Igniting full-length isoform analysis in single-cell and spatial RNA-seq data with FLAMESv2

Changqing Wang[1,2*✉], Yair D.J. Prawer[3*✉], Oliver Voogd[1], Jakob Schuster[1], Ricardo De Paoli-Iseppi[3], Anran Li[3], Jeannette Hallab[3], Luyi Tian[4,5], Hongke Peng[4,5], Margaux David[6], Mei R.M. Du[1], Silvia Velasco[7], Maria G. Garone[7], Xueyi Dong[1,2], Kathleen Zeglinski[1], Chiara Pavan[8], Kevin C.L. Law[8], Kwaku Dad Abu-Bonsrah[8,9], Cameron P.J. Hunt[8], Clare L. Parish[8], Quentin Gouil[10,11], Rachel Thijssen[12], Nadia M. Davidson[1,2], Matthew E. Ritchie[1,2†✉], Michael B. Clark[3†✉], and Yupei You[1,2†✉]

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
[2]Department of Medical Biology, The University of Melbourne, Parkville, Victoria, Australia
[3]Department of Anatomy and Physiology, The University of Melbourne, Parkville, Victoria, Australia
[4]Guangzhou National Laboratory, Guangzhou, China
[5]GMU-GIBH Joint School of Life Sciences, Guangzhou Medical University, Guangzhou, China
[6]Laboratory for Neuroimmunology, Department of Neurosciences, Leuven Brain Institute, KU Leuven, Leuven, Belgium
[7]Murdoch Children's Research Institute, Stem Cell Biology, Parkville, Australia
[8]The Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, Australia
[9]Department of Paediatrics, The University of Melbourne, Parkville, Victoria, Australia
[10]Olivia Newton-John Cancer Research Institute, Heidelberg, Victoria, Australia
[11]School of Cancer Medicine, La Trobe University, Bundoora, Victoria, Australia
[12]Amsterdam University Medical Centre, Amsterdam, The Netherlands
[*]These authors contributed equally as co-first
[†]These authors contributed equally as co-last authors

Long-read single-cell RNA-sequencing enables the profiling of RNA isoform expression and alternative splicing at the single cell level. However, diverse single-cell technologies and sparse isoform data demand flexible, accurate, processing and analysis tools. Here we introduce FLAMESv2, a highly modular and protocol-agnostic R/Bioconductor package for long-read single-cell RNA-seq data processing. FLAMESv2 supports a wide range of single-cell and spatial experimental protocols, is highly configurable and scalable, allowing seamless multi-sample analysis and provides versatile visualisation and analysis outputs. We demonstrate FLAMESv2 compatibility with both droplet-based and combinatorial barcoding single-cell methods, as well as spatial transcriptomics workflows. Applying FLAMESv2 to *in vitro* differentiation of stem cells into neurons, we identify cell-types, differentiation trajectories, expression of annotated and novel isoforms and isoform expression diversity and heterogeneity within individual cells. FLAMESv2 provides a comprehensive, flexible approach to analysing long-read single-cell RNA-sequencing, unlocking this powerful methodology for RNA isoform characterisation.

single-cell | bulk | gene expression | long-read sequencing | ONT | PacBio

Correspondence: *wang.ch@wehi.edu.au, sefi.prawer@unimelb.edu.au, mritchie@wehi.edu.au, michael.clark@unimelb.edu.au, you.yu@wehi.edu.au*

RNA-sequencing (RNA-seq) has transformed the molecular investigation of genes, cells, tissues, diseases and genetic variants. More recently, single-cell RNA-sequencing (scRNA-seq), has become a powerful and popular method to measure gene expression within individual cells and characterise cellular heterogeneity, identify novel cell types and map developmental trajectories [1, 2]. Rapid advances in scRNA-seq technology have exponentially increased the number of cells that can be profiled, with millions of cells now able to be examined within a study [3–5]. Such progress has driven the proliferation of scRNA-seq, including the generation of large-scale atlases of healthy, developing and diseased tissues [3, 5]. Building on scRNA-seq, spatial transcriptomics methodologies aim to profile gene expression at high resolution, enabling cellular expression profiles to be mapped within tissues. This can provide deeper insights into tissue architecture, cellular expression micro-environments and cell-cell communication [6–9]. However, the restriction of most single-cell and spatial methods to gene-level counts prevents the full-elucidation of the transcriptome.

Beyond gene-level expression, genes commonly express multiple RNA products (isoforms) through processes such as alternative transcriptional initiation, splicing and transcriptional termination. Isoform expression varies between cell types and more than >90% of human genes are reported to undergo alternative splicing [10]. Alternative isoforms from a gene can have modified, or even opposing, functions and expression switches between isoforms play key roles in development, disease and cell-type specific functions [11–13]. However, while standard, short-read sequencing based single-cell and spatial transcriptomic techniques accurately quantify gene expression, they only sequence the 5' or 3' ends of expressed RNAs, preventing the comprehensive identification and quantification of alternative RNA isoforms. In contrast, long-read (LR) methodologies can sequence entire RNA transcripts in single reads, enabling the accurate profiling of the different RNA isoforms expressed by genes.

By combining long-read sequencing with scRNA-seq [14–19] or spatial transcriptomics [20, 21], researchers can now achieve isoform-resolved analysis at the single-cell or spatial resolution. This rapidly evolving field has spurred the development of numerous innovative sequencing protocols. However, the analysis landscape for long-read single-cell and spatial data remains highly fragmented. Many bioinformatic tools are tied to specific protocols and data types [18, 19, 22–

24], while even the most flexible existing tools are tailored to a limited list of known protocols [25–27], limiting reproducibility, analytical flexibility, application to new sequencing techniques and comparisons between studies that used different methodologies. In addition, isoform-level single-cell data is even more sparse than at the gene level, necessitating the use of accurate isoform-quantification tools specifically designed for long-read data, which have only recently become available [28]. These limitations, together with the increasing level of interest in generating these types of data, has created the need for a modular, protocol-agnostic framework that can integrate diverse data types and analysis strategies.

Here we introduce FLAMESv2, an enhanced and expanded R/Bioconductor version of our previous software [17] that offers improved flexibility, performance, speed and usability. FLAMESv2 is compatible with a wide range of data-types, including both single-cell and spatial long-read protocols. It can be applied to data generated using long reads only, or in combination with short reads. FLAMESv2 also provides convenient in-built functions for data visualization and analysis. We demonstrate the performance of FLAMESv2 on *in vitro* stem-cell to neuron differentiation datasets, identifying cell-type specific expression of known and novel RNA isoforms and characterising isoform diversity within individual cells. FLAMESv2 transforms a fragmented analysis landscape into a modular, unified pipeline for isoform-level single-cell and spatial analysis.

Results

FLAMESv2: Methodology Overview

The FLAMESv2 pipeline identifies and quantifies the expression of genes and isoforms from single-cell and spatial sequencing experiments. FLAMESv2 performs six major steps: barcode demultiplexing, genome alignment, gene quantification, isoform identification, transcriptome alignment and transcript quantification (Figure 1). Briefly, sequenced reads are assigned to individual cells (and to spatial coordinates, if applicable), and mapped to the reference genome. Next, gene quantification is performed which includes UMI deduplication, a step to collapse duplicate molecules into a single read for accurate quantification. Next, FLAMESv2 identifies isoforms and generates an experiment-specific transcriptome reference, to which the deduplicated reads are re-aligned. Finally, transcript quantification is performed to measure the expression level of each isoform in each cell. See Methods for more detailed explanation of each step.

FLAMESv2 provides high performance pipelines and extensive visualization options for diverse data types

FLAMESv2 supports a diverse range of input data types. To maintain compatibility with the growing range of modalities amenable to long-read sequencing, FLAMESv2 supports inputs from a wide range of protocols (Figure 1, Table S1), including widely used single-cell protocols (e.g., 10x 3' and 5' kits, LR-Split-seq [18]), spatial transcriptomics protocols (10x Visium), and bulk protocols, including mini-bulk protocols using well-barcodes (e.g., TIRE-seq [29]). The different protocols yield reads with different structures, hence instead of designing specific support for individual protocols, FLAMESv2 allows users to describe their read structure via the flexiplex [30] package. Another major advance in FLAMESv2's input flexibility is its ability to operate with or without matched short-read data. This makes the pipeline broadly accessible to researchers using only long-read sequencing, while still supporting the use of matched short-read data if available. This flexibility is achieved via the inclusion of the state-of-the art long-read demultiplexing packages flexiplex [30] and BLAZE [31] (Table 1).

FLAMESv2 offers configurable, modularised pipelines with incorporated tools. The six steps of the FLAMESv2 pipeline are highly modular. Each step produces and accepts standard file formats, ensuring interoperability with alternative software and enabling flexible integration into diverse analysis workflows. This enables users to choose between built-in algorithms and other supported tools using the FLAMESv2 configuration file. In addition to BLAZE and flexiplex, field-leading tools including bambu [32] and oarfish [28] can be configured for isoform identification and quantification respectively. Furthermore, many steps of the pipeline can now be selectively included or excluded, offering flexibility for different analysis needs. For example, excluding the isoform identification step will result in quantification of reference transcripts only. The modular nature of FLAMESv2 also permits re-configuring and resuming a completed or crashed pipeline, avoiding the need to re-run jobs from scratch. To facilitate deployment on high performance computing clusters, steps can be submitted as jobs through the cluster's job scheduler, allowing specific resource configuration for individual jobs to optimize resource usage.

Support for multi-sample study designs. The FLAMESv2 multi-sample pipeline ensures uniform processing of data from each sample, including consistent naming and annotation of novel isoforms across samples.

Speed enhancements. Typical long-read scRNA-seq experiments now produce tens of millions to billions of reads, emphasising the importance of fast and efficient software. We have optimised FLAMESv2 for speed. We illustrate this by processing four single-cell and single-nuclei samples from the LongBench study [33], totalling more than 300 GB of compressed FASTQ data, within 24 hours (Figure S1). Other core functions have been re-implemented with significantly improved speed, for example, a 10-fold improvement in single-cell mutation quantification (Figure S2).

Enhanced data visualization options. To allow convenient data visualization, we have added functions to create publication-ready figures. These span analysis tasks from quality control to spatial isoform expression plots (Figure 1). For example, users can process and visualize isoform expression and genotypes at specific loci on a UMAP. We demonstrated this function using peripheral blood mononuclear cells (PBMCs) from a chronic lymphocytic leukemia (CLL) patient [17], highlighting differential transcript usage

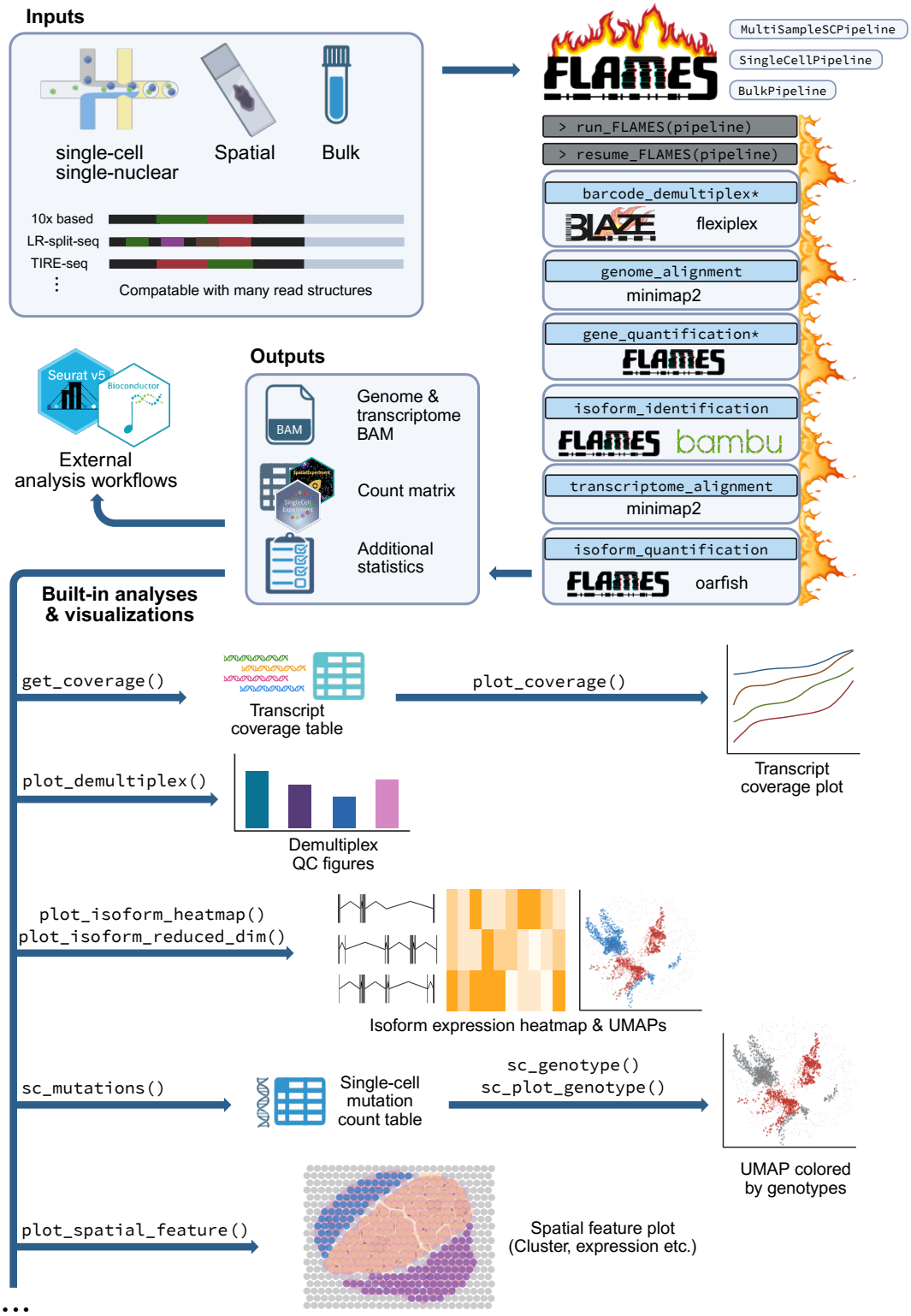

Fig. 1. `FLAMESv2` pipeline and visualization functions `FLAMESv2` supports inputs from a wide range of experimental protocols, including single-cell (sc), single-nuclear (sn), spatial, and bulk long-read sequencing. The workflow accommodates diverse read structures and can process either individual or multiple samples. Users can perform analyses through the single-cell pipeline which is designed to handle sc, sn and spatial experiments, or through the dedicated bulk pipeline (asterisks (*) indicate steps not included in the bulk pipeline). The pipeline consists of six main stages (barcode demultiplexing, genome alignment, gene quantification, isoform identification, transcriptome alignment and isoform quantification). Outputs include genome and transcriptome BAM files, gene and isoform count matrices and summary statistics. FLAMESv2 allows flexible tool selection at key steps, enabling customization for different experimental needs. In addition to core processing, FLAMESv2 provides built-in quality control and visualization functions for data exploration and publication-ready output, including coverage plots, isoform heatmaps, genotype-annotated UMAPs, and spatial feature maps.

of *RPS24* between healthy and cancer cell populations (Figure 2A) and the presence of the venetoclax resistant *BCL2 G101V* allele in a subset of cells (Figure 2B).

Measuring isoform diversity at single cell resolution. FLAMESv2 includes functionality to measure isoform diversity in individual cells. The find_diversity function provides a direct measure of whether intra-cellular and inter-cellular expression is dominated by a single isoform or distributed across multiple isoforms of a gene. We demonstrate the use and biological implications of this metric in Section "Profiling within-cell diversity using Shannon's entropy."

Demonstrating FLAMESv2 compatibility with diverse input data types

Combinatorial barcoding-based long-read scRNA-seq: To demonstrate flexible input support beyond 10x based protocols, we reprocessed LR-split-seq data [18], using only the long reads (Figure S3). Four cell clusters from three samples were expected (Figure S4) [18]. As shown in the UMAP derived from FLAMESv2's gene count matrix (Figure 2C), FLAMESv2 retrieved more cells than the LR-split-seq custom pipeline (including all previously identified barcodes), and the four expected clusters showed clear separation.

Spatial data: FLAMESv2 also supports 10x Visium long-read spatial data, where reads have a similar structure to 10x Chromium 3' technology. Using a FLAMESv2 helper function to attach spatial coordinates to the single-cell results, long-read spatial datasets can be analysed and visualised. To demonstrate this functionality, mouse olfactory bulb spatial long-read transcriptomic data from Lebrigand *et al.* (2023) [20] was re-processed and clustered. We successfully recapitulated the previously observed spatial distribution of different cell types and spatial-specific isoform switching of *Plp1* (Figure 2D-F).

FLAMESv2 enables benchmarking of sequencing protocols with a unified pipeline.

The extensive variety of inputs, quality metrics and visualisations enables convenient technology and protocol benchmarking with FLAMESv2. We showcase this using the LongBench data, which includes Oxford Nanopore (ONT) and PacBio single-cell and single-nuclei long-read transcriptomics data from a mixture of eight lung cancer cell lines [33]. Using FLAMESv2, we found the 10x 3' single-cell kit produced more (90%) usable (demultiplexed) reads than the single-nuclei kit (64%), as shown in Figure 2G. Of the demultiplexed reads, we observed a higher percentage with an identified TSO adapter at the 5' end in PacBio Kinnex libraries (99%) than ONT libraries (92%), indicating more PacBio reads contain the 5' end of the sequenced cDNA (Figure S5). We then compared read coverage across transcripts of different lengths for the different sequencing platforms, revealing decreased coverage and stronger 3' bias as transcript length increased (Figure 2H). Lastly, we compared single-cell and single-nuclei isoform quantification to matched bulk RNA-seq samples. We re-processed all LongBench ONT bulk, single-cell and single-nuclei long-read samples and performed pseudo-bulk aggregation on the single-cell and single-nuclei libraries for each cell line. We then calculated their transcript count correlation with the bulk libraries. We found single-cell libraries had higher Pearson correlations with bulk data (0.72 on average) than single-nuclei (0.59 on average, Figure 2I).

FLAMESv2 enhances isoform discovery and quantification accuracy in long-read single-cell data

To demonstrate the improved capabilities of FLAMESv2, we applied the single-sample pipeline to long-read scRNA-seq data generated from cortical projection neurons differentiated *in vitro* from induced pluripotent stem cells (iPSCs). We collected cells at day 55 of the differentiation and sequenced the sample using one PromethION flow cell, generating approximately 46 million pass reads across ~340 cells - (~130,000 reads per cell) (Table 2). For comparison, we also processed the dataset using FLAMESv1.

FLAMESv2 showed marked improvements over the original version across key metrics, identifying more unique genes (29,013 vs 5,578) and isoforms (93,721 vs 15,430), assigned a higher fraction of reads to cells (77% vs 71%) and used nearly three times more reads for isoform quantification (14.6 vs 5.1 million) (Table 2).

To assess the quality of isoform discovery, we analysed structural classifications using SQANTI [34, 35]. Over 95% of isoforms detected by FLAMESv2 were classified as full-splice matches (FSM) to known isoforms. In contrast, FLAMESv1 reported a much higher proportion of novel isoforms (45%), with only 27% classified as FSM (Figure S6). This inflated novel fraction likely reflects a higher false positive rate, driven at least in part by internal priming events that FLAMESv1 assigns as novel due to the lack of dedicated solutions to address internal priming events. Together, these improvements reflect enhanced quantification accuracy and transcript classification in FLAMESv2.

FLAMESv2 enables seamless downstream analyses

FLAMESv2 outputs are compatible with widely used downstream analysis tools, enabling flexible and efficient exploration of long-read RNA-seq data. By supporting formats such Seurat [36], Scanpy [37] and SingleCellExperiment (SCE) [38] objects, FLAMESv2 readily integrates into established workflows for gene and isoform-level expression profiling, cell-type annotation and differential transcript usage analysis. We illustrated this flexibility by analysing the day 55 neuronal culture using Seurat and exploring isoform-level expression across major neuronal populations. UMAP projections based on both gene- and isoform-level expression clearly resolved the three major cell types expected at this stage of neural differentiation: radial glia progenitors, deep-layer projection (excitatory) neurons and GABAergic neurons [39–41] (Figure 3A). The similarity between gene- and isoform-level clustering indicates that both quantification approaches effectively capture cell-type-specific expression patterns.

We confirmed cell identities using module scores for canonical markers (Figure 3B) and identified differ-

Feature	FLAMESv1	FLAMESv2
Input support	10x 3' protocol only	Extended protocol support
Long read only support	✗	✓
Putative barcode calling	✗ (matching short-read required)	✓ (via BLAZE or flexiplex)
Multi-sample support	✗	✓
Gene-level quantification	✗	✓
Novel isoform discovery	Built-in method only	Built-in method / Bambu
Isoform-level quantification	Built-in counting method only	Built-in counting / oarfish EM estimation
Modular pipeline structure	✗	✓
Visualizations	✗	✓

Table 1. Comparison of FLAMESv1 and FLAMESv2. ✓ indicates supported features; ✗ indicates not supported. FLAMESv2 incorporates support for isoform-level quantification, modular architecture, and multi-sample workflows, enhancing scalability and interpretability of long-read scRNA-seq data.

entially expressed genes and isoforms with Seurat's `FindAllMarkers`, followed by GO and KEGG pathway enrichment analysis (Figure 3C). These analyses confirmed cluster identities and uncovered biologically relevant programs, including mitotic and chromosome replication pathways in the radial glia progenitors, as well as neuronal development and glutamatergic or GABAergic synapse formation in neuronal populations. Once cell type classification was confirmed, we explored gene and isoform diversity within each population. While the number of genes and isoforms expressed was broadly concordant across cell types, we observed fewer genes, isoforms and novel isoforms in the GABAergic population (~3,000, 11,000 and 750 respectively) likely a result of smaller cell numbers and reduced sensitivity to detect low-abundance features (Figure 3D). Of the 2,855 novel isoforms detected, 1,773 were ubiquitously expressed across cell types, while 176 were cell-type specific (Sup Figure S7). Over 60% of genes expressed two or more isoforms (Figure 3E), and some, such as *MIR9-1HG* and *SNHG1*, exhibited extreme complexity with over 75 unique isoforms reported (Figure 3F).

Next we explored canonical vs non-canonical isoform expression across cell types. Interestingly, radial glial cells exhibited a 5% increase in canonical isoform expression compared to neuronal populations, which may suggest that more mature neuronal cell types exhibit more diverse usage of gene isoforms (Figure 3G). Finally, we assessed the distribution of novel isoform expression across individual cells. All cells in our dataset expressed at least one novel isoform, with a near-normal distribution centred around a median of 156 novel isoforms per cell (Figure 3H). This highlights the cell-to-cell heterogeneity in novel isoform expression and the ability of `FLAMESv2` to examine lowly expressed and rare transcriptomics features. Together, these results highlight how `FLAMESv2` improves transcript discovery and quantification and provides a flexible framework for analysis.

Profiling the isoform landscape during neuronal differentiation with the `FLAMESv2` multi-sample pipeline

LR scRNA-seq tools need to be able to process growing numbers of samples and cells, enabling the analysis of RNA isoforms over development or in disease. Developing and mature brain tissues exhibit extensive alternative isoform usage [42–44], making models of developing brain well suited to demonstrating the performance of `FLAMESv2` in investigating isoform trajectories across multiple samples and time points.

We explored isoform expression during neurogenesis using a well-established protocol for differentiating iPSCs into neurons over an 80 day time-course [39–41] (Figure S8). We collected eight samples for single-cell processing spanning 4-time points: stem cells (D0), Day 25 (D25), Day 55 (D55) and Day 80 (D80). These time points were selected as they encompass key developmental stages, including neural induction and neuronal progenitor expansion (D25), early neuron development (D55) and neuron maturation and upper layer neuron emergence (D80) [41]. We dissociated the cells and processed them using the 10x 3' gene expression assay followed by sequencing on the ONT PromethION sequencer (Figure 4A), We generated ~949.3M pass reads (q score > 10) that were used for downstream analysis (Table S2)

We processed the reads using the `FLAMESv2` multi-sample pipeline, which handled all eight samples and the large number of sequencing reads in a single run. Performing isoform identification across all samples in a single pass provides greater read depth to support both known and novel transcripts, enabling improved isoform identification and quantification. In addition, multi-sample mode enables consistent isoform quantification and naming of novel isoforms across all samples, making it possible to investigate isoform expression dynamics across the time series. `FLAMESv2` identified 12,958 cells across the eight samples, 9,648 of which were retained after preprocessing in `Seurat` (see methods). We identified 1,705 to 3,099 high quality cells from each time point (Table S2, Figure S9). Across the dataset, cells contained a median of 33,993 counts assigned at the gene level and 21,272 at the isoform level. We integrated all samples using `Harmony` [45], yielding coherent batch-corrected profiles across each sample and time point.

We identified the cell types presented in the integrated dataset. As expected, we identified a large population of stem cells at Day 0 and radial glial progenitors at Day 25, including a population of outer radial glia cells expressing *HOPX*, *FAM107A* and *PTPRZ1* (Figure 4B, Figure S10).

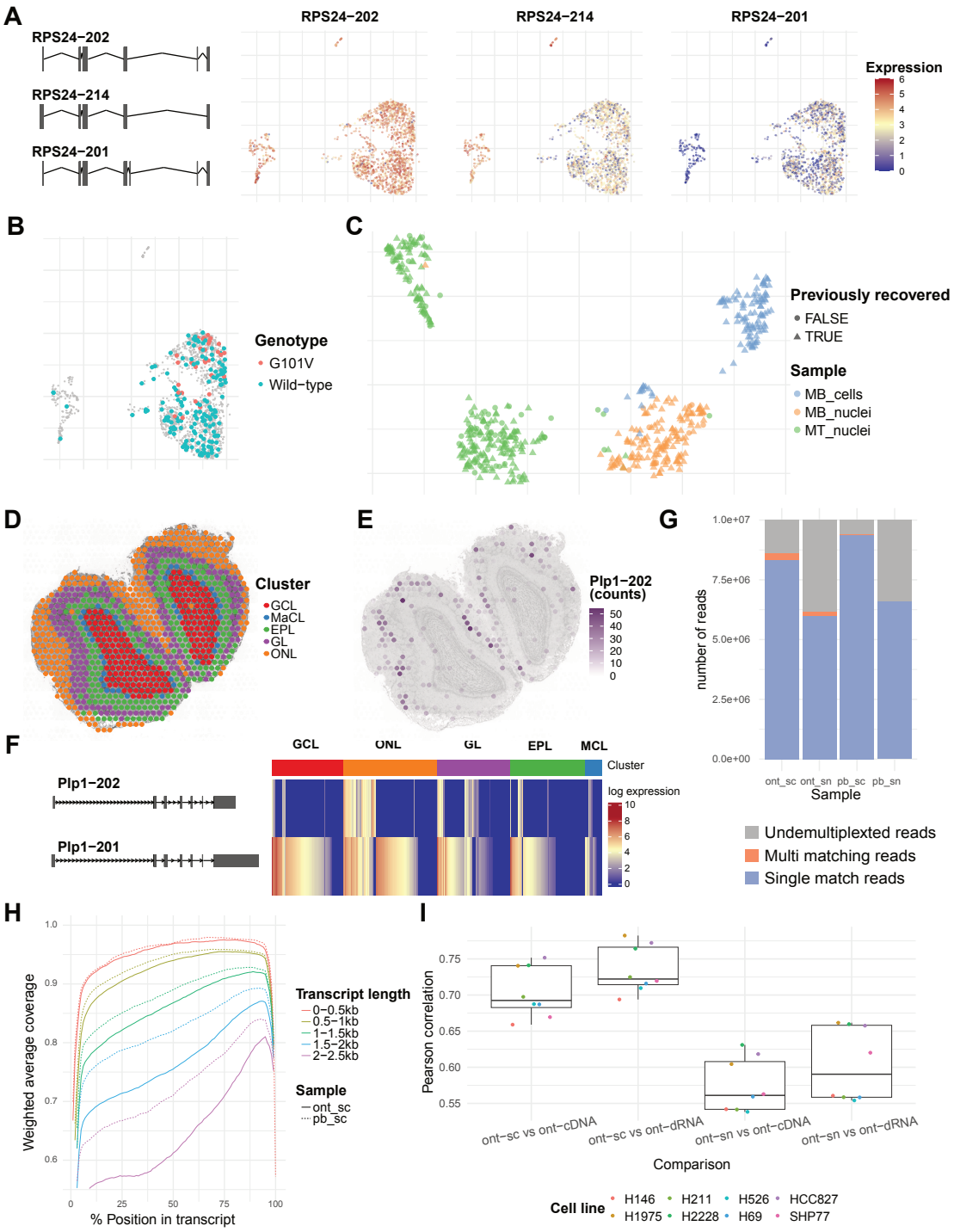

Fig. 2. Analyzing existing datasets with `FLAMESv2` (**A**) UMAPs of CLL2 data from Tian *et al.* showing expression of *RPS24* isoforms created with `FLAMESv2` using UMAP embeddings from Tian *et al.* [17] (**B**) UMAP of CLL2 data coloured by *BCL2 G101V* allele created with `FLAMESv2` using UMAP embeddings from Tian *et al.* (**C**) UMAP of LR-split-seq data from Rebboah *et al.* from re-processing with `FLAMESv2`, coloured by sample name from matching short-read barcodes [18]. (**D-F**) Re-analysis and visualization of Lebrigand *et al.* (2023) [20] Visium Nanopore data with `FLAMESv2`. GCL: Granule Cell Layer; MaCL: Mitral Cell Layer; EPL: Outer Plexiform Layer; GL: Glomerular Layer; ONL: Olfactory Nerve Layer. (**D**) Spatial plot of annotated clusters. (**E**) Spatial plot of *Plp1-202*. (**F**) Heatmap of *Plp1* isoform expression. (**G-I**) `FLAMESv2` analyses of LongBench data. Single-cell (sc), single-nuclei (sn), Nanopore (ont), PacBio (pb). (**G**) `FLAMESv2` demultiplexing results, 10 million reads were used for each sample. (**H**) Average sequencing coverage along transcript body (**I**) Pearson correlation of gene expression counts between ONT bulk (direct RNA, PCR-cDNA) and ONT pseudobulk (single nucleus and single cell) data for each cell line.

Next, we focused on excitatory glutamatergic neurons corresponding to the deep and middle cortical layers (layers VI-IV). Following the approach in Figure 3B, we surveyed canonical deep layer glutamatergic markers (*EOMES*, *TBR1*, *SLC17A7/VGLUT1* and *SLC17A6/VGLUT2*). Only a small cluster of 121 cells expressed these marker genes (Figure S10). Instead, expression of known inhibitory neuron markers such as *DLX1*, *DLX2*, *GAD1* and *GAD2*, showed high expression across Day 55 and Day 80 samples (Figure 4C). These data indicate that the differentiation predominately

Metric	FLAMESv1	FLAMESv2
Cells identified	337	339
Number of unique genes	5,578	29,013
Number of isoforms	15,430	93,721
Median genes per cell	2,390	7,182
Median isoforms per cell	3,334	10,812
Median reads per cell	12,922	31,923
Number of novel isoforms*	5,834	2,855
Incomplete transcripts (ISM)	236	0
Usable reads (reads assigned to a cell)	71%	77%
Reads used to quantify isoforms**	5,135,544	14,586,085

Table 2. Comparison of FLAMES versions across key metrics. Genes identified include protein coding and non-coding. * Novel isoforms not classified as FSM or ISM by SQANTI3. ** Summed counts in the isoform counts file. Reads used to quantify isoforms have been demultiplexed and deduplicated. Novel Bambu genes have been excluded from the `FLAMESv2` metrics.

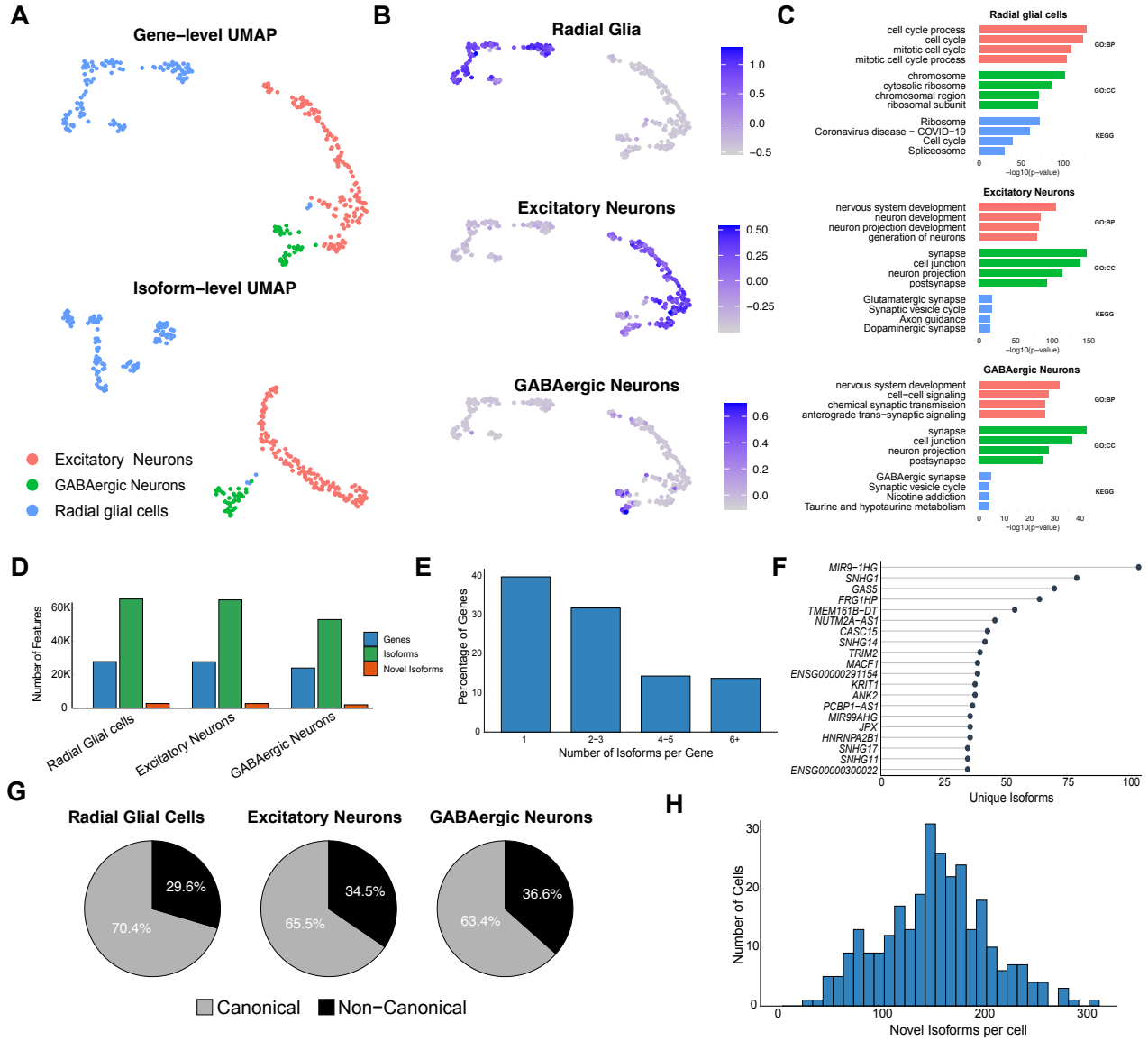

Fig. 3. FLAMESv2 enables gene and isoform level analysis, capturing both annotated and novel isoforms across cell types. (**A**) UMAPs generated from gene-level (top) and isoform-level (bottom) expression matrices reveal three major populations: radial glial cells, excitatory neurons, and GABAergic neurons. (**B**) UMAPS overlaid with module score show relative enrichment of the selected gene set per cell and confirm cell type identities. Positive values indicate higher-than-expected expression of the gene set, while negative values indicate lower-than-expected expression. Canonical marker genes used for each gene set: radial glia (*VIM, HES1, HES5, SOX2, SLC1A3, FABP7, HOPX, NOTCH1*), excitatory neurons (*SLC17A7, SLC17A6, CAMK2A, GRIA2, TBR1, FEZF2, BCL11B, SYN1, NRGN*), and GABAergic neurons (*GAD1, GAD2, SST, PVALB, DLX1, DLX2, NPY, TAC1*). **C**) Gene ontology enrichment analysis of cell-type specific genes, highlighting biological processes enriched in radial glia (cell cycle, proliferation), excitatory neurons (synapse organization, glutamatergic signalling), and GABAergic neurons (GABAergic synapse, inhibitory signalling). (**D**) Bar plots showing the number of genes, isoforms, and novel isoforms detected across the three major cell types. (**E**) Number of identified isoforms per gene. (**F**) Top 20 genes ranked by number of expressed isoforms. (**G**) Proportion of canonical versus non-canonical isoform expression across cell types. Novel isoforms are included in the non-canonical category. (**H**) Histogram showing the distribution of novel isoforms per cell, highlighting that all cells have at least one novel isoform.

produced a population of inhibitory neurons. Previous studies have shown variability in differentiation outcomes to be a common occurrence, including the generation of more ventral or caudal neurons such as inhibitory interneurons [46, 47]. Previously, enrichment techniques such as RNA Capture-seq have been required to properly identify such neuronal subtypes, however these were clearly defined using LR RNA-seq and `FLAMESv2` [47].

A clear progression of cell maturation was observed along the time course. At Day 0, the cultures are dominated by stem cells. By Day 25, radial glial and progenitor populations emerge (98% of cells), alongside a small proportion of neurons (2%). At Days 55 and 80, the vast majority of cells adopt neuronal identities, with a minor population of progenitor and radial glia cells remaining (2% of cells) (Figure 4D, Figure S11). We validated these dynamics by reconstructing developmental trajectories using `Monocle3` [48]. Pseudotime analysis recapitulated the anticipated progression from stem cells through radial glial cells to more mature neuronal states, consistent with known neuronal differentiation (Figure 4D). Integrating multiple samples and mapping developmental trajectories enabled us to examine both genes and isoforms whose expression changes correlate with pseudotime. We selected the 5,000 most variable genes and 10,000 most variable isoforms, then computed their correlation with pseudotime using the `Monocle3:graphtest` function. This revealed 557 genes and 216 isoforms whose expression is significantly associated with developmental progression. Notably, all of these features were also detected as differentially expressed, suggesting they capture not only discrete expression changes but also gradual, continuous shifts within the developing populations. Figure 4D highlights six isoforms whose expression was most closely correlated with pseudotime. Such patterns are particularly informative for uncovering subtle regulatory programs underlying differentiation, beyond what conventional differential expression (DE) analysis alone would reveal.

Next, we explored changes in isoform usage between cell types or states, commonly referred to as isoform switching. Using `IsoformSwitchAnalyzeR` and `DTUrtle` [49, 50], we identified 18,490, and 15,794 isoforms exhibiting significant switching between time points respectively. Of these, 11,163 were common across both methods (Table S3). Both methods identified a switching event in the gene *PKM* (*ENSG00000067225*), whose gene level expression is ubiquitous, yet shows distinct cell-type–specific isoform regulation (FDR < 0.01) (Figure 4E). Within our dataset *PKM* expresses 16 isoforms, with global expression dominated by the canonical transcript (*ENST00000335181*, 51.2% of expression) and a novel isoform (*BambuTx13041*, 31.6% of expression). The novel *PKM* isoform lacks canonical exon 9 (chr15:72202454-72202620) encoding part of the second pyruvate-kinase domain, suggesting functional divergence from the canonical transcript (Figure 4E). Isoform-level expression of *PKM* follows a clear, developmental-specific pattern in which progenitor populations are enriched for the canonical isoform, whereas neural populations express the novel *BambuTx13041* isoform (Figure 4E and S12). These results demonstrate that the `FLAMESv2`, multi-sample mode provides a robust framework for probing isoform-level regulation during complex processes, underscoring the package's versatility and broad applicability.

Profiling within cell diversity using Shannon's entropy

Single-cell studies have traditionally focused on differences in gene or isoform expression between distinct populations (inter-cell type expression differences). Here, we leveraged our long-read data to ask two complementary questions: do individual cells express multiple isoforms of a gene (intra-cellular isoform diversity)? And do individual cells of the same cell type express the same set of isoforms in similar proportions, or is there appreciable isoform-level heterogeneity even among phenotypically similar cells (inter-cellular isoform heterogeneity)?

To address this, we developed the `find_diversity` function in the `FLAMESv2` package. This function calculates normalised Shannon's entropy for each gene in each cell, providing a direct measure of whether expression is dominated by a single isoform or distributed across multiple isoforms (Figure **5**A). Because single-cell data are inherently sparse, filtering is required to ensure reliable entropy estimates. Full details of the strategy are provided in the Methods section. Importantly, all parameters are user-adjustable, allowing the function to be applied across datasets with varying sequencing depth and cell numbers.

We applied `find_diversity` to 117 radial glial cells from our deeply sequenced Day 55 dataset (median depth: 28,540 isoform counts per cell). This approach enabled robust estimation of entropy across 435 genes. A majority of genes displayed moderate-to-high isoform diversity (24% low, 30% medium, 46% high; Figure **5**B), showing that most genes we tested are expressing multiple isoforms within individual cells (intra-cellular isoform diversity is common). Figure **5**C highlights genes spanning the spectrum of isoform diversity. Low-diversity genes such as *VIM*, which had a median entropy of 0, had nearly all expression from a single isoform (Figure **5**D). At the other extreme, highly diverse genes such as *RPL19*, (median entropy = 0.98) expressed three isoforms in all cells (Figure **5**E).

An advantage of per-cell entropy measurements is the ability to also capture heterogeneity in isoform proportions between cells of the same population. For example, *RPL19* commonly expressed isoforms in similar proportions across cells, reflecting a stable multi-isoform expression pattern (Figure **5**E). In contrast, *MAP1B* had a high median entropy (0.91) but displayed variability between individual cells. While four isoforms consistently contributed to its overall expression, their proportions differed greatly between cells, revealing inter-cellular heterogeneity in isoform composition within a cell type (Figure **5**F).

By providing gene-level diversity scores at single-cell resolution, the `find_diversity` function adds a powerful capability to the `FLAMESv2` package. This novel and complementary approach enables the investigation of isoform ex-

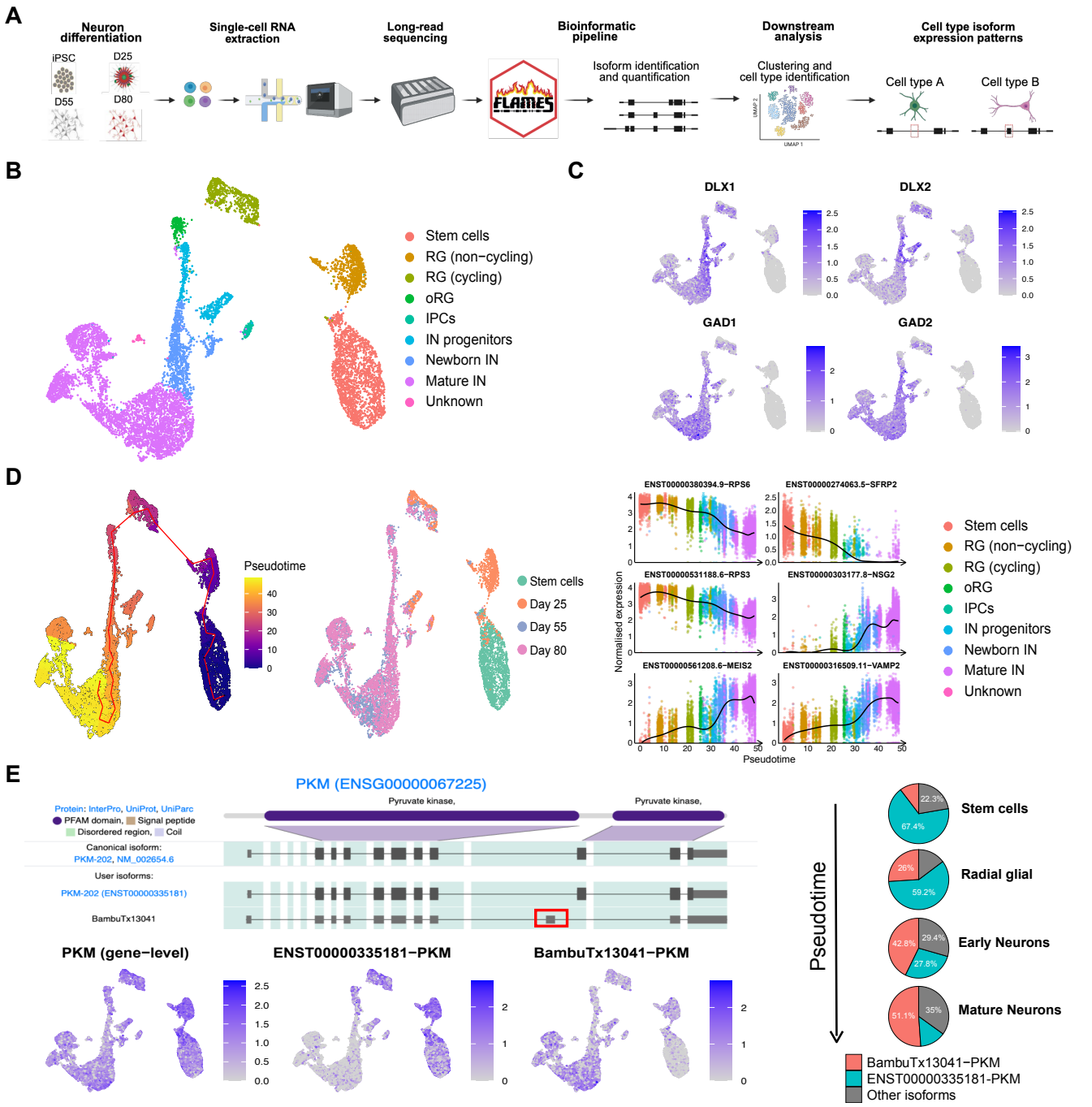

Fig. 4. Multi sample analysis workflow in using `FLAMESv2`. **(A)** Experimental overview. Single-cell preparation and long-read scRNA-seq workflow using FLAMESv2. **(B)** Gene level UMAP of 8 integrated samples with annotated cell types (**RG** = Radial glia, **oRG** = Outer radial glia, **IPCs** = Intermediate progenitor cells **IN** = Inhibitory neurons). **(C)** UMAPs showing inhibitory marker expression across integrated samples. **(D)** Left: UMAP coloured by pseudotime; middle: UMAP coloured by collection day showing concordance between sampling time and inferred maturation; right: representative isoforms whose expression is strongly associated with pseudotime (three down-regulated and three up-regulated), with points coloured by cell type to illustrate the relationship between cell maturity and isoform abundance. y axis units are normalised expression values generated from `Seurat` function `LogNormalize`. **(E)** Top: IsoVis [51] image highlighting two *PKM* isoforms (canonical and a novel isoform; red box marks an exon absent from the canonical transcript). Known ORFs are denoted via thicker boxes. Bottom: UMAPs of gene-level *PKM* and isoform-specific expression demonstrate cell-type-specific isoform usage across progenitor and neuronal populations. Right: pie charts summarizing isoform proportions across broad cell classes, showing enrichment of the novel *PKM* isoform with increasing maturation.

pression patterns between cells and states, and highlights genes with consistently high or variable diversity that may warrant further investigation.

Discussion

Long-read scRNA-seq enables isoform-resolved analysis, revealing isoform usage across complex systems and its roles in development and disease. However, the analysis ecosystem remains fragmented. Many pipelines are tightly coupled

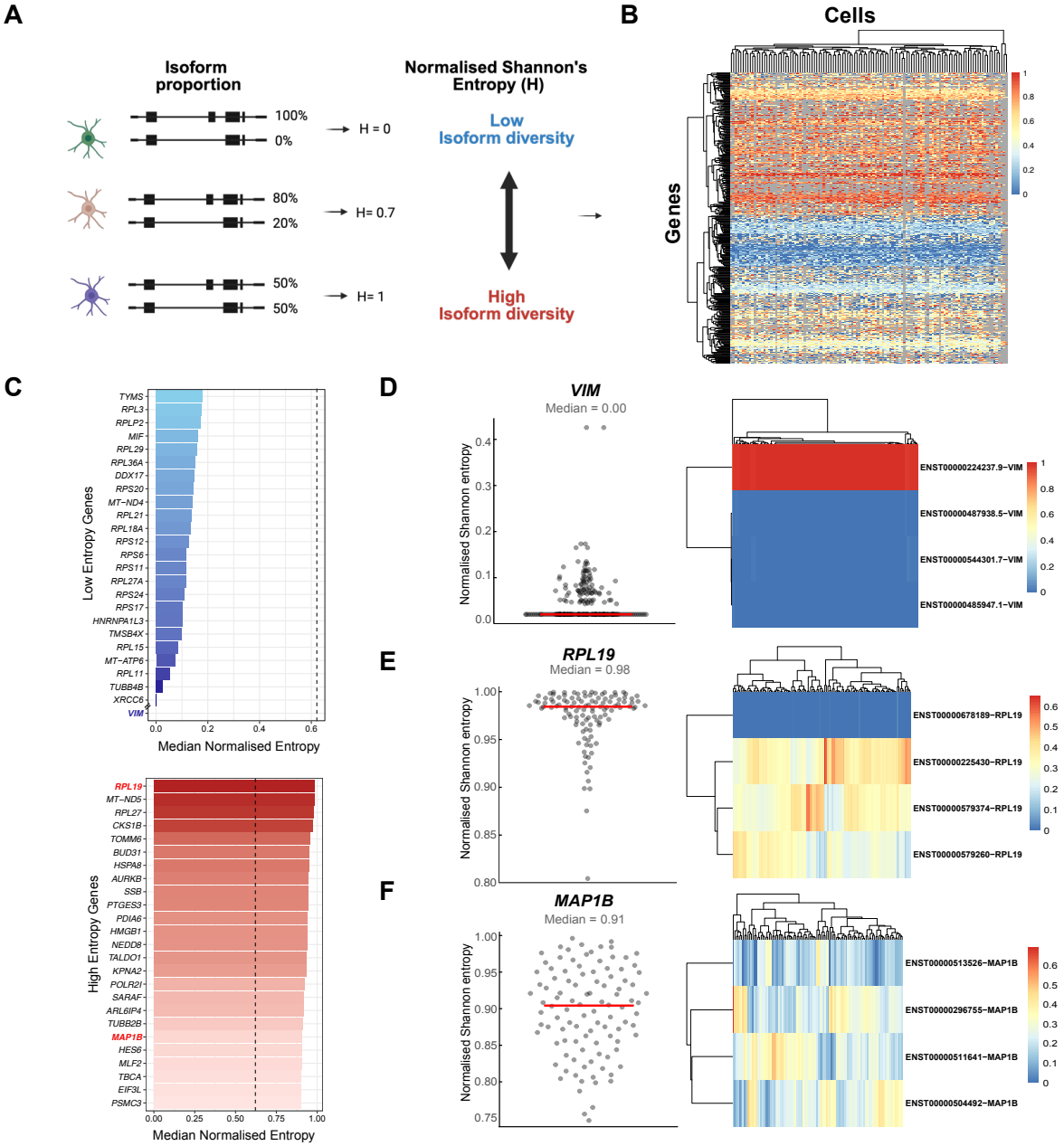

Fig. 5. Single-cell profiling of isoform diversity using Shannon's entropy. (A) Schematic illustrating how normalised Shannon's entropy (H) captures isoform diversity. Genes dominated by a single isoform have low entropy (H $\approx$ 0), while those with expression more evenly distributed across multiple isoforms approach 1. **(B)** Heatmap of normalised entropy values for 435 genes across 117 radial glial cells from the Day 55 single-sample dataset. Blue indicates low normalised entropy, red indicates high normalised entropy, and grey marks cells where entropy could not be calculated for that gene due to insufficient counts. **(C)** Genes with the lowest (top) and highest (bottom) median entropy. In addition to *VIM* (shown at the bottom of the low-entropy panel), 21 genes had a median entropy of 0 but are omitted here for clarity. The dotted line in both plots indicates the median entropy values across all 435 genes. Coloured gene names represent genes shown in panels D-F. **(D-F)** Gene entropy examples. Scatter plots (left) show per-cell entropy values. Each cell is shown by a dot. Red line represents median entropy for the gene. Isoform proportion heatmap (right) shows expression proportion of each isoform in each cell. **(D)** Example of a low-diversity gene, *VIM*. **(E)** Example of a high-diversity gene, *RPL19*. **(F)** Example of a gene with heterogeneous diversity, *MAP1B*.

to specific wet-lab protocols, restricting their broader applicability, while more flexible programs rely on pre-defined demultiplexing solutions designed specifically for a limited set of supported protocols, or lack desirable features such as short-read-free data processing, SNV identification or data visualisation (Table S1).

We sought to address this issue by developing an adaptable and flexible pipeline that supports multiple data types. To achieve this, we developed FLAMESv2 and demonstrate it can accurately process standard 10x Genomics data, as well as alternate protocols such as LR-split-seq and spatial datasets despite substantial differences in read structure.

While FLAMESv2 supports common analyses such as isoform identification and quantification, it is neither feasible nor desirable to hard-code every possible analysis option into the pipeline. Accordingly, FLAMESv2 has been built to be modular, with standards-based inputs/outputs to maximize compatibility for custom steps. For example, the isoform-discovery step is tool agnostic, meaning any method that uses BAM files as input and outputs a GTF can be utilised, allowing users to swap discovery tools or benchmark them side-by-side while keeping downstream steps constant. FLAMESv2 also integrates with the Bioconductor ecosystem and returns standard Bioconductor objects, such as SingleCellExperiment, for compatibility with these and other analysis packages. This design echoes other standardised and modular workflows such as Seurat [36].

FLAMESv2's flexibility lowers the barrier to isoform-resolved biological discovery and the rich information embedded in long-read single-cell datasets. We demonstrated this potential using neuronal differentiations, where FLAMESv2 quantified genes, known and novel isoforms; reconstructed developmental trajectories; and resolved cell types. FLAMESv2 detected time-ordered isoform programs, supporting the idea that developmental shifts are also encoded at the isoform level rather than by gene abundance alone [11, 52, 53]. These isoforms provide targets for further investigation and functional follow-up. For example, we identified an isoform switch in *PKM*, where the novel *BambuTx13041* isoform encodes a distinct open reading frame that lacks part of the second pyruvate kinase domain. This domain-level rewiring likely reflects a functional shift in *PKM* function [54, 55].

Beyond differential expression, long-read scRNA-seq can quantify isoform diversity and FLAMESv2 computes a per-gene per-cell entropy score. Consistent with prior bulk studies, low-entropy genes were typically short, housekeeping loci optimized for efficient production of a single product - and high-entropy genes with richer alternative splicing profiles often related to specific developmental signatures [56, 57]. These analyses also demonstrated how individual cells can express multiple alternative isoforms of a gene and that isoform expression differs between both cells and cell types. As with any single-cell metric, data sparsity and ambiguous read alignments are potential limitations and it is recommended to restrict these analysis to genes and cells with sufficient counts for accurate isoform quantification.

Future development of FLAMES could further improve its performance and usability, including adding native support for empty-droplet detection and ambient RNA removal. Currently users who want decontamination must rerun parts of the pipeline with a curated background allow-list. An optional RNA-clean module that outputs empty-droplet/background count matrices and removes empty droplets would simplify workflows and improve quantification. In addition, while FLAMES is compatible with the 10x Visium spatial transcriptomics kits, future testing will be required to confirm compatibility with the recently released Visium HD platform. The rapid advance of the long-read single-cell field also creates the need for standardised systems to benchmark experimental and computational methods. This has been difficult due to a lack of single-cell controls, such as synthetic standards, which would provide a ground truth for evaluating isoform discovery and quantification. Within this landscape, modular pipelines, where tools for individual steps can be swapped in and out, while all others are held constant, will be especially valuable.

FLAMESv2 significantly improves on the FLAMES toolkit for investigating gene and isoform biology at single-cell resolution. The updated pipeline introduces substantial improvements in input support, modularity, core functionality and user experience. Users can now identify and quantify more genes and isoforms, run the pipeline with novel sequencing protocols, choose to implement the protocol with or without paired short-read data, calculate isoform diversity in single cells and visualize their data. These significant enhancements provide a faster and more versatile package dedicated to exploring genes and isoforms in single cell and spatial datasets. We anticipate that FLAMESv2 will be instrumental in building isoform-resolved cell atlases and in pinpointing isoforms that control development and disease.

Methods

FLAMESv2 implementation

An overview of FLAMESv2 pipelines are shown in Figure 1A. The first step is demultiplexing reads. The BLAZE and flexiplex packages were incorporated into FLAMESv2 via basilisk [58] and Rcpp [59] respectively to support demultiplexing. We have updated BLAZE and flexiplex to output barcodes and Unique Molecular Identifiers (UMIs) as HTSlib tags in FASTQ files for consistency. Both packages have been modified to perform parallel gzip compression for the FASTQ output to optimize speed and storage. In many protocols, a sequence can be found at one end of the read to indicate it is full length, for example, the template switch oligo (TSO) from protocols based on 10x 3' kits or the poly(dT) sequence in 5' kits. Depending on the user's specification, FLAMESv2 can trim these sequences off using cutadapt [60]. Users may also specify to retain only reads that contain such sequences to reduce false discoveries of novel isoforms arising from truncated reads.

The demultiplexed reads are aligned to the reference genome using minimap2 [61] and the alignment files are used for

isoform identification and optional gene quantification. Next, the isoform identification step creates a transcriptome assembly, after which reads are re-aligned to this assembly using `minimap2` [61]. If novel isoform discovery is not required, users can skip this step, and the pipeline will automatically use the reference annotation instead. Isoform quantification is performed using the transcriptome alignments. `FLAMESv2` retains the option of using isoform quantification implemented in `FLAMESv1`, which is relatively conservative and assigns counts only from reads that can be confidently assigned to a single isoform, while the `Oarfish` quantification option probabilistically quantifies multi-mapping reads. Finally, a `SingleCellExperiment` object [38] containing the annotations and count matrices is returned.

Visualization of read coverage along transcript body To visualize read coverage and the effect of read length on coverage, `FLAMESv2` provides a transcript coverage plot as shown in Figure 2H. For each transcript, the coverage at a specific genomic coordinate is calculated as the number of reads covering that coordinate divided by the total number of reads aligned to that transcript. Read coverage is then sampled along the transcript body, and average coverage is calculated for different transcript length ranges. To minimise the impact of misaligned reads, `FLAMESv2` uses the `filter_coverage` function to remove transcripts showing dramatic coverage change inside the transcript body (Figure S13–S14), as such patterns often reflect incorrectly mapped reads introducing artificial transcript boundaries. Specifically, the `filter_coverage` function calculates the convolution of the coverage with a step function and removes transcripts with convolutions exceeding a user-specified threshold (with the default being no more than 30% change in coverage within a sliding window of 4% of the transcript length). Coverages of individual transcripts are also available to the user to allow for further investigation with the `get_coverage` function.

Re-processing of previous datasets

LR-split-seq data re-processing: Barcode identification was performed with `FLAMESv2`, using the barcode knee-plot shown in Supplementary Figure S3. A threshold of 100 UMI counts was used and 606 barcodes retained. The `FLAMESv2` pipeline was then executed using the selected barcodes. The output gene count matrix was used to generate a UMAP since Rebboah *et al* (2021) used a long-read gene count matrix to generate their UMAP [18]. Matching short-read data from the same split-seq library was used to associate the cell barcodes with sample IDs. Of the 606 cells identified by `FLAMESv2` 546 cells are present in the short-read data. Previous studies show some differences in identified barcodes between long- and short-read are expected [31].

Longbench data analysis: To compare demultiplexing and read coverage between single-cell and single-nuclei samples, 10 million reads from each sample were processed with `FLAMESv2`. To identify correlations between bulk and single-cell/nuclei samples, all 24 LongBench long-read bulk samples were processed together using the `FLAMESv2` bulk pipeline, whereas all 4 long-read single-cell and single-nuclei samples were processed using the single-cell multi-sample pipeline. Alignment was performed across all 4 samples in parallel (each process was given 64 cores) as this step constitutes the primary computational bottleneck even with the highly optimised `minimap2` software [61]. For single-cell and single-nuclei samples, isoform counts were aggregated by cell-line, using LongBench's short-read derived cell-line annotation. The aggregated counts were then used to identify expression correlations with bulk samples.

Single-cell Sample Preparation and Sequencing

Human pluripotent stem cells and neuronal differentiation: Human-induced pluripotent stem cell (hiPSC) lines RM3.5 [62] and KOLF2.1J [63] were maintained under xenogeneic conditions as described by Niclis *et al.* [64]. Cortical differentiation was performed following the protocol of Gantner *et al.* [39] as shown in Figure S8. For single-sample differentiations, RM3.5 cells were used, with culture, dissociation and sequencing methods as previously described in You *et al.* (2023) [31]. For multi-sample time-course experiments, KOLF2.1J cells were used. From day 39 onwards, cultures were passaged with Accutase approximately every 10 days to reduce ganglion-like aggregation and ensure efficient single-cell dissociation. Apart from this modification, conditions followed those described in [39] and [31] respectively.

Preparation of KOLF2.1J single-cell suspensions for scRNA-seq: A total of eight samples were collected, comprising two biological replicates at each differentiation stage (Day 0, Day 25, Day 55, and Day 80). To minimize potential batch effects, differentiations were run in parallel and were staggered by approximately 30 days. This design allowed for the simultaneous collection of two samples from different timepoints (e.g., Day 25 and Day 55). To disassociate cultures and prepare a single-cell suspension for 10x processing, we implemented a modified version of the "Dissociation of neuronal culture to single-cells for scRNA-seq" protocol [65]. Briefly, cultures were gently washed twice with 300 µl PBS 1x -/- to remove debris, before dissociation buffer was added. For iPSCs, the buffer contained Accutase (Thermo Fisher, Cat. #A1110501). For all other samples, the buffer comprised 2.5 ml Accutase, 2.5 ml PBS 1x, and one vial of papain (Worthington Biochemical Corporation, Cat. #LK003176). Cells were incubated at 37°C and samples were inspected every 10 minutes to ensure that breaks in the cell monolayer could be observed. Once the cell monolayer had lifted from the well, gentle trituration was performed to break apart cell clumps. The dissociation buffer was then inactivated using Wash Buffer 1 consisting of 15 mL DMEM/F12 + GlutaMAX, 15 µL ROCK inhibitor (Y-27632, 10 mM stock, Tocris Bioscience), and one vial DNase (250 µL) (Worthington Biochemical Corporation Catalog #LK003170), followed by additional gentle trituration to promote formation of a single-cell suspension. To remove any remaining cell debris the cell suspension was passed through a Flowmi™ strainer (Flowmi; Cat. No. 64709–60). Cell suspensions were washed 4 times in Wash Buffer 2 (PBS 1x plus BSA at 0.05% and Rock in-

hibitor (10 mM)) to remove ambient RNA contamination. Each wash was followed by centrifugation at 1200 rpm for 3 min at 4 C° with the supernatant discarded. Cells were counted manually using a hemocytometer, and viability assessed with trypan blue stain (Thermo Fisher Scientific Cat. No. 15250061). Finally, cells were resuspended in Wash Buffer 2 at a concentration of 700-1200 cells/µl as per the 10x Genomics Chromium Single-cell 3' gene expression protocol (v3.1) user guide.

10X single-cell processing and cDNA amplification: Single-cell suspensions were loaded onto the 10x Genomics Chromium Controller with a target recovery of 5,000 cells per sample, following the manufacturer's protocol (Chromium Next GEM Single Cell 3' Reagent Kits v3.1, CG000204 Rev D). Library preparation was carried out according to the FLT-seq protocol [66] with the reverse transcription extension time increased to 2 h to promote full-length cDNA synthesis. Following GEM generation, emulsions were split 80:20 and the 20% fraction (approximately 1,000 cells) was processed for Nanopore long-read sequencing. To minimize short fragment amplification, the cDNA was cleaned twice using 0.6× SPRI selection prior to amplification. Long-read cDNA was amplified according to the FLT-seq protocol using 12 PCR cycles followed by a final 0.6× SPRI select clean to remove primer dimer.

Nanopore single-cell library preparation and sequencing: Full-length cDNA generated using the FLT-seq protocol described above was prepared with the SQK-LSK114 Ligation Sequencing Kit (Oxford Nanopore Technologies, ONT) with minor modifications: incubation times for end-preparation and A-tailing were extended by 10 min, and all AMPureXP cleanup steps were performed at a 0.6× bead ratio. Libraries were sequenced on PromethION flow cells (FLO-PRO114M) with 25 fmol input. Fast5 files were generated using MinKnow v22.12.5 and basecalled with Guppy v6.3.8 (dna_r10.4.1_e8.2_400bps_sup.cfg). As per You *et al.* (2023), the Day 55 single-sample cDNA library was prepared using the SQK-LSK110 kit and sequenced on FLO-PRO002 flow cells. Fast5 files were generated using MinKnow v22.03.4 and basecalled with Guppy v6.0.7 (dna_r10.4_e8.1_sup.cfg).

Analysing long-read scRNA-seq data of differentiating stem cells

Data analysis using `FLAMES`: To generate gene and isoform count data we used `FLAMESv2` using the following configuration settings on both the Day 55 single-sample and the multi-sample time-course data. All gene biotypes - coding and non coding were quantified. Isoform discovery was performed with `Bambu` using a Novel Discovery Rate (NDR) threshold of 0.75, while isoform quantification was carried out with the default `Oarfish` model. Reads were aligned to the hg38.analysisSet reference genome with all alternative contigs removed (retaining only chromosomes 1–22, X, Y, and M). We used the GENCODE v47 comprehensive annotation [67]. Cell barcodes were identified using long-read data alone and read demultiplexing was performed with `BLAZE`. All other parameters were set to default. For all `FLAMESv1` analysis we relayed on the original implementation of read demultiplexing using a barcode allow list from matched short-reads and the original implementation of isoform discovery and quantification steps. All other `FLAMESv1` settings were the default.

Long-read scRNA-seq data processing and analysis using Seurat: Multi and single sample gene and isoform count matrices generated with `FLAMESv2` were processed in `Seurat` v5.1.0 [36]. The full workflow closely follows the `FLAMESv2` analysis tutorial (https://sefi196.github.io/FLAMESv2_LR_sc_tutorial/). In short, each sample was independently filtered on the gene count data to remove low-quality cells and doublets. Cells were retained if they expressed at least 3,000 genes and had mitochondrial content below 10%. We performed doublet detection using `DoubletFinder` v2.0.4 [68], setting the doublet rate = 0.039. For multi-sample analyses, gene-level data were integrated using `Harmony` v1.2.1 [45]. Isoform data was added as a new assay to the gene level Seurat object ensuring only cells that passed gene level QC were retained in the isoform assay. The isoform assay was filtered to remove isoforms expressed in less than 1% of cells. In addition, all novel Bambu genes and their corresponding isoforms were removed and not analysed in these datasets. Trajectory inference was performed with `Monocle3` v1.3.7 [48] and marker genes were identified using Seurat's `FindAllMarkers` function with the logfc.threshold set to 0.25. Gene ontology and pathway analysis were carried out with `gprofiler2` v0.2.3 [69], using a custom background set of all genes or isoforms expressed in the filtered Seurat object. Isoforms were structurally classified with `SQANTI v5.2.1` [34, 35], with analysis restricted to those present in the filtered Seurat object (i.e., low-abundance isoforms excluded during Seurat filtering were not classified). All other parameters were kept at default unless otherwise specified. Canonical isoform classification was carried out using `BioMart` v2.6.2 [70].

Differential Transcript Usage: We performed differential transcript usage (DTU) analysis using `IsoformSwitchAnalyzeR v2.4.0` and `DTUrtle` v1.0.3 [49, 50]. For IsoformSwitchAnalyzeR, isoform counts were aggregated from the Seurat object using the `AggregateExpression` function, stratified by sample ID and time point (two samples per time point). Genes and isoforms were retained for testing if they reached a minimum of 30 gene level and 20 isoform level counts. Significance was defined as isoform q-value < 0.01 with a differential isoform fraction (dIF) cutoff of 0.25. For `DTUrtle`, we analysed the same Seurat object using the single-cell filtering option provided by the `DTUrtle` package and compared all time points (e.g. Day25 cells vs Day55 cells). Only significant hits consistently detected by both methods were considered DTU.

Measuring cell-cell isoform variability using Shannon's entropy

To mitigate the sparsity of single cell data, we limited our isoform variability measurement to highly expressed genes and in cells in which these genes were moderately to highly expressed. Specifically, we measured isoform variability on genes that had ≥ 100 counts and had at least two isoforms expressed in the data. For each gene, we tested isoform variability in cells with ≥ 10 counts for that gene. Isoforms were filtered to include only those expressed in $\geq 5\%$ of cells. Additionally, within each cell, isoforms were ranked in decreasing order of abundance and filtered to include only those that cumulatively contributed up to 95% of the total gene expression. Isoforms used for variability calculation were then defined as the intersection of these 95%-cumulative isoform sets across all cells. We then used Shannon's entropy [71] to quantify isoform variability for a given gene, similar to previous studies [72–74], but computed at the per cell level as:

$$H_{gj} = -\sum_{i \in I_g} p_{ij} \log_2 p_{ij}$$

Where:

- I_g denotes the set of isoforms retained after filtering for a given gene g.
- H_{gj} is the Shannon entropy for cell j, quantifying isoform diversity for gene g.
- p_{ij} is the relative abundance of isoform i in cell j, calculated as:
$$p_{ij} = \frac{c_{ij}}{\sum_{i \in I_g} c_{ij}},$$
, where c_{ij} is the count of isoform i in cell j.

Next, we calculate normalised entropy values $\tilde{H}_{gj}$ with a scale of 0-1:

$$\tilde{H}_{gj} = \frac{H_{gj}}{\log_2(|I_g|)}$$

, where $|I_g|$ denotes the number of isoforms in I_g.
Genes were retained only if normalised entropy could be computed in at least 20% of cells. Genes were excluded from analysis if they contained isoforms with identical transcript structures (for example, that only differ in the annotated ORF). For the diversity results reported, we quantified isoforms using the `oarfish` coverage model.

Data Availability

Cortical differentiation fastq data are publicly available to download from ENA:PRJEB100552.

Code Availability

`FLAMESv2` is available from Bioconductor: https://bioconductor.org/packages/FLAMES. Code for re-analysis of previous data and Figure 2 is available at https://github.com/ChangqingW/FLAMES_figures and code for data analysis of cortical differentiations is available at https://github.com/Sefi196/FLAMESv2_data_analysis.

Author Contributions

C.W. and Y.Y. developed the `FLAMESv2` software with contributions from Y.D.J.P., O.V., J.S., L.T. and H.P. Y.D.J.P., J.H., A.L., and C.P.J.H. performed cell differentiations with assistance from C.L.P., K.C.L.L., K.D.A.-B. and C.P. Y.D.J.P. and R.D.P.-I. performed sequencing. Q.G. and R.T. generated data. Y.D.J.P., M.G.G. and S.V. performed cell type annotations. C.W., Y.D.J.P., M.D., M.R.M.D., X.D. and K.Z. performed data analysis. N.M.D. created research software and supervised the research. M.E.R., M.B.C. and Y.Y. planned and supervised the research. C.W., Y.D.J.P., M.E.R., M.B.C. and Y.Y. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by funding from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (Grant No. 2019-002443 to M.E.R.), Australian National Health and Medical Research Council (NHMRC) Investigator Grants (GNT2007996 to Q.G., GNT2016547 to N.M.D, GNT1196841 to M.B.C. and GNT2017257 to M.E.R.), the Australian Research Council (Discovery Project No. 200102460 to M.B.C. and 200102903 to M.E.R.), PhD Fellowship of the Research Foundation-Flanders (FWO-Vlaanderen, 11F9623N/11F9625N to M.D.) the Genomics Innovation Hub, Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS.

Declaration of interests

Y.D.J.P., R.D.P., A.L., Q.G., N.M.D., M.B.C. and Y.Y. have received support from Oxford Nanopore Technologies (ONT) to present their findings at scientific conferences. However, ONT played no role in the study design, execution, analysis, or publication of this research.

Acknowledgements

The authors are grateful for the feedback obtained during code review of our R package when it was submitted to Bioconductor. The authors thank Dr. William Skarnes at The Jackson Laboratory for providing the KOLF2.1 cell line. This project was run as part of the 10x Genomics Millenium Science Start Single Cell Fellowship. The authors thank Dr. Catherine King and Dr. Gerry Ma (10x Genomics) and Dr. Paul Gooding (Millennium Science) for their training and assistance in the preparation of 10x libraries. The authors thank Michaela Sacco for her helpful suggestions and comments on the manuscript. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This facility was established with the assistance of

LIEF Grant LE170100200. Figure 1 and 4A was created under license with BioRender.com.

Ethics approval and consent to participate

All research activities involving iPSC lines were performed under institutional ethics approval from the University of Melbourne HREC, ethics IDs: 1239208, 12374 and 28478.

Bibliography

1. Li, X. & Wang, C.-Y. From Bulk, Single-Cell to Spatial RNA Sequencing. *International Journal of Oral Science* **13,** 36. ISSN: 2049-3169. doi:10.1038/s41368-021-00146-0 (Nov. 15, 2021).
2. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential Scaling of Single-Cell RNA-seq in the Past Decade. *Nature Protocols* **13,** 599–604. ISSN: 1750-2799. doi:10.1038/nprot.2017.149 (Apr. 2018).
3. He, S. *et al.* Single-Cell Transcriptome Profiling of an Adult Human Cell Atlas of 15 Major Organs. *Genome Biology* **21,** 294. ISSN: 1474-760X. doi:10.1186/s13059-020-02210-0 (Dec. 7, 2020).
4. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6** (ed Gingeras, T. R.) e27041. ISSN: 2050-084X. doi:10.7554/eLife.27041 (Dec. 5, 2017).
5. Rood, J. E. *et al.* The Human Cell Atlas from a Cell Census to a Unified Foundation Model. *Nature* **637,** 1065–1071. ISSN: 1476-4687. doi:10.1038/s41586-024-08338-4 (Jan. 2025).
6. Chen, A. *et al.* Spatiotemporal Transcriptomic Atlas of Mouse Organogenesis Using DNA Nanoball-Patterned Arrays. *Cell* **185,** 1777–1792.e21. ISSN: 0092-8674, 1097-4172. doi:10.1016/j.cell.2022.04.003 (May 12, 2022).
7. Stickels, R. R. *et al.* Highly Sensitive Spatial Transcriptomics at Near-Cellular Resolution with Slide-seqV2. *Nature Biotechnology* **39,** 313–319. ISSN: 1546-1696. doi:10.1038/s41587-020-0739-1 (3 Mar. 2021).
8. Rodriques, S. G. *et al.* Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution. *Science* **363,** 1463–1467. doi:10.1126/science.aaw1219 (Mar. 29, 2019).
9. *Visium Platform Page* https://www.10xgenomics.com/platforms/visium.
10. Wang, E. T. *et al.* Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **456,** 470–476. ISSN: 1476-4687. doi:10.1038/nature07509 (Nov. 2008).
11. Weyn-Vanhentenryck, S. M. *et al.* Precise Temporal Regulation of Alternative Splicing during Neural Development. *Nature Communications* **9,** 2189. ISSN: 2041-1723. doi:10.1038/s41467-018-04559-0 (1 June 6, 2018).
12. Boggs, R. T., Gregor, P., Idriss, S., Belote, J. M. & McKeown, M. Regulation of Sexual Differentiation in D. Melanogaster via Alternative Splicing of RNA from the Transformer Gene. *Cell* **50,** 739–747. ISSN: 0092-8674. doi:10.1016/0092-8674(87)90332-1 (Aug. 28, 1987).
13. Wang, G.-S. & Cooper, T. A. Splicing in Disease: Disruption of the Splicing Code and the Decoding Machinery. *Nature Reviews Genetics* **8,** 749–761. ISSN: 1471-0064. doi:10.1038/nrg2164 (10 Oct. 2007).
14. Gupta, I. *et al.* Single-Cell Isoform RNA Sequencing Characterizes Isoforms in Thousands of Cerebellar Cells. *Nature Biotechnology* **36,** 1197–1202. ISSN: 1546-1696. doi:10.1038/nbt.4259 (Dec. 2018).
15. Singh, M. *et al.* High-Throughput Targeted Long-Read Single Cell Sequencing Reveals the Clonal and Transcriptional Landscape of Lymphocytes. *Nature Communications* **10,** 3120. ISSN: 2041-1723. doi:10.1038/s41467-019-11049-4 (1 July 16, 2019).
16. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High Throughput Error Corrected Nanopore Single Cell Transcriptome Sequencing. *Nature Communications* **11,** 4025. ISSN: 2041-1723. doi:10.1038/s41467-020-17800-6 (Aug. 12, 2020).
17. Tian, L. *et al.* Comprehensive Characterization of Single-Cell Full-Length Isoforms in Human and Mouse with Long-Read Sequencing. *Genome Biology* **22,** 310. ISSN: 1474-760X. doi:10.1186/s13059-021-02525-6 (Nov. 11, 2021).
18. Rebboah, E. *et al.* Mapping and Modeling the Genomic Basis of Differential RNA Isoform Expression at Single-Cell Resolution with LR-Split-seq. *Genome Biology* **22,** 286. ISSN: 1474-760X. doi:10.1186/s13059-021-02505-w (Oct. 7, 2021).
19. Philpott, M. *et al.* Nanopore Sequencing of Single-Cell Transcriptomes with scCOLOR-seq. *Nature Biotechnology* **39,** 1517–1520. ISSN: 1546-1696. doi:10.1038/s41587-021-00965-w (12 Dec. 2021).
20. Lebrigand, K. *et al.* The Spatial Landscape of Gene Expression Isoforms in Tissue Sections. *Nucleic Acids Research,* gkad169. ISSN: 0305-1048. doi:10.1093/nar/gkad169 (Mar. 17, 2023).
21. Boileau, E. *et al.* Full-Length Spatial Transcriptomics Reveals the Unexplored Isoform Diversity of the Myocardium Post-MI. *Frontiers in Genetics* **13.** ISSN: 1664-8021. https://www.frontiersin.org/articles/10.3389/fgene.2022.912572 (2023) (2022).
22. *Epi2me-Labs/Wf-Single-Cell* EPI2ME Labs, Apr. 30, 2024. https://github.com/epi2me-labs/wf-single-cell (2024).
23. Wang, Q. *et al.* Single-Cell Transcriptome Sequencing on the Nanopore Platform with ScNapBar. *RNA* **27,** 763–770. ISSN: 1355-8382, 1469-9001. doi:10.1261/rna.078154.120. PMID: 33906975 (Jan. 7, 2021).
24. Volden, R. *et al.* Improving Nanopore Read Accuracy with the R2C2 Method Enables the Sequencing of Highly Multiplexed Full-Length Single-Cell cDNA. *Proceedings of the National Academy of Sciences* **115,** 9726–9731. ISSN: 0027-8424, 1091-6490. doi:10.1073/pnas.1806447115. PMID: 30201725 (Sept. 25, 2018).
25. Trull, A., Worthey, E. A. & Ianov, L. Scnanoseq: An Nf-Core Pipeline for Oxford Nanopore Single-Cell RNA-sequencing. *Bioinformatics,* btaf487. ISSN: 1367-4811. doi:10.1093/bioinformatics/btaf487 (Sept. 4, 2025).
26. Fu, Y. *et al.* Single Cell and Spatial Alternative Splicing Analysis with Nanopore Long Read Sequencing. *Nature Communications* **16,** 6654. ISSN: 2041-1723. doi:10.1038/s41467-025-60902-2 (July 19, 2025).
27. Xu, Z., Qu, H.-Q., Kao, C., Hakonarson, H. & Wang, K. "Single-Cell Omics for Transcriptome CHaracterization (SCOTCH): Isoform-Level Characterization of Gene Expression through Long-Read Single-Cell RNA Sequencing". In: *bioRxiv.* (Apr. 30, 2024). doi:10.1101/2024.04.29.590597.
28. Jousheghani, Z. Z. & Patro, R. "Oarfish: Enhanced Probabilistic Modeling Leads to Improved Accuracy in Long Read Transcriptome Quantification". In: *bioRxiv.* (Mar. 1, 2024). doi:10.1101/2024.02.28.582591.
29. O'Keeffe, P. *et al.* TIRE-seq Simplifies Transcriptomics via Integrated RNA Capture and Library Preparation. *Scientific Reports* **15,** 15385. ISSN: 2045-2322. doi:10.1038/s41598-025-98282-8 (May 2, 2025).
30. Cheng, O. *et al.* Flexiplex: A Versatile Demultiplexer and Search Tool for Omics Data. *Bioinformatics* **40,** btae102. ISSN: 1367-4811. doi:10.1093/bioinformatics/btae102 (Mar. 1, 2024).
31. You, Y. *et al.* Identification of Cell Barcodes from Long-Read Single-Cell RNA-seq with BLAZE. *Genome Biology* **24,** 66. ISSN: 1474-760X. doi:10.1186/s13059-023-02907-y (Apr. 6, 2023).
32. Chen, Y. *et al.* Context-Aware Transcript Quantification from Long-Read RNA-seq Data with Bambu. *Nature Methods* **20,** 1187–1195. ISSN: 1548-7105. doi:10.1038/s41592-023-01908-w (8 Aug. 2023).
33. You, Y. *et al.* "Benchmarking Long-Read RNA-sequencing Technologies with LongBench: A Cross-Platform Reference Dataset Profiling Cancer Cell Lines with Bulk and Single-Cell Approaches". In: *bioRxiv.* (Sept. 12, 2025). doi:10.1101/2025.09.11.675724.
34. Keil, N., Monzó, C., McIntyre, L. & Conesa, A. Quality Assessment of Long Read Data in Multisample lrRNA-seq Experiments Using SQANTI-reads. *Genome Research* **35,** 987–998. ISSN: 1088-9051, 1549-5469. doi:10.1101/gr.280021.124. PMID: 40032587 (Jan. 4, 2025).
35. Pardo-Palacios, F. J. *et al.* SQANTI3: Curation of Long-Read Transcriptomes for Accurate Identification of Known and Novel Isoforms. *Nature Methods* **21,** 793–797. ISSN: 1548-7105. doi:10.1038/s41592-024-02229-2 (May 2024).
36. Hao, Y. *et al.* Dictionary Learning for Integrative, Multimodal and Scalable Single-Cell Analysis. *Nature Biotechnology* **42,** 293–304. ISSN: 1546-1696. doi:10.1038/s41587-023-01767-y (Feb. 2024).
37. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biology* **19,** 15. ISSN: 1474-760X. doi:10.1186/s13059-017-1382-0 (Feb. 6, 2018).
38. Amezquita, R. A. *et al.* Orchestrating Single-Cell Analysis with Bioconductor. *Nature Methods* **17,** 137–145. ISSN: 1548-7105. doi:10.1038/s41592-019-0654-x (2 Feb. 2020).
39. Gantner, C. W. *et al.* FGF-MAPK Signaling Regulates Human Deep-Layer Corticogenesis. *Stem Cell Reports* **16,** 1262–1275. ISSN: 2213-6711. doi:10.1016/j.stemcr.2021.03.014. PMID: 33836146 (May 11, 2021).
40. Shi, Y., Kirwan, P., Smith, J., Robinson, H. P. C. & Livesey, F. J. Human Cerebral Cortex Development from Pluripotent Stem Cells to Functional Excitatory Synapses. *Nature Neuroscience* **15,** 477–486. ISSN: 1546-1726. doi:10.1038/nn.3041 (Mar. 2012).
41. Shi, Y., Kirwan, P. & Livesey, F. J. Directed Differentiation of Human Pluripotent Stem Cells to Cerebral Cortex Neurons and Neural Networks. *Nature Protocols* **7,** 1836–1846. ISSN: 1750-2799. doi:10.1038/nprot.2012.116 (Oct. 2012).
42. Zhang, X. *et al.* Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. *Cell* **166,** 1147–1162.e15. ISSN: 0092-8674, 1097-4172. doi:10.1016/j.cell.2016.07.025. PMID: 27565344 (Aug. 25, 2016).
43. Melé, M. *et al.* The Human Transcriptome across Tissues and Individuals. *Science* **348,** 660–665. doi:10.1126/science.aaa0355 (May 8, 2015).
44. Su, C.-H., D, D. & Tarn, W.-Y. Alternative Splicing in Neurogenesis and Brain Development. *Frontiers in Molecular Biosciences* **5.** ISSN: 2296-889X. doi:10.3389/fmolb.2018.00012 (Feb. 12, 2018).
45. Korsunsky, I. *et al.* Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony. *Nature Methods* **16,** 1289–1296. ISSN: 1548-7105. doi:10.1038/s41592-019-0619-0 (Dec. 2019).
46. Strano, A., Tuck, E., Stubbs, V. E. & Livesey, F. J. Variable Outcomes in Neural Differentiation of Human PSCs Arise from Intrinsic Differences in Developmental Signaling Pathways. *Cell Reports* **31.** ISSN: 2211-1247. doi:10.1016/j.celrep.2020.107732. PMID: 32521257 (June 9, 2020).
47. Pokhilko, A. *et al.* Targeted Single-Cell RNA Sequencing of Transcription Factors Enhances the Identification of Cell Types and Trajectories. *Genome Research* **31,** 1069–1081. ISSN: 1088-9051, 1549-5469. doi:10.1101/gr.273961.120. PMID: 34011578 (Jan. 6, 2021).
48. Qiu, X. *et al.* Reversed Graph Embedding Resolves Complex Single-Cell Trajectories. *Nature Methods* **14,** 979–982. ISSN: 1548-7105. doi:10.1038/nmeth.4402 (Oct. 2017).
49. Vitting-Seerup, K. & Sandelin, A. IsoformSwitchAnalyzeR: Analysis of Changes in Genome-Wide Patterns of Alternative Splicing and Its Functional Consequences. *Bioinformatics* **35,** 4469–4471. ISSN: 1367-4803. doi:10.1093/bioinformatics/btz247 (Nov. 1, 2019).
50. Tekath, T. & Dugas, M. Differential Transcript Usage Analysis of Bulk and Single-Cell RNA-seq Data with DTUrtle. *Bioinformatics* **37,** 3781–3787. ISSN: 1367-4803. doi:10.1093/bioinformatics/btab629 (Nov. 5, 2021).
51. Wan, C. Y. *et al.* IsoVis – a Webserver for Visualization and Annotation of Alternative RNA Isoforms. *Nucleic Acids Research* **52,** W341–W347. ISSN: 0305-1048. doi:10.1093/nar/gkae343 (July 5, 2024).
52. Joglekar, A. *et al.* Single-Cell Long-Read Sequencing-Based Mapping Reveals Specialized Splicing Patterns in Developing and Adult Mouse and Human Brain. *Nature Neuro-*

science **27,** 1051–1063. ISSN: 1546-1726. doi:10.1038/s41593-024-01616-4 (June 2024).
53. Patowary, A. *et al.* Developmental Isoform Diversity in the Human Neocortex Informs Neuropsychiatric Risk Mechanisms. *Science* **384,** eadh7688. doi:10.1126/science.adh7688 (May 24, 2024).
54. Kejiou, N. S. *et al.* Pyruvate Kinase M (PKM) Binds Ribosomes in a Poly-ADP Ribosylation Dependent Manner to Induce Translational Stalling. *Nucleic Acids Research* **51,** 6461–6478. ISSN: 0305-1048. doi:10.1093/nar/gkad440 (July 7, 2023).
55. He, P. *et al.* PKM2 Is a Key Factor to Regulate Neurogenesis and Cognition by Controlling Lactate Homeostasis. *Stem Cell Reports* **20,** 102381. ISSN: 2213-6711. doi:10.1016/j.stemcr.2024.11.011 (Jan. 14, 2025).
56. Zhao, F., Yan, Y., Wang, Y., Liu, Y. & Yang, R. Splicing Complexity as a Pivotal Feature of Alternative Exons in Mammalian Species. *BMC Genomics* **24,** 198. ISSN: 1471-2164. doi:10.1186/s12864-023-09247-y (Apr. 12, 2023).
57. Ameri, A. J. & Lewis, Z. A. Shannon Entropy as a Metric for Conditional Gene Expression in Neurospora Crassa. *G3 Genes|Genomes|Genetics* **11,** jkab055. ISSN: 2160-1836. doi:10.1093/g3journal/jkab055 (Apr. 1, 2021).
58. Lun, A. T. Basilisk: A Bioconductor Package for Managing Python Environments. *Journal of Open Source Software* **7,** 4742. ISSN: 2475-9066. doi:10.21105/joss.04742 (Nov. 4, 2022).
59. Eddelbuettel, D. & Balamuta, J. J. Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician* **72,** 28–36. ISSN: 0003-1305. doi:10.1080/00031305.2017.1375990 (Jan. 2, 2018).
60. Martin, M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.journal* **17,** 10–12. ISSN: 2226-6089. doi:10.14806/ej.17.1.200 (1 May 2, 2011).
61. Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **34,** 3094–3100. ISSN: 1367-4803. doi:10.1093/bioinformatics/bty191 (Sept. 15, 2018).
62. Kao, T. *et al.* GAPTrap: A Simple Expression System for Pluripotent Stem Cells and Their Derivatives. *Stem Cell Reports* **7,** 518–526. ISSN: 2213-6711. doi:10.1016/j.stemcr.2016.07.015. PMID: 27594589 (Sept. 13, 2016).
63. Pantazis, C. B. *et al.* A Reference Human Induced Pluripotent Stem Cell Line for Large-Scale Collaborative Studies. *Cell Stem Cell* **29,** 1685–1702.e22. ISSN: 1934-5909, 1875-9777. doi:10.1016/j.stem.2022.11.004. PMID: 36459969 (Dec. 1, 2022).
64. Niclis, J. C. *et al.* Efficiently Specified Ventral Midbrain Dopamine Neurons from Human Pluripotent Stem Cells Under Xeno-Free Conditions Restore Motor Deficits in Parkinsonian Rodents. *Stem Cells Translational Medicine* **6,** 937–948. ISSN: 2157-6564. doi:10.5966/sctm.2016-0073 (Mar. 1, 2017).
65. Jerber, J., Haldane, J., Steer, J., Pearce, D. & Patel, M. Dissociation of Neuronal Culture to Single Cells for scRNA-seq (10x Genomics). doi:dx.doi.org/10.17504/protocols.io.bh32j8qe (July 23, 2020).
66. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343,** 776–779. doi:10.1126/science.1247651 (Feb. 14, 2014).
67. Mudge, J. M. *et al.* GENCODE 2025: Reference Gene Annotation for Human and Mouse. *Nucleic Acids Research* **53,** D966–D975. ISSN: 1362-4962. doi:10.1093/nar/gkae1078 (Jan. 6, 2025).
68. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems* **8,** 329–337.e4. ISSN: 2405-4712, 2405-4720. doi:10.1016/j.cels.2019.03.003. PMID: 30954475 (Apr. 24, 2019).
69. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. "Gprofiler2 – an R Package for Gene List Functional Enrichment Analysis and Namespace Conversion Toolset g:Profiler". In: *F1000Research*. (Nov. 17, 2020). doi:10.12688/f1000research.24956.2.
70. Durinck, S. *et al.* BioMart and Bioconductor: A Powerful Link between Biological Databases and Microarray Data Analysis. *Bioinformatics* **21,** 3439–3440. ISSN: 1367-4803. doi:10.1093/bioinformatics/bti525 (Aug. 15, 2005).
71. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27,** 379–423. ISSN: 1538-7305. doi:10.1002/j.1538-7305.1948.tb01338.x (1948).
72. Ritchie, W., Granjeaud, S., Puthier, D. & Gautheret, D. Entropy Measures Quantify Global Splicing Disorders in Cancer. *PLOS Computational Biology* **4,** e1000011. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1000011 (Mar. 14, 2008).
73. Jones, E. F., Haldar, A., Oza, V. H. & Lasseigne, B. N. Quantifying Transcriptome Diversity: A Review. *Briefings in Functional Genomics* **23,** 83–94. ISSN: 2041-2657. doi:10.1093/bfgp/elad019 (Mar. 1, 2024).
74. Wu, S. & Schmitz, U. "SclsoX: A Multidimensional Framework for Measuring Isoform-Level Transcriptomic Complexity in Single Cells". In: *bioRxiv*. (May 12, 2025). doi:10.1101/2025.04.28.650897.