MPAQT: A novel data integration framework for isoform quantification with short-read and long-read RNA-seq

Michael Apostolides

Department of Human Genetics, Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec

May 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

TABLE OF CONTENTS

ABSTRACT

Background: Transcript quantification is an ongoing problem in biomedical research, which is not fully solved. RNA-sequencing with short reads (SRs) is currently the leading approach due to low cost, high depth, and many available software tools for downstream analysis. Short reads, however, are often unable to resolve complex splicing events among highly similar transcripts. On the other hand, long reads (LRs) provide full-length transcript sequences, allowing unambiguous assignment of reads to transcripts, but usually at lower depth due to high cost. New computational methods are needed for joint analysis of SR and LR data to leverage the capabilities that are unique to each approach.

Methodology: We introduce MPAQT (Multi-Platform Aggregation and Quantification of Transcripts), a novel statistical framework that takes advantage of the high depth of SRs and the high accuracy and unambiguity of LRs. MPAQT's generative model explicitly connects the transcript abundance profile of a sample to the expected SR and LR distribution, allowing maximum-likelihood estimation of transcript abundances from SR data alone or the combination of SR and LR data.

Results: Using various experimental and simulated benchmarking datasets, we show that MPAQT quantifies transcripts more accurately than other leading tools such as kallisto, salmon, and RSEM; this improvement remains true at both gene level and transcript level. Using SR data alone, we show that MPAQT captures quantification information from transcripts with low expression often missed by other tools. When combining both SRs and LRs, we show MPAQT improves quantification of select transcripts, compared to when only SR data are used; transcripts with improved quantification are often from longer genes with more exons, have more splicing variants, and are enriched for neuronal differentiation and brain-related processes. Finally, we analyzed human embryonic stem cells (hESCs) undergoing *in vitro* differentiation toward cortical neurons using paired SR and LR data. We highlight MPAQT's improved quantification of transcript abundance changes accompanying neuronal differentiation, including isoform switch events not captured with SR data alone. Differentially quantified transcripts tend to be similar, differing by one or two exons; LRs can detect such small differences due to their complete transcript coverage. MPAQT's ability to integrate SR and LR data, and its improved quantification of transcripts from longer genes with more exons and splicing variants, make it a novel tool to

study transcript quantification in tissues with complex splicing patterns such as the brain and cancer.

RÉSUMÉ

Contexte: La quantification des ARNs (acides ribonucléiques) est un problème qui persiste dans la recherche biomédicale et qui n'est pas encore résolu. Le RNA-Seq (séquençage de l'ARN) avec lectures de séquence courtes (LCs) est actuellement la méthode de prédilection en raison de son faible coût, sa grande profondeur, et la disponibilité des outils logiciel pour les analyses subséquentes. Cependant, les LCs ne sont pas toujours capables de résoudre les événements d'épissage complexes parmi les transcrits très similaires, alors que les lectures longues (LLs) fournissent des séquences de transcrits complètes permettant d'assigner précisément les lectures aux transcrits, bien qu'avec une faible profondeur en raison du coût élevé. De nouvelles méthodes informatiques doivent être développées afin d'effectuer l'analyse conjointe des données de LCs et LLs pour profiter des capacités propres a chaque methode.

Méthode: Nous présentons MPAQT (*Multi-Platform Aggregation and Quantification of Transcripts*), un nouveau cadre statistique qui tire profit de la grande profondeur des LCs, de la haute précision et de la faible ambiguïté des LLs. Le modèle génératif de MPAQT relie spécifiquement le profil d'abondance d'un échantillon à la distribution attendue des LCs et LLs, ce qui permet l'estimation, selon la méthode du maximum de vraisemblance, de l'abondance des transcrits provenant des données LC uniquement, ou de la combinaison des données de LCs et LLs.

Résultats : En utilisant divers ensembles de données expérimentales et simulées, nous démontrons que MPAQT parvient à quantifier les transcrits de façon plus précise que les autres outils analogues tels que kallisto, salmon et RSEM. Cette amélioration persiste au niveau des gènes et des transcrits. En utilisant des données de LCs uniquement, nous démontrons que MPAQT est capable de quantifier les transcrits qui ont une faible expression, ces dernier étant fréquemment omis par les autres outils. En combinant les LCs et LLs, nous démontrons que MPAQT améliore la quantification de certains transcrits, comparé à l'utilisation des LCs seulement. Les transcrits qui sont les mieux quantifiés proviennent souvent de gènes plus longs, contenant plus d'exons, produisant plus de variants d'épissage, et qui sont enrichis pour les processus biologiques liés au cerveau et différenciation neuronale. Finalement, nous avons fait l'analyse de cellules souches embryonnaires humaines différenciés *in vitro* en neurones corticaux, en utilisant des données de LCs et LLs. Nous soulignons la quantification des changements d'abondance supérieure des

transcrits liés à la différenciation neuronale effectuée par MPAQT, y compris pour les événements de changement d'isoforme non détectés avec les données de LC uniquement. Les transcrits différentiellement quantifiés tendent à être similaires, différent seulement d'un ou deux exons. Les LLs sont en mesure de détecter ces différences minimes grâce à leur couverture complète des transcrits. L'habileté de MPAQT à intégrer les données LC et LL, et sa quantification améliorée des transcrits qui proviennent de gènes plus longs et comportant plus d'exons et de variants d'épissage, en font un outil prometteur pour l'étude de la quantification des transcrits dans les tissus avec des motifs d'épissage complexes, tels que le cerveau et le cancer.

ACKNOWLEDGEMENTS

I would like to acknowledge Gabrielle Perron, Elby Mackenzie and Justine Desrochers for their edits and feedback on the French abstract, Ali Saberi for his contributions to installation and running LR analysis tools, my supervisory committee, Drs. Guillaume Bourque and Jacek Majewski for their feedback on my work, and my supervisor, Dr. Hamed Najafabadi for his guidance and mentorship.

DEDICATION

This work is dedicated to all beings great and small. May all beings be at ease.

CONTRIBUTIONS OF AUTHORS

The computational methods were developed by Hamed Najafabadi and Michael Apostolides. Hamed Najafabadi conceptualized and implemented the statistical framework. Michael Apostolides performed benchmarking, data analysis, and software implementation. Michael Apostolides prepared all presentation materials and wrote the thesis, with contributions from and edits by Hamed Najafabadi. Albertas Navickas, Benedict Choi, and Hani Goodarzi (UCSF) performed the experiments for differentiation of neurons and generated the raw short-read and long-read data.

ABBREVIATIONS

BAM	Binary alignment map
CAGE	Cap analysis of gene expression
CCS	Circular consensus sequencing
cDNA	Complimentary DNA
CPM	Counts per million
DE	Differential expression
EC	Equivalence class
FL	Full length
FSM	Full splice match
GFF	General feature format
hESCs	Human embryonic stem cells
HGNC	HUGO Gene Nomenclature Committee
HiFi	High fidelity
IGV	Integrative genomics viewer
logFC	Log fold change
LR	Long read
MAQC	MicroArray Quality Control
mRNA	Messenger RNA
ONT	Oxford Nanopore Technologies
OU	Observation Unit
Pacbio	Pacific Biosciences
RC	Read Class
RMSD	Root Mean Square Deviation
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RT-qPCR	Reverse transcription-quantitative polymerase chain reaction
SR	Short read
TMM	Trimmed mean of M values
TPM	Transcripts per million
TSS	Transcription start site
TTS	Transcription termination site

LIST OF FIGURES AND TABLES

1 INTRODUCTION

1.1 Introduction to isoforms and isoform switches

Most human protein-coding genes encode multiple isoforms, which are transcripts with different sequences that are produced from the same locus. This sequence variation can occur during transcription by alternative TSSs (transcription start sites) or TTSs (transcription termination sites), often causing differing 5' or 3' untranslated regions (UTRs) or modifying the start/end of the transcript open-reading frame. Alternative splicing during pre-mRNA processing is another major source of variation among transcript isoforms, resulting in inclusion of different exons and/or retention of introns in different isoforms [1]. **Figure 1.1**a-b schematically shows the mechanisms that contribute to transcript variation.

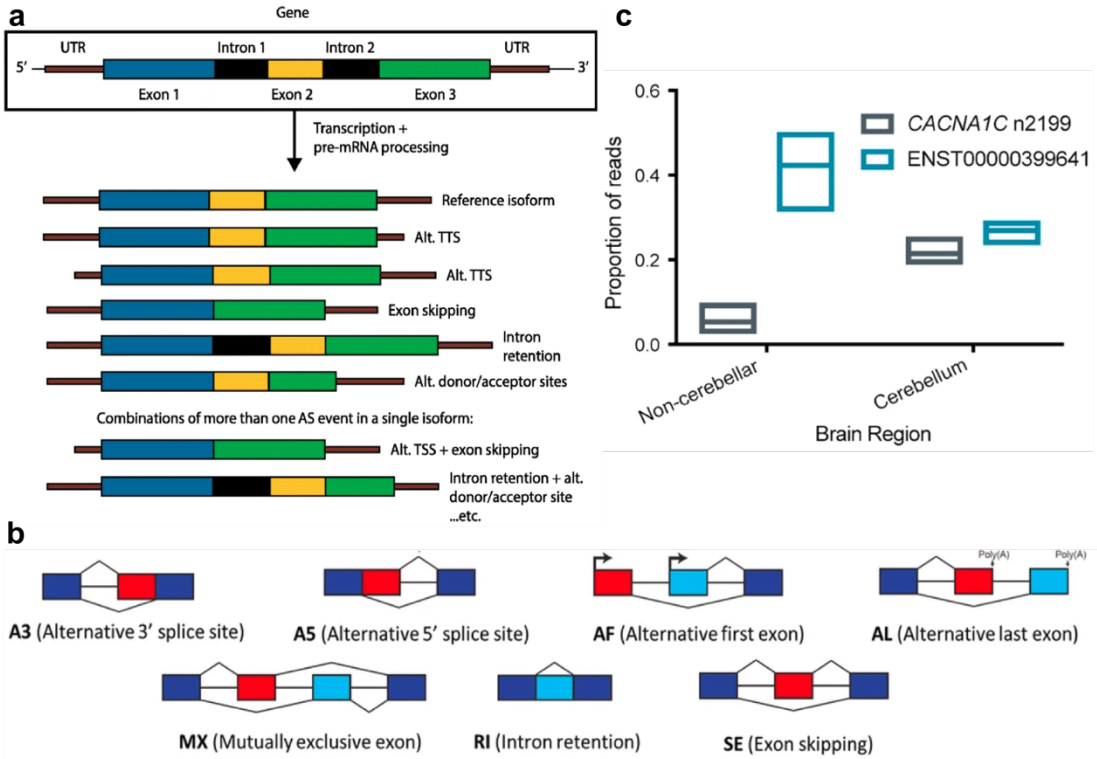


Figure 1.1: Sources of transcript variation, and an example of an isoform switch event. **(a)** Schematic of mechanisms contributing to transcript variation (from [1]). (b) Another schematic showing sources of transcript variation, including examples of specific splicing patterns commonly seen in transcripts (from [2]) (c) Example of an isoform switch event in the *CACNA1C* gene (from [3]). Images are unmodified from their original sources, and have been reproduced under the Creative Commons Attribution 4.0 International License [http://creativecommons.org/licenses/by/4.0/].

An isoform switch is defined as a change in the most highly expressed isoform between cell states or tissue types [3]. The dominant isoform can change between conditions, or alternatively, a lower-abundance isoform can increase in abundance to approximately match the abundance of the dominant isoform [3]. **Figure 1.1**c shows an example isoform switch for the *CACNA1C* gene, where a low abundance isoform from non-cerebellar brain regions becomes approximately equal to the most abundant isoform in the cerebellum.

Understanding transcript variation and isoform switches are important for understanding cellular function and disease. For example, proteins produced from different isoforms of the same gene can have different functions and cellular localizations, and can even have opposite functions to one another, particularly in diseases such as cancer [4]. Quantification of relative abundances of isoforms and identification of isoform switches is also of great interest in studying biological processes related to the nervous system, where genes often have complex splicing patterns producing isoforms with different functions in different cell types [3, 5].

Isoform switches can have important biological consequences. For example, altering the 3' UTR can affect mRNA stability, since factors that regulate mRNA stability often recognize and bind to elements within the 3' UTR (see [6] for an example), affecting the amount of mRNA available for creating protein products. Particularly in neurons, it has been shown that 3' UTRs contain regulatory motifs that can affect localization to different cellular compartments, mRNA stability, and regulation of translation [7]. Differential exon usage by alternative splicing can also produce protein isoforms with different functions. For example, alternative splicing of the Tau protein affects its N-terminal projection region and microtubule-binding domains, producing two isoforms in a 1:1 balance in adult human brains: 4-repeat (4R) and 3-repeat (3R) tau [8]. Disruption of the balance of 4R:3R Tau has been found in many brain diseases, including Alzheimer's disease [8], highlighting the importance of detecting isoform switch events when comparing healthy and diseased tissues. These findings suggest that characterization and quantification of isoform switches in tissues with complex splicing patterns is integral to understanding their role in normal cell function and disease.

1.2 Overview of RNA-seq

Transcript quantification is an ongoing problem in biomedical research which, despite decades of methodological advances, is still not fully solved. Various approaches have been developed over the past decades for gene expression quantification including microarrays, qPCR and RNA sequencing (RNA-seq) (see [9] for a timeline of transcript profiling methods). For high-throughput transcript quantification, RNA-seq is currently the leading approach due to its throughput, ability to detect transcripts with low expression (sensitivity), and suitability for a variety of downstream analyses such as fusion detection, isoform quantification, and differential expression analysis [10].

Current methods for RNA-seq can broadly be divided into short-read (SR) and long-read (LR) sequencing. The most prominent platforms for SR sequencing are Illumina sequencing and Ion Torrent sequencing, both of which use a "sequencing by synthesis" approach [11]. First, RNA is reverse transcribed into cDNA during library preparation, followed by clonal amplification and sequencing [11]. In Illumina sequencing, DNA fragments bind to flow cells, "bridging" PCR is used to amplify the sequences, and fluorescent nucleotides are used for synthesis, during which different fluorescent signals corresponding to each nucleotide are captured [11]. In Ion Torrent sequencing, DNA fragments bind to beads, followed by emulsion PCR to amplify the DNA. A, T, G and C nucleotides are added in sequence to the emulsion, and if the correct nucleotide is added, DNA polymerase incorporates it into the sequence and releases a hydrogen ion, which is detected by the sensor [11].

Current SR sequencing technologies can produce reads up to 600 bp, but most often the read length is limited to around 100-200 bp, which falls short of the actual length of most RNA molecules—for example, 81% of isoforms have length greater than 500 bp based on GENCODE annotations (median = 1543 bp, mean = 2121 bp) [12]. This length limitation makes quantifying and assigning reads to different isoforms challenging. However, SR RNA-seq has been around for longer, has many publicly available data and analysis tools, is relatively inexpensive, and has a higher sequencing depth than LR (and, therefore, is less noisy).

On the other hand, LR technology is revolutionizing biomedical research due to unambiguous mapping of full-length reads to transcripts and the potential for discovery of novel transcripts that can contain novel splice junctions, TSSs and TTSs. As a result, LR technology development is a rich area of active research [13, 14]. Two of the leading approaches for LR sequencing include those provided by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio).

Illumina has also recently released a LR technology called “Infinity”, which uses existing Illumina machines [15], whereas PacBio and ONT are uniquely aimed at providing long sequencing reads.

PacBio has been able to produce HiFi CCS (high fidelity circular consensus sequence) reads with 99.9% base accuracy, a comparable accuracy to Illumina SR RNA-seq, and has read lengths ranging from 5kb-25kb, allowing for resolution of full-length transcripts [12]. PacBio SMRT sequencing works by sequencing cDNA: adapters are first added at either end of the full-length cDNA, followed by creation of a circular sequence. A DNA polymerase then passes over the cDNA multiple times, creating a continuous long read containing many subreads. During synthesis, fluorescence signal is continuously recorded, which can then be used to generate a CCS from the subreads [12]. However, at comparable costs, PacBio currently produces 1-2 orders of magnitude fewer reads compared to Illumina—this low read depth creates challenges for precise quantification of transcripts, particularly for lower-abundance transcripts. Low depth can also cause dropouts, making transcripts with low expression indistinguishable from those with no expression.

PacBio sequencing requires conversion of RNA to cDNA, causing modified base information to be lost. In comparison, ONT provides the possibility of directly sequencing single RNA molecules without cDNA conversion. ONT works by passing single-stranded RNA through a nanopore, where different nucleotides cause different resistances within the pore, producing electrical currents known as “squiggles”, allowing for conventional and modified bases to be identified [12]. Then, a base caller (usually Guppy, although several others are available for different functions) is used to decide on the sequence using these squiggles as input [12]. ONT sequencing produces reads ranging from 500bp to a record of 2.3 Mb; however, it currently suffers from a high error rate of ~5%. In addition, pore blockage and fragmentation can result in truncated reads, leading to biased coverage toward the 3′ end [12].

Overall, the rapid advancement of LR technologies will allow for novel research in the biomedical sciences, highlighting the need for computational methods that leverage the unique benefits of LR data.

1.3 Overview of computational methods for SR and LR analysis

SR RNA-seq quantification methods can generally be divided into alignment-based and alignment-free methods. Alignment-based methods such as RSEM [16] use an aligner to first align

reads to the reference transcriptome, then use the aligned counts to obtain gene and transcript expression estimates. Alignment-based methods tend to have longer running times and require more RAM, but output a BAM file containing unambiguously mapped reads, allowing for read visualization in genome browsers such as IGV [17].

Alignment-free methods such as Salmon [18] and kallisto [19] use "quasi-mapping" and "pseudoalignment", respectively, to obtain sets of transcripts/isoforms which are compatible with each read, then use a statistical model to estimate transcript abundances based on these mappings. Alignment-free methods tend to have shorter running times and often outperform alignment-based methods. Kallisto uses the concept of an equivalence class (EC) for its pseudoalignment algorithm, and uses EC mappings as the basis for transcript quantification. Transcripts with high similarity (e.g., isoforms that have common exons) produce similar sequencing reads in a SR RNA-seq experiment. As a result, short reads can be grouped into equivalent read sets based on the set of transcripts that can produce them. Pseudoalignment methods such as kallisto [19] use SR data to quantify the ECs (i.e. assign the reads to each EC); the EC quantities are then used as the basis for transcript quantification.

Compared to SR RNA-seq, LR RNA-seq offers new benefits, such as discovery of new transcripts that are not currently annotated in existing databases or reference transcriptomes and that may be specific to the cell type being studied. Additionally, reads can be unambiguously assigned to known or novel transcripts due to long read length, allowing for quantification of isoforms that are not easily distinguishable from each other using SR data.

For processing of IsoSeq LR sequencing data, several tools exist. In this work, we use the tools Isoseq3 and SQANTI3 [20, 21]; so, we will focus on introducing these tools here. The IsoSeq pipeline (Pacific Biosciences) can be used to generate circular consensus sequence (CCS) reads. Then, a series of commands can be used to remove primer sequences, poly-A tails and concatemers, cluster similar reads, align them to the genome, and collapse them, producing a transcriptome file containing long read counts for each transcript (see methods for tool details). SQANTI [21] can then be used for isoform classification and quality control, artifact filtering, and annotation. Quality control measures implemented in SQANTI include annotation and description of isoforms, allowing isoform inspection and identification of library preparation issues. Descriptors provided by SQANTI also allow artifact filtering (transcripts that may be false positives). SQANTI also produces a report with a series of plots and metrics about the data, which

can help to understand data quality and characteristics, as described in SQANTI documentation [20]. Filtered GFF and FASTA files are produced at the end, which can be used as input to other tools.

SQANTI uses several external data sources for quality control. One such data type is CAGE peak data, from the refTSS database, which is a transcriptional start site (TSS) database from human and mouse that combines several public resources [22]. This CAGE peak reference dataset provides TSS locations, which SQANTI uses to give the distance and direction of the PacBio TSS's from the reference TSS in the CAGE peak data [22]. Another supporting dataset is splice junction coverage from Intropolis [23]. Intropolis contains exon-exon junctions from over 21k human RNA-seq samples [23], which SQANTI uses for quality control of the LR-defined transcriptome produced by the IsoSeq pipeline [21]. SQANTI can also take as input a polyA site and polyA motif list file, which allows for annotation of isoforms with the distance to nearest polyA site/motif, as well as a tappAS annotation file containing functional annotations of isoforms from a reference transcriptome [21].

1.4 Motivation for development of MPAQT

Overall, a survey of the literature suggests there is no tool that can combine LR and SR data for general-use transcript quantification. FLAIR is a tool to identify, correct and perform alternative splicing analysis of noisy long reads [24]. However, FLAIR cannot combine SR and LR data for quantification, but it can use SR data for splice site identification and improving splice junction boundary confidence [24]. IDP-denovo is a tool for de novo transcriptome assembly without a reference genome, annotation and quantification of isoforms, and integrates LR and SR data [25]. However, it is primarily for transcriptome assembly, as its performance is benchmarked against transcriptome assembly tools, but not against quantification tools [25]. TAPIS performs a variety of tasks on IsoSeq LR data, and can include SR data, but does not provide a transcript quantification feature [26]. The new StringTie version uses both LR and SR RNA-seq to improve transcriptome assembly [27]. Although a quantification feature is mentioned, performance is not explicitly benchmarked or compared to other known RNA-seq quantification tools, and its focus is on transcriptome assembly [27].

None of these tools provide a principled statistical framework for transcript quantification from joint analysis of SR and LR data, nor have they been benchmarked for such a purpose. Combining SR and LR RNA-seq could give better estimates of RNA abundance, which motivated us to develop MPAQT to combine SR and LR data.

The importance of such an approach can be seen by examination of SR and LR data generated from the same RNA sample. As I will describe in the Results section, we have analyzed such SR and LR data from a novel dataset obtained from cells undergoing *in vitro* differentiation toward neuronal lineage, and observed various examples of isoform mis-quantification by commonly used SR quantification tools. One such example is CHL1, a transmembrane cell adhesion molecule that regulates neuronal differentiation [28] (Gene Ontology [29, 30] term GO:0030182), is expressed in the brain, known to promote neurite outgrowth and branching [31], and has documented splicing changes linked to aging and disease [32]. Upon processing LR data and comparing the results to SR quantification tool predictions, we observed significant differences between SR and LR quantifications. Both kallisto, an alignment-free method, and RSEM, an alignment-based method, predict that transcript "ENST00000397491" is the most abundant isoform after neuronal differentiation, while "ENST00000256509" has lower or zero expression (**Figure 1.2**). However, from the LR data we can see that "ENST00000256509" is in fact the dominant isoform. This, and other similar examples, show that methods which use SR data alone are insufficient to quantify transcript abundances. Given the critical role of alternative splicing in the functional diversity of genes in the brain [5], quantification of brain-related isoforms using joint analysis of LR and SR data has the potential to reveal new isoforms that are involved in cell differentiation and function.

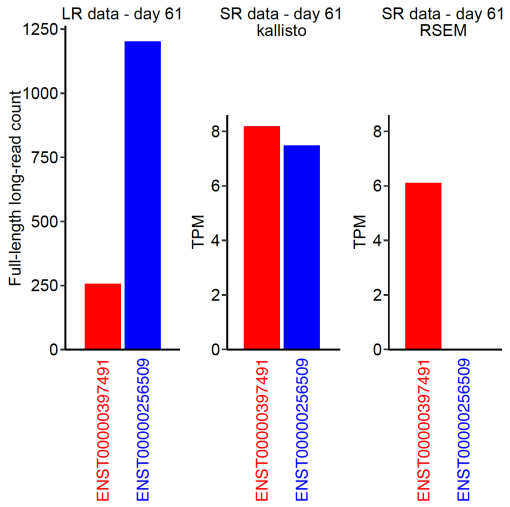


Figure 1.2: Spotlight on CHL1, demonstrating mis-assignment of reads to the correct transcript by both kallisto and RSEM. Integration of FL counts into a quantification framework has the potential to correct such false negative transcript quantifications.

1.5 Aims and hypothesis

We hypothesize that a statistical framework that combines SR and LR data will allow for more accurate isoform quantification than using SR data alone. Our goal is to develop and comprehensively benchmark a statistical method that quantifies transcript isoforms by joint analysis of SR and LR data. We use this method in the context of neuronal differentiation to identify isoform switch events that accompany this process. Neurons have a diverse alternative splicing landscape, which is known to contribute to neuron differentiation and cell fate determination [5]. Therefore, elucidation of the isoform landscape of neurons is important to understanding brain development and disease. Accordingly, the specific aims of this project are as follows:

Aim 1. Developing a new statistical framework to quantify isoforms by integrating SR and LR sequencing data: We have developed MPAQT (Multi-Platform Aggregation and Quantification of Transcripts), which integrates quantification information from SR and LR data using a novel statistical framework. Based on our benchmarking, MPAQT outperforms leading transcript quantification tools RSEM [16], salmon [18], and kallisto [19] on SR data alone, on SR+LR data,

and at the gene level for ~15k protein-coding genes with RT-qPCR measurements. We are excited by this discovery, as it can have broad applications to genomics research.

Aim 2. Application of this framework to neuronal differentiation cell lines to validate the method and identify isoform changes specific to different cell states: Human embryonic stem cells (hESCs) are a powerful model for studying cellular processes involved in development and differentiation. We study hESCs before and after differentiation into cortical neurons to gain a better picture of isoform switches during neuronal lineage differentiation. Using a combination of SR and LR sequencing, we quantify and identify isoforms expressed in immature neurons and in terminally differentiated cortical neurons *in vitro*, using MPAQT. This allows for an accurate map of isoform quantification and switching in neurons, which can potentially lead to identification of novel isoform-specific regulatory events involved in cellular fate determination in this lineage.

2 MATERIALS AND METHODS

2.1 MPAQT generative model

I start by discussing the concept behind MPAQT, our novel method for Multi-Platform Aggregation and Quantification of Transcripts. I then discuss the implementation of this model for analysis of SR RNA-seq data, as well as its expansion to jointly analyze both SR and LR data.

At the core of MPAQT is a generative model that connects the latent transcript abundances to the expected counts of "observation units" (OUs). Here, an observation unit is any entity that we can directly quantify from RNA-seq reads, and can be defined depending on the technology. For example, in long-read sequencing data, in which most reads can each be unambiguously assigned to one transcript, we can simply define each transcript as one OU, resulting in a one-to-one relationship between the transcripts and the OUs. In SR RNA-seq data, in which most reads can be mapped to multiple transcripts, the relationship between transcripts and OUs is more complicated. For example, we can define each equivalence class (EC) as one OU. In this case, each transcript may be connected to multiple OUs, and each OU may be connected to multiple transcripts. MPAQT's generative model explicitly connects the abundance of each transcript to the expected (mean) count of each OU.

Figure 2.1 schematically shows this generative model, demonstrating how, starting from the same counts for OUs, different transcription abundances can be inferred depending on the parameters of the underlying generative model. Therefore, to quantify the transcript abundances given the OU counts, we need to know the probability of observing a read that gets assigned to each OU, given a specific transcript as the origin of that read. We discuss in later sections how we obtained these probabilities for SR RNA-seq data (in which each OU is one equivalence class) and long-read RNA-seq data (in which each OU is one transcript).

Consider an RNA-seq dataset produced from sequencing a mixture of transcripts from the set T, with each transcript $t \in T$ having the relative abundance f_t ($\Sigma_{t \in T} f_t = 1$). Similar to previous works [19], we define the effective length l_t to be a transcript-specific normalization factor such that $f_t l_t = \mathrm{P}(t)$, where $\mathrm{P}(t)$ is the probability of selecting reads (or fragments) from transcript t ($\Sigma_{t \in T}$ $\mathrm{P}(t) = 1$). In other words, $\mathrm{P}(t)$ is the expected proportion of reads in the dataset that originated from transcript t.

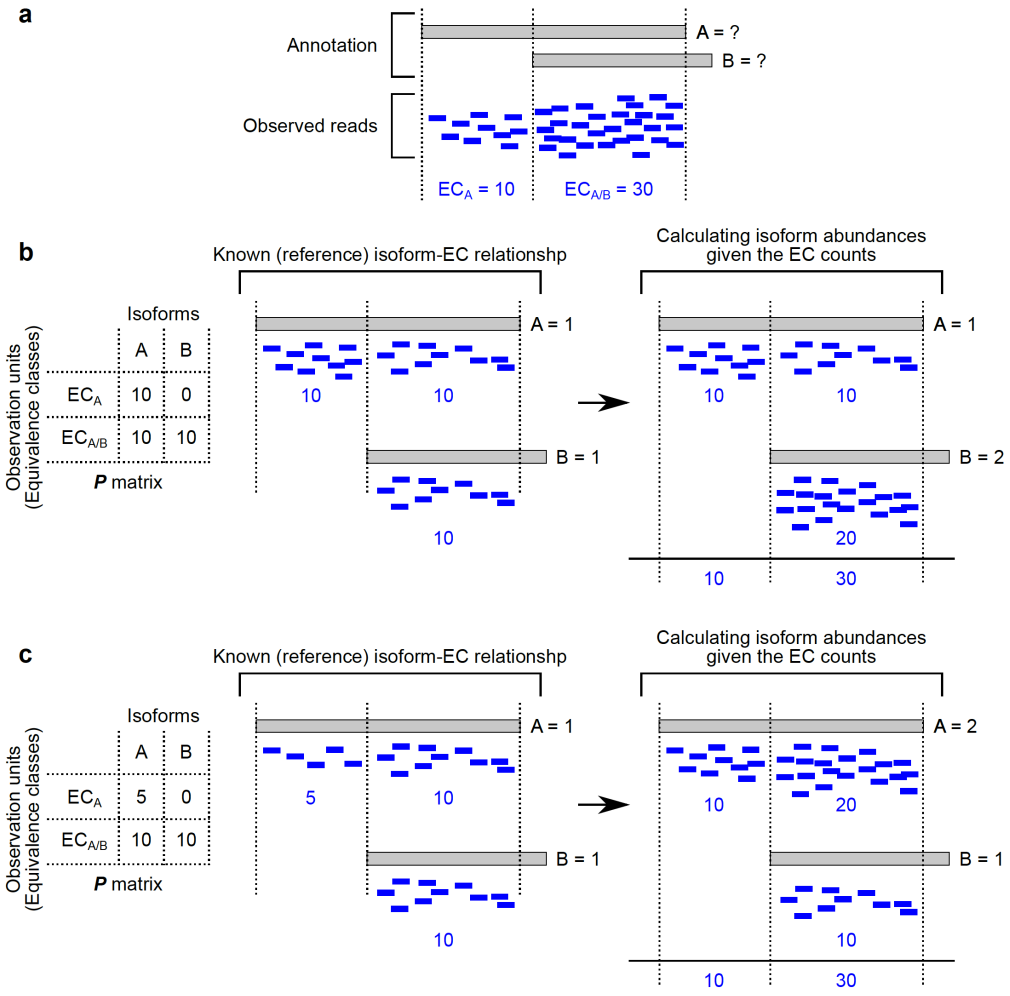


Figure 2.1: Outline of MPAQT's statistical framework. **(a)** Use of relationship between expected EC counts (provided in matrix **P**) and measured read counts to infer isoform abundance. This schematic shows a toy example of a cellular system with two transcripts, A and B, which can generate reads that are classified into two ECs, EC_A and $EC_{A/B}$. The observed counts for this sample are $EC_A = 10$ and $EC_{A/B} = 30$. The inferred transcript abundances depend on the parameters of the generative process, i.e., the expected proportion of reads generated by 1 unit of each transcript. **(b-c)** Examples of quantification inferences depending on the known reference isoform-EC relationships in matrix **P**.

Each read from this dataset is then assigned to one of the observation units from the set U. For a read that originated from a given transcript t, the probability of being assigned to a given observation unit $u \in U$ is represented by $P(u|t)$. In other words, $P(u|t)$ is the probability of a read mapping to u conditional on that read having been selected from t. It then follows that:

$$\mathrm{P}(u) = \sum_{t \in T} \mathrm{P}(t)\mathrm{P}(u|t) = \sum_{t \in T} f_t l_t \mathrm{P}(u|t)$$

In the above, P(u) is the probability of selecting reads that map to the observation unit u (i.e., the expected proportion of reads in the dataset that map to u; $\Sigma_{u \in U}$ P(u)=1).

After sequencing and read-OU assignments, what we observe for each OU is n_u ($n_u \in \mathbb{Z}_{\geq 0}$), i.e., the number of reads mapping to the observation unit u ($\Sigma_{u \in U} n_t = N$, where N is the total library size). Following previous work [33], conditional on the total library size N, we can assume a multinomial distribution for $\boldsymbol{n}=(n_1,\ldots,n_{|U|}) \in \mathbb{Z}_{\geq 0}^{|U|}$. However, given that N is often large, it is possible to approximate each n_u as an independent Poisson variable for computational tractability [33]:

$$n_u \sim \mathrm{Pois}(\lambda_u)$$

$$\lambda_u = \mathrm{P}(u) \times N = N \sum_{t \in T} f_t l_t \mathrm{P}(u|t)$$

The Poisson assumption is also consistent with early studies that demonstrated Poisson distribution of the technical variability in RNA-seq data [34].

Let's define $p_{u,t} = l_t \mathrm{P}(u|t)$. For each transcript t and each observation unit u, the value of $p_{u,t}$ is independent of the transcript abundances, and instead is a function of the transcript sequence, potential biases introduced by the sequencing technology, and potential biases/errors introduced by the process for read-OU mapping. We represent all the values $p_{u,t}$ for all observation units $u \in U$ and all transcripts $t \in T$ using the matrix $\boldsymbol{P} \in \mathbb{R}_{\geq 0}^{|U| \times |T|}$. The procedure for obtaining this matrix is described in later sections, but for now we consider the matrix $\boldsymbol{P}$ to be known. It follows that:

$$\boldsymbol{\lambda} = \left(\lambda_1, \ldots, \lambda_{|U|}\right)^{\mathsf{T}} = \boldsymbol{P}\boldsymbol{\beta}$$

$$\boldsymbol{\beta} = N\boldsymbol{f}$$

$$\boldsymbol{f} = \left(f_1, \ldots, f_{|T|}\right)^{\mathsf{T}}$$

Here, $\boldsymbol{\lambda} \in \mathbb{R}_{+}^{|U|}$ is the column vector of expected counts for observation units U, and $\boldsymbol{\beta} \in \mathbb{R}_{+}^{|T|}$ is the column vector of relative abundances for transcripts T, multiplied by the library size N. We further assume a multivariate log-normal prior distribution for $\boldsymbol{\beta}$, following previous body of work showing that, on the log scale, mRNA levels have a normal distribution (e.g., see [35])

$$\log \boldsymbol{\beta} \sim \mathcal{N}(\mu \mathbf{1}_T, \sigma^2 \boldsymbol{I})$$

Here, μ and σ^2 are hyperparameters that will be selected using an empirical Bayes approach, as discussed in later sections. $\mathbf{1}_T$ is a $|T|$-vector of 1's, and $\boldsymbol{I}$ is the $|T| \times |T|$ identity matrix. Our

objective is to infer the maximum a posteriori (MAP) estimate of $\boldsymbol{\beta}$, which can then be used to estimate the relative abundance (and TPM) of each transcript *t*:

$$\widehat{f_t} = \frac{\widehat{\beta_t}}{\sum_{t \in T} \widehat{\beta_t}}$$

$$\widehat{\mathrm{TPM}_t} = \widehat{f_t} \times 10^6$$

To summarize, the equations above form the following generative model, which we will fit to the observed OU counts:

$$\log \boldsymbol{\beta} \sim \mathcal{N}(\mu \mathbf{1}_T, \sigma^2 \boldsymbol{I})$$

$$\boldsymbol{\lambda} = \boldsymbol{P}\boldsymbol{\beta}$$

$$n_u \sim \mathrm{Pois}(\lambda_u)$$

2.2 Joint modeling of sequencing data from multiple platforms

Consider *K* sequencing datasets produced from the same RNA mixture (with the same transcript abundances) using different platforms (which can potentially include short and long-read sequencing technologies). The reads from each dataset *k* can be mapped to OUs in the dataset-specific set U_k ($k \in \{1,\ldots,K\}$). Therefore, the overall OU set *U* for the combination of the *K* datasets consists of the union of all subsets U_k:

$$U = \bigcup_{k=1}^{K} U_k$$

Each dataset *k* also has its own transcript-OU mapping probability matrix $\boldsymbol{P}_k \in \mathbb{R}_{\geq 0}^{|U_k| \times |T|}$, and its own library size N_k. However, since the datasets are generated from the same RNA mixture, they share the same vector of transcript abundances $\boldsymbol{f} \in \mathbb{R}_+^{|T|}$. Using the same procedure as in the previous section, we can see that the expected read count vector $\boldsymbol{\lambda} \in \mathbb{R}_+^{|U|}$ can be obtained as follows:

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{P}_1 N_1 \\ \vdots \\ \boldsymbol{P}_K N_K \end{bmatrix} \boldsymbol{f}$$

For compatibility with the previous section (and to be able to use the same prior as in the previous section), we can rewrite the equation above using $\boldsymbol{\beta}$ instead of $\boldsymbol{f}$, with $\boldsymbol{\beta}$ defined as $\boldsymbol{f}$ multiplied by the library size for the first dataset, N_1:

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{P}_1 \\ \boldsymbol{P}_2 s_1 \\ \vdots \\ \boldsymbol{P}_K s_K \end{bmatrix} \boldsymbol{\beta}$$

$$\boldsymbol{\beta} = N_1 \boldsymbol{f}$$

$$s_k = \frac{N_k}{N_1}$$

To summarize, we model the observed counts $\boldsymbol{n}$ (concatenated across the K datasets) as:

$\log \boldsymbol{\beta} \sim \mathcal{N}(\mu \mathbf{1}_T, \sigma^2 \boldsymbol{I})$

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{P}_1 \\ \boldsymbol{P}_2 s_1 \\ \vdots \\ \boldsymbol{P}_K s_K \end{bmatrix} \boldsymbol{\beta}$$

$$n_u \sim \text{Pois}(\lambda_u)$$

2.3 Obtaining the matrix *P* for short-read RNA-seq data

We obtain the matrix $\boldsymbol{P}$ by simulation from a reference transcriptome in which all transcripts have exactly equal abundances, using the Rsubread "simReads" function [36], followed by read-EC assignment using kallisto. First**,** the **kallisto bus** command is used to generate a bus file which contains the EC mappings for each read, and **bustools text** is used to convert it to .txt format. Downstream scripts are used to count reads assigned to each EC and to generate a matrix containing the transcript of origin of each read. These scripts are available at https://github.com/csglab/MPAQT.

To make sure that $\boldsymbol{P}$ accurately approximates the EC probabilities, we simulate 24 replicates of 100 million 75 base pair single-ended reads, for a total of 2.4 billion reads. Since each simulated read is tagged with its transcript of origin (t) and is mapped to a unique EC (u), we can calculate the proportion of reads that originate from transcript t and map to EC u, i.e., $p_{u,t}$. In practice, however, we do not calculate the proportions $p_{u,t}$, but instead directly use the read count $m_{u,t}$. Since $m_{u,t}$ is proportional to $p_{u,t}$, it only affects the scale of the inferred values of $\boldsymbol{\beta}$, which will be corrected when $\boldsymbol{\beta}$ is converted to TPM values.

2.4 The matrix *P* for long-read RNA-seq data

In the simplest scenario, where each long read is unambiguously assigned to one transcript (from the transcript set T) and read counts are proportional to the transcript abundances (i.e., no biases exist, such as length or GC-bias, among others), the matrix $\boldsymbol{P}$ for long-read sequencing data is simply the $|T|\times|T|$ identity matrix $\boldsymbol{I}$. However, both these assumptions can be violated in real-life applications. Particularly, as discussed in the Results section, we have found that substantial length and GC-biases exist in PacBio Sequel II data. Therefore, I will discuss here how these biases can be incorporated in the matrix $\boldsymbol{P}$.

Let's define the vector $\boldsymbol{p}$ to represent the main diagonal elements of matrix $\boldsymbol{P}$ for a long-read dataset, so that:

$$\boldsymbol{P} = \mathrm{diag}(\boldsymbol{p})$$

Our goal is to model $\boldsymbol{p}$ as a function of the covariate set D, representing potential sources of bias. These covariates (across the transcript set T) form the matrix $\boldsymbol{C}$ ($\boldsymbol{C}\in\mathbb{R}^{|T|\times|D|}$). We model $\boldsymbol{p}$ as:

$$\log \boldsymbol{p} = \boldsymbol{C\gamma}$$

Here, $\boldsymbol{\gamma}$ is a column vector of coefficients representing the effect of the covariates on the propensity of the transcripts to be captured by long-read sequencing ($\boldsymbol{\gamma}\in\mathbb{R}^{|D|}$). The log-link ensures that $\boldsymbol{p}$ is restricted to the domain $\mathbb{R}_{\geq 0}{}^{|T|}$. The MAP estimate of $\boldsymbol{\gamma}$ is obtained during model fitting. In other words, we jointly fit $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to the observed data.

We note that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are interdependent and together form an underspecified system. For example, we can easily see that:

$$\boldsymbol{\lambda} = \boldsymbol{P\beta} = \mathrm{diag}[\exp(\boldsymbol{C\gamma})]\boldsymbol{\beta} = \exp(\boldsymbol{C\gamma}) \circ \boldsymbol{\beta} = \boldsymbol{I}\exp(\boldsymbol{C\gamma}) \circ \boldsymbol{\beta}$$

Here, $\circ$ represents the Hadamard product, and 'exp' represents the element-wise exponential. The equation above suggests that, for example, we can replace $\boldsymbol{P}$ with the identity matrix $\boldsymbol{I}$, and multiply each element of $\boldsymbol{\beta}$ with the corresponding element of $\exp(\boldsymbol{C\gamma})$, without changing the expected OU counts that are predicted by the model. This redundancy can also be understood in terms of the difficulty in separating biological from technical sources of bias. For example, if we see higher read counts for high-GC transcripts, is it because these transcripts are truly expressed at higher levels, or is it a bias introduced by the sequencing procedure?

This issue, however, is alleviated when short-read RNA-seq data are combined with long-read RNA seq data: the matrix $\boldsymbol{P}$ for short-read RNA-seq data is pre-determined, allowing for a unique

MAP solution to be found for the model parameters even when $\boldsymbol{\gamma}$ (and, subsequently, the matrix $\boldsymbol{P}$ for long-read data) are a priori unknown and need to be inferred from data. This can be interpreted as short-read RNA-seq data providing an implicit reference against which MPAQT learns the sources of bias in the long-read RNA-seq data.

2.5 Inference of model parameters

2.5.1 Inferring $\boldsymbol{\beta}$

We start by showing how the maximum-likelihood estimate (MLE) of $\boldsymbol{\beta}$ can be obtained in the absence of a log-normal prior on $\boldsymbol{\beta}$. We will then show the inference of the maximum a posteriori (MAP) estimate of $\boldsymbol{\beta}$ with a log-normal prior. In the absence of a prior, we have:

$$\boldsymbol{\lambda} = \boldsymbol{P}\boldsymbol{\beta}$$
$$n_u \sim \mathrm{Pois}(\lambda_u)$$

The MLE can be obtained by minimizing the negative log-likelihood function:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{u \in U} (\lambda_u - n_u \log \lambda_u)$$

We use a sequential coordinate-wise descent (SCD) approach to iteratively solve each element β_t of $\boldsymbol{\beta}$ ($t \in T$). Consider the current estimate β_t^i, where i denotes the last iteration of the optimization algorithm. Let's assume that in the next iteration (i+1), the updated estimate for β_t will be different from the current estimate by $\delta^{[i+1]}$; i.e., the next estimate will be $\beta_t^{[i+1]} = \beta_t^{[i]} + \delta^{[i+1]}$. If the vector of the current predicted OU abundances is $\boldsymbol{\lambda}^{[i]}$, then the next set of predicted fragment abundances is given by:

$$\lambda_u^{[i+1]} = \lambda_u^{[i]} + \delta^{[i+1]} p_{u,t}$$

Therefore:

$$\beta_t^{[i+1]} = \beta_t^{[i]} + \delta^{[i+1]}$$
$$\delta^{[i+1]} = \arg\min_{\delta} \sum_{u \in U} \left[\lambda_u^{[i]} + \delta p_{u,t} - n_u \log\left(\lambda_u^{[i]} + \delta p_{u,t}\right)\right]$$

To solve for $\delta^{[i+1]}$, we take the derivative of the negative log-likelihood function with respect to δ and set it to zero:

$$\frac{d}{d\delta} \sum_{u \in U} \left[\lambda_u^i + \delta p_{u,t} - n_u \log\left(\lambda_u^i + \delta p_{u,t}\right)\right] = 0$$
$$\sum_{u \in U} p_{u,t} - \frac{n_u p_{u,t}}{\lambda_u^i + \delta p_{u,t}} = 0$$

To find the root of this function, we use Newton's method to update δ in each iteration:

$$f(\delta) = \sum_{u \in U} \left[p_{u,t} - \frac{n_u p_{u,t}}{\lambda_u^i + \delta p_{u,t}} \right] = \sum_{u \in U} p_{u,t} \left(1 - \frac{n_u p_{u,t}}{\lambda_u^i + \delta p_{u,t}} \right)$$

$$\delta^{[i+1]} = \delta^{[i]} - \frac{f\left(\delta^{[i]}\right)}{f'(\delta^{[i]})}$$

$$\delta^{[i+1]} = \delta^{[i]} - \frac{\sum_{u \in U} p_{u,t} \left(1 - \frac{n_u}{\lambda_u^i + \delta^{[i]} p_{u,t}} \right)}{\sum_{u \in U} n_u \left(\frac{p_{u,t}}{\lambda_u^i + \delta^{[i]} p_{u,t}} \right)^2}$$

Note that since $\delta^{[i+1]}$ is calculated with respect to the current value of $\beta_t^{[i]}$, then $\delta^{[i]}$ must also be calculated relative to $\beta_t^{[i]}$, which means that $\delta^{[i]} = \beta_t^{[i]} - \beta_t^{[i]} = 0$. Therefore:

$$\delta^{[i+1]} = - \frac{\sum_{u \in U} p_{u,t} \left(1 - \frac{n_u}{\lambda_u^i} \right)}{\sum_{u \in U} n_u \left(\frac{p_{u,t}}{\lambda_u^i} \right)^2}$$

This provides an iterative procedure where each β_t is updated by adding the value of $\delta^{[i+1]}$ from the equation above, followed by updating each $\lambda_u^{[i]}$ (for $u \in U$) by adding $\delta^{[i+1]} p_{u,t}$ to it.

Now, we will modify the equations above to show how the MAP estimate of $\boldsymbol{\beta}$ can be obtained when a log-normal prior is placed on $\boldsymbol{\beta}$:

$$\log \boldsymbol{\beta} \sim \mathcal{N}(\mu \mathbf{1}_T, \sigma^2 \boldsymbol{I})$$

$$\boldsymbol{\lambda} = \boldsymbol{P\beta}$$

$$n_u \sim \text{Pois}(\lambda_u)$$

In this case, the negative log-likelihood function also includes a regularization term that acts to shrink the logarithm of each β_t toward μ:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2\sigma^2} \sum_{t \in T} (\log \beta_t - \mu)^2 + \sum_{u \in U} (\lambda_u - n_u \log \lambda_u) \right]$$

Following the same method as above, we can see that for each transcript t, its abundance in iteration i+1 can be updated as:

$$\beta_t^{[i+1]} = \beta_t^{[i]} + \delta^{[i+1]}$$

$$\delta^{[i+1]} = \arg\min_{\delta} \left[\frac{1}{2\sigma^2} \left[\log\left(\beta_t^{[i]} + \delta \right) - \mu \right]^2 + \sum_{u \in U} \left[\lambda_u^{[i]} + \delta p_{u,t} - n_u \log\left(\lambda_u^{[i]} + \delta p_{u,t} \right) \right] \right]$$

Again, using Newton's method and following the same method as above, we can see that:

$$\delta^{[i+1]} = -\frac{\frac{\log \beta_t^{[i]} - \mu}{\sigma^2 \beta_t^{[i]}} + \sum_{u \in U} p_{u,t} \left(1 - \frac{n_u}{\lambda_u^i}\right)}{\frac{-\log \beta_t^{[i]} + \mu + 1}{\sigma^2 \left(\beta_t^{[i]}\right)^2} + \sum_{u \in U} n_u \left(\frac{p_{u,t}}{\lambda_u^i}\right)^2}$$

In practice, we have found that in the presence of a log-normal prior, Newton's method occasionally overshoots for some transcripts in some of the early iterations of the optimization algorithm. We detect such overshoot events by examining whether $\delta^{[i+1]}$, as calculated by Newtons's method using the equation above, is outside the range between the value obtained from the MLE estimate and the value that would make the logarithm of $\beta_t^{[i+1]}$ equal to the mean of the prior. When this occurs, we minimize the negative log-likelihood function using the 'optimize' function in R, which uses "a combination of golden section search and successive parabolic interpolation" [37]. This procedure, in practice, resolves the overshoot problem in a few iterations, so that in the subsequent iterations Newton's method can be used without any overshoots.

2.5.2 The prior distribution for $\boldsymbol{\beta}$

We use an adaptive prior, which is iteratively updated based on the distribution of all values β_t (for all $t \in T$). In other words, after each iteration i, we update the prior mean μ to be the mean of $\log(\beta)$, and prior variance σ^2 to be the variance of $\log(\beta)$.

2.5.3 The sources of bias in long-read RNA-seq data ($\boldsymbol{\gamma}$)

As discussed in section 2.4, we model the submatrix $\boldsymbol{P}_k$ matrix for long-read data as:

$$\boldsymbol{P}_k = \mathrm{diag}[\exp(\boldsymbol{C\gamma})]$$

where k is the index of the dataset containing long-read counts, and $\boldsymbol{C}$ is a $|T| \times |D|$ matrix representing the value of variables D, the potential sources of bias, across $|T|$ transcripts. Therefore, we can model the OU counts observed in dataset k as:

$$\lambda_{u \in U_k} = \beta_{t(u)} \sum_{d \in D} \exp\left(c_{u,d} \gamma_d\right)$$

$$n_{u \in U_k} \sim \mathrm{Pois}\left(\lambda_{u \in U_k}\right)$$

Here, $t(u)$ represents the transcript that corresponds to OU u (note that there is a one-to-one relationship between OUs in the long-read data and the transcripts). $c_{u,d}$ represents the element (u,d) of matrix $\boldsymbol{C}$. At each iteration i, we update $\boldsymbol{\gamma}$ by maximizing the model likelihood using the 'glm' function in R with a Poisson error distribution and log-link, with the long-read counts as the dependent variable, $\boldsymbol{C}$ as the independent variables, and $\beta^{[i]}$ as the offset.

2.6 Cortical neuron differentiation and RNA-seq data generation

2.6.1 Cortical neuron differentiation

The hESC SOX10::GFP bacterial artificial chromosome reporter line (in the H9 background) was used for neural differentiation according to the protocol adapted from a study of brain organoids [38]. In brief, the hESC line was maintained in feeder-free conditions with the E8 medium. Neural differentiation was initiated when the cells reached 90-100% confluency. From days 0-11, the cells were maintained in neural induction medium (10 μM SB431542 and 100 nM LDN193189 in E6 medium) with medium change every two days. From day 12, the cells were fed the cortical neuron medium (10 ng/mL GDNF, 100 μM ascorbic acid, 1x Glutagro, 1x N2 supplement, 1x B27 without vitamin A in neurobasal medium) with medium change every other day until rosette structures became visible. Then, neurons were detached using Accutase and replated on poly-L-ornithin/fibronection/laminin-coated plates. Neurons were maintained in cortical neuron medium with medium change every other day. On days 22-24, neurons were checked for the presence of axonal projections and 10 μM DAPT was included in the cortical neuron medium until the projections appeared. From day 30, neurons were considered mature with the medium feeding frequency reduced to 1-2 times per week.

2.6.2 RNA extraction, short-/long-read RNA-seq library prep, and sequencing

Cells were harvested at days 0, 41, and 61, followed by RNA extraction using Zymo Quick-RNA Microprep kit according to the manufacturer's protocol. SIRV set 4 (Lexogen) was spiked at 1% in the hESC and differentiated neuron-derived RNA samples. Short-read RNA-seq libraries were prepared using the SMARTer Stranded Total RNA-Seq Kit v3. Libraries were sequenced on a NextSeq 550 sequencer (2x75 bp paired-end). PacBio Iso-seq libraries from the same RNA samples were generated using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module, PacBio Iso-Seq Express Oligo Kit and SMRTbell express template prep kit 2.0 according to the manufacturer's protocol. The libraries were sequenced on a PacBio Sequel IIe.

2.7 Processing of short-read RNA-seq data

All SR data, including those generated from the neuronal differentiation model (above), those obtained from publicly available data, and simulated data (below) were processed using RSEM

[16] (version 1.3.3, following aligment with bowtie2 version 2.4.2), salmon [18] (version 1.3.0, with the --validateMappings and --gcBias flags), kallisto [19] (version 0.48.0) and MPAQT.

The MAQC data was taken from GEO accession GSE83402, and single-end samples MAQCA_1 (4 technical replicates: SRR3670977, SRR3670978, SRR3670979, SRR3670980) and MAQCB_1 (4 technical replicates: SRR3670985, SRR3670986, SRR3670987, SRR3670988) were processed with the above SR quantification tools (RSEM, salmon, kallisto and MPAQT). Technical replicates for each sample were combined during quantification. Differential expression was calculated as the logarithm of fold-change (logFC) between MAQCB_1 and MAQCA_1.

Simulated datasets for benchmarking, in the form of paired-end FASTQ files, were generated from ground truth TPM values using the simReads function from the Rsubread R package [36]. Rsubread takes as input the number of reads to simulate and a list of transcripts with their desired TPMs. We generated two simulated datasets using two different sets of ground truth TPMs. For the first dataset, ground truth TPMs were sampled from an exponential distribution (using rexp R function with default rate=1). For the second dataset, we first used kallisto to quantify transcript abundances from RNA-seq data of the MDA-MB-231 cancer cell line (GEO entries GSM4886854, GSM4886855) [39] and then used the resulting TPMs as the ground truth for read simulation. Since kallisto was used to generate the truth TPMs, and since kallisto predicts 0 TPM for many low-abundance transcripts, a pseudocount of 0.05 (basically equivalent to zero expression) is added to truth TPMs to prevent outliers dominating the correlations. For the rexp.sim dataset, three simulated “replicates” were generated, and one sample was generated for the MDA-MB-231.SIM dataset, each with 30 million paired-end reads of 75 bp. These samples were processed with RSEM, salmon, kallisto and MPAQT, as described above.

SR sequencing data for the neuronal differentiation samples were processed using the paired-end options for above tools.

2.8 Processing of long-read RNA-seq data

2.8.1 ISOSEQ

The IsoSeq pipeline (Pacific Biosciences) was used to process the neuronal differentiation LR data and generate circular consensus sequence (CCS) reads, which were stored in uBAM (unaligned BAM) format. Next, lima (PacBio) was used to remove primer sequences. IsoSeq3 ‘refine’ command was used to remove poly-A tails and concatemers (reads which are attached

end-to-end), followed by the ‘cluster’ command to cluster reads that represent the same transcript (i.e. make them adjacent). The ‘align’ command of pbmm2 (PacBio) was then used to align reads to the reference genome, followed by the Isoseq3 ‘collapse’ command to condense the data into a transcriptome (fasta + GFF) and provide an abundance file containing full length counts (FL counts).

2.8.2 SQANTI

The quality control script was used (sqanti3_qc.py) using SQANTI3 with supporting data types (CAGE peak, polyA motif list, polyA peaks file, and Intropolis splice junctions), removing low-quality transcripts according to SQANTI’s quality criteria. Next, the rules filter script (sqanti3_RulesFilter.py) was run to further filter transcripts based on the following criteria: If a transcript is a full splice match (FSM), then it is kept unless the 3' end is unreliable (intrapriming). If a transcript is not a FSM, then it is kept only if all of below are true: (1) 3' end is reliable; (2) the transcript does not have a junction that is labeled as RT Switching; and (3) all junctions are canonical. Finally, the full-length (FL) LR counts from SQANTI3 output that had the same “associated_transcript” were combined, providing transcript counts for input to MPAQT and for use in benchmarking.

2.9 Differential expression analysis of neuronal differentiation samples

First, the STAR aligner (version 2.5.0c) [40] was used to generate the “ReadsPerGene.out.tab” file, which gives us read counts per gene. Then, data was cleaned by filtering out genes with low counts, since low counts are an unreliable source of gene expression quantification information. Specifically, counts per million (CPM) was calculated using edgeR [41], and a gene was kept if at least 3 samples have a CPM > 1. GENCODE identifiers were mapped to HGNC symbols and gene descriptions using the biomaRt R package [42].

Next, data was normalized using trimmed mean of M values (TMM) normalization, which estimates scale factors between samples, allowing for us to have relative RNA levels from our RNA-seq data [43]. This was done using the calcNormFactors function from edgeR [41], converting raw library sizes into effective library sizes.

Prior to DE analysis, the CPM data is log2 transformed. Then, the lmFit [44] function from the limma [45] package is used to fit a linear model for each gene. Then, the eBayes function (limma) [44] is used to run the empirical Bayes method, which computes DE, with trend=TRUE

set to specify RNAseq data. Then, the Benjamini-Hochberg (BH) adjustment method is used to adjust the p-values to decrease the false discovery rate.

2.10 Term enrichment for DE analysis of neuronal differentiation samples

After separating genes into three groups (upregulated, downregulated, and not-DE), we used these groups for functional enrichment analysis. We identified enriched terms using g:Profiler through the gprofiler2 R package [46], retaining highly significant ($p<5\times10^{-5}$) and small terms ($n<500$), which are referred to as "filtered terms" from this point onwards. I performed term size filtering because terms with high number of genes are often less specific in terms of the biological process implicated. To quantify proportions of enriched terms related to neuronal differentiation, counts of neuron-related terms were calculated for each enrichment analysis and for each of "up", "down" and "total" gene sets (**Supplementary Table 1**). Terms that matched the strings "synap", "neur", "axon", and "dendr" were considered neuron-related.

2.11 Spike-ins

We used spike-ins to benchmark MPAQT.SR's performance against kallisto and salmon. Since the reference genome is needed to run these tools, and since spike-ins are synthetic RNAs, they needed to be added to the reference genome. Each spike-in was added in as its own separate chromosome, and 1000 base pairs of "N" spacer nucleotides were added on either side of the spike-in sequence. This allowed for the pseudo-aligners to map reads to the spike-in sequences provided in the Lexogen FASTA file. We used the SIRV-Set 4 from Lexogen (https://www.lexogen.com/sirvs/sirv-sets/) , which contains 114 spike-in transcripts. We used 107 in this analysis, since SIRV 403-410 were not included in reference FASTA provided by Lexogen.

2.12 Reference transcriptome and genome version

For all analyses, reference transcriptome and genome annotations from GENCODE [47] v38 was used, corresponding to human genome assembly GRCh38.p13.

2.13 Data and code availability

All processed data generated as part of this study are available at https://github.com/csglab/MPAQT/tree/main/data . Raw data will be deposited in GEO. MPAQT is available at https://github.com/csglab/MPAQT, including the code to create the "index" matrix ***P*** for short-read RNA-seq data, pre-built indexes for GENCODE v38, as well as the scripts for statistical analysis and joint quantification of short- and long-read RNA-seq data.

3 RESULTS

3.1 Benchmarking: gene-level quantification

MPAQT can perform transcript quantification using either SR data alone or SR+LR data. To examine whether MPAQT might be broadly applicable to gene expression quantification, we began by benchmarking its performance against that of three leading SR analysis tools, salmon [18], kallisto [19], and RSEM [16], for gene-level quantification using SR data alone. For this purpose, we used data from the RNA Sequencing Quality Control (SEQC) project [48] led by the MicroArray Quality Control (MAQC) consortium. Our dataset consists of single-end RNA-seq data for two MAQC samples [49]: MAQCA (Universal Human Reference RNA, pool of 10 cell lines) and MAQCB (Human Brain Reference RNA). Each MAQC dataset is accompanied by RT-qPCR expression measurements for 18,080 protein-coding genes in the form of Cq-values (representing the number of PCR cycles before a signal is seen for a given gene; higher Cq-values correspond to lower abundances). Of these, 14,956 genes have Cq-values between 11 and 32, a range deemed reliable in the original report [49]. For this subset, the ground truth differential expression (DE) was calculated as the difference of RT-qPCR Cq-values between MAQCA and MAQCB (representing $\log_2$ fold-change of expression), which was then compared to log fold-change of TPM (transcripts per million) calculated from SR RNA-seq data by different tools.

We observed excellent agreement between log fold-changes inferred by MPAQT and the ground truth (Pearson r=0.91, **Figure 3.1**a). In contrast, for all other existing tools, we saw distinct outliers which are visibly separated from the remainder of the data points (**Figure 3.1**a). **Figure 3.1**b shows kallisto's outliers, isolated from other, well-behaving genes. Interestingly, MPAQT preserves differential expression information for these outlier genes (**Figure 3.1**b). A similar trend is observed for outliers from salmon and RSEM (**Supplementary Figure 1**). These outliers represent ~2-3% of genes in our data, with 154 outliers shared among kallisto, salmon, and RSEM (**Figure 3.1**c). The outliers correspond to genes with higher mean and median Cq-values (**Figure 3.1**d), suggesting they are transcripts with low expression.

Additionally, an extended filtering range was used (Cq-value between 8-35) to assess MPAQT's performance on noisier qPCR measurements, allowing for an additional 1148 genes to be included in the analysis (**Supplementary Figure 2**). Surprisingly, MPAQT's performance remains comparable to the more conservative filtering range (Cq-value between 11-32) used in the original report [49], whereas the number of outliers for the other SR tools approximately double

(**Supplementary Figure 2**). These findings suggest that what was previously perceived as RT-qPCR noise due to low correlation with RNA-seq-based measurements [49] may in fact be due to poor performance of SR quantification tools, and demonstrate MPAQT's ability to provide more accurate gene-level quantification, particularly for genes with low expression .

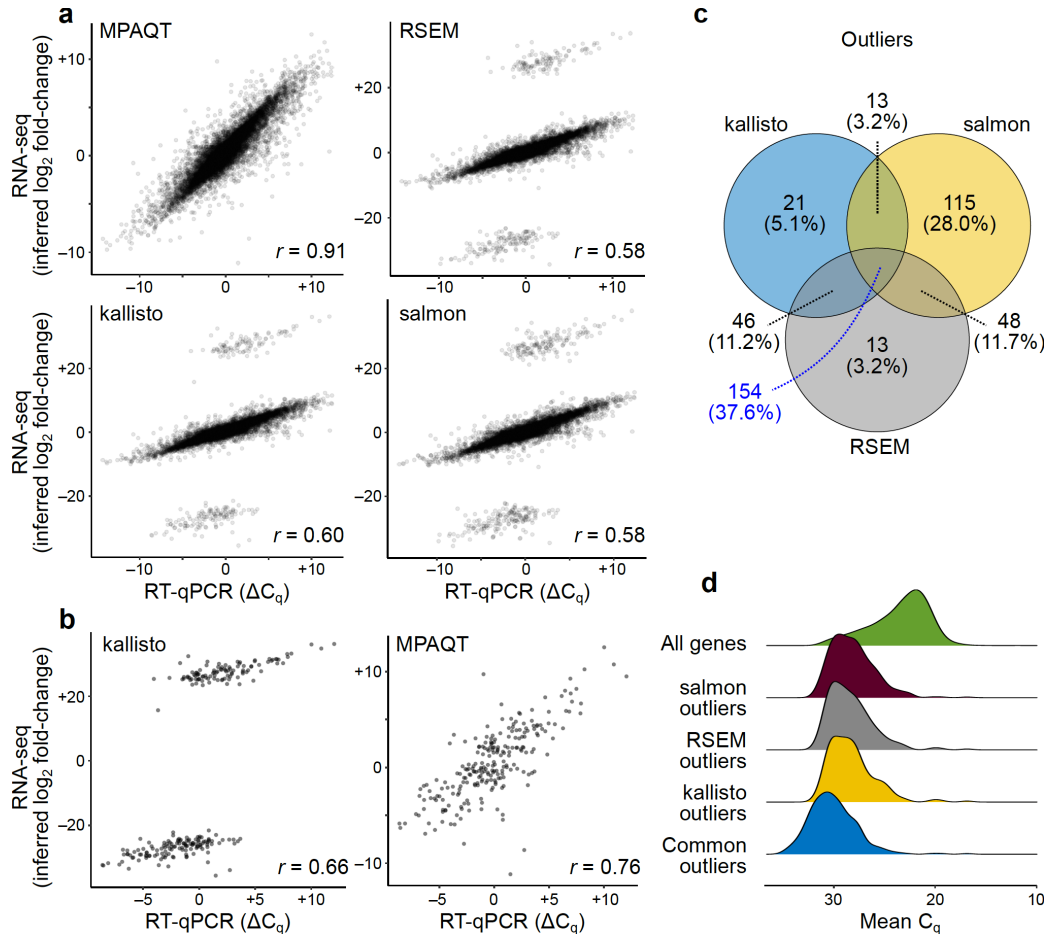


Figure 3.1: Benchmarking the performance of MPAQT for inference of gene-level abundances from SR data. **(a)** Inferred logFC, calculated from TPM predictions by MPAQT, RSEM, kallisto, and salmon, plotted against qPCR difference. **(b)** Left: kallisto's outliers, isolated from other genes. Right: MPAQT's performance on kallisto's outliers (234 genes). **(c)** Venn diagram showing overlap of outliers among kallisto, salmon, and RSEM. **(d)** Density curves of mean of MAQCA and MAQCB Cq-values for each tool's outliers and for all genes. For benchmarking results with the expanded filtering range, see **Supplementary Figure 2** .

3.2 Benchmarking: isoform-level quantification

Since no real datasets exist for benchmarking quantification at the isoform level (potentially due to technical difficulties of isoform-specific RT-qPCR analysis), we used simulated data to benchmark the ability of MPAQT and other existing tools for isoform-level quantification, starting with SR data alone. As described in the Methods section, we used two different simulated datasets, one with ground truth TPM values sampled randomly from an exponential distribution, and the other with ground truth TPM values modeled after measurements from real RNA-seq data. In both simulations, we observed that MPAQT substantially outperforms the other tools in terms of Pearson correlation and Root Mean Square Deviation (RMSD) (**Figure 3.2** and **Table 3.1**). The amount of variance in the ground truth log-TPMs that is captured by MPAQT (i.e., R^2 between MPAQT inferences and ground truth) is ~13%-32% higher than the next best method (13% for the data simulated for TPMs that are modeled after real RNA-seq measurements, and 32% for the data simulated for TPMs that are exponentially distributed).

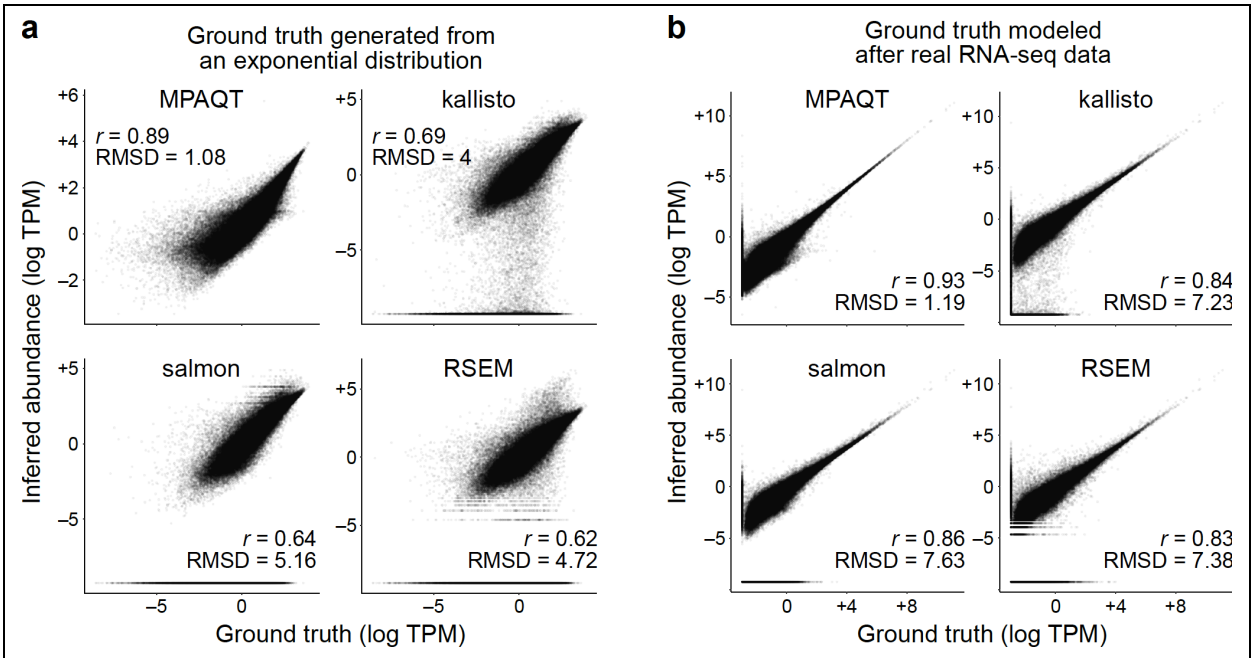


Figure 3.2: Benchmarking the performance of MPAQT and three other tools at the transcript level using two types of simulated samples. **(a)** Simulated samples with TPMs generated from an exponential distribution. **(b)** A simulated sample modeled after real sample MDA-MB-231, generated using TPMs from kallisto.

Table 3.1 Table of RMSD and Pearson correlation values for MPAQT and three other tools, based on three replicates of RNA-seq data simulated from an exponential distribution. The mean of the three replicates of the test statistic is shown, with the standard error of mean in brackets.

Name	RMSD	Pearson *r*
kallisto (SR)	4.00 (1.73×10^{-2})	0.69 (1.70×10^{-3})
salmon (SR)	5.17 (3.33×10^{-3})	0.64 (1×10^{-3})
RSEM (SR)	4.71 (3.33×10^{-3})	0.63 (1.36×10^{-3})
MPAQT (SR)	1.08 ($<1\times10^{-5}$)	0.89 (4.67×10^{-4})
MPAQT (LR)	1.04 (3.33×10^{-3})	0.90 (4.18×10^{-4})

Together, these simulation experiments suggest that, even without LR data, MPAQT outperforms kallisto, salmon and RSEM in transcript quantification from SR data alone (**Figure 3.2**).

Next, we set out to examine whether LR data can further improve MPAQT's estimates of transcript abundances. Again, due to lack of appropriate real data for benchmarking, we used simulated data. We used the same ground truth TPM sets that we created for SR data simulation, and generated a simulated LR dataset for each replicate in the form of transcript counts sampled from independent Poisson distributions, with the ground truth TPM of each transcript as the mean of its Poisson distribution (adjusted to obtain ~200K transcript counts per sample). Then, the combination of simulated SR data (from above) and LR counts was used as input to MPAQT.

We observed a small improvement in MPAQT's overall performance when simulated LR data were included in the analysis, compared to inference from SR data alone (**Figure 3.3**a). Due to the low coverage of simulated LR data, the small overall improvement in performance is expected. However, when MPAQT's inferences from SR data alone are directly compared to those from SR+LR data, we can identify a subset of transcripts whose quantified abundances differ substantially between the two measurements (**Figure 3.3**b). For this subset, we see substantial improvement in SR+LR data in terms of agreement with ground truth (Pearson correlation 0.67-0.71 for SR+LR, compared to –0.26 to –0.18 for SR data alone, **Figure 3.3**b and **Supplementary Figure 3**a). The next section provides a more detailed description of this subset of transcripts.

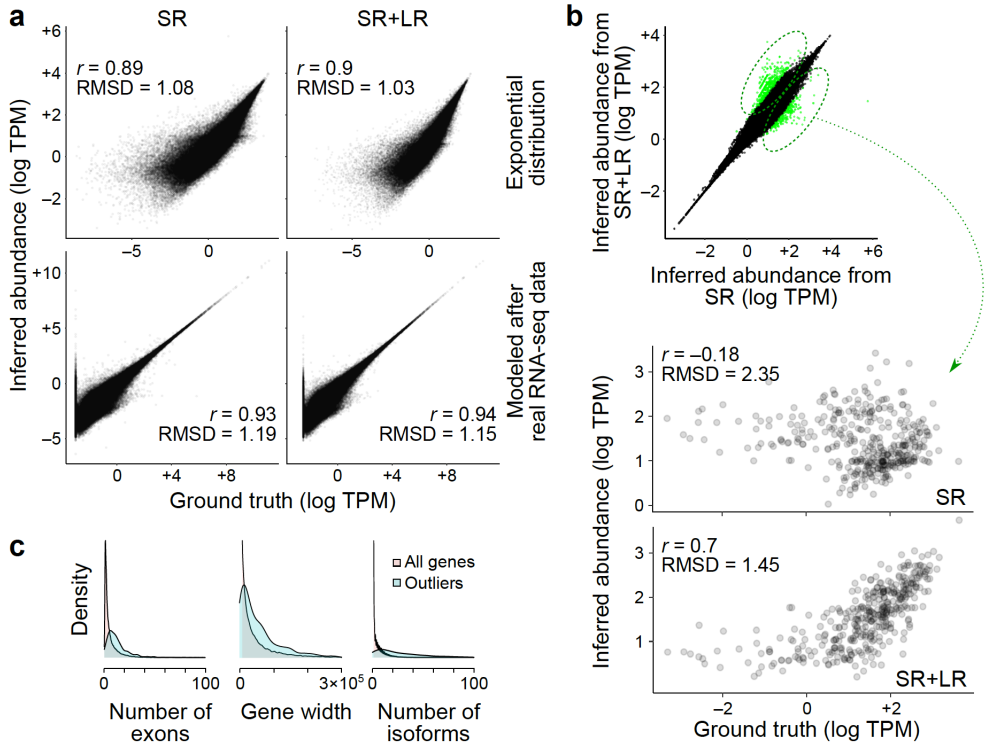


Figure 3.3: MPAQT quantifications improve with addition of LR data for a subset of transcripts with complex splicing patterns. **(a)** Performance of MPAQT using SR data alone (left) or SR+LR data (right), on simulated data generated from ground truth TPMs with an exponential distribution (top) or modeled after real RNA-seq data (bottom). **(b)** Top: Comparison of SR vs. SR+LR inferences for data simulated from an exponential distribution. Transcripts with significantly different inferences are highlighted (outlier analysis based on Mahalanobis distance, with p-value cutoff 10^{-10}). Bottom: Comparison of inferred vs. ground truth TPM for transcripts that are significantly differentially quantified (i.e., transcripts highlighted in the top panel). **(c)** Distributions of number of exons, gene width, and isoform numbers for genes encoding the transcripts that are differentially quantified between SR and SR+LR analyses. The distributions for all genes are also shown, for comparison.

3.3 Investigation of transcripts with improved LR-based quantification

As mentioned above, for a subset of transcripts, we observed substantial differences between SR-only and SR+LR quantifications. Specifically, after removing short RNAs (i.e., transcripts shorter than 250bp, including miRNAs, Y RNAs, snRNAs and snoRNAs), which are unlikely to be captured by either SR or LR RNA-seq in real-life applications, we identified 490 transcripts for which inclusion of LR data had a significant effect in at least one simulated replicate (**Supplementary Figure 3**). In all replicates, the MPAQT inferences that utilized both SR and LR

data correlated strongly with the ground truth for these transcripts, whereas SR-only inferred abundances showed a weak, negative correlation with ground truth (**Supplementary Figure 3**).

We observed a significant overlap among the three replicates for these transcripts (**Supplementary Figure 3**b), with ~30% identified in more than one replicate, suggesting that these transcripts may have intrinsic features that make them sensitive to the presence/lack of LR data. Interestingly, we found that the genes encoding these transcripts have more exons, larger widths, and more isoforms (**Figure 3.3**e) than the total set of genes. Furthermore, pathway enrichment analysis revealed that the GO Biological Process "nervous system development" is recurrently enriched among the LR-sensitive genes in all replicates ($P=1.8\times10^{-5}$, 5.4×10^{-5}, and 6.9×10^{-5} for the three replicates, respectively, based on g:Profiler [46]). This finding is consistent with previous reports showing that that genes preferentially expressed in the nervous system tend to be longer, have more exons, and exhibit more complex splicing patterns compared to other tissues [3, 50]. These findings suggest that joint analysis of SR and LR data using MPAQT will be particularly impactful in investigating isoforms involved in nervous system development, as they are inaccurately quantified with SR data alone.

3.4 Measuring transcript isoform abundances during neuronal differentiation

The analyses presented above suggest that transcriptome profiling using a combination of SR and LR sequencing can substantially improve isoform quantification for genes related to neuronal differentiation. Therefore, to investigate the landscape of differential isoform usage during neuronal cell differentiation, we analyzed human embryonic stem cells (hESCs) undergoing *in vitro* differentiation toward cortical neurons (**Supplementary Figure 4**), by joint SR and LR RNA-seq from cells collected at days 0, 41, and 61 since the start of growth in neural induction medium (see Methods for details).

To confirm that neuronal differentiation is occurring as expected, we started by differential expression analysis using SR data, at the gene level, between days 0 and 41, and also between days 41 and 61 (see Methods for details of DE analysis). Between days 0 and 41, we observed a large number of downregulated genes (12,000), whereas only 1,828 genes were upregulated (FDR≤0.05, $|\log_2$ fold-change$|\geq0.6$). This observation is in line with undifferentiated hESCs reducing expression of many genes as they undergo differentiation into the more specialized neuronal cells. By contrast, we see roughly an equal number of genes upregulated and downregulated between days 41 and 61 (1805 and 1865, respectively; **Supplementary Data Table 1**). Furthermore,

functional enrichment analysis confirmed the enrichment of neuron-related terms among genes up-regulated between days 0 and 41 as well as genes up-regulated between days 41-61. Specifically, about ~45% of up-regulated terms are neuron-related, compared to only ~2% of down-regulated terms (**Supplementary Table 1**).

Next, to further evaluate the performance of MPAQT in isoform quantification, we used it to analyze the SR data of each sample. Two lines of evidence from this analysis further support the superior performance of MPAQT compared to other tools. First, MPAQT's SR-based quantifications show a slight but reproducible improvement, compared to quantifications provided by other tools, for synthetic mRNAs that were spiked in the samples at known concentrations (**Figure 3.4**a). Secondly, MPAQT's SR-based quantifications of gene isoforms are more consistent with full-length LR counts obtained from the same sample, compared to SR-based quantifications provided by other tools (**Figure 3.4**b and **Table 3.2**).

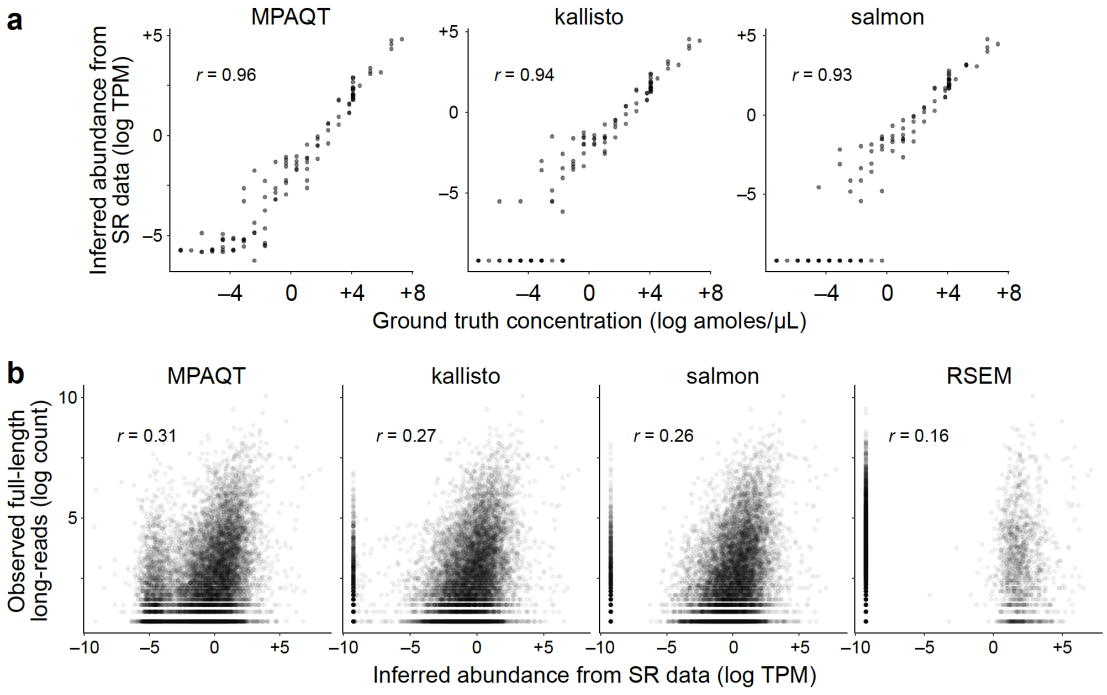


Figure 3.4: Benchmarking of MPAQT and other tools using two internal controls in neuronal differentiation samples: LR counts and spike-in RNAs. **(a)** Performance of MPAQT, kallisto, and salmon based on 107 spike-in transcripts as truth measurements across all six neuronal differentiation samples. Each dot represents one spike-in RNA in one of six samples. **(b)** Comparison of SR-based abundances and full-length LR counts for genomic transcripts. Each data points represents one transcript in one of six neuronal differentiation samples. Data points with LR count of zero were removed to allow log-scale plotting of the y-axis.

Table 3.2: Pearson correlation between log(LR) counts and log(TPM) for MPAQT, kallisto, salmon and RSEM across the six neuronal differentiation samples.

Sample	MPAQT	kallisto	salmon	RSEM
Day 0 – rep 1	0.2706	0.2389	0.2338	0.1709
Day 0 – rep 2	0.3088	0.2605	0.2628	0.1771
Day 41 – rep 1	0.2706	0.2384	0.2281	0.1532
Day 41 – rep2	0.3137	0.2706	0.2598	0.1601
Day 61 – rep1	0.3125	0.2735	0.2646	0.1498
Day 61 – rep2	0.2888	0.2442	0.2323	0.1269

We note, however, that we generally see a low correlation between SR quantifications and LR counts (**Table 3.2**), suggesting the presence of potential biases in LR RNA-seq data. To investigate such biases, we asked whether the deviation between LR and SR data can be explained by transcript length, transcript GC content, or the transcript biotype. As shown in **Figure 3.5**a, all three variables showed a significant effect on LR counts, with longer transcripts, GC-poor transcripts, and protein-coding mRNAs more likely to be captured by LR sequencing. Subsequently, once we account for the biases introduced by these factors, we see a significant improvement in the agreement between SR and LR data (**Figure 3.5**a). Importantly, the same biases can be replicated when we compare the LR counts of spike-in RNAs to their ground truth concentrations (**Figure 3.5**b), including the significant enrichment of long and GC-poor transcripts in LR sequencing (the effect of transcript biotype cannot be modeled for synthetic RNAs). We note, however, that the LR biases might be experiment-, instrument-, and/or protocol-specific. Therefore, as described in the Methods section, we included the ability to account for LR biases in the MPAQT statistical model, allowing it to learn sample-specific sources of bias and incorporate them in its framework for integration of LR and SR data.

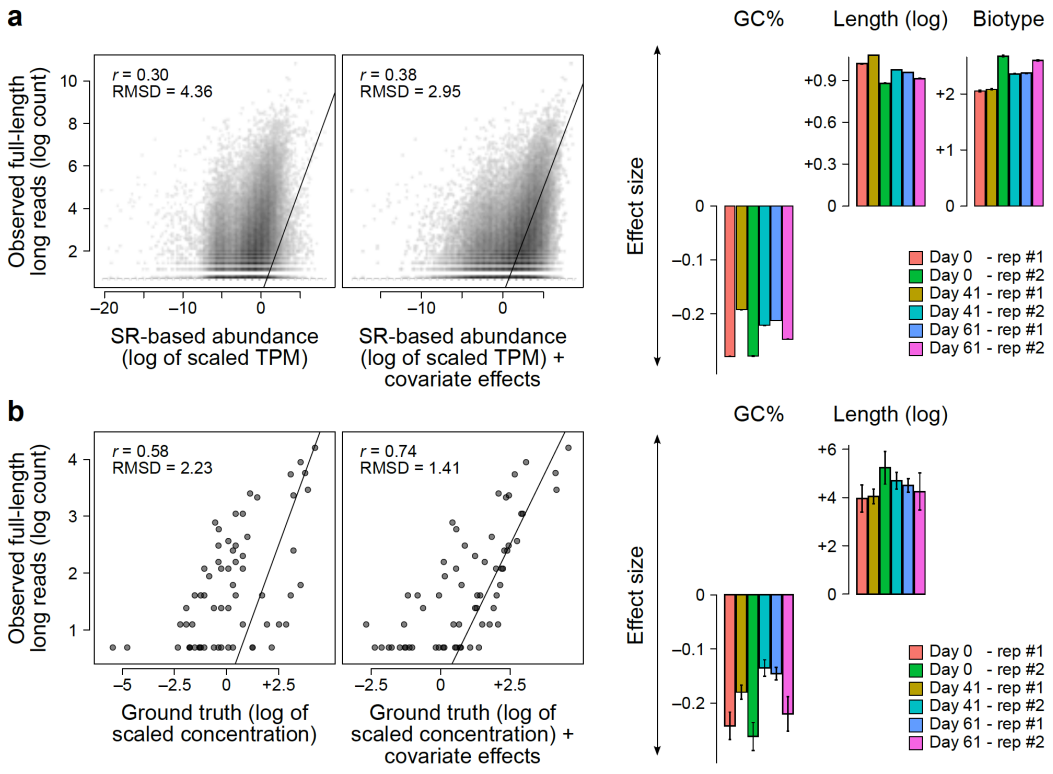


Figure 3.5 Full-length long-read counts are affected by transcript properties. **(a)** Left: The scatterplots show the relationship between full-length LR count and SR-based abundance (left) or SR-based abundance after adding the effect of covariates (transcript GC%, length, and biotype). The SR-based abundances are scaled separately for each sample and each model to maximize the likelihood of LR counts. Each dot in each scatterplot shows one transcript in one sample (six samples combined in each plot). Right: The estimated effect size of each covariate on LR counts. **(b)** Similar to (a), showing the relationship between LR counts and ground truth concentration for spiked-in RNA controls (combination of six samples), and the variable coefficients for models that are fitted to spiked-in RNA LR counts.

3.5 MPAQT quantification of isoforms involved in neuronal differentiation by combined analysis of SR and LR data

We used MPAQT to analyze the LR and SR data obtained from neuronal differentiation samples at days 0, 41, and 61. We found 6309 transcripts whose inferred abundances based on joint analysis of LR and SR data deviated substantially from abundances inferred from SR data in at least one of the three time points (Mahalanobis distance p-value < 0.01). Even at a substantially stricter cutoff (Mahalanobis distance p-value < 10^{-10}), there were still 2459 transcripts whose

LR+SR and SR-only quantifications differed significantly in at least one time point (**Figure 3.6**a). These transcripts were from genes with larger gene widths, more exons, and more isoforms (**Supplementary Figure 6**), similar to the transcripts that, in our simulations (**Supplementary Figure 3**), could be quantified more accurately using a combination of LR and SR data. Although at this stage we have no ground truth TPM measurements for these transcripts in our neuronal differentiation samples, their similar properties to the LR-sensitive transcripts identified in our simulations suggest that they may represent transcripts with improved quantification after inclusion of LR data.

As expected, the transcripts with potentially improved quantification are enriched for genes specifically expressed in human prefrontal cortex (Enrichr [51] p-value 2.7×10^{-8}). For example, *CNTN1*, which encodes a contactin involved in neuronal differentiation, encodes several isoforms, one of which has an inferred abundance almost equal to zero based on SR data, while it becomes the second most dominant isoform in day 61 once LR data is considered (isoform B in figure **Figure 3.6**b). *NFE2L2* is another example of a gene involved in neuronal differentiation, encoding a transcript that becomes the dominant isoform at day 61 only once LR data are considered (isoform A in **Figure 3.6**b). This isoform differs from the other two most highly expressed isoforms of this gene in its first exon at the 5' end, which corresponds to a different 5'UTR.

Figure 3.6b also shows additional examples of isoform switch events that are identified by MPAQT based on combination of LR and SR data but not with SR data alone. These include *VEZT*, a gene that encodes the protein vezatin, an acetylcholine receptor binding protein required for formation of neuromuscular synapses [52], and *SLC16A9*, a gene encoding a membrane-spanning solute carrier with a role in neurotransmission in neurological and neurodegenerative disorders [53]. More examples can be found in **Supplementary Figure 7**.

These examples demonstrate the importance of using LR data to adjust relative abundances of transcript isoforms in genes with complex splicing patterns, since SR data alone cannot always accurately capture relative abundances. Addition of LR data allows for the most abundant isoform to be adjusted, and for isoform switches to be detected which would be otherwise missed. In many of the above cases, transcripts differ in their 5' and 3' UTR lengths, or by a single exon, which are small differences that can be difficult to capture using SR data, but for which full-length long reads allow unambiguous detection. Although the above genes have literature support for involvement in brain-related processes or disease, involvement of specific isoforms in these processes has not

been investigated. Analysis of LR data with MPAQT thus opens the door to investigation of specific isoform functionality in neurons with its improved isoform quantification.

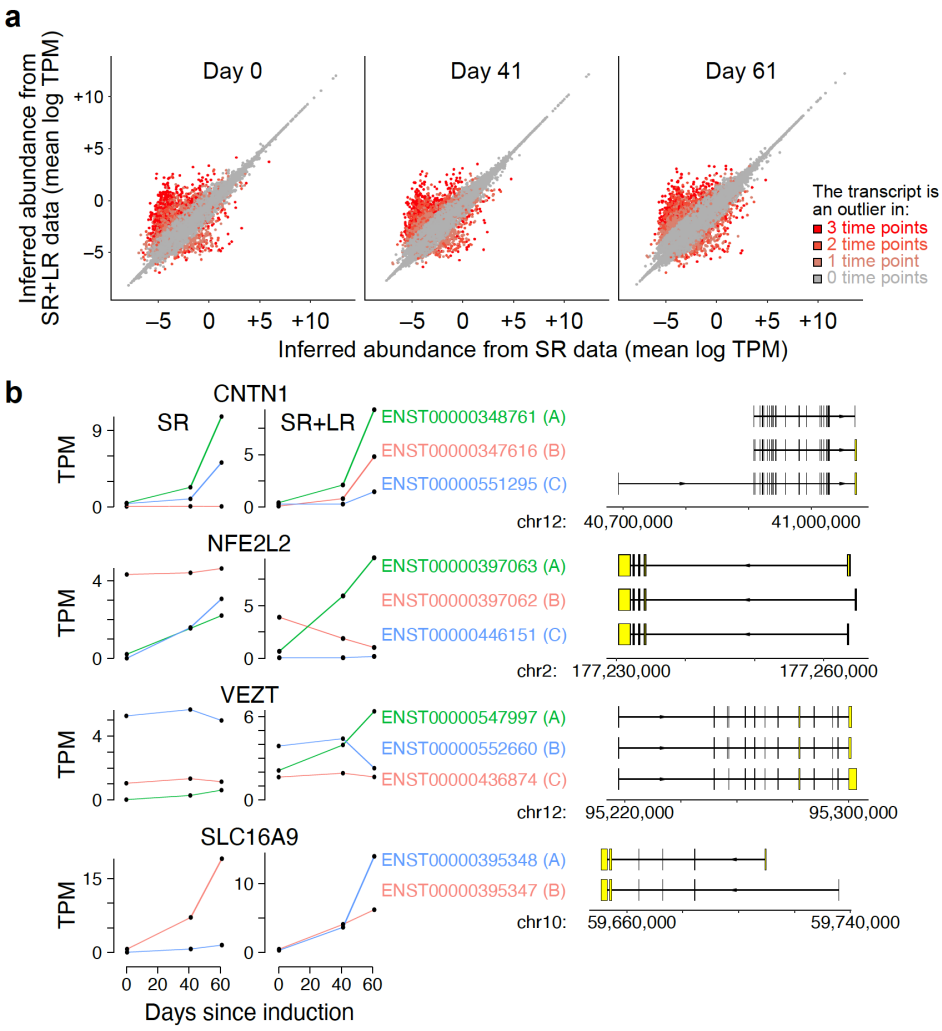


Figure 3.6: Examples of transcripts from neuronal differentiation samples that are differentially quantified by MPAQT when LR data is added. **(a)** Comparison of inferred TPMs based on SR data alone (x-axis) vs. SR+LR data (y-axis) in each of the three time points during neuronal differentiation. For each measurement, the mean of two replicates is used. Each data point is one transcript, with the dot color representing the number of time points in which the inferred abundance of the transcript differs significantly between SR and SR+LR measurements (Mahalanobis distance p-value $<10^{-10}$). **(b)** Four genes with differentially quantified transcripts are examined: *CNTN1*, *NFE2L2*, *VEZT* and *SLC16A9*. Additional examples are shown in **Supplementary Figure 7**.

4 DISCUSSION

In this section, I will start by summarizing our findings and discussing their implications, including our observations from benchmarking MPAQT and applying it to a neuronal differentiation model. I will then discuss the future directions to address some of the limitations of MPAQT and expand its applications.

4.1 Findings

As we showed in different benchmarking experiments, MPAQT shows improved performance both as a SR analysis tool and a data integration tool to combine SR and LR data. Interestingly, MPAQT's ability to improve SR-based quantifications appear to be concentrated on low-expression genes, where other SR tools often provide inaccurate quantifications. This discrepancy between existing SR-based quantification tools and RT-qPCR measurements may have contributed to the perception that RT-qPCR measurement are noisy for such genes [49], whereas MPAQT shows excellent agreement between SR- and RT-qPCR-based measurements even for such lowly expressed genes.

Several potential mechanisms may explain the superior performance of MPAQT compared to existing methods. Even though MPAQT's quantification relies on observation units (OUs) whose counts are provided by existing tools such as kallisto, its statistical framework to infer transcript abundances from the OU counts differs substantially. First, its reliance on simulated data to construct a reference matrix of OU-transcript associations allows it to take into account the errors introduced by the pseudo-alignment algorithms in mapping reads to OUs. In other words, MPAQT is implicitly aware of the probability that a read gets assigned to the wrong OU by a specific tool if its reference $\boldsymbol{P}$ matrix is produced by the same tool. We note, however, that we have not found evidence for widespread read-OU misassignments by kallisto (data not shown), and therefore this mechanism may have only a minor contribution to the increased performance of MPAQT. Secondly, unlike existing frameworks that maximize the likelihood over the reads or OUs that are "observed", MPAQT also considers not observing particular reads as informative. For example, if a transcript is expected to produce reads from two OUs, but only one OU has non-zero counts in a given sample, this lack of observation provides valuable transcript abundance information, which is currently ignored by kallisto or similar methods. This mechanism can be tested, for example, by

removing OUs that have zero observed counts in the benchmarking experiments, to study how the performance of MPAQT is affected.

While MPAQT provides improved quantifications from SR data compared to state-of-the-art, we believe its real strength lies in its unique ability to combine SR and LR data in a statistically principled way. As discussed in the Results section, integration of simulated LR and SR data appears to specifically improve the quantification of transcripts that are encoded by longer genes with more exons and more known isoforms. Such genes with complex isoform splicing patterns are largely enriched for neuron differentiation and development pathways [3, 5]. This suggests that MPAQT's ability to integrate SR and LR data is particularly well-suited to improving quantification of genes involved in neuronal differentiation. This is supported by our analysis of SR and LR data generated from hESC cells undergoing differentiation toward cortical neurons, where we found thousands of transcript isoforms whose inferred abundances are substantially affected by inclusion of LR data. Consistent with our simulations, these transcripts were from larger genes with more exons and more isoforms.

We note that most of the isoforms that we manually examined, which were differentially quantified between SR and SR+LR analyses, showed high similarity to at least one other annotated isoform, with their 5' and/or 3' UTRs being a major exception. For example, *NFE2L2*, *XRCC5*, *ZNF512*, and *SLC16A9* isoforms have differences in their 5' UTRs, isoforms of *CHL1* and *VEZT* have alternative 3' UTRs, and isoforms of *GNG2* and *CCDC82* have notable differences in both 5' and 3' UTRs (**Figure 3.6**b and **Supplementary Figure 7**). In many transcripts, 5' UTRs contain cis-regulatory elements that regulate translation, contributing to variation in corresponding protein levels [54]. Secondary structures can inhibit translation of the mRNA, and RNA-binding proteins can bind to motifs within the 5'UTR, modulating the translation efficiency of the transcript [54]. Additionally, upstream open-reading frames (uORFs) and upstream AUG codons, which may be located within the 5'UTR, can act as decoys, stalling the ribosome and inhibiting translation of the downstream ORF. There are also many potential regulatory elements in the 5'UTR that are uncharacterized [54].

Similarly, 3' UTRs are also known to contain cis-regulatory elements [54]; changing 3'UTR length, for example, can alter the binding of microRNAs (miRNAs) and RNA-binding proteins (RBPs) and, subsequently, affect transcript processing and localization, which eventually translates to downstream signaling modifications related to neuronal differentiation [55]. Through

interactions with miRNAs and RBPs, 3' UTRs have been shown to regulate transcript degradation, translation, and cellular localization, and can determine co-translational protein complex formation [56, 57]. These properties of 5' and 3' UTRs highlight the importance of accurate quantification of isoforms with different UTR sequences for studying neuron growth and differentiation. SR quantification tools seem to have trouble quantifying isoforms that differ in their 5' or 3' UTR lengths, likely due to high transcript similarity, leaving few transcript-specific reads to distinguish transcript quantities. By augmenting SR data with LR data, MPAQT can correct such mis-quantifications. In addition to UTR differences, LRs can also facilitate the detection and quantification of the presence or absence of cassette exons. For example, one isoform of CHL1 differs from its most similar isoform by a single cassette exon (**Supplementary Figure 7**). Therefore, we anticipate that combining LR and SR data will also facilitate the study of alternative splicing events.

We note, however, that current approaches for generation of LR data may have specific biases that are still not well understood. For example, through comparison of SR and LR data, as well as comparison of ground truth concentrations of spike-in RNA molecules with their LR counts, we uncovered a substantial bias toward longer transcripts and transcripts with low GC content. The bias toward longer transcripts may be explained by the size selection steps during library preparation. Furthermore, the GC bias may also result from the impact of GC content on cDNA synthesis (e.g., GC-rich transcripts may form resistant structures that prevent efficient reverse transcription), or PCR-based cDNA amplification. However, we note that such factors also exist in the protocols for SR library preparation, whereas we did not find any significant bias in SR-based quantifications of spike-in transcripts. Therefore, additional experiments and analyses are required to identify the sources of this bias in LR data. Nonetheless, MPAQT can uniquely model these biases and account for them during the integration of SR and LR data.

4.2 Comparison of MPAQT to similar tools

A number of LR specific tools exist which process LR sequencing or improve its quantification such as TALON [58], SQANTI [21], FLAIR [24], LIQA [59], and FLAMES [60]. Such LR specific tools can be used as input to MPAQT if they provide LR counts for transcripts. However, due to the low depth of LR sequencing, direct benchmarking between MPAQT and these

tools would not be very informative since LR sequencing provides information on a smaller subset of the transcriptome. MPAQT provides quantification information for the entire transcriptome using both LR and SR RNA-seq data, whereas LR-specific tools only provide sparse quantification information for select transcripts which are captured by LR sequencing. In other words, since LR sequencing tends to be lower depth than SR sequencing, LR analysis tools provide quantification for only transcripts captured by LR sequencing, which tend to be the most highly abundant transcripts. MPAQT uses SR data as the basis for its quantification, and improves quantification of transcripts which have corresponding LR counts. This results in improved quantification of the transcriptome overall.

Tools to detect alternative splicing events such as rMATS [61], Cufflinks [62], and MISO [63] are also similar, but focus on alternative splicing instead of whole transcriptome quantification. As such, they cannot be directly compared to MPAQT.

4.3 Future Directions and Limitations

Some of the immediate next steps to follow up on the work presented here will include experimental validation and functional annotation of the transcripts that, based on analysis of neuronal differentiation samples, are differentially quantified after including LR data. We identified thousands of such isoforms, only a few of which were highlighted in the Results section. Isoform-specific RT-qPCR measurements can be used to validate the MPAQT inferences from SR+LR data. Furthermore, a more thorough evaluation of functional domains and regulatory regions in these transcripts could reveal the potential functional impacts of these isoforms, and direct the design of experiments for their functional characterization, with the potential to uncover novel biology related to neuronal differentiation.

Investigating the performance of MPAQT for analysis of data from other platforms that we did not study in this work is another interesting line of inquiry. In this work, PacBio IsoSeq was used for long-read sequencing. However, MPAQT is agnostic to the type of LR technology. Therefore, it can be used to analyze data from other LR technologies, such as Oxford Nanopore or Illumina Infinity, to improve SR quantification results. In fact, MPAQT has the potential to integrate data from multiple LR technologies for the same sample. This multi-platform integration could offer benefits, since biases present in one type of LR data could be offset by other LR data sources.

Furthermore, we envision the MPAQT framework to be extendable to analysis of single-cell RNA-seq libraries that have been sequenced by both SR and LR technologies, an area that is a subject of active research by different groups. Most of the widely used single-cell technologies, such as 10x Chromium, currently do not provide full-length transcript coverage as most of the reads correspond to transcript 3' ends. This bias substantially hinders the ability to quantify individual isoforms and study splicing events at the single-cell level. Characterizing the 10x cDNA libraries using a combination of SR sequencing (to obtain the depth needed to resolve cellular heterogeneity) and LR sequencing (to obtain the unambiguity needed to quantify individual isoforms) can potentially resolve this issue.

In a recent preprint [64], the authors used SR single-cell 10x transcriptomics data to characterize cell populations in the mouse brain, giving high sequencing depth for each cell, allowing for identification of brain cell types, which resulted in 395 cell clusters. The authors created a sc-LR method called ScISOr-Seq2, and obtained both ONT and Pacbio HiFi barcoded long reads for the 395 cell clusters determined using SR data. They then investigated full-length isoforms on three axes: multiple adult brain regions, cell subtypes, and developmental timepoints. Using a combination of SR and LR data, they were able to uncover novel biology, including the discovery that cells of the same type which are present in different brain regions, and cells of the same type at different points along the developmental axis, have biologically relevant isoform differences. We believe that MPAQT's statistically principled framework for combining SR and LR data, and its demonstrated performance in improved isoform quantification, could be of use in this field of research for improving our understanding of differential isoform usage at the single cell level in the brain.

For MPAQT to be applicable to single-cell data, two modifications will be needed: (a) we will need to obtain a reference OU-transcript mapping matrix ($\boldsymbol{P}$) that properly reflects the 3' bias of single-cell SR data; (b) we will need to scale up the inference algorithm to allow analysis of thousands of cells at the same time. We believe both these aims are achievable.

We note that, in this work, we focused on quantification of known (annotated) isoforms. However, PacBio LR data provides an opportunity to discover novel transcripts, and a method is needed for integration of these transcripts into the reference transcriptome. Tissues such as cancer and the brain have a high degree of alternative splicing and novel isoforms, making analysis of these samples using SR data alone insufficient. Due to the existence of cancer cell line-specific

isoforms that result from aberrant splicing [65], a tissue-specific reference transcriptome generated from LR sequencing is needed to identify and quantify these novel isoforms (**Supplementary Figure 8**) [65]. For example, in a study of gastric cancer promoter diversity, kallisto was used to quantify SR data using a GTF from SQANTI2 as a reference transcriptome [2], but such LR GTFs suffer from transcript dropouts due to low LR sequencing depth. As such, integration of LR GTFs with a known reference is needed. Furthermore, existing studies [2] that use LR data for transcript discovery and SR data for quantification ignore the valuable information that can be gained about isoform abundances by combining SR and LR data.

One example that highlights the importance (and the complexities) of analysis of novel transcripts is *PRUNE2*, a gene involved in Alzheimer's Disease susceptibility with well characterized functions in the brain [66]. In our neuronal differentiation samples, when only the known transcripts are used for quantification, the most highly expressed known isoform in days 41 and 61 is ENST00000424866, although its expression seems to increase at day 61 (**Supplementary Figure 9**). However, once we allow the identification of novel transcripts using LR data, two novel isoforms PB.10420.4 and PB.10619.7 seem to be the most highly expressed transcripts at days 41 and 61, respectively, with almost all PRUNE2 long reads being assigned to these two novel transcripts. These two novel isoforms seem to differ from one another only slightly in their 3'UTR lengths (**Supplementary Figure 9**). The most abundant isoform in day 61, PB.10619.7, has no equivalent transcript in the GENCODE reference (**Supplementary Figure 9**), highlighting potential false quantifications that may arise from relying only on a reference annotation.

To enable quantification of novel transcripts, however, MPAQT will need to use sample- or dataset-specific reference matrices for OU-transcript mapping. Like other transcript quantification tools, which require initial indexing of the reference transcriptome to create a tool-specific reference, MPAQT's reference matrix $\boldsymbol{P}$ is generated specifically for each given transcriptome annotation file. Nuances of PacBio's GTF metadata need to be accounted for to ensure that a sample's novel transcriptome is as complete and non-redundant as possible. Among the challenges in this process is the task of equating novel transcripts between samples within a given dataset, since IsoSeq gives sample-specific identifiers. However, mapping transcripts between samples is required when doing DE or isoform switch analysis. Importantly, there is a trade-off between removing redundant (similar) isoforms and keeping unique ones when novel transcripts across

multiple samples need to be merged: if we are too stringent, important transcripts may be lost, but if we are too lax, we may include redundant, similar transcripts. There is also the challenge of removing redundancy between the known and novel transcripts. For example, SQANTI provides "associated transcript" for many transcripts, although the transcript in the reference (known) transcriptome may not be identical to SQANTI's annotation, adding another layer of complication.

A recent tool, Bambu [67], has been developed to improve quantification of known and novel transcripts. Bambu considers that novel transcripts may be missing from the reference, that most transcripts aren't expressed in each sample, and that there are many false positive transcripts. Bambu works by creating read classes (RCs) which are similar in concept to ECs. Bambu then combines RCs across samples and calculates novel discovery rate (NDR), a metric below which RCs are considered novel transcripts. Bambu then estimates expression levels of each transcript using both uniquely assigned reads and reads assigned to multiple transcripts. It may be useful to use the concept of RCs to improve MPAQT's framework, or alternatively use Bambu output as input to MPAQT to improve LR quantification values. Concepts from Bambu could also be useful for improving generation of matrix $\boldsymbol{P}$ for the LR sequencing data, which assumes that LRs are unambiguously mapped. Indeed, Bambu assigns RCs to transcripts, which allows for inexact matches and accounts for alignment errors. Additionally, Bambu's ability to create a sample-specific reference transcriptome with both known and novel transcripts could be useful for input to MPAQT. However, Bambu relies on LR data alone to create the reference transcriptome, so does not fully solve the need to integrate the known transcriptome with novel transcripts identified with LR data, especially considering the low depth of LRs which results in transcript dropouts. Indeed, the authors of Bambu do not address the fact that many LR samples are low depth, and that transcripts of lower expression can be biologically relevant.

One limitation of MPAQT is the need to generate matrix $\boldsymbol{P}$. In addition to requiring large amounts of memory, the running time of each of the 24 replicates is around two hours. Use of a job scheduler that allows for parallelization of this task, in our experience, reduces the time need to create the matrix $\boldsymbol{P}$ to around three hours, but if replicates are computed in series the running time is close to two days. There may be a more efficient way to generate this reference matrix, such as mapping reads to ECs on the fly instead of writing FASTQ files to disk (writing intermediate files to disk can slow down running time), or modification of the existing software implementation to be more computationally efficient.

Overall, the research area of using LR sequencing to better quantify and discover isoforms is a new and expanding field. MPAQT provides, to the best of our knowledge, the first principled statistical framework to quantify isoforms using both LR and SR data, allowing for characterization of tissues with high splicing diversity not previously possible. Further development of MPAQT to better quantify and identify isoforms, and to integrate different sequence data types, will allow for more detailed characterization of tissue-specific isoform expression patterns, enabling the characterization of disease-relevant biological processes at a resolution previously inaccessible.

5 REFERENCES

1. Arzalluz-Luque A, Conesa A: **Single-cell RNAseq for the study of isoforms-how is that possible?** *Genome Biol* 2018, **19:**110.
2. Huang KK, Huang J, Wu JKL, Lee M, Tay ST, Kumar V, Ramnarayanan K, Padmanabhan N, Xu C, Tan ALK, et al: **Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer.** *Genome Biol* 2021, **22:**44.
3. Clark MB, Wrzesinski T, Garcia AB, Hall NAL, Kleinman JE, Hyde T, Weinberger DR, Harrison PJ, Haerty W, Tunbridge EM: **Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain.** *Mol Psychiatry* 2020, **25:**37-47.
4. Belluti S, Rigillo G, Imbriano C: **Transcription Factors in Cancer: When Alternative Splicing Determines Opposite Cell Fates.** *Cells* 2020, **9**.
5. Su CH, D D, Tarn WY: **Alternative Splicing in Neurogenesis and Brain Development.** *Front Mol Biosci* 2018, **5:**12.
6. Vlasova IA, Tahoe NM, Fan D, Larsson O, Rattenbacher B, Sternjohn JR, Vasdewani J, Karypis G, Reilly CS, Bitterman PB, Bohjanen PR: **Conserved GU-rich elements mediate mRNA decay by binding to CUG-binding protein 1.** *Mol Cell* 2008, **29:**263-270.
7. Tushev G, Glock C, Heumuller M, Biever A, Jovanovic M, Schuman EM: **Alternative 3' UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments.** *Neuron* 2018, **98:**495-511 e496.
8. Naseri NN, Wang H, Guo J, Sharma M, Luo W: **The complexity of tau in Alzheimer's disease.** *Neurosci Lett* 2019, **705:**183-194.
9. Cieslik M, Chinnaiyan AM: **Cancer transcriptome profiling at the juncture of clinical translation.** *Nat Rev Genet* 2018, **19:**93-109.
10. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A: **A survey of best practices for RNA-seq data analysis.** *Genome Biol* 2016, **17:**13.
11. Hu T, Chitnis N, Monos D, Dinh A: **Next-generation sequencing technologies: An overview.** *Hum Immunol* 2021, **82:**801-811.
12. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q: **Opportunities and challenges in long-read sequencing data analysis.** *Genome Biol* 2020, **21:**30.
13. Au KF: **The blooming of long-read sequencing reforms biomedical research.** *Genome Biol* 2022, **23:**21.
14. Wright DJ, Hall NAL, Irish N, Man AL, Glynn W, Mould A, Angeles AL, Angiolini E, Swarbreck D, Gharbi K, et al: **Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes.** *BMC Genomics* 2022, **23:**42.
15. **High-performance long-read assay enables contiguous data with N50 of 6–7 kb on existing Illumina platforms** [https://www.illumina.com/science/genomics-research/articles/infinity-high-performance-long-read-assay.html]
16. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12:**323.
17. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29:**24-26.

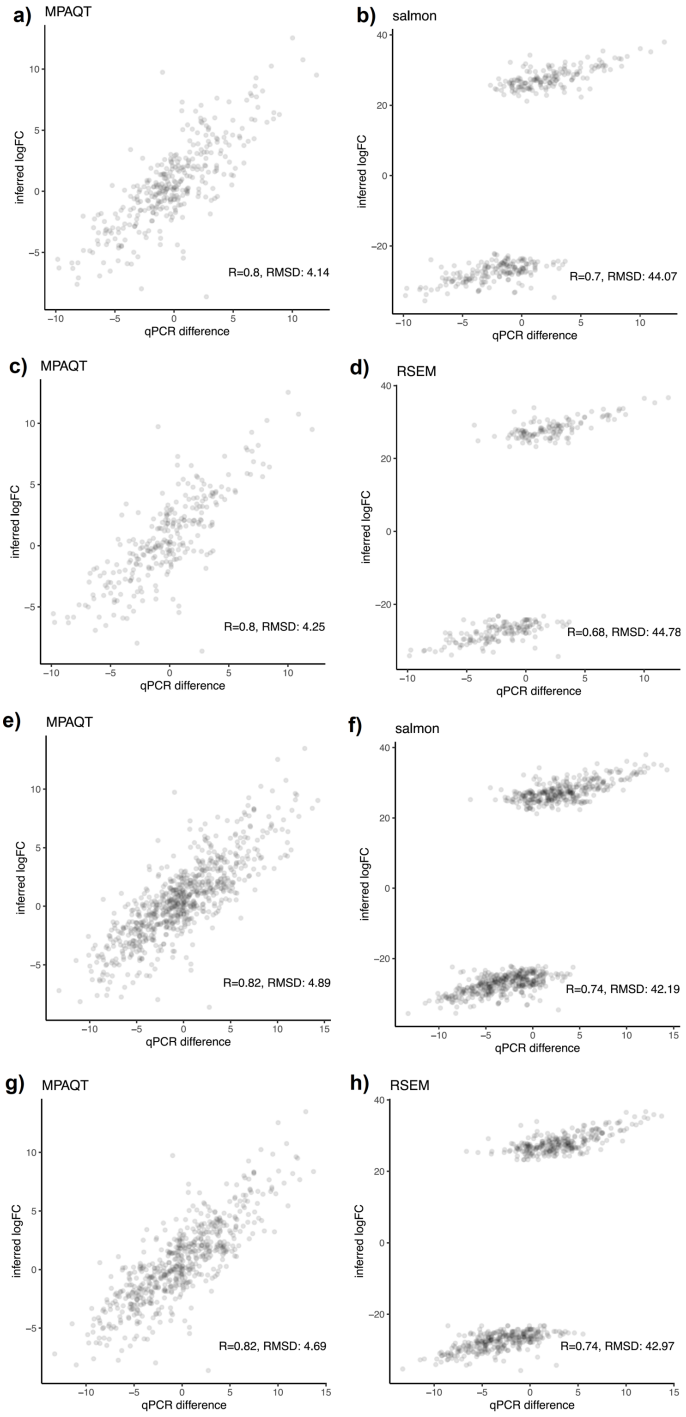
18. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: **Salmon provides fast and bias-aware quantification of transcript expression.** *Nat Methods* 2017, **14:**417-419.
19. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol* 2016, **34:**525-527.
20. **SQANTI3 Github page** [https://github.com/ConesaLab/SQANTI3]
21. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al: **SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification.** *Genome Res* 2018, **28:**396-411.
22. Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, Kasukawa T: **refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites.** *J Mol Biol* 2019, **431:**2407-2422.
23. Nellore A, Jaffe AE, Fortin JP, Alquicira-Hernandez J, Collado-Torres L, Wang S, Phillips RA, III, Karbhari N, Hansen KD, Langmead B, Leek JT: **Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive.** *Genome Biol* 2016, **17:**266.
24. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN: **Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns.** *Nat Commun* 2020, **11:**1438.
25. Fu S, Ma Y, Yao H, Xu Z, Chen S, Song J, Au KF: **IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing.** *Bioinformatics* 2018, **34:**2168-2176.
26. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy AS: **A survey of the sorghum transcriptome using single-molecule long reads.** *Nat Commun* 2016, **7:**11706.
27. Shumate A, Wong B, Pertea G, Pertea M: **Improved transcriptome assembly using a hybrid of long and short reads with StringTie.** *PLoS Comput Biol* 2022, **18:**e1009730.
28. Guseva D, Jakovcevski I, Irintchev A, Leshchyns'ka I, Sytnyk V, Ponimaskin E, Schachner M: **Cell Adhesion Molecule Close Homolog of L1 (CHL1) Guides the Regrowth of Regenerating Motor Axons and Regulates Synaptic Coverage of Motor Neurons.** *Front Mol Neurosci* 2018, **11:**174.
29. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
30. Gene Ontology C: **The Gene Ontology resource: enriching a GOld mine.** *Nucleic Acids Res* 2021, **49:**D325-D334.
31. Chawla G, Lin CH, Han A, Shiue L, Ares M, Jr., Black DL: **Sam68 regulates a set of alternatively spliced exons during neurogenesis.** *Mol Cell Biol* 2009, **29:**201-213.
32. Tollervey JR, Wang Z, Hortobagyi T, Witten JT, Zarnack K, Kayikci M, Clark TA, Schweitzer AC, Rot G, Curk T, et al: **Analysis of alternative splicing associated with aging and neurodegeneration in the human brain.** *Genome Res* 2011, **21:**1572-1582.
33. Townes FW, Hicks SC, Aryee MJ, Irizarry RA: **Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model.** *Genome Biol* 2019, **20:**295.

34. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509-1517.
35. Bengtsson M, Stahlberg A, Rorsman P, Kubista M: **Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels.** *Genome Res* 2005, **15:**1388-1392.
36. Liao Y, Smyth GK, Shi W: **The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads.** *Nucleic Acids Res* 2019, **47:**e47.
37. Brent RP: *Algorithms for minimization without derivatives.* Mineola, N.Y: Dover; 2013.
38. Tian A, Muffat J, Li Y: **Studying Human Neurodevelopment and Diseases Using 3D Brain Organoids.** *J Neurosci* 2020, **40:**1186-1193.
39. Fish L, Khoroshkin M, Navickas A, Garcia K, Culbertson B, Hanisch B, Zhang S, Nguyen HCB, Soto LM, Dermit M, et al: **A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements.** *Science* 2021, **372**.
40. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.
41. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
42. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc* 2009, **4:**1184-1191.
43. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11:**R25.
44. Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK: **Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression.** *Ann Appl Stat* 2016, **10:**946-963.
45. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015, **43:**e47.
46. 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.** *F1000Res* 2020, **9**.
47. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, et al: **Gencode 2021.** *Nucleic Acids Res* 2021, **49:**D916-D923.
48. Consortium SM-I: **A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium.** *Nat Biotechnol* 2014, **32:**903-914.
49. Everaert C, Luypaert M, Maag JLV, Cheng QX, Dinger ME, Hellemans J, Mestdagh P: **Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data.** *Sci Rep* 2017, **7:**1559.
50. Zylka MJ, Simon JM, Philpot BD: **Gene length matters in neurons.** *Neuron* 2015, **86:**353-355.

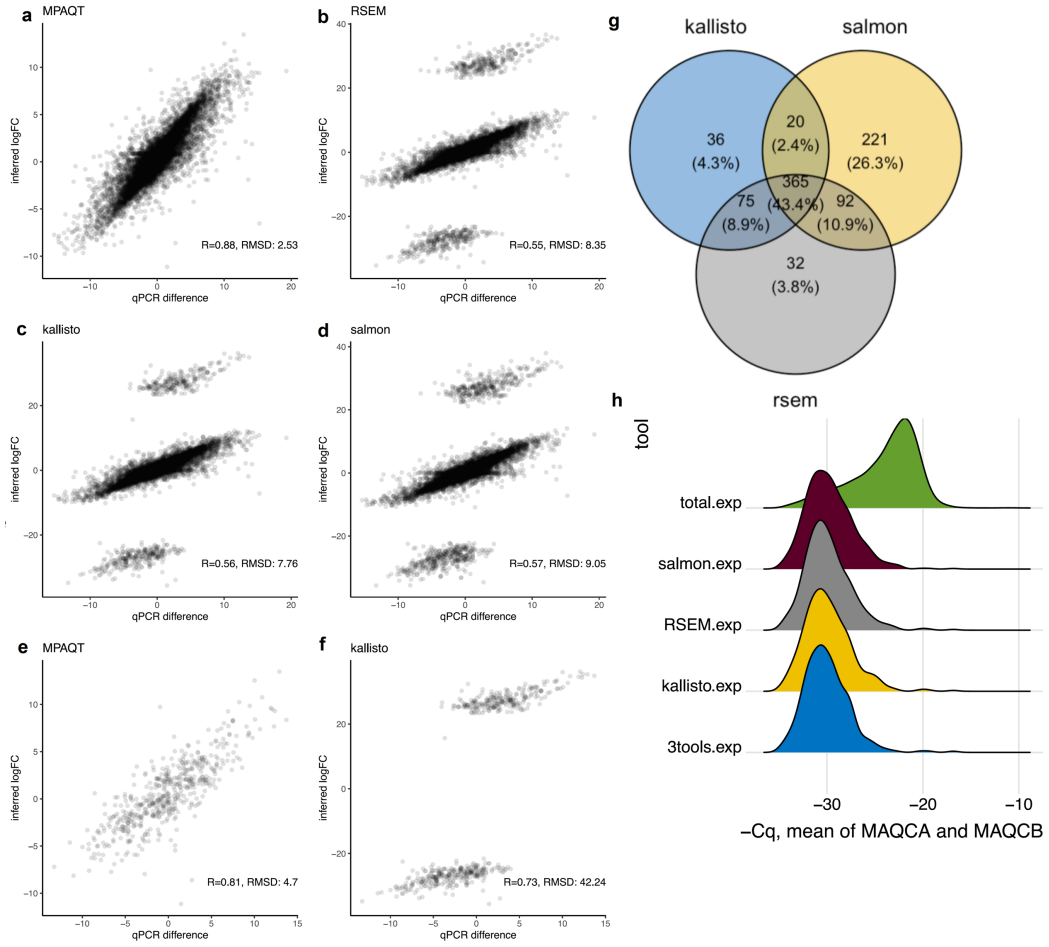
51. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.** *BMC Bioinformatics* 2013, **14:**128.
52. Koppel N, Friese MB, Cardasis HL, Neubert TA, Burden SJ: **Vezatin is required for the maturation of the neuromuscular synapse.** *Mol Biol Cell* 2019, **30:**2571-2583.
53. Ayka A, Sehirli AO: **The Role of the SLC Transporters Protein in the Neurodegenerative Disorders.** *Clin Psychopharmacol Neurosci* 2020, **18:**174-187.
54. Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, Suresh U, Burns SC, Penalva LO: **Before It Gets Started: Regulating Translation at the 5' UTR.** *Comp Funct Genomics* 2012, **2012:**475731.
55. Ognibene M, Pezzolo A: **Ezrin interacts with the tumor suppressor CHL1 and promotes neuronal differentiation of human neuroblastoma.** *PLoS One* 2020, **15:**e0244069.
56. Mayr C: **Regulation by 3'-Untranslated Regions.** *Annu Rev Genet* 2017, **51:**171-194.
57. Mendonsa S, von Kugelgen N, Dantsuji S, Ron M, Breimann L, Baranovskii A, Lodige I, Kirchner M, Fischer M, Zerna N, et al: **Massively parallel identification of mRNA localization elements in primary cortical neurons.** *Nat Neurosci* 2023.
58. Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, Matheos D, Zeng W, Williams B, Trout D, et al: **A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification.** *bioRxiv* 2020**:**672931.
59. Hu Y, Fang L, Chen X, Zhong JF, Li M, Wang K: **LIQA: long-read isoform quantification and analysis.** *Genome Biol* 2021, **22:**182.
60. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C, et al: **Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing.** *Genome Biol* 2021, **22:**310.
61. Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y: **rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.** *Proc Natl Acad Sci U S A* 2014, **111:**E5593-5601.
62. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7:**562-578.
63. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation.** *Nat Methods* 2010, **7:**1009-1015.
64. Joglekar A, Hu W, Zhang B, Narykov O, Diekhans M, Balacco J, Ndhlovu LC, Milner TA, Fedrigo O, Jarvis ED, et al: **Single-cell long-read mRNA isoform regulation is pervasive across mammalian brain regions, cell types, and development.** *bioRxiv* 2023.
65. Oka M, Xu L, Suzuki T, Yoshikawa T, Sakamoto H, Uemura H, Yoshizawa AC, Suzuki Y, Nakatsura T, Ishihama Y, et al: **Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer.** *Genome Biol* 2021, **22:**9.
66. Li S, Itoh M, Ohta K, Ueda M, Mizuno A, Ohta E, Hida Y, Wang MX, Takeuchi K, Nakagawa T: **The expression and localization of Prune2 mRNA in the central nervous system.** *Neurosci Lett* 2011, **503:**208-214.

67. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Goke J: **Context-aware transcript quantification from long-read RNA-seq data with Bambu.** *Nat Methods* 2023.
68. Watakabe A, Ohsawa S, Ichinohe N, Rockland KS, Yamamori T: **Characterization of claustral neurons by comparative gene expression profiling and dye-injection analyses.** *Front Syst Neurosci* 2014, **8:**98.
69. Quan W, Li J, Jin X, Liu L, Zhang Q, Qin Y, Pei X, Chen J: **Identification of Potential Core Genes in Parkinson's Disease Using Bioinformatics Analysis.** *Parkinsons Dis* 2021, **2021:**1690341.
70. Lee SJ, Kwon S, Gatti JR, Korcari E, Gresser TE, Felix PC, Keep SG, Pasquale KC, Bai T, Blanchett-Anderson SA, et al: **Large-scale identification of human cerebrovascular proteins: Inter-tissue and intracerebral vascular protein diversity.** *PLoS One* 2017, **12:**e0188540.

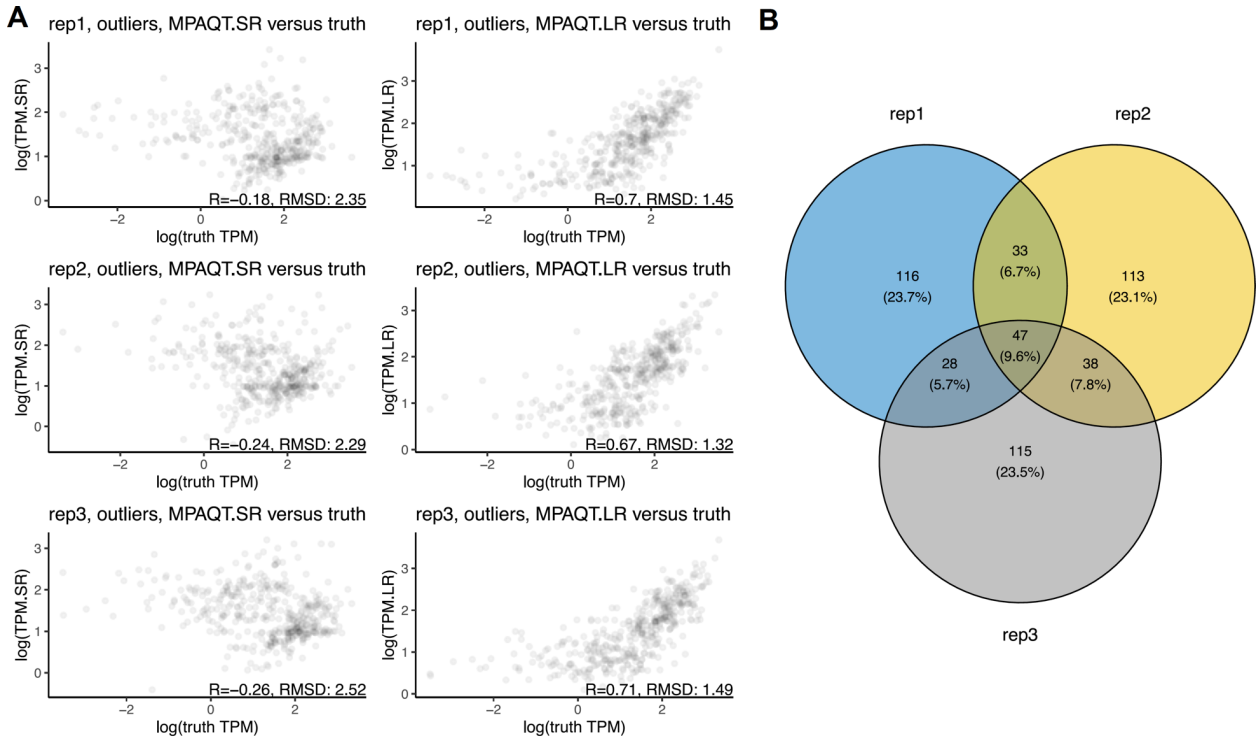
6 SUPPLEMENTARY FIGURES



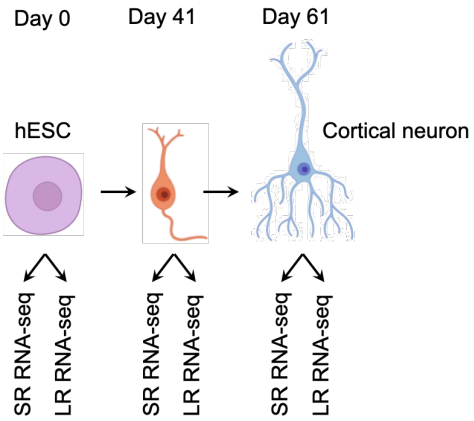
Supplementary Figure 1: MPAQT's performance on salmon and RSEM's outliers for two filtering ranges. **(a-d)** Filtering range 32 > Cq > 11; **(e-h)** Filtering range 35 > Cq > 8.



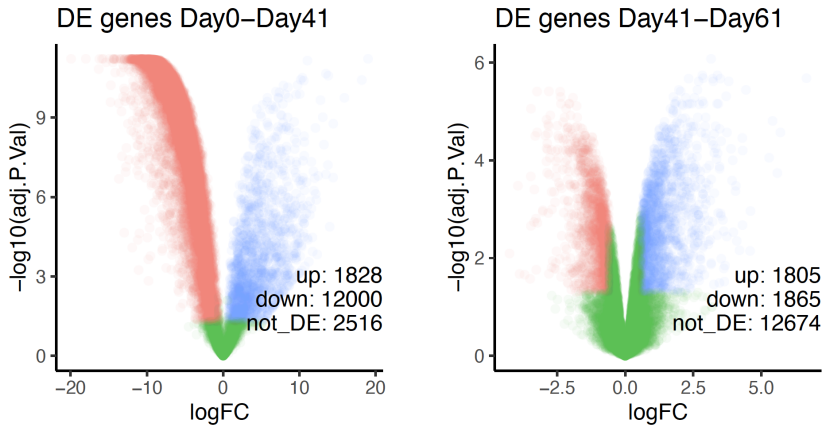
Supplementary Figure 2: Widening range for filtering Cq values (35 > Cq > 8) shows comparable MPAQT performance, with doubling of the number of outliers for the other 3 tools. The number of genes remaining after filtering increases from 14,956 to 16,104. **(a-h)** Figure caption is identical to **Figure 3.1** except for the extended filtering range.



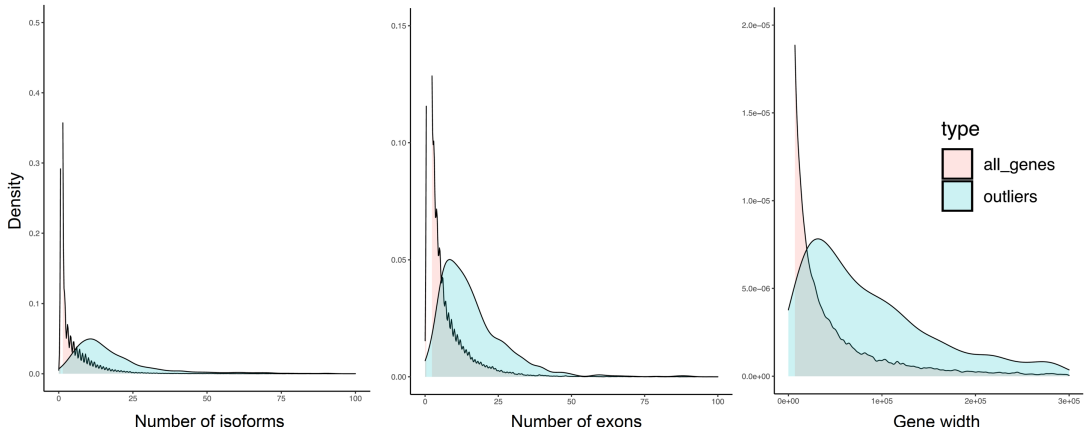
Supplementary Figure 3: Characterization of transcripts differentially quantified by MPAQT upon addition of LR data. Mahalanobis distance was used with p-value=10^{-10} to extract differentially quantified transcripts between MPAQT (LR+SR) and MPAQT (SR) measurements, leaving around 650 outliers for each replicate. **(a)** Quantifications of outliers from genes >250bp correlate better with truth TPMs in MPAQT (LR+SR) than MPAQT (SR) across all three replicates. **(b)** Intersection among the outliers of the three samples, for filtered genes >250bp. All panels correspond to reads simulated from TPM values generated from an exponential distribution.



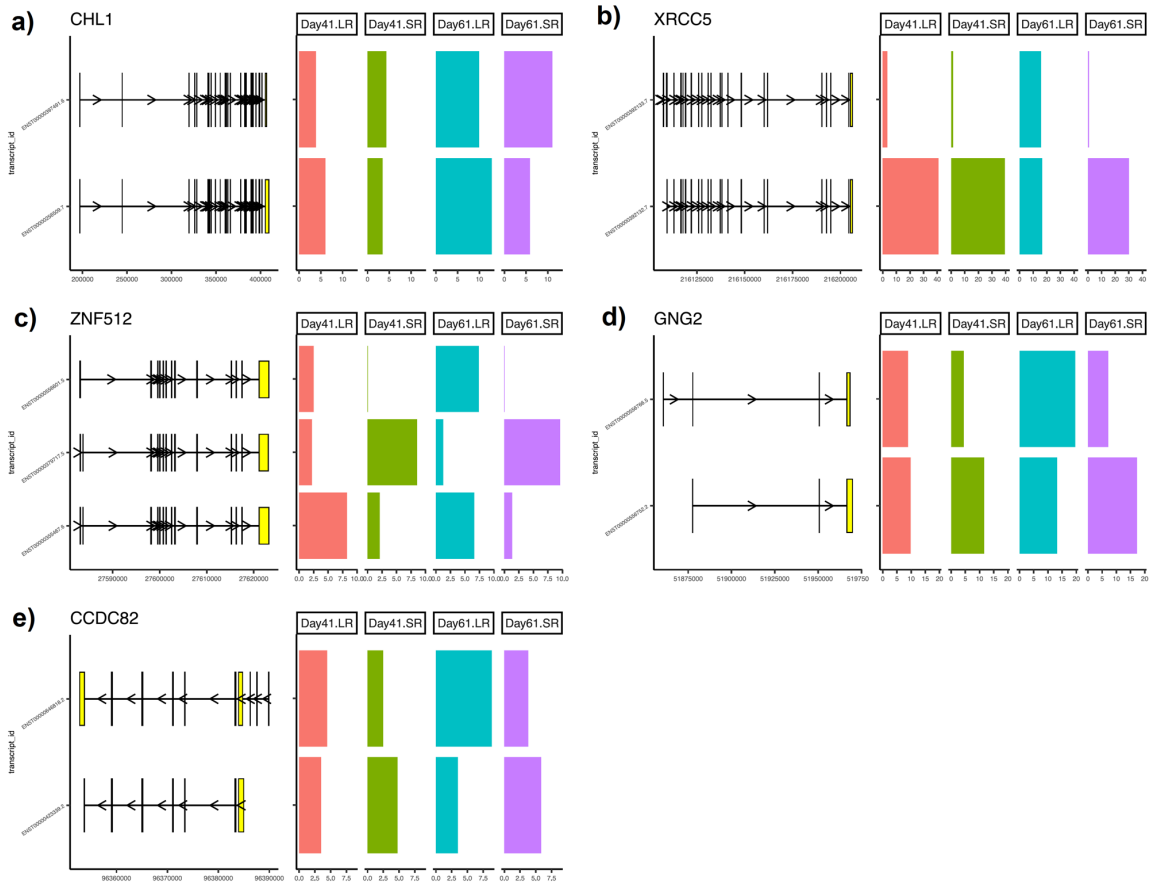
Supplementary Figure 4: Schematic of neuronal differentiation. Two replicates were collected for each sample, followed by RNA-seq of each replicate using either short-read or long-read sequencing.



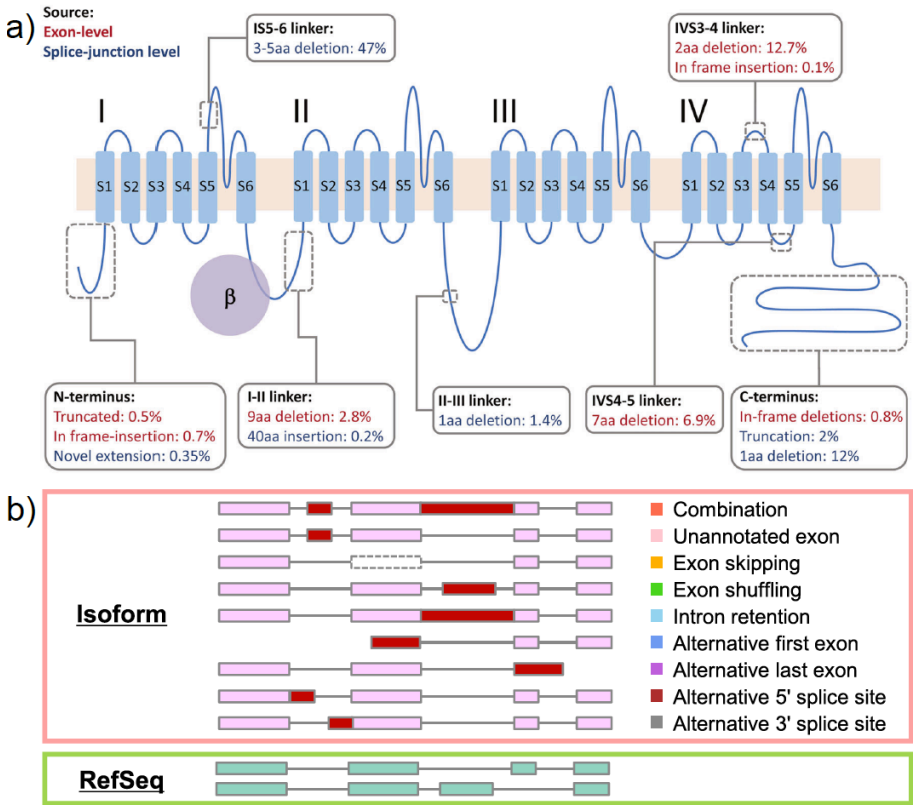
Supplementary Figure 5: Volcano plots of upregulated (blue), downregulated (red) and non-DE (green) genes between days 0 and 41 (left) and days 41 and 61 (right).



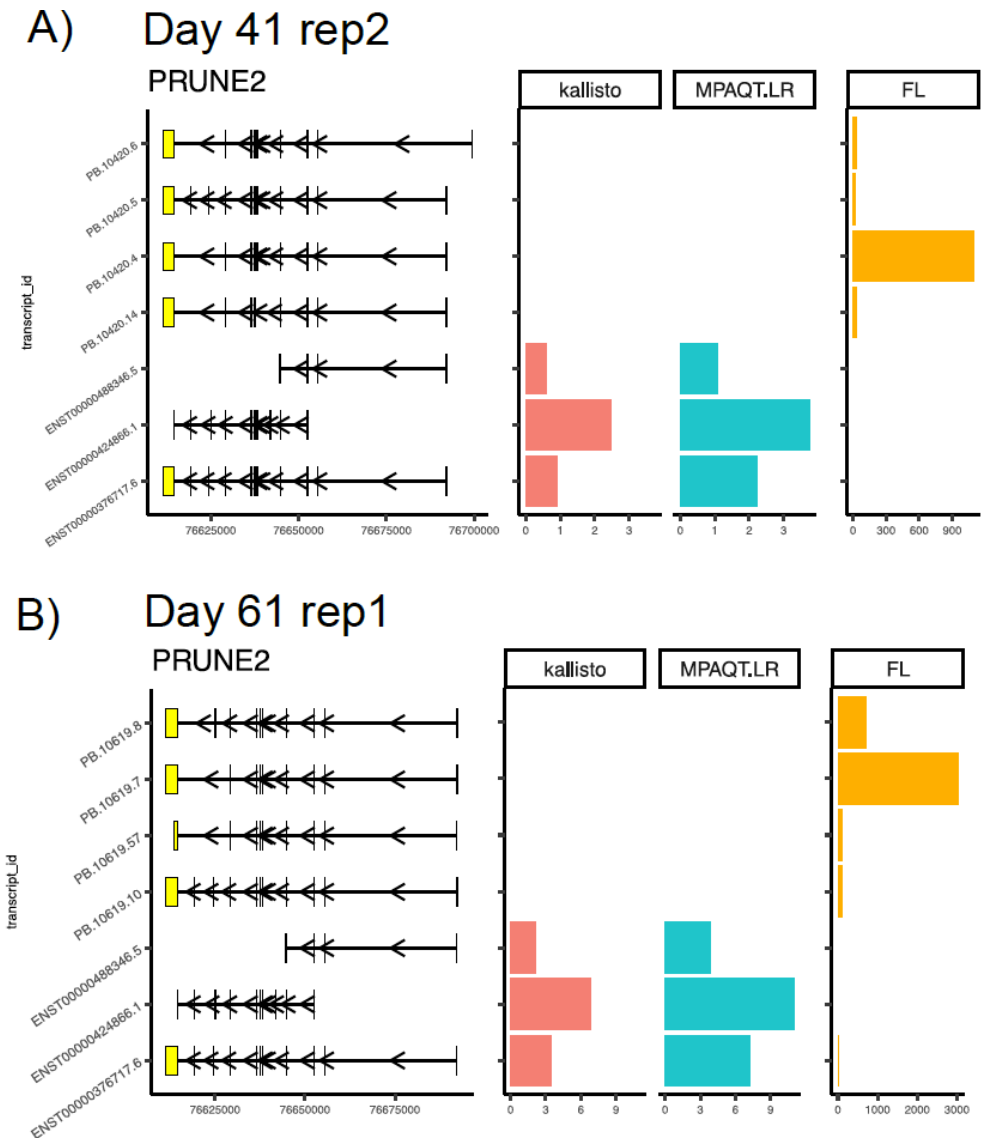
Supplementary Figure 6: Characteristics of differentially quantified transcripts found using Mahalanobis distance when comparing MPAQT (LR+SR) and MPAQT (SR) inferences at day 61. Genes of differentially quantified transcripts have more isoforms, more exons, and larger gene width.



Supplementary Figure 7: Additional examples of isoform switches and differential quantification in neuronal differentiation samples possible only with LR data added to MPAQT. **(a)** CHL1, **(b)** XRCC5, **(c)** ZNF512, **(d)** GNG2, and **(e)** CCDC82. Some of these genes have proposed connections to the biology of nervous system development and disease. For example, GNG2 has brain-specific expression *[68]*, and a bioinformatic analysis of core genes involved in Parkinson's disease found GNG2 to be the most highly connected upregulated hub gene, and may play a key role in pathogenesis of Parkinson's disease *[69]*. Despite brain-related literature support, its function in neuronal development remains unclear. Addition of LR data shows ENST00000556766 is upregulated at Day 61 and differs from another isoform at both the 5' and 3' ends . Another example is CCDC82, a brain vascular marker from the coiled-coil domain-containing (CCDC) family of proteins *[70]*. MPAQT (LR + SR) detects doubling of isoform ENST00000646818 between days 41 and 61, uncovering a novel isoform switch event.



Supplementary Figure 8: The brain and cancer are two tissue types which exhibit a high degree of splicing diversity. **(a)** Domain structure diversity of CACNA1A, image from [3] **(b)** Example of cancer splicing diversity as compared to reference transcriptome RefSeq, image from [65]. Images are unmodified from their original sources, and are reproduced under the Creative Commons Attribution 4.0 International License [http://creativecommons.org/licenses/by/4.0/].



Supplementary Figure 9: PRUNE2 novel transcripts PB.10420.4 and PB.10619.7, the most abundant transcripts according to LR data at days 41 and 61, respectively, differ from known transcripts. Quantifications using only the GENCODE reference incorrectly infer that 3 known transcripts are the most abundant. The addition of FL LR counts to MPAQT does not correct this erroneous quantification, since novel transcripts are not included in the reference transcriptome. **(a)** Day 41, replicate 2; **(b)** Day 61, replicate 1.

7 SUPPLEMENTARY TABLES

Supplementary Table 1: Counts of neuron-related terms for up-regulated and down-regulated genes. Columns "up" and "down" refer to the total number of terms for each of these categories, whereas "up, neuro" and "down, neuro" are the counts of string matches to neuron-related terms for the up-regulated and down-regulated genes, respectively. The columns "total" and "total, neuro" refer to the total number of enriched terms for both up-regulated and down-regulated genes, for all terms and neuron-related terms, respectively.

Contrast	Up	Up, neuro	Down	Down, neuro	Total	Total, neuro
Day 41 – hESC	148	66	1532	26	1532	70
Day 61 – hESC	166	71	1582	26	1538	77
Day 61 – Day 41	53	24	531	10	338	31

8 SUPPLEMENTARY DATA TABLES

Supplementary Data Table 1: Differential expression output containing the following fields: "hgnc_symbol" (HGNC gene symbol), "logFC" (log fold-change), "AveExpr" (average expression across samples), "t" (t-statistic for comparison between conditions), "P.Value" (the associated p-value), "adj.P.Val" (FDR-adjusted P-value), "B" (B-statistics, which is the log-odds that the gene is differentially expressed), and "gene_id" (Ensembl Gene ID). Two subtables are available online:

(**a**) Comparison of day 41 vs. day 0 during cortical neuron differentiation, available from:

[https://github.com/csglab/MPAQT/tree/main/data/supplementary_data_table_1a.topfit.SOX10_Day41-hESC.xlsx]

(**b**) Comparison of day 61 vs. day 41, available from:

[https://github.com/csglab/MPAQT/tree/main/data/supplementary_data_table_1b.topfit.SOX10_Day61-SOX10_Day41.xlsx] .