genes detected by MARVEL. MARVEL identified 141 significantly enriched pathways among the differentially spliced genes, including pathways related to RNA splicing, gene translation and regulation, and ribonucleoprotein complex formation (Figure 3W). Both RNA splicing and ribonucleoproteins have been shown to regulate stem cell self-renewal and differentiation by modulating protein translation (20,64).

To understand the functional consequences of alternative splicing of differentially spliced genes, MARVEL predicted nonsense-mediated decay (NMD) for a given alternative splicing event and investigated the relationship between gene expression and alternative splicing-related NMD. We observed RI as alternative splicing events that affected protein-coding transcripts with the highest rate (86%) of introducing premature terminal codons (PTCs), followed by A5SS (46%), A3SS (41%) and SE (35%) (Figure 3X). Only RI-mediated NMD led to a significant decrease in gene expression when iPSCs were differentiated into endoderm cells (Figure 3Y). This is consistent with a previous study that showed a decrease in gene expression by RI-mediated NMD but not NMD mediated by other splicing event types (65). Genes subjected to alternative splicing-related NMD and were concurrently down-regulated in endoderm cells included *BUB3*, *HSPA4*, *EIF5*, *RPL22L1*, *DDX39B, SRRM1* and the splicing factor *SRSF10* (Figure 3Z). BUB3 is essential for mitotic spindle checkpoint function during cellular proliferation and differentiation (66), whereas HSPA4 represents a class of heat-shock proteins (HSPs) targeting misfolded proteins for degradation (67). EIF5 and RPL22L1 are involved in transcription and translation (68,69). DDX39B, SRRM1 and SRSF10 regulate RNA splicing (70–72). Taken together, MARVEL links NMD-related splicing changes to gene expression changes to enable prioritization of candidate spliced genes for downstream functional studies (25).

To further validate the computational results of MARVEL, we analyzed a single-cell full-length transcriptome dataset constituting seven cell populations during mouse endothelial-to-hematopoietic stem cell (HSC) transition with experimentally validated splicing events (32). MARVEL identified an average of 5,614 expressed alternative splicing events per single cell (Supplementary Figure S11A). T1 pre-HSCs expressed the highest number of splicing events (Supplementary Figure S11B). Overall, SE was the most prevalent splicing event, followed by RI, A3SS, A5SS, AFE, ALE and MXE (Supplementary Figure S11C–I). Stratification of splicing distributions into modalities for each cell population revealed included and excluded to be the most common splicing patterns (Supplementary Figure S12A). Notably, 42% and 35% of splicing events demonstrated modality change from AECs to HECs and from HECs to T1 pre-HSCs, respectively (Supplementary Figure S12B and C). We next focused on splicing events whose modality changed from others (bimodal/middle/multimodal/no modality) to included modality as previously described (32). We observed 2950 HEC-initiated included splicing events defined as splicing events that transited from bimodal/middle/multimodal/no modality in AECs to included modality in HECs (Supplementary Figure S12D). We further observed 1956 T1 pre-HSC-initiated included splicing events defined as splicing events that transited from bimodal/middle/multimodal/no modality in AECs and HECs to included modality in T1 pre-HSCs. Notably, 852 splicing events retained their included modality from HECs to adult HSCs, and these splicing events were defined as HEC-persistently included splicing events. Pathway enrichment analysis revealed pathways associated with RNA metabolism, transcription, translation, and cell cycle to be enriched among genes that constituted the HEC- and T1 pre-HSC-initiated, and HEC-persistently included splicing events (Supplementary Figure S12E). Examples of HEC- and T1 pre-HSC-initiated, or HEC-persistently included splicing events were *Sec31a*, *Zfpl1*, *Coro1a*, *Mpdu1*, *Ntmt1*, *Clk1*, *Ttll4* and *E130309D02Rik* (Supplementary Figure S12F and G), which have been experimentally validated in HECs and T1 pre-HSCs using FISH previously (32). In general, MARVEL reproduced and showed similar results as described in the original study (32).

MARVEL application for analyzing 10x genomics scRNA-seq dataset

MARVEL also facilitates single-cell alternative splicing analysis for droplet-based scRNA-seq library preparation methods, such as 10x Genomics. To demonstrate the utility of MARVEL, we analyzed scRNA-seq data from the cell differentiation of iPSCs into 10-day-old cardiomyocytes (30). Differential splice junction analysis identified 575 and 243 splice junctions, comprising 539 genes, significantly up or downregulated, in cardiomyocytes relative to iPSCs (Figure 4A; Supplementary Table S4). Differentially spliced genes were enriched in muscle and actin-myosin filament sliding, stem cell differentiation, energy production, and WNT signaling pathways (Figure 4B). Examples of differentially spliced genes included *MYH10*, *ATP5F1C* and *CBX1* (Figure 4C–F) (73). MYH10 is required for proper functioning of the epicardial and formation of coronary vessels (74). *ATP5F1C* encodes a subunit of mitochondrial ATP synthase required for energy production (75). CBX proteins play a role in chromatin remodeling and neuron development (76).

We identified 539 differentially spliced genes among 818 differentially spliced junctions. 222 (41%) genes were con-

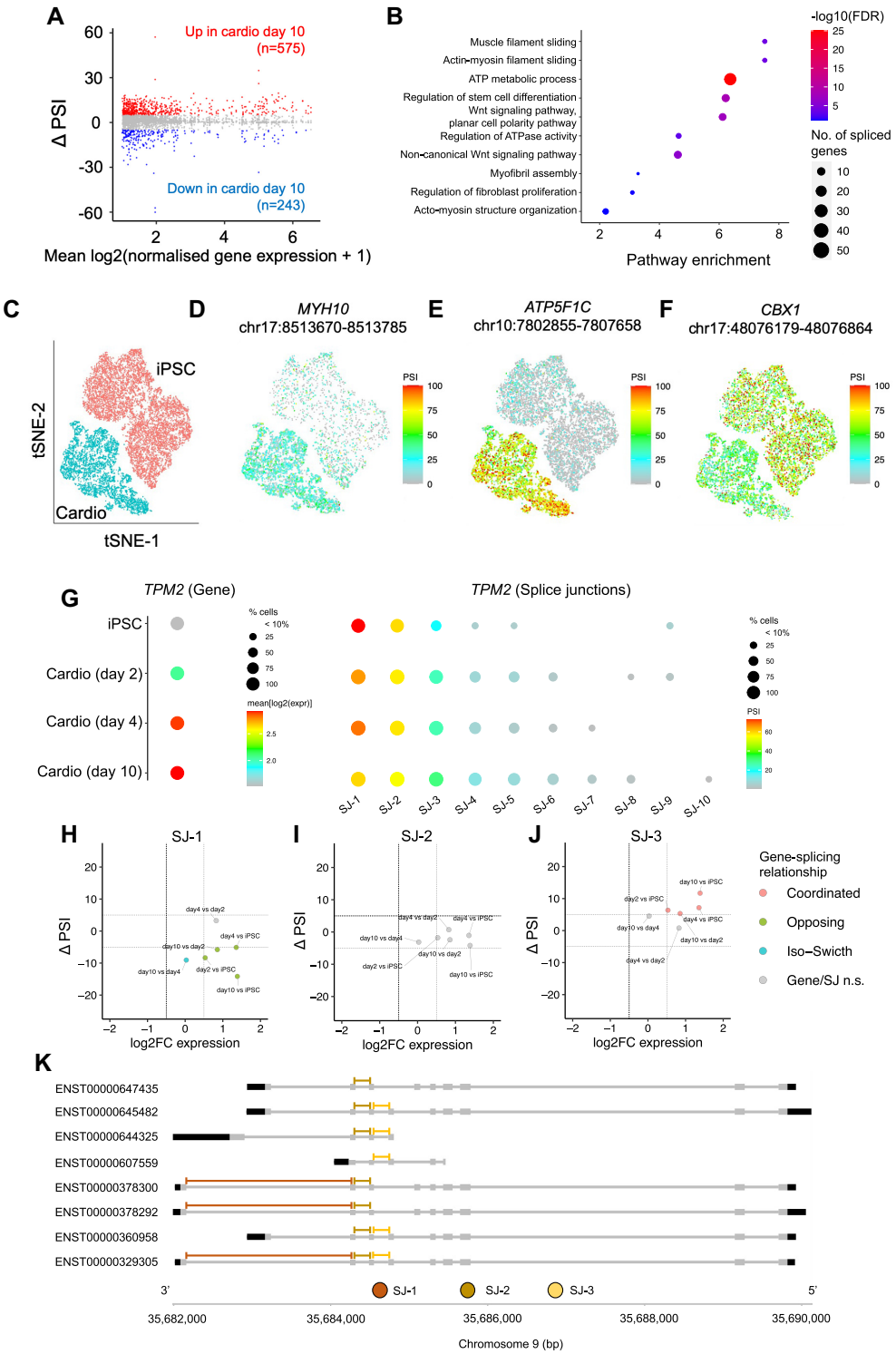

Figure 4. Application of MARVEL on iPSCs differentiated to cardiomyocytes. (**A**) Mean PSI difference in splice junction usage in cardiomyocytes relative to iPSCs. X-axis represents the mean normalized gene expression value across all cells from both populations. Y-axis represents the difference in mean PSI values. Blue denotes splice junctions down-regulated in cardiomyocytes relative to iPSCs ($\Delta PSI < -5$ and $P < 0.05$). Red denotes splice junctions up-regulated in cardiomyocytes relative to iPSCs ($\Delta PSI > 5$ and $P < 0.05$). Grey denotes splice junctions not differentially spliced between the two cell populations ($-5 < \Delta PSI < 5$ or $P \geq 0.05$). (**B**) Enrichment scores, FDR values, and gene set sizes of pathways enriched among differentially spliced genes. (C–F) tSNE embeddings generated using 1160 highly variable genes and consisting of 11 244 iPSCs and 5937 day-10 cardiomyocytes annotated with (**C**) cell types and muscle-related genes, (**D**) *MYH10*, (**E**) *ATP5F1C* and (**F**) *CBX1*. (**G**) *TPM2* gene expression and splice junction usage across all cardiomyocyte developmental stages. Splice junctions were arranged in decreasing expression from left to right. (**H–J**) Changes in splice junction usage relative to changes in gene expression levels when comparing more mature cardiomyocytes to less mature cardiomyocytes and iPSCs for the top three highly expressed splice junctions. (**K**) The gene browser shows the relative position of the top three highly expressed splice junctions on *TPM2* protein-coding transcripts.

currently differentially expressed based on Wilcoxon rank-sum test. Using MAST, we identified 232 (43%) differentially spliced genes to be concurrently differentially expressed, and 219 of them overlapped with identified genes from Wilcoxon rank-sum test ($P < 2.2e{-}16$; Supplementary Figure S13A). Next, we investigated gene expression changes relative to splice junction usage changes. Of the 539 genes that were differentially spliced, 19% and 15% of these genes exhibited changes in the same (coordinated) or opposite (opposing) direction relative to splice junction usage, respectively (Supplementary Figure S13B). Genes with a coordinated or opposing relationship with splice junction usage were *VIM* and *UQCRH* (Supplementary Figure S13C–F). *VIM* plays an essential role in maintaining muscle cytoarchitecture and is a reliable marker for muscle cell regeneration (77), while *UQCRH* participates in cardiac muscle contraction (78). More than half (59%) of differentially spliced genes exhibited isoform switching, i.e. differential splice junction usage in the absence of differential gene expression changes, such as the *RBM39* gene that was differentially spliced but not differentially expressed (Supplementary Figure S13G–I). RBM39 is an RNA-binding protein involved in alternative splicing and genetic mutations in *RBM39* are associated with muscle myopathies (79). Last, 7% of genes exhibited a complex relationship with its splice junction usage. Examples of genes with a complex relationship with splice junction usage are *TPM1* and *TPM2* (Supplementary Figure S13J–O). *TPM1* and *TPM2* gene expression were up-regulated in cardiomyocytes relative to iPSCs (Supplementary Figure S13J and M). While one of the splice junctions in both genes exhibited higher expression in cardiomyocytes (Supplementary Figure S13K and N), the other splice junctions exhibited higher expression in iPSCs (Supplementary Figure S13L and O). *TPM1* and *TPM2* are members of the tropomyosin family of highly conserved actin-binding proteins involved in striated and smooth muscle contraction. Genetic mutations in *TPM1* are associated with cardiac hypertrophy, while genetic mutations in *TPM2* were previously reported in patients with congenital myopathy (80,81).

Most differentially spliced genes, 438 (82%), occurred in the absence (iso-switch) or opposite to splice junction usage changes (opposing relationship) or have a complex relationship with splice junction usage changes. Therefore, most splice junction usage changes cannot be inferred directly from gene expression changes alone.

To further illustrate the intricate relationship between splicing and gene expression profile, we characterized the overall splice junction usage of *TPM2* relative to its corresponding gene expression changes across the developmental stages of cardiomyocytes. We chose *TPM2* for demonstration because its splice junction usage showed a complex relationship relative to its gene expression changes when comparing day-10 cardiomyocytes to iPSCs (Supplementary Figure S13N and O). *TPM2* expression increased from iPSCs to more mature cardiomyocytes (Figure 4G). On the other hand, splice junction-1 (SJ-1; chr9:35682164–35684245) decreased from iPSCs relative to mature cardiomyocytes (Figure 4H), while SJ-2 (chr9:35684316–35684487) usage was relatively consistent across all developmental stages (Figure 4I). Similar to gene expression changes, SJ-3 (chr9:35684551–35684731) usage increased from iPSCs to mature cardiomyocytes (Figure 4J). The overall splice junction usage across developmental stages decreased from SJ-1 to SJ-2 and SJ-3. We hypothesized that this is due to the 3′-bias inherent in scRNA-seq datasets generated from 3′-bias library preparation methods. To this end, we implemented a gene browser visualization function in MARVEL to inspect the specific location of splice junctions of interest relative to the transcripts. Indeed, SJ-1 was located on the most 3′-end of the transcripts, followed by SJ-2 and SJ-3 (Figure 4K). Therefore, the end-bias inherent in single-cell library preparation methods would be taken into account during single-cell alternative splicing analysis.

Lastly, we assessed the ability of MARVEL to scale to a large 10x genomics dataset. We analyzed 104 559 cells derived from brain tissues of 15 ASD patients and 16 controls (47), consisting of 17 cell populations (Supplementary Figure S14A). Among neuronal cell types, 691 and 1903 splice junctions were significantly spliced-in and -out, respectively, in ASD patients relative to controls (Supplementary Figure S14B). Among non-neuronal cell types, 297 and 173 splice junctions were significantly spliced-in and -out, respectively, in ASD patients relative to controls (Supplementary Figure S14C and Supplementary Table S5). Notably, *SYT1* gene, a canonical marker gene for excitatory neurons (47), was differentially spliced. Specifically, the splice junction chr12:78865110–78977798 of this gene was significantly spliced-out in ASD patients relative to controls (Supplementary Figure S14B and D). We further evaluated if differentially spliced genes between ASD patients and controls were enriched in previously reported ASD-related genes. Of the 602 differentially spliced genes, 49 overlapped ($P = 5.6e{-}24$) with ASD-related genes reported by the Simons Foundation Autism Research Initiative (SFARI) database (82) (Supplementary Figure S14E). Pathway enrichment analysis of differentially spliced genes showed RNA splicing processes and pathways associated with the nervous system, including synaptic, axonic and dendritic processes, and tau-protein kinase activity (Supplementary Figure S14F). Overall, among the 17 cell populations, L5/6-CC showed the highest number of differentially spliced junctions (Supplementary Figure S14G). The median processing time for differential splicing analysis between ASD patients and controls for a given cell population was ∼1.5 min (Supplementary Figure S14H). The total running time to complete the differential splicing analysis across all 17 cell populations was ∼36 min, computed on MacBook Pro with 2 GHz Quad-Core Intel Core i5 processor and 16GB 3733 MHz LPDDR4X memory.

DISCUSSION

We have developed MARVEL to address key issues in single-cell alternative splicing analysis and enable transcriptome-wide characterization of the alternative splicing dynamics in scRNA-seq datasets. We benchmarked MARVEL against the existing alternative splicing analysis tools and demonstrated the utility of MARVEL using for datasets generated from the plate- and droplet-based methods.

While MARVEL is largely an application-oriented R package, it also introduces several approaches for single-cell alternative splicing analysis. First, MARVEL identifies and adjusts false bimodal modality classification of PSI distributions that are attributed to PCR amplification bias (14), to enable more reliable modality classification of splicing patterns. Second, MARVEL combines Anderson–Darling and D Test Statistics (DTS) (44) together with the bimodal-adjusted algorithm for differential splicing analysis to identify differentially spliced exons between two cell populations. Third, MARVEL incorporates a permutation-based approach to identify differentially spliced junctions between two pseudo-bulk populations for scRNA-seq data generated from droplet-based methods. Fourth, MARVEL detects and quantifies biologically-relevant alternative first and last exon (AFE and ALE) splicing events. To date, most splicing analysis tools focus on SE, MXE, RI, A5SS, and A3SS (19,20,39). Fifth, MARVEL enables splicing-associated nonsense-mediated decay (NMD) prediction. There are no comprehensive R packages to date that implement splicing-related NMD prediction. We summarized the available features of MARVEL compared to other single-cell alternative splicing analysis tools shown in Supplementary Table S1.

MARVEL employed a splice junction-based approach to estimate the PSI directly from splice junction reads that reflect true biological phenomena (20). PSI values estimated by using probabilistic frameworks, such as BRIE, have shown bias in PSI estimation and underestimated cell-to-cell heterogeneity at low coverage (18,83,84). We showed that MARVEL had a better cell-to-cell and cell-to-bulk correlation of PSI values in homogenous cell lines than the Bayesian approach. It is noteworthy that Expedition and MARVEL demonstrated similar cell-to-cell and cell-to-bulk correlation of PSI values because both software utilized a splice junction-based approach for PSI quantification.

Most single-cell alternative splicing analysis tools constrained PSI quantification to only an exon-skipping splicing event. Nevertheless, other splicing event types also contribute to the cellular phenotype. For example, aberrant intron retention in cancer leads to abnormal proteins presented on the tumor surface as neoantigens, which may be amenable to immunotherapy (26). Alternative 3′ splice sites are preferentially mis-spliced by mutant splicing factor *SF3B1* (85,86). Therefore, MARVEL has been developed to include PSI quantification for all main exon-level splicing event types, comprising SE, MXE, RI, A5SS, A3SS, AFE and ALE. MARVEL also requires less processing time and lower memory usage than other tools.

PSI values reflect the percentage of splice junction reads supporting the alternative exons and are therefore represented by any values between 0 and 100. Song *et al.* previously introduced the concept of 'modality' to categorize the PSI distribution for a given alternative splicing event into discrete categories (20). The classes of modalities were included, excluded, bimodal, middle, and multimodal. MARVEL introduces primary and dispersed sub-modalities for included and excluded modalities. We showed that ∼50% of included and excluded comprised of primary sub-modality, whereas another ∼50% showed dispersed sub-modality, suggesting that MARVEL could increase the current repertoire of modality classes and provide a finer distinction between the different PSI distributions.

A significant proportion of bimodality may have been misclassified (14). We tabulated a catalog of true and false bimodal alternative splicing events previously validated using qPCR, smFISH, and inferred mRNA counts (14,16,20,29). We identified key features that distinguished true from false bimodality. These features were incorporated into MARVEL to identify and adjust for the false bimodal class, leading to more accurate modality classification and modality change detection between different cell populations.

Current approaches for differential alternative splicing analysis in single cells include a comparison of two cells at a time or detection of modality changes between cell populations. For the former approach, comparing all possible cell pairs is impractical when the number of cells becomes large (18). For the latter approach, changes in splicing patterns are defined on modality changes across different cell populations, such as included to excluded modality change (20). This approach may miss changes in splicing patterns that do not involve any modality change. MARVEL incorporated the statistical framework Anderson–Darling and D Test Statistic (44) combined with the bimodal-adjusted modality assignment to enable unbiased evaluation of the differences in PSI distribution across different cell populations. We showed that 90% of differential alternative splicing events identified by MARVEL, when iPSCs were differentiated into endoderm cells, demonstrated no explicit change in PSI modality. These events would have been missed based on the original modalities proposed by Song *et al.* (20).

Current single-cell analysis tools offer only gene or alternative splicing analysis exclusively (18,20,87). Nevertheless, alternative splicing and gene expression changes may be intricately linked. MARVEL integrates alternative splicing and gene expression to study the relationship between alternative splicing and gene expression changes across different cell populations. For example, comparative analysis between iPSCs and endoderm cells demonstrated that only about 23% of differentially spliced genes showed gene expression changes that occurred in the same direction as the corresponding PSI changes. The remaining differentially spliced genes occurred in the opposite or the absence of gene expression changes. This reaffirms the complex relationship underlying gene expression and alternative splicing.

Current single-cell alternative splicing analysis tools fall short in providing context to understand the functional consequence of alternative splicing. Alternative splicing represents one of many mechanisms by which gene expression is regulated. We incorporated nonsense-mediated decay (NMD) prediction as a functional annotation feature, in addition to gene ontology analysis, into MARVEL. MARVEL can predict whether the insertion of a given alternative exon subjects the corresponding isoforms to NMD or not. It compares gene expression levels between genes that are

subjected to NMD. We showed that increased intron retention decreased gene expression levels when iPSCs were differentiated into endoderm cells. This is reminiscent of the complex, in this case opposing relationship between alternative splicing and gene expression changes. This finding is in line with a previous report demonstrating that only intron retention, but not other splicing event types, was associated with decreased gene expression levels using long-read RNA-sequencing in *SF3B1*-mutated chronic lymphocytic leukemia patients (65).

We extended MARVEL's framework to enable integrated gene and alternative splicing analysis in the dataset generated from a droplet-based platform. MARVEL was able to identify differential splice junction usage enriched in muscle-, neuron-, and heart-related pathways in iPSCs differentiated to cardiomyocytes. Moreover, only 19% of differentially spliced genes demonstrated the same directional changes in splice junction usage and gene expression. This is consistent with the intricate relationship between alternative splicing changes and gene expression changes revealed by the plate-based analysis. Lastly, MARVEL enables single-cell visualization of splice junction usage on linear or non-linear dimensionality reduction to verify differential splicing junction usage across different cell populations.

Due to the technical differences between plate- and droplet-based library preparation protocols, MARVEL implements different strategies to address the computational challenges arising from RNA-sequencing datasets. MARVEL computes PSI values at the exon-level for plate-based sequencing data, whereas MARVEL computes PSI values at the splice junction-level for droplet-based sequencing data. This is because the sequencing coverage is more uniformly distributed across the isoforms in plate-based compared to droplet-based sequencing data (24,50,88). MARVEL creates a pseudo-bulk for each cell population prior to differential splicing analysis of any two cell populations for droplet-based sequencing data. This is because the dropout rate (the number of cells whereby the splice junction is not detected) in droplet-based is higher than in plate-based sequencing data (50,51). Lastly, MARVEL leverages on the exon-level information available in plate-based splicing analysis for NMD prediction.

Our study provides a comprehensive computational framework to characterize alternative splicing dynamics at single-cell resolution. As far as we are aware, MARVEL is the only single-cell alternative splicing computational tool to enable alternative splicing analysis on scRNA-seq data generated from the plate- and droplet-based library preparation methods. MARVEL supports the integration of gene-level expression and alternative splicing analysis. For alternative splicing analysis, MARVEL provides end-to-end features to characterize single-cell alternative splicing landscape, starting from alternative splicing event validation, percent spliced-in quantification, modality assignment and correction, differential splicing analysis, to the functional annotation using gene ontology and nonsense-mediate decay prediction. We anticipate MARVEL to be prospectively applied to single-cell datasets generated from various settings (e.g. health and disease states) to reveal novel biological insights.

DATA AVAILABILITY

MARVEL is available on the Comprehensive R Archive Network (CRAN): https://cloud.r-project.org/web/packages/MARVEL/index.html. The software tutorial containing the pre-processed data and codes to reproduce the figures related to the application of MARVEL on plate- and droplet-based RNA-sequencing data are available at https://wenweixiong.github.io/MARVEL_Plate.html and https://wenweixiong.github.io/MARVEL_Droplet.html respectively. All data sources included in this study are publicly available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the WIMM Centre for Computational Biology for testing the MARVEL package and providing useful input. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

FUNDING

The Clarendon Fund; Oxford-Radcliffe Scholarship in conjunction with WIMM Prize PhD Studentship (to W.X.W.); Medical Research Council (MRC) Senior Clinical Fellowship; CRUK Senior Cancer Research Fellowship (to A.J.M.); Oxford-Bristol Myers Squibb (BMS) Fellowship (to S.T.); National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC).
Conflict of interest statement. None declared.

REFERENCES

1. Roy,A., Wang,G., Iskander,D., O'Byrne,S., Elliott,N., O'Sullivan,J., Buck,G., Heuston,E.F., Wen,W.X., Meira,A.R. *et al.* (2021) Transitions in lineage specification and gene regulatory networks in hematopoietic stem/progenitor cells over human development. *Cell Rep.*, **36**, 109698.
2. Popescu,D.M., Botting,R.A., Stephenson,E., Green,K., Webb,S., Jardine,L., Calderbank,E.F., Polanski,K., Goh,I., Efremova,M. *et al.* (2019) Decoding human fetal liver haematopoiesis. *Nature*, **574**, 365–371.
3. Aizarani,N., Saviano,A., Sagar, Mailly,L., Durand,S., Herman,J.S., Pessaux,P., Baumert,T.F. and Grun,D. (2019) A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*, **572**, 199–204.
4. Eze,U.C., Bhaduri,A., Haeussler,M., Nowakowski,T.J. and Kriegstein,A.R. (2021) Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.*, **24**, 584–594.
5. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The Human Cell Atlas. *Elife*, **6**, e27041.
6. Giustacchini,A., Thongjuea,S., Barkas,N., Woll,P.S., Povinelli,B.J., Booth,C.A.G., Sopp,P., Norfo,R., Rodriguez-Meira,A., Ashley,N. *et al.* (2017) Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.*, **23**, 692–702.

7. Psaila,B., Wang,G., Rodriguez-Meira,A., Li,R., Heuston,E.F., Murphy,L., Yee,D., Hitchcock,I.S., Sousos,N., O'Sullivan,J. *et al.* (2020) Single-cell analyses reveal megakaryocyte-biased hematopoiesis in myelofibrosis and identify mutant clone-specific targets. *Mol. Cell*, **78**, 477–492.
8. Louka,E., Povinelli,B., Rodriguez-Meira,A., Buck,G., Wen,W.X., Wang,G., Sousos,N., Ashley,N., Hamblin,A., Booth,C.A.G. *et al.* (2021) Heterogeneous disease-propagating stem cells in juvenile myelomonocytic leukemia. *J. Exp. Med.*, **218**, e20180853.
9. Patel,A.P., Tirosh,I., Trombetta,J.J., Shalek,A.K., Gillespie,S.M., Wakimoto,H., Cahill,D.P., Nahed,B.V., Curry,W.T., Martuza,R.L. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.
10. Ma,L., Hernandez,M.O., Zhao,Y., Mehta,M., Tran,B., Kelly,M., Rae,Z., Hernandez,J.M., Davis,J.L., Martin,S.P. *et al.* (2019) Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. *Cancer Cell*, **36**, 418–430.
11. Mathys,H., Davila-Velderrain,J., Peng,Z., Gao,F., Mohammadi,S., Young,J.Z., Menon,M., He,L., Abdurrob,F., Jiang,X. *et al.* (2019) Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, **570**, 332–337.
12. Mazin,P.V., Khaitovich,P., Cardoso-Moreira,M. and Kaessmann,H. (2021) Alternative splicing during mammalian organ development. *Nat. Genet.*, **53**, 925–934.
13. Wen,W.X., Mead,A.J. and Thongjuea,S. (2020) Technological advances and computational approaches for alternative splicing analysis in single cells. *Comput. Struct. Biotechnol. J.*, **18**, 332–343.
14. Buen Abad Najar,C.F., Yosef,N. and Lareau,L.F. (2020) Coverage-dependent bias creates the appearance of binary splicing in single cells. *Elife*, **9**, e54603,
15. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
16. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.
17. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
18. Huang,Y. and Sanguinetti,G. (2017) BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol.*, **18**, 123.
19. Huang,Y. and Sanguinetti,G. (2021) BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. *Genome Biol.*, **22**, 251.
20. Song,Y., Botvinnik,O.B., Lovci,M.T., Kakaradov,B., Liu,P., Xu,J.L. and Yeo,G.W. (2017) Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol. Cell*, **67**, 148–161.
21. Hu,Y., Wang,K. and Li,M. (2020) Detecting differential alternative splicing events in scRNA-seq with or without Unique Molecular Identifiers. *PLoS Comput. Biol.*, **16**, e1007925.
22. Liu,S., Zhou,B., Wu,L., Sun,Y., Chen,J. and Liu,S. (2021) Single-cell differential splicing analysis reveals high heterogeneity of liver tumor-infiltrating T cells. *Sci. Rep.*, **11**, 5325.
23. Wen,W.X., Mead,A.J. and Thongjuea,S. (2020) VALERIE: visual-based inspection of alternative splicing events at single-cell resolution. *PLoS Comput. Biol.*, **16**, e1008195.
24. Picelli,S., Bjorklund,A.K., Faridani,O.R., Sagasser,S., Winberg,G. and Sandberg,R. (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, **10**, 1096–1098.
25. Shiozawa,Y., Malcovati,L., Galli,A., Sato-Otsubo,A., Kataoka,K., Sato,Y., Watatani,Y., Suzuki,H., Yoshizato,T., Yoshida,K. *et al.* (2018) Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. *Nat. Commun.*, **9**, 3649.
26. Smart,A.C., Margolis,C.A., Pimentel,H., He,M.X., Miao,D., Adeegbe,D., Fugmann,T., Wong,K.K. and Van Allen,E.M. (2018) Intron retention is a source of neoepitopes in cancer. *Nat. Biotechnol.*, **36**, 1056–1058.
27. Park,S.M., Ou,J., Chamberlain,L., Simone,T.M., Yang,H., Virbasius,C.M., Ali,A.M., Zhu,L.J., Mukherjee,S., Raza,A. *et al.* (2016) U2AF35(S34F) promotes transformation by directing aberrant ATG7 pre-mRNA 3′ end formation. *Mol. Cell*, **62**, 479–490.
28. Brooks,A.N., Choi,P.S., de Waal,L., Sharifnia,T., Imielinski,M., Saksena,G., Pedamallu,C.S., Sivachenko,A., Rosenberg,M., Chmielecki,J. *et al.* (2014) A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS One*, **9**, e87361.
29. Linker,S.M., Urban,L., Clark,S.J., Chhatriwala,M., Amatya,S., McCarthy,D.J., Ebersberger,I., Vallier,L., Reik,W., Stegle,O. *et al.* (2019) Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity. *Genome Biol.*, **20**, 30.
30. Ou,M., Zhao,M., Li,C., Tang,D., Xu,Y., Dai,W., Sui,W., Zhang,Y., Xiang,Z., Mo,C. *et al.* (2021) Single-cell sequencing reveals the potential oncogenic expression atlas of human iPSC-derived cardiomyocytes. *Biol. Open*, **10**, bio053348.
31. Falcao,A.M., van Bruggen,D., Marques,S., Meijer,M., Jakel,S., Agirre,E., Samudyata, Floriddia,E.M., Vanichkina,D.P., Ffrench-Constant,C. *et al.* (2018) Disease-specific oligodendrocyte lineage cells arise in multiple sclerosis. *Nat. Med.*, **24**, 1837–1844.
32. Wang,F., Tan,P., Zhang,P., Ren,Y., Zhou,J., Li,Y., Hou,S., Li,S., Zhang,L., Ma,Y. *et al.* (2022) Single-cell architecture and functional requirement of alternative splicing during hematopoietic stem cell formation. *Sci. Adv.*, **8**, eabg5369.
33. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, **17**, 10–12.
34. Veeneman,B.A., Shukla,S., Dhanasekaran,S.M., Chinnaiyan,A.M. and Nesvizhskii,A.I. (2016) Two-pass alignment improves novel splice junction quantification. *Bioinformatics*, **32**, 43–49.
35. 1000 Genome Project Data Processing Subgroup, Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
36. Zhou,F., Li,X., Wang,W., Zhu,P., Zhou,J., He,W., Ding,M., Xiong,F., Zheng,X., Li,Z. *et al.* (2016) Tracing haematopoietic stem cell formation at single-cell resolution. *Nature*, **533**, 487–492.
37. Kovaka,S., Zimin,A.V., Pertea,G.M., Razaghi,R., Salzberg,S.L. and Pertea,M. (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.*, **20**, 278.
38. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
39. Shen,S., Park,J.W., Lu,Z.X., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5593–E5601.
40. Broseus,L. and Ritchie,W. (2020) Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput. Struct. Biotechnol. J.*, **18**, 501–508.
41. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szczesniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.
42. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
43. Razali,N. and Yap,B.W. (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling test. *J. Stat. Model. Anal.*, **2**, 21–33.
44. Dowd,C. (2020) A new ECDF two-sample test statistic. arXiv doi: https://arxiv.org/abs/2007.01360, 02 July 2020, preprint: not peer reviewed.
45. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L. *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (N Y)*, **2**, 100141.
46. Yu,G., Wang,L.G., Han,Y. and He,Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
47. Velmeshev,D., Schirmer,L., Jung,D., Haeussler,M., Perez,Y., Mayer,S., Bhaduri,A., Goyal,N., Rowitch,D.H. and Kriegstein,A.R. (2019) Single-cell genomics identifies cell type-specific molecular changes in autism. *Science*, **364**, 685–689.

48. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
49. Wang,G., Wen,W.X., Mead,A.J., Roy,A., Psaila,B. and Thongjuea,S. (2022) Processing single-cell RNA-seq datasets using SingCellaR. *STAR Protoc.*, **3**, 101266.
50. Kaminow,B., Yunusov,D. and Dobin,A. (2021) STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv doi: https://doi.org/10.1101/2021.05.05.442755, 05 May 2021, preprint: not peer reviewed.
51. Efremova,M., Vento-Tormo,M., Teichmann,S.A. and Vento-Tormo,R. (2020) CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.*, **15**, 1484–1506.
52. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
53. Hagemann-Jensen,M., Ziegenhain,C., Chen,P., Ramskold,D., Hendriks,G.J., Larsson,A.J.M., Faridani,O.R. and Sandberg,R. (2020) Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.*, **38**, 708–714.
54. Westoby,J., Herrera,M.S., Ferguson-Smith,A.C. and Hemberg,M. (2018) Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. *Genome Biol.*, **19**, 191.
55. Schmitz,U., Pinello,N., Jia,F., Alasmari,S., Ritchie,W., Keightley,M.C., Shini,S., Lieschke,G.J., Wong,J.J. and Rasko,J.E.J. (2017) Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biol.*, **18**, 216.
56. Hatle,K.M., Gummadidala,P., Navasa,N., Bernardo,E., Dodge,J., Silverstrim,B., Fortner,K., Burg,E., Suratt,B.T., Hammer,J. *et al.* (2013) MCJ/DnaJC15, an endogenous mitochondrial repressor of the respiratory chain that controls metabolic alterations. *Mol. Cell. Biol.*, **33**, 2302–2314.
57. Bleichert,F. and Baserga,S.J. (2010) Ribonucleoprotein multimers and their functions. *Crit. Rev. Biochem. Mol. Biol.*, **45**, 331–350.
58. Kim,E.Y., Barefield,D.Y., Vo,A.H., Gacita,A.M., Schuster,E.J., Wyatt,E.J., Davis,J.L., Dong,B., Sun,C., Page,P. *et al.* (2019) Distinct pathological signatures in human cellular models of myotonic dystrophy subtypes. *JCI Insight*, **4**, e122686.
59. Loh,K.M., Ang,L.T., Zhang,J., Kumar,V., Ang,J., Auyeong,J.Q., Lee,K.L., Choo,S.H., Lim,C.Y., Nichane,M. *et al.* (2014) Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell*, **14**, 237–252.
60. Leone,S., Bar,D., Slabber,C.F., Dalcher,D. and Santoro,R. (2017) The RNA helicase DHX9 establishes nucleolar heterochromatin, and this activity is required for embryonic stem cell differentiation. *EMBO Rep.*, **18**, 1248–1262.
61. White,L.S., Soodgupta,D., Johnston,R.L., Magee,J.A. and Bednarski,J.J. (2018) Bclaf1 promotes maintenance and self-renewal of fetal hematopoietic stem cells. *Blood*, **132**, 1269.
62. Peng,X., Shen,X., Chen,X., Liang,R., Azares,A.R. and Liu,Y. (2015) Celf1 regulates cell cycle and is partially responsible for defective myoblast differentiation in myotonic dystrophy RNA toxicity. *Biochim. Biophys. Acta*, **1852**, 1490–1497.
63. Liu,Q., Wang,G., Lyu,Y., Bai,M., Jiapaer,Z., Jia,W., Han,T., Weng,R., Yang,Y., Yu,Y. *et al.* (2018) The miR-590/Acvr2a/Terf1 axis regulates telomere elongation and pluripotency of mouse iPSCs. *Stem Cell Rep.*, **11**, 88–101.
64. Sampath,P., Pritchard,D.K., Pabon,L., Reinecke,H., Schwartz,S.M., Morris,D.R. and Murry,C.E. (2008) A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell. Stem. Cell.*, **2**, 448–460.
65. Tang,A.D., Soulette,C.M., van Baren,M.J., Hart,K., Hrabeta-Robinson,E., Wu,C.J. and Brooks,A.N. (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.*, **11**, 1438.
66. Kalitsis,P., Earle,E., Fowler,K.J. and Choo,K.H. (2000) Bub3 gene disruption in mice reveals essential mitotic spindle checkpoint function during early embryogenesis. *Genes Dev.*, **14**, 2277–2282.
67. Mohamed,B.A., Barakat,A.Z., Zimmermann,W.H., Bittner,R.E., Muhlfeld,C., Hunlich,M., Engel,W., Maier,L.S. and Adham,I.M. (2012) Targeted disruption of Hspa4 gene leads to cardiac hypertrophy and fibrosis. *J. Mol. Cell Cardiol.*, **53**, 459–468.
68. Mathews,M.B. and Hershey,J.W. (2015) The translation factor eIF5A and human cancer. *Biochim. Biophys. Acta*, **1849**, 836–844.
69. O'Leary,M.N., Schreiber,K.H., Zhang,Y., Duc,A.C., Rao,S., Hale,J.S., Academia,E.C., Shah,S.R., Morton,J.F., Holstein,C.A. *et al.* (2013) The ribosomal protein Rpl22 controls ribosome composition by directly repressing expression of its own paralog, Rpl22l1. *PLos Genet.*, **9**, e1003708.
70. Fleckner,J., Zhang,M., Valcarcel,J. and Green,M.R. (1997) U2AF65 recruits a novel human DEAD box protein required for the U2 snRNP-branchpoint interaction. *Genes Dev.*, **11**, 1864–1872.
71. Shkreta,L., Delannoy,A., Salvetti,A. and Chabot,B. (2021) SRSF10: an atypical splicing regulator with critical roles in stress response, organ development, and viral replication. *RNA*, **27**, 1302–1317.
72. Jimenez-Vacas,J.M., Herrero-Aguayo,V., Montero-Hidalgo,A.J., Gomez-Gomez,E., Fuentes-Fayos,A.C., Leon-Gonzalez,A.J., Saez-Martinez,P., Alors-Perez,E., Pedraza-Arevalo,S., Gonzalez-Serrano,T. *et al.* (2020) Dysregulation of the splicing machinery is directly associated to aggressiveness of prostate cancer. *EBioMedicine*, **51**, 102547.
73. Park,I., Han,C., Jin,S., Lee,B., Choi,H., Kwon,J.T., Kim,D., Kim,J., Lifirsu,E., Park,W.J. *et al.* (2011) Myosin regulatory light chains are required to maintain the stability of myosin II and cellular integrity. *Biochem. J.*, **434**, 171–180.
74. Ridge,L.A., Mitchell,K., Al-Anbaki,A., Shaikh Qureshi,W.M., Stephen,L.A., Tenin,G., Lu,Y., Lupu,I.E., Clowes,C., Robertson,A. *et al.* (2017) Non-muscle myosin IIB (Myh10) is required for epicardial function and coronary vessel formation during mammalian development. *PLoS Genet.*, **13**, e1007068.
75. Jabs,E.W., Thomas,P.J., Bernstein,M., Coss,C., Ferreira,G.C. and Pedersen,P.L. (1994) Chromosomal localization of genes required for the terminal steps of oxidative metabolism: alpha and gamma subunits of ATP synthase and the phosphate carrier. *Hum. Genet.*, **93**, 600–602.
76. Sawai,A., Pfennig,S., Bulajic,M., Miller,A., Khodadadi-Jamayran,A., Mazzoni,E.O. and Dasen,J.S. (2022) PRC1 sustains the integrity of neural fate in the absence of PRC2 function. *Elife*, **11**, e72769.
77. Soglia,F., Mazzoni,M., Zappaterra,M., Di Nunzio,M., Babini,E., Bordini,M., Sirri,F., Clavenzani,P., Davoli,R. and Petracci,M. (2019) Distribution and expression of vimentin and desmin in Broiler Pectoralis major affected by the growth-related muscular abnormalities. *Front Physiol.*, **10**, 1581.
78. Zong,Y. and Li,X. (2021) Identification of causal genes of COVID-19 using the SMR method. *Front Genet.*, **12**, 690349.
79. Nordin,A., Larsson,E. and Holmberg,M. (2012) The defective splicing caused by the ISCU intron mutation in patients with myopathy with lactic acidosis is repressed by PTBP1 but can be derepressed by IGF2BP1. *Hum. Mutat.*, **33**, 467–470.
80. Jongbloed,R.J., Marcelis,C.L., Doevendans,P.A., Schmeitz-Mulkens,J.M., Van Dockum,W.G., Geraedts,J.P. and Smeets,H.J. (2003) Variable clinical manifestation of a novel missense mutation in the alpha-tropomyosin (TPM1) gene in familial hypertrophic cardiomyopathy. *J. Am. Coll. Cardiol.*, **41**, 981–986.
81. Citirak,G., Witting,N., Duno,M., Werlauff,U., Petri,H. and Vissing,J. (2014) Frequency and phenotype of patients carrying TPM2 and TPM3 gene mutations in a cohort of 94 patients with congenital myopathy. *Neuromuscul. Disord.*, **24**, 325–330.
82. Abrahams,B.S., Arking,D.E., Campbell,D.B., Mefford,H.C., Morrow,E.M., Weiss,L.A., Menashe,I., Wadkins,T., Banerjee-Basu,S. and Packer,A. (2013) SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism*, **4**, 36.
83. Liu,W. and Zhang,X. (2020) Single-cell alternative splicing analysis reveals dominance of single transcript variant. *Genomics*, **112**, 2418–2425.
84. Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.
85. Lee,S.C., North,K., Kim,E., Jang,E., Obeng,E., Lu,S.X., Liu,B., Inoue,D., Yoshimi,A., Ki,M. *et al.* (2018) Synthetic lethal and

convergent biological effects of cancer-associated spliceosomal gene mutations. *Cancer Cell*, **34**, 225–241.
86. Schischlik,F., Jager,R., Rosebrock,F., Hug,E., Schuster,M., Holly,R., Fuchs,E., Milosevic Feenstra,J.D., Bogner,E., Gisslinger,B. *et al.* (2019) Mutational landscape of the transcriptome offers putative targets for immunotherapy of myeloproliferative neoplasms. *Blood*, **134**, 199–210.
87. Qiu,X., Hill,A., Packer,J., Lin,D., Ma,Y.A. and Trapnell,C. (2017) Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods*, **14**, 309–315.
88. Ramskold,D., Luo,S., Wang,Y.C., Li,R., Deng,Q., Faridani,O.R., Daniels,G.A., Khrebtukova,I., Loring,J.F., Laurent,L.C. *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.