Detecting differential transcript usage in complex diseases with SPIT

Beril Erdogdu[1,2], Ales Varabyou[1,3], Stephanie C. Hicks[1,4,5], Steven L. Salzberg[1,2,3,4,6], Mihaela Pertea[1,2,3,6]

Affiliations:

[1]Center for Computational Biology, Johns Hopkins University; Baltimore, MD, United States
[2]Department of Biomedical Engineering, Johns Hopkins School of Medicine and Whiting School of Engineering; Baltimore, MD, United States
[3]Department of Computer Science, Johns Hopkins University; Baltimore, MD, United States
[4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, MD, USA
[5]Malone Center for Engineering in Healthcare, Johns Hopkins University, MD, USA
[6]Department of Genetic Medicine, Johns Hopkins School of Medicine; Baltimore, MD, United States

Abstract

Differential transcript usage (DTU) plays a crucial role in determining how gene expression differs among cells, tissues, and different developmental stages, thereby contributing to the complexity and diversity of biological systems. In abnormal cells, it can also lead to deficiencies in protein function, potentially leading to pathogenesis of diseases. Detecting such events for single-gene genetic traits is relatively uncomplicated; however, the heterogeneity of populations with complex diseases presents an intricate challenge due to the presence of diverse causal events and undetermined subtypes. SPIT is the first statistical tool that quantifies the heterogeneity in transcript usage within a population and identifies predominant subgroups along with their distinctive sets of DTU events. We provide comprehensive assessments of SPIT's methodology in both single-gene and complex traits and report the results of applying SPIT to analyze brain samples from individuals with schizophrenia. Our analysis reveals previously unreported DTU events in six candidate genes.

Introduction

Alternative splicing enables eukaryotic cells to produce a diverse batch of transcripts and, consequently, proteins from a single gene. While for some genes these distinct transcripts (isoforms) may be used interchangeably, many protein-coding genes have a dominant isoform that is favored in expression across the healthy individuals of a human population.[1] Predominant expression of alternative isoforms may subject these genes to changes and potential errors in their function.[2] Differential transcript usage (DTU) analysis is conducted using RNA-Seq data to search for systematic differences in the expression ratios of isoforms that may explain changes in phenotype between cell types, tissues, or populations[2, 3].

Isoform abundance is often tissue-specific, and DTU (also called isoform switching) may result in proteins with distinct functions, which in turn may play different roles in the cell.[2-6] There is also a growing interest in the effects of DTU in complex human diseases. Instances of DTU have been associated with DNA repair, numerous human cancer types, heart failure, and psychiatric diseases such as autism, schizophrenia, and bipolar disorder.[7-9] State-of-the-art DTU analysis tools provide a framework to detect cases where the isoform proportions are consistent within and significantly different between any two groups of samples. However, transcriptomic profiles within populations comprising individuals affected by a complex disease are rarely consistent due to a multiplicity of causal events and disease subgroups; i.e., a cohort of patients diagnosed with the same disease might actually have several distinct underlying genetic disorders.[10] Therefore, a DTU analysis method that measures and accounts for the structured heterogeneity within complex disease populations is still needed.

We present SPIT, a statistical tool that identifies subgroups within populations at the transcript level and compares their isoform abundance measures. Using both simulated and real RNA-Seq data from human heart tissue, we show that SPIT improves specificity rates compared to the state-of-the-art tools with similar sensitivity, and detects DTU events exclusive to subgroups as well as DTU events shared amongst all case samples. Downstream of DTU analysis, SPIT uses detected DTU events to provide insight into potentially hierarchical subgrouping patterns present in complex disease populations using hierarchical clustering.

Within the SPIT algorithm, subgroups with divergent abundances for each transcript are detected using a kernel density estimator, after which the distributions are compared via a nonparametric Mann-Whitney U test. SPIT provides a conservative approximation of the biological and technical variability within datasets with its SPIT-Test module, significantly reducing false-discovery rates. Rather than estimating the expression variability per transcript, SPIT-Test samples a null distribution of minimal U statistic p-values based on the control group and assumes that, for each transcript, the minimal U statistic p-value is drawn from the same underlying distribution when there is no real disease association independent of biological or technical variability.

We applied SPIT to search for DTU events associated with schizophrenia, a psychiatric disorder canonically recognized as a heritable complex disease with an undetermined number of subtypes.[11-13] Genetic causes of schizophrenia have long been studied, however, a clear consensus on the level of genetic liability or the acting set of causal events has not been reached to this day. Whole genome, exome and RNA sequencing studies suggest that a wide range of both common and rare genetic variations, including single-nucleotide polymorphisms (SNPs), copy-number variations (CNVs), ultra-rare coding variants (URVs), and alternative splicing events, may contribute to the pathogenesis of schizophrenia.[9, 14-16] After analyzing RNA-Seq data from the dorsolateral prefrontal cortex (DLPFC) of 146 schizophrenia patients and 208 controls, SPIT identified six candidate genes that had statistically significant DTU events associated with schizophrenia. Previously-reported disease associations for these candidate genes include neurodegenerative and psychiatric disorders such as Alzheimer's disease, bipolar disorder, schizophrenia, major depressive disorder, attention-deficit hyperactivity disorder, and autism spectrum disorder. No previous report had identified DTU events in any of these genes.

SPIT is open-source software freely available at *https://github.com/berilerdogdu/SPIT*. Additionally, a user-friendly Google Colaboratory configuration and step-by-step guide are provided at *https://colab.research.google.com/drive/1u3NpleqcAfNz_0EAgO2UHItozd9PsF1w?usp=sharing*.

Results

A demonstration on simulated data

A DTU event is defined as a significant difference in the proportions of isoforms contributing to the overall expression of a locus between individual or groups of samples. We are particularly interested in cases where there is a clearly dominant isoform in healthy individuals, where DTU can potentially disrupt cellular function and cause anomalies.

We describe a modeled DTU case with artificially generated data in order to exemplify such DTU events, and to demonstrate the key steps of the SPIT algorithm. Consider a locus from which two distinct isoforms, Isoform 1 and Isoform 2, are transcribed as represented in Figure 1.a. Suppose that the protein translated from Isoform 1 is a functional protein, whereas Isoform 2 gets translated into a dysfunctional, aberrant protein. Consequently, the primary expression profile of this locus in a healthy individual is expected to be Isoform 1. Figure 1.b shows the relative abundances of Isoform 1 and Isoform 2 for four individuals with varying levels of expression at the locus. The left panel of Figure 1.b demonstrates a clear example of DTU between Individual 1 and Individual 2, with Isoform 1 dominant for Individual 1 and Isoform 2 dominant for Individual 2. The right panel of Figure 1.b illustrates why changes in overall expression at the gene/locus or transcript/isoform level are not sufficient indicators of DTU, as illustrated for the same isoforms in Individuals 3 and 4, where overall expression changes but the relative proportion of the isoforms remains the same.

Figure 1

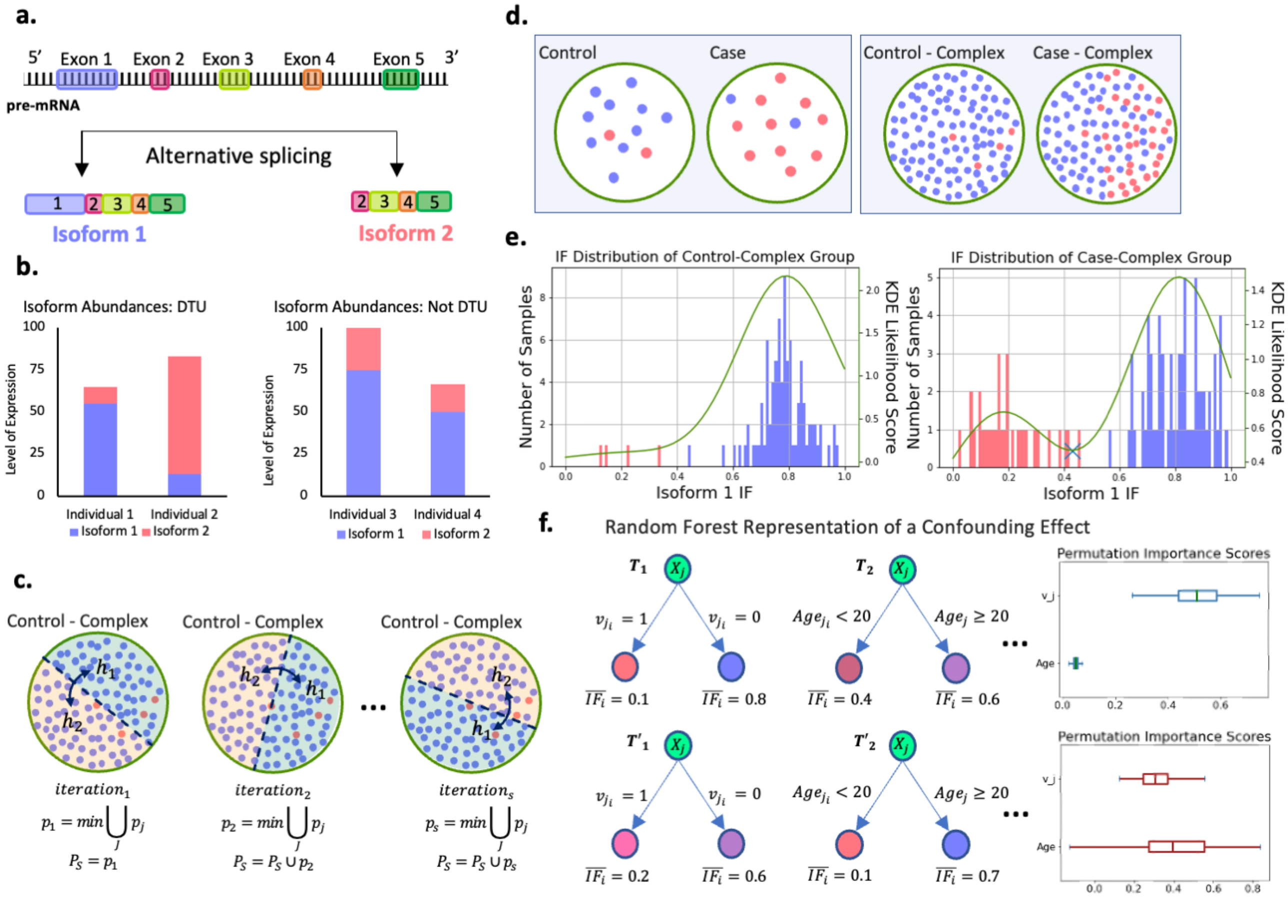

Figure 1: DTU detection demonstration a. Gene locus going though alternative splicing to produce Isoform 1 and Isoform 2. **b.** Left panel: Isoform abundances in a sample case of DTU between individuals 1 and 2. Right panel: Isoform abundances in a sample case without DTU but with changes in overall expression between individuals 3 and 4. **c. Three** SPIT-Test iterations demonstrated with random splits of the Control-Complex group. Samples (dots) are color coded based on their dominant isoforms for the locus in c-f, with blue=isoform 1 and red=isoform 2. **d.** Left panel: Conventional DTU analysis assumption with no structured heterogeneity in either group. Right panel: Heterogeneity structure in complex disease samples, where a subset of cases share the same genetic abnormality (Case-Complex). **e.** Corresponding isoform fraction (IF) distributions for the samples represented in groups Control-Complex and Case-Complex. **f.** Random forest regression representation when there is not a significant confounding effect in the DTU transcript (Upper panel) vs. when there is a clear confounding effect by the covariate "age" (Lower panel). Corresponding permutation importance scores for age and v_j are shown on the right.

DTU analysis usually entails comparing two groups of samples rather than individuals. In the interest of brevity, suppose for any given individual, either Isoform 1 or Isoform 2 is significantly dominant for the locus in our model DTU case, and note that each individual is color-coded based on their dominant isoform in Figure 1.c-e. Small sample sizes are quite common in RNA-Seq experiments[17], and the left panel of Figure 1.d represents a typical experiment setup for DTU analysis with 12 samples in each group. If a DTU event between Isoform 1 and Isoform 2 has a causal link to a disease, the left panel of Figure 1.d depicts the expected scenario for a simple genetic disease where the disease is caused by a single or a small set of genes. In this scenario, one assumes that all or nearly all controls have normal gene expression patterns, while the cases all share a distinct but abnormal gene or transcript expression pattern that has caused them to be placed in the disease cohort.

In contrast, the causal set of genes or events are not expected to be shared amongst all individuals affected by a complex disorder. The idea that the majority of complex disorders are likely polygenic, and that distinct combinations of causal events might lead to similar pathogenesis in different patient groups is widely accepted.[18] When focusing on a particular causal event such as the DTU case between Isoform 1 and Isoform 2, this implies that only a subgroup of patients within the case group are likely to have this event among their causal factors, as depicted in the right panel of Figure 1.d. By segregating this subgroup from the remaining case group, we gain the capability to detect a DTU event that might have otherwise gone unnoticed, and to differentiate potential subclusters of the disease group based on shared DTU events.

In order to do so, we compare the distributions of isoform fractions (IFs) between the two groups, which refers to the proportion of total expression attributed to each isoform. Figure 1.e shows the IF levels for Isoform 1 in both Control-Complex and Case-Complex groups, which is expectedly high for individuals with Isoform 1 as the dominant isoform at the locus, and low for individuals with Isoform 2 as the dominant isoform. By fitting a kernel density estimator (KDE)[19-21] on the IF distributions, we can search for bimodality, which if found indicates a separation within the groups themselves. The right panel of Figure 1.e demonstrates the clear partition of the Case-Complex subgroups by a global minimum marked with a cross on the KDE curve. We should note that SPIT does not presuppose the existence of a partition in populations and still detects any shared DTU events in the absence of bimodality.

Partitioning of subgroups

The transcript counts are transformed into IFs for each sample as follows:
(1)

$$IF_{i,j} = {}^{t_{i,j}}\big/_{\sum_{G_j} t_{i,j}}$$

where $IF_{i,j}$ is the isoform fraction for transcript j in sample i, $t_{i,j}$ is the transcript count for transcript j in sample i, and G_j stands for the set of all transcripts that belong to the same gene as transcript j. We fit a KDE with Gaussian kernel[19-21] (details on bandwidth selection are described in the Methods section on parameter fitting) on the two vectors of $IF_{I_c,j}$, where I_c stands for the samples in groups $c \in \{\text{case, control}\}$. If the $IF_{I_{case},j}$ distribution is bimodal, indicating a significant stratification of two subgroups based on the dominance status of transcript j, we observe this as a global minimum of the KDE (Figure 1.e). While we acknowledge the possibility of observing a similar divergence within the control group due to technical or biological variability, our primary objective is to identify subgroups within the case samples for potential associations with disease status. The KDE on control group is utilized for flagging the most significant candidate DTU genes as described in the Methods section.

There are several advantages to detecting subgroups based on density estimation, the most important of which is the ability to avoid an underlying distribution assumption for the data set, which can be challenging for RNA-Seq driven data even after multiple normalization steps.[22] Furthermore, while outlier samples can alter the shape of a KDE, they have a relatively negligible impact on the global minima/maxima as long as appropriate smoothing is applied.[21] Unlike k-means or hierarchical clustering methods, there is not a hyperparameter that fundamentally effects whether or not clusters are detected in the data, and the choice of the bandwidth parameter (h) works in our advantage to account for overdispersion by oversmoothing (see Methods section on parameter fitting).

In the presence of a global minimum in the case group at $IF_{i,j} = m_{case}$, we define the left tails of the case and control IF_j distributions as the samples that fall to the left of point m_{case}, and the right tails as the samples that fall to the right:

(2)

$$l_{case} = \{i \in I_{case} \mid IF_{i,j} \leq m_{case}\} \text{ and } r_{case} = \{i \in I_{case} \mid IF_{i,j} > m_{case}\},$$
$$l_{control} = \{i \in I_{control} \mid IF_{i,j} \leq m_{case}\} \text{ and } r_{control} = \{i \in I_{control} \mid IF_{i,j} > m_{case}\}.$$

To search for candidate DTU events in l_{case} and r_{case} independently, the left tails of the case and control IF_j distributions are compared internally, as are the right tails, using the non-parametric Mann-Whitney U test. I.e. $\bigcup_{i \in l_{case}} IF_{i,j}$ is compared with $\bigcup_{i \in l_{control}} IF_{i,j}$, while $\bigcup_{i \in r_{case}} IF_{i,j}$ is compared with $\bigcup_{i \in r_{control}} IF_{i,j}$. This analysis determines whether the samples in l_{case} could have been drawn from the left-tail control samples with $IF_{i,j} \leq m_{case}$, or if they exhibit significant differences. Likewise, the same rationale applies for the right tails.

In the absence of a global minimum, a Mann-Whitney U test is conducted between the entire groups of I_{case} and $I_{control}$.

Estimating dispersion with SPIT-Test and detecting DTU

Although the use of non-parametric statistical tests can help control the false discovery rate (FDR) in differential analyses, the effectiveness of several competing methods is notably diminished when the input data is overdispersed with outliers[23], a common characteristic of RNA-Seq data[24]. This prevalent phenomenon suggests that we are not capable of precisely estimating dispersion for each individual transcript or gene, in addition to not being able to adequately correct for the vast number of hypotheses being tested. To overcome this challenge, we choose to estimate a single null distribution for the minimal Mann-Whitney U-statistic p-values, and assume that these observed minimal p-values reflect the upper threshold of dispersion in the input dataset.

The true null distribution P_S of the minimal U-statistic p-values represents the lowest expected p-values when there is no real association between the phenotype of interest, such as a disease, and the changes in isoform dominance among individuals or groups. To estimate $\widehat{P}_S$, SPIT-Test takes advantage of the control group in which disease association is absent, yet individual differences due to biological, technical, or other confounding factors can be observed. As illustrated in Figure 1.c, SPIT-Test is an iterative process which randomly splits the control group in half, and identifies the most significant difference in isoform fractions between the two halves. Later on, the candidate DTU events between the case and control groups are compared, in terms of their significance, to the observed differences between random halves of the control group.

The following steps are performed at each iteration s:

1. Randomly split the control samples into two sets of equal size, namely $h_{k,s}$ where $k \in \{1, 2\}$ represents each half for iteration s.
2. Select a random split point o_s, to define the left and right tails of each half as:
   $l_{h_{1,s}} = \{i \ \in I_{h_{1,s}} \left| IF_{i,j} \leq o_s \right\}$ and $r_{h_{1,s}} = \{i \ \in I_{h_{1,s}} \left| IF_{i,j} > o_s \right\}$,
   $l_{h_{2,s}} = \{i \ \in I_{h_{2,s}} \left| IF_{i,j} \leq o_s \right\}$ and $r_{h_{2,s}} = \{i \ \in I_{h_{2,s}} \left| IF_{i,j} > o_s \right\}$
3. For each transcript j, conduct a Mann-Whitney U test between the sets of $l_{h_{1,s}}$ and $l_{h_{2,s}}$, yielding a Mann-Whitney U-statistic p-value $p_{j_{l,s}}$. Similarly, conduct a Mann-Whitney U test between the sets of $r_{h_{1,s}}$ and $r_{h_{2,s}}$, yielding $p_{j_{r,s}}$.
4. Assign $p_{j,s} = \min(\ p_{j_{l,s}},\ p_{j_{r,s}})$ to each transcript j for iteration s.
5. Among the U-statistic p-values assigned to all transcripts, store $p'_s = \min \bigcup_J p_{j,s}$. In order to avoid excessive influence from outlier transcripts, we only sample p' once from the same transcript throughout all iterations. In other words, in iteration s we consider transcripts from which $p'_{s_1,\ldots,s_{n-1}}$ have not been sampled.
6. $\hat{P}_S = \hat{P}_S \cup p'_s$.

SPIT-Test estimates dispersion on a global scale, assuming that any transcript could have been subject to the highest observed level of dispersion. Therefore, for an arbitrary transcript j, $\hat{P}_S$ is considered as an empirical null distribution of the minimal U-statistic p-value. This approach emulates the min-P and max-T procedures[25], and is employed to set a p-value threshold, $p'_{threshold}$, based on $\hat{P}_S$ that determines the set of candidate DTU transcripts between case and control samples as:

(3)

$$p'_{threshold} = \left(\kappa * \left|\hat{P}_S\right|\right)^{\text{th}} \text{ smallest } p\text{-value in } \hat{P}_S,$$

where κ is a user-set parameter. For instance, if $\kappa = 0.1$ for 1000 iterations, the threshold would be the 100th smallest p-value. SPIT-Test deviates from a traditional permutation test in its randomization steps 1 and 2, and its exclusion of the case samples due to the potential presence of unknown subgroups. Although κ cannot directly translate into a target family-wise error rate (FWER), we experimentally show that smaller values of κ achieve remarkable control over FWER.

DTU simulation and evaluation

Simulated RNA-Seq reads are conventionally used to evaluate differential analysis tools, as we lack knowledge of ground truth in real data. However, research has consistently shown that simulated reads do not accurately represent the overdispersion levels in real RNA-Seq experiments, leading to underestimation of FDR.[23, 26] In order to obtain a more accurate assessment of SPIT's performance, we make use of both simulated and real RNA-Seq data. In these two types of evaluation sets, we compare the true positive rate (TPR) and FDR outcomes of SPIT, and the state-of-the-art tools *DEXSeq*[27] and *DRIMSeq*[28] used together with the stage-wise adjustment tool *stageR*.[29]

Evaluation with simulated RNA-Seq reads

We borrow the DTU simulation with the largest sample sizes from the "Swimming Downstream" pipeline by Love *et al*.[30] as our test dataset with simulated RNA-Seq reads. (Please see the corresponding Methods section for details.) This dataset simulates a large number of (> 1500) DTU events in relatively homogenous populations, resembling the scenario depicted in the left panel of Figure 1.d. While dispersion is incorporated into the transcript expression patterns, there are no subgroups or divergence in the DTU events.

The TPR and FDR at the gene level are reported for each tool in Figure 2.b, where both *DEXSeq* and *DRIMSeq* have 3 outcomes corresponding to *stageR* target overall FDR (OFDR) values $0.01, 0.05, 0.1$. For SPIT, we report 5 outcomes corresponding to setting hyperparameter $\kappa = 0.2, 0.4, 0{,}6, 0.8$ and 1 on 1000 iterations. Although the tuning of target OFDR for *stageR* and κ for SPIT are not directly comparable, lower values of both parameters lead to more conservative behavior, allowing better control over FDR and often yielding decreased TPR.

TPR and FDR outcomes of *DEXSeq* and *DRIMSeq* were consistent with the "Swimming Downstream" evaluation. Both tools yielded high sensitivity levels while *DEXSeq* maintained a slightly better control over FDR. On the same simulated dataset, SPIT yielded a comparable yet slightly lower TPR value while always keeping the FDR lower than 0.05. Different values of the hyperparameter κ did not result in noticeable differences in TPR or FDR on this dataset.

Evaluation with real RNA-Seq reads

To form the basis of our test dataset with real RNA-Seq reads, we quantified Illumina reads of 235 normal heart (left-ventricle) samples obtained from the Genotype-Tissue Expression (GTEx) project[31]. Figure 2.a shows the mean-standard deviation plots of the two datasets, revealing a significantly higher level of dispersion in the GTEx dataset compared to the "Swimming Downstream" dataset of simulated RNA-Seq reads.

Next, we conducted 20 separate experiments in each of which we compared random halves of the GTEx dataset after introducing 100 simulated DTU events into one of the halves (please see the corresponding Methods section for details). In an effort to model the expected heterogeneity in a complex disease group, we distributed the 100 DTU events between 5 subgroups in such a way that some DTU events are shared between the subgroups while some are exclusive (see Figure 2.c for an example). For the rest of the paper we'll refer to any such subgroup that shares the same DTU events as a "splicotype" group.

In any random partition of real RNA-Seq samples into two groups, it is not certain that there are no actual DTU events beyond the ones we introduced. Therefore, the TPR and FDR measures for the GTEx experiments are only estimates. Our hypothesis in evaluating these experiments was that if any method consistently detected additional DTU events between random partitions of a healthy sample group, the discoveries were either noise or else due to biological variance that are not of interest. As such, we present the mean estimated FDR and TPR values of 20 experiments for SPIT and *DRIMSeq*+*stageR* pipeline in Figure 2.b with error bars indicating the minimum and maximum FDR/TPR values obtained. Additionally, in Figure 1.d, we show the individual Venn diagrams representing the overlap between the *DRIMSeq*+*stageR* pipeline and SPIT results with the simulated DTU genes for the first experiment out of the 20 conducted. Venn diagrams for the remaining experiments showing similar results are provided in Supplementary Fig.1.

Due to its generalized linear model (GLM) fitting step, *DEXSeq* requires significant compute time for large sample sizes. After running for 168 hours on 24 cores and 256 GB RAM, dispersion estimation for the first experiment remained unfinished. Therefore, we only compare *DRIMSeq* + *stageR* and SPIT results for the GTEx experiments.

In line with the "Swimming Downstream" evaluation, we applied *DRIMSeq* followed by *stageR* with target OFDR values of 0.01, 0.05, and 0.1 to the GTEx experiments. Because the SPIT pre-filtering process is included in the DTU simulation, we performed *DRIMSeq* + *stageR* analysis on the SPIT-filtered counts and bypassed *DRIMSeq* filters.

In contrast to the TPR and FDR values obtained with the simulated "Swimming Downstream" dataset, the *DRIMSeq* + *stageR* pipeline yielded a wider range of estimated TPR and FDR values on the GTEx experiments. For the GTEx experiments, the *DRIMSeq* + *stageR* pipeline produced lower TPR and notably higher FDR estimates for all target OFDR values (0.01, 0.05, and 0.1), with a more significant difference in performance between each OFDR value. We also note the wide error bars in the pipeline, indicating a large range of performance across all 20 experiments. This variability could be attributed to the distinct biological differences between the random partitions in each experiment or to the level of heterogeneity introduced in the simulation through varying compositions of shared DTU events between random splicotypes.

As with the "Swimming Downstream" dataset, we report TPR and FDR estimates for SPIT obtained by setting hyperparameter $\kappa = 0.2, 0.4, 0{,}6, 0.8$ and 1. For input datasets with large number of control samples ($n \geq 32$), SPIT offers an optional cross-validation procedure to estimate the optimal value κ^* based on inferred dispersion, which is detailed in the Methods section on parameter fitting. In Figure 2.b, the TPR and FDR obtained using the estimated κ^* is represented by a triangle, which for this dataset is 0.6. Overall, the estimated TPR and FDR levels for SPIT remained comparable to the values obtained for the "Swimming Downstream" dataset with a slight increase in both TPR and FDR. The gain in sensitivity is expected for SPIT with large sample sizes since it uses the Mann-Whiney U test when comparing any two sets of IF values. While the optimal κ^* parameter still has an estimated FDR < 0.05, SPIT's control over FDR also diminished with real RNA-Seq reads compared to the simulated test set. A clear increase in both TPR and FDR was observed for $\kappa = 0.8$ and $\kappa = 1$, which are included to demonstrate the effects of using radically large values for hyperparameter κ. The range represented by the error bars in Figure 2.b is smaller for SPIT compared to that of *DRIMSeq*, which indicates higher consistency across all 20 experiments.

Upon detecting the DTU events for any given dataset, SPIT outputs a binary matrix M of DTU events that marks the presence (1) or absence (0) of a DTU event at the gene level for any sample in the case group relative to the control group. We show that using SPIT's output matrix M, we are able to cluster the case samples into their separate splicotype groups based on their shared events by applying hierarchical clustering. The chosen distance metric calculates the proportion of unique events between any two samples relative to the total number of DTU events. As shown in Figure 2.c, SPIT perfectly captures the five clusters that were artificially created. Clustering on the first experiment is shown in Figure 2.c based on the SPIT output with κ^*; the remaining experiments are shown in Supplementary Fig.1.

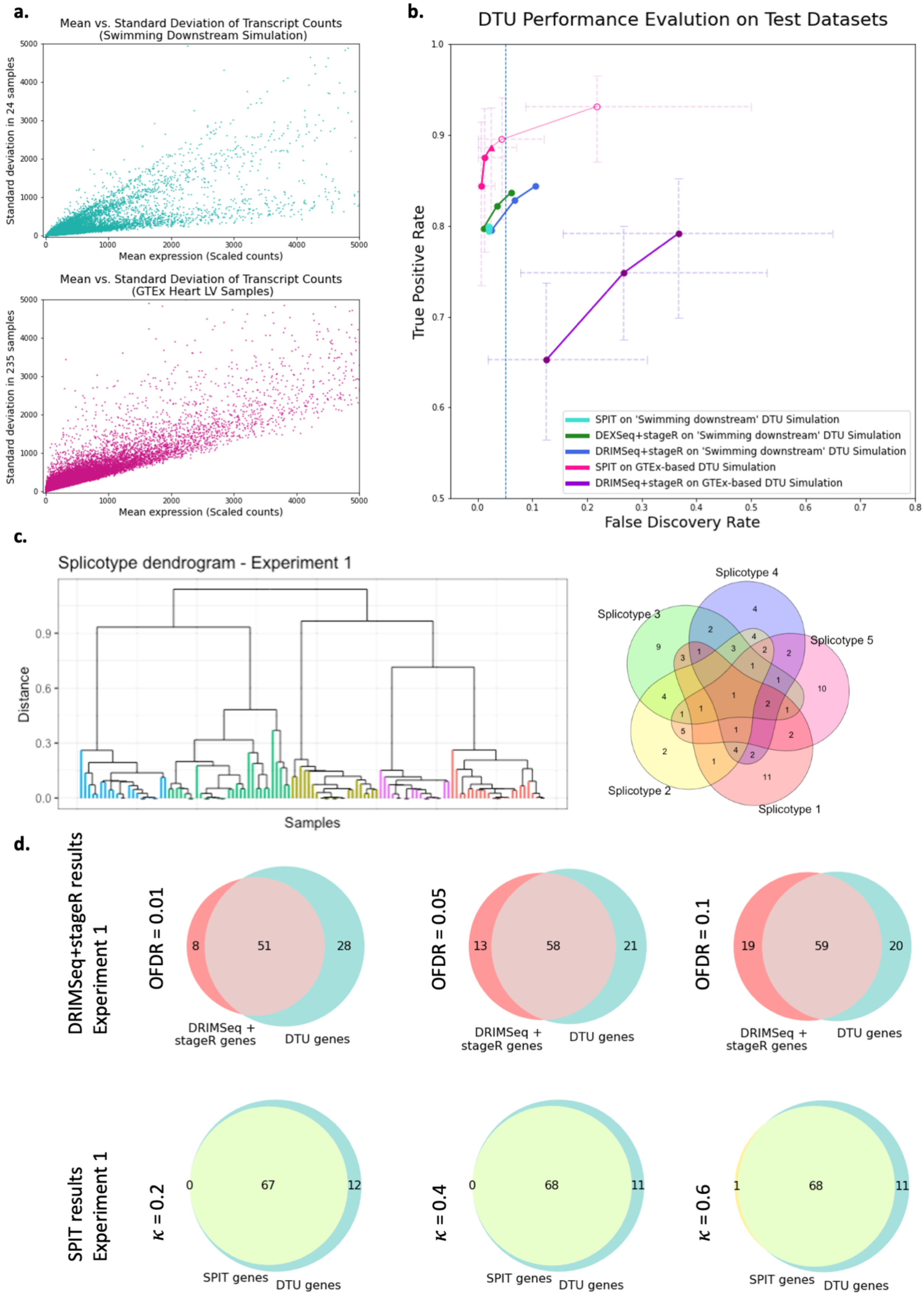
a.
Mean vs. Standard Deviation of Transcript Counts
(Swimming Downstream Simulation)
Standard deviation in 24 samples
Mean expression (Scaled counts)
0
1000
2000
3000
4000
5000
Mean vs. Standard Deviation of Transcript Counts
(GTEx Heart LV Samples)
Standard deviation in 235 samples
Mean expression (Scaled counts)
b.
DTU Performance Evalution on Test Datasets
True Positive Rate
False Discovery Rate
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
SPIT on 'Swimming downstream' DTU Simulation
DEXSeq+stageR on 'Swimming downstream' DTU Simulation
DRIMSeq+stageR on 'Swimming downstream' DTU Simulation
SPIT on GTEx-based DTU Simulation
DRIMSeq+stageR on GTEx-based DTU Simulation
c.
Splicotype dendrogram - Experiment 1
Distance
0.9
0.6
0.3
0.0
Samples
Splicotype 4
Splicotype 3
Splicotype 5
Splicotype 2
Splicotype 1
4
2
9
3
4
2
1
3
1
2
1
10
4
1
1
2
1
1
5
1
2
2
4
2
1
11
d.
DRIMSeq+stageR results
Experiment 1
OFDR = 0.01
8
51
28
DRIMSeq +
stageR genes
DTU genes
OFDR = 0.05
13
58
21
DRIMSeq +
stageR genes
DTU genes
OFDR = 0.1
19
59
20
DRIMSeq +
stageR genes
DTU genes
SPIT results
Experiment 1
κ = 0.2
0
67
12
SPIT genes
DTU genes
κ = 0.4
0
68
11
SPIT genes
DTU genes
κ = 0.6
1
68
11
SPIT genes
DTU genes

Figure 2:a. Mean vs. standard deviation of the transcript counts are plotted for the Swimming Downstream and GTEx experiment samples to represent relative dispersion levels. **b.** Gene-level DTU-performance evaluation on both Swimming Downstream and GTEx test datasets. Radical values of $\kappa = 0.8$ and $\kappa = 1$ are included (as unfilled circles) to show the effects of hyperparameter adjustment. **c.** The DTU event sharing Venn diagram for the first experiment in the GTEx simulations (Right), and the corresponding final subcluster dendrogram based on the SPIT DTU matrix (Left). The subclusters are color coded based on their distinct sets of simulated DTU events (splicotypes). **d.** Overlap of DRIMSeq+stageR pipeline (Top) and SPIT (Bottom) results with simulated DTU genes in the first experiment.

Detecting known tissue-dependent DTU events

As a positive control experiment, we next investigated a set of four tissue-dependent DTU events that had been previously confirmed individually by various studies and also collectively validated by Reyes & Huber in 2018[32]. For this analysis, we utilized samples from the GTEx dataset (Supplementary Table 1) that were aligned as part of the CHESS 3 project.[33] Figure 3 visually illustrates differentially expressed transcripts between tissues at each locus. All transcriptional landscape were created using the sashimi plot module in TieBrush after aggregating read alignments from all samples in each tissue. SPIT results on all four DTU events are detailed below.

SLC25A3

The mitochondrial phosphate transporter gene *SLC25A3* exhibits a phenomenon known as “mutually exclusive exons”[3], which refers to the observation that specific exons within the gene are spliced into distinct isoforms but they are not simultaneously present within the same isoform. We compared 497 samples of heart tissue and 380 samples of colon tissue from the GTEx dataset, and SPIT was able to confirm that one of these isoforms, which is recognized as the primary expression preference in heart and skeletal muscle, is indeed more prevalent in heart tissue samples (Figure 3.a).

ANK3

Together with two more ankyrin genes, *ANK3* plays a crucial role in generating a diverse array of ankyrin proteins in mammals. Tissue-specific splicing of *ANK3* has previously been shown in skeletal muscle and tibial nerve tissues[32, 34]. A total number of 480 muscle and 339 nerve tissue samples from GTEx were analyzed using SPIT, confirming the presence of an isoform switch characterized by alternative start sites and distinct patterns of exon splicing (Figure 3.b).

MEF2C

MEF2 transcription factors are significant in regulating cell differentiation and expression, and they undergo tissue-specific alternative splicing, adding to their functional diversity. *MEF2C* in humans has two mutually exclusive exons, one of which is shown to be more prevalent in skeletal muscle[35]. We compared 480 muscle tissue samples from GTEx with 361 thyroid samples using SPIT and were able to detect the isoform switching as a significant DTU event (Figure 3.c).

MYO1C

Myosin IC encodes a protein of the myosin family, which serves multiple cellular functions including vesicle transportation, transcription and DNA repair[36, 37]. The presence of a tissue-dependent transcription start site in *Myosin IC* has been demonstrated, leading to splicing of an alternative first exon[36], which SPIT successfully detects upon comparing 497 heart and 199 pancreas samples from GTEx (Figure 3.d).

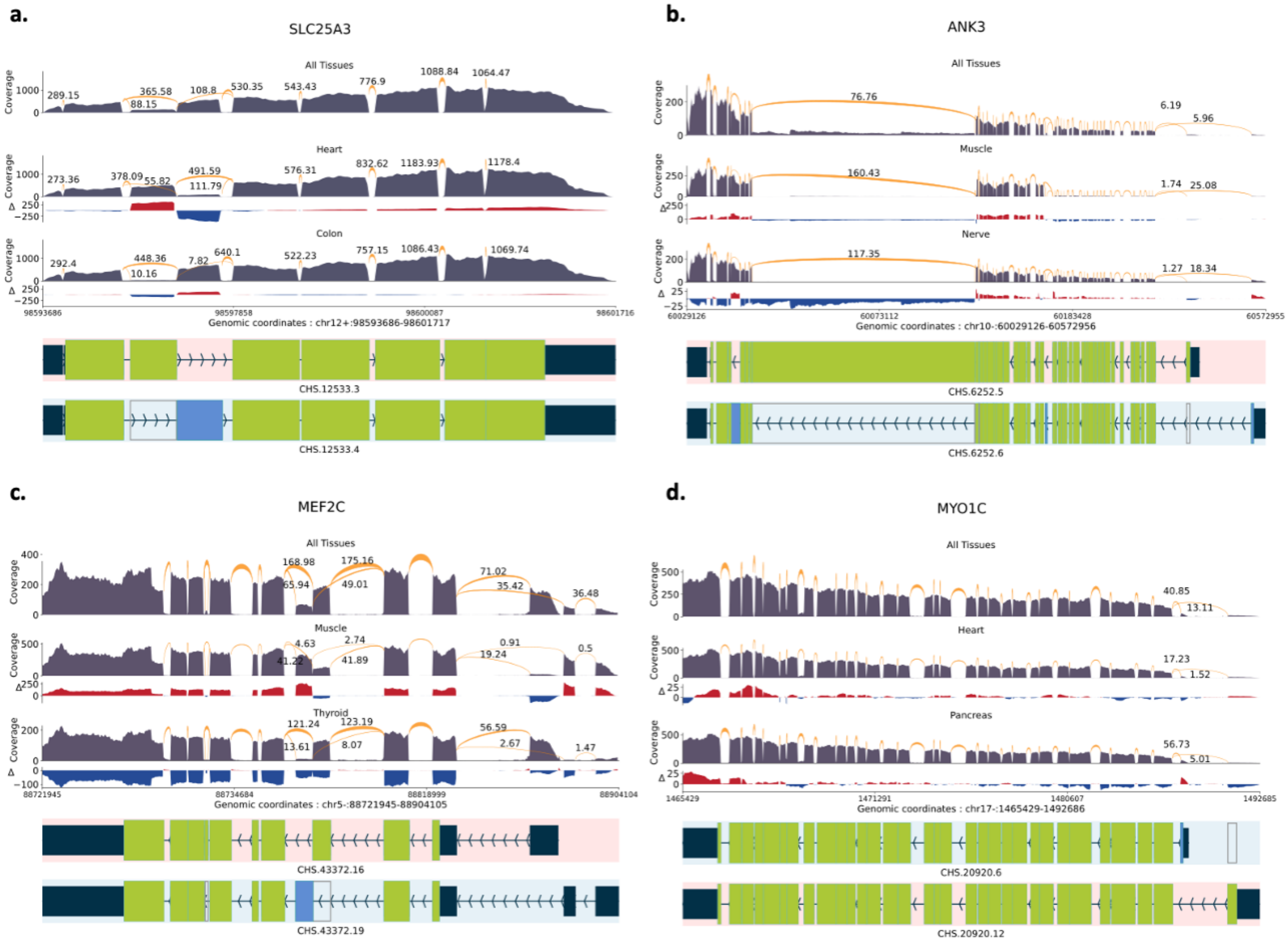

Figure 3: Sashimi plots with normalized coverage and junction values from GTEx samples of CHESS 3 project. Only the relevant isoforms and junction values are included for readability. The normalized coverage values for each tissue were subtracted from the normalized coverage of the entire GTEx dataset, and the results were illustrated as the Δ track. **a.** *SLC25A3* DTU event between heart and colon tissues. **b.** *ANK3* DTU event between muscle and nerve tissues. **c.** *MEF2C* DTU event between muscle and thyroid tissues. **d.** *MTO1C* DTU event between heart and pancreas tissues.

Schizophrenia application

After evaluating its performance, we explore the application of SPIT in identifying DTU genes associated with schizophrenia, where we expect a divergence in the causal mechanisms underlying pathogenesis for individual or groups of patients. We obtained RNA-Seq samples of post-mortem DLPFC tissue from a total of 354 adult brains, which were sequenced by the Lieber Institute for Brain Development.[38] After applying various quality filtering criteria that are described in detail in the Methods section, we selected 146 schizophrenia samples and 208 control samples for comparison in our analysis (Supplementary Table 2).

The parameter-fitting process was applied to the control samples, resulting in $\kappa^* = 0.4$. Prior to confounding analysis, SPIT detected 135 potential DTU events between the case and control samples. The binary DTU matrix for these 135 transcripts was then inputted to the confounding control module of SPIT which is described in the Methods section. Covariates considered for all samples included sex, race, age, batch identification, and RNA integrity number (RIN) which highly correlates with RNA degradation.[39] 129 candidate transcripts were eliminated based on their permutation importance scores, leaving a final set of six DTU transcripts in six genes (Figure 4.c). The SPIT-Chart for this analysis (Figure 4.a) shows the relationship between the median *p*-values obtained from 1000 iterations of SPIT-Test and the *p*-values resulting from comparing control and schizophrenia samples for transcripts.

Amongst the six candidate genes, four (*BDH2*, *CLDND1*, *GAS8*, *TRIP4*) displayed DTU events in all schizophrenia samples, while the other two genes (*LARP4*, *NVL*) showed significant DTU events in specific subgroups. Figure 4.b depicts the clustering of schizophrenia samples based on identified DTU events, revealing a partitioning into four subgroups in this dataset. We present short descriptions of the functions and associations of the six candidate genes below.

GAS8 (Growth Arrest Specific 8): A multi-tissue study examined SNPs for enrichment of expression quantitative trait loci (eQTL) across 11 genome-wide association studies (GWAS) focused on schizophrenia and affective disorders (including bipolar disorder, major depressive disorder, autism spectrum disorder, and attention-deficit hyperactivity disorder)[40]. The study identified *GAS8* amongst genes affected by the high-confidence cis-eQTLs in multiple brain regions, and reported its cross-disorder associations as well as specific associations with bipolar disorder.

NVL (Nuclear VCP Like): This gene is a member of the AAA family (ATPases associated with diverse cellular activities) and encodes for two proteins with recognized distinct functions, *NVL1* and *NVL2*[41], the latter of which is involved in regulating ribosome biogenesis in eukaryotes[42]. There is a growing body of evidence suggesting correlations between disrupted ribosome synthesis and aging, as well as neurodegenerative diseases like Alzheimer's disease and Parkinson's disease[43-48]. In the subset of schizophrenia samples where *NVL* is implicated in DTU, we observed that the *NVL1* isoform was preferred, potentially indicating perturbed ribosomal synthesis (Supplementary Figure 3).

LARP4 (La Ribonucleoprotein 4): The protein encoded by this gene enables RNA-binding activity and plays a critical role in translation regulation[49]. *LARP4* has been found to show differential expression between the unaffected siblings and first-degree relatives of schizophrenia patients compared to unaffected individuals unrelated to the patients[50].

BDH2 (3-Hydroxybutyrate Dehydrogenase 2): This gene is responsible for encoding a siderophore that plays a crucial role in maintaining iron balance within cells, offering protection against oxidative stress[51]. Studies have indicated a significant downregulation of *BDH2* in response to inflammation and endoplasmic reticulum (ER) stress[52]. Disrupted iron homeostasis and ER stress have long been associated with neurodegenerative diseases like Alzheimer's disease and Huntington's disease[53, 54]. Recent studies report *BDH2* to be directly implicated in Alzheimer's disease progression[55].

TRIP4 (Thyroid Hormone Receptor Interactor 4): The protein encoded by this gene is one of the four components of the activating signal cointegrator 1 (ASC-1) complex. Mutations in ASC-1 components have been described as shared anomalies between the neurodegenerative diseases Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA)[56]. Mutations in *TRIP4* and *ASCC1*, another component of ASC-1 complex, are widely recognized as a cause of SMA[57, 58].

CLDND1 (Claudin Domain Containing 1): This gene encodes transmembrane proteins of tight junctions, which play a role in regulating the permeability of brain endothelial cells[59]. *CLDND1* has been linked to Alzheimer's disease[60], with one study indicating a potential correlation specifically with a subgroup of the condition[61].

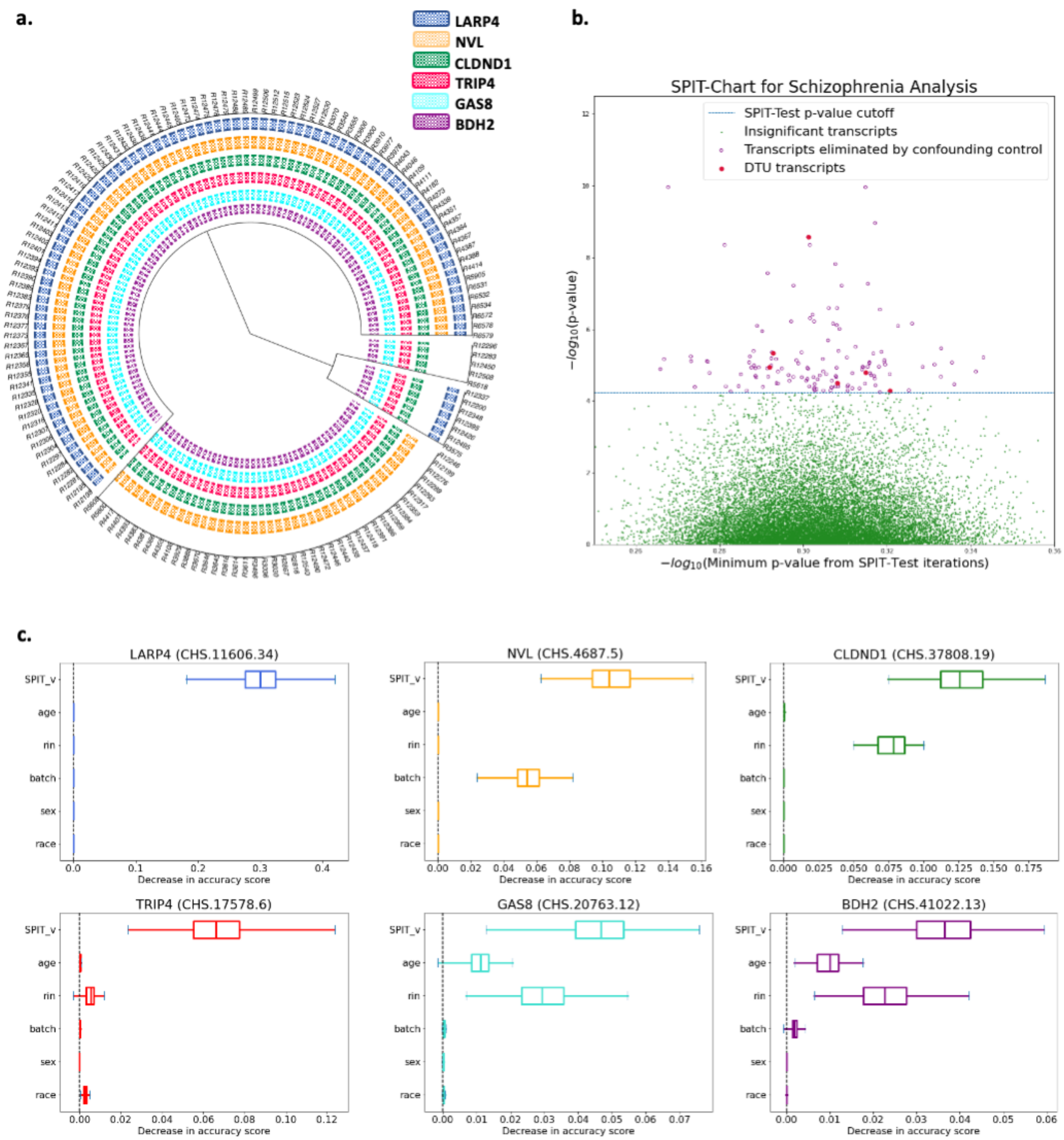

Figure 4: a. Dendrogram representation of hierarchical clustering applied on the SPIT DTU matrix for schizophrenia samples **b. SPIT-Chart for the schizophrenia analysis**: For each transcript that passed the initial filtering steps, the median p-value that has been observed through 1000 iterations of the SPIT-Test ($\text{median}\ (\bigcup_S p_{j,s})$) is plotted on the x-axis, and the p-value observed in the actual comparison of the schizophrenia samples to the controls is plotted on the y-axis, both on $-log_{10}$ scale. **c.** Box plots of permutation importance scores (generated from 100 permutations) of the SPIT output vector and provided covariates for the final 6 DTU genes.

Discussion

Transcriptomic profiles in populations with complex diseases can exhibit inherent complexity where differentially expressed events are not necessarily shared among all individuals affected by the specific disorder. Consequently, applying the same statistical assumptions for these populations as those used for simple genetic disorders can lead to misleading results in differential analyses. SPIT is the first DTU tool built to accommodate and detect structured heterogeneity within populations. Through DTU simulations built on GTEx samples, we show that SPIT not only achieves improved sensitivity and specificity in detecting DTU genes in heterogeneous populations, but also successfully captures the specific DTU events for the prevalent subpopulations present.

Our results on the “Swimming Downstream” dataset by Love *et al*. also demonstrate that SPIT is equally effective on relatively homogeneous populations, and proves to be applicable for diverse scenarios, including simple genetic disorders, tissue-to-tissue comparisons, and other types of DTU studies. SPIT consistently maintains notably low false discovery rates regardless of the level of dispersion in the datasets.

In addition to simulated experiments, we present four previously confirmed tissue-specific DTU cases that SPIT successfully detected in GTEx samples, as well as six novel DTU associations with schizophrenia. However, to establish any causal link between these six candidate DTU events and schizophrenia, a much more comprehensive investigation is needed, which is beyond the scope of this paper.

Methods

Pre-filtering

The main input SPIT requires is transcript-level count data from an RNA-Seq quantification tool, a mapping file that assigns gene names to each of the transcripts, and any metadata for the samples. Pre-filtering the transcripts before DTU analysis has been shown to improve performance for state-of-the-art tools[30, 62], which also holds true for SPIT. The default behavior of SPIT involves the stringent pre-filtering steps listed below which build upon the *DRIMSeq* filtering criteria:

1. Each transcript must have a Counts per million (CPM) value of at least 10 in at least n_{small} samples, where n_{small} is a user-set parameter that defines the smallest sample size presumed for the subgroups within populations.
2. Each transcript must have a positive read count in at least a fraction p_r of the samples in both the case and control groups, respectively. p_r is a user-set parameter and defaults to 0.20.
3. Each gene must have a read count of at least g_c in at least g_n samples, where g_c and g_n are user-set parameters and default to 10.

4. Each transcript must have an IF value larger than f in at least n_{small} samples, where f is a user-set parameter and defaults to 0.1.
5. After the filtering steps above, there must remain at least 2 transcripts for each gene.
6. The control group must have a consistently dominant isoform for each gene. This criterion is met for a gene when the same isoform of the gene has the largest IF in at least a fraction p_d of the control samples, where p_d is a user-set parameter and defaults to 0.75.

As is the case for any filtering criteria prior to differential analyses, these steps may inadvertently exclude genuine DTU cases and lower sensitivity. Thus, while these steps are included and recommended in the SPIT pipeline, any or all of them can be excluded from the analysis by the user. Supplementary Figure 2 outlines the application of this filtering pipeline on the schizophrenia samples discussed in the Results section.

Test set with simulated RNA-Seq reads: "Swimming Downstream"

Love *et al*. simulated DTU events in 1,500 genes by swapping Transcript Per Million (TPM) abundance values between two isoforms. In an additional 1,500 genes, they simulated differential transcript expression (DTE) by altering the abundance value of a single isoform by a fold change between 2 and 6. For these DTE genes, if the differentially expressed transcript is not the only isoform, they were also considered DTU cases as the relative isoform abundances were also expected to change. We include both types of these DTU events in our analysis.

Love *et al*. conducted four experiments with various sample sizes in the case and control groups ($n = 3$ vs. $3, n = 6$ vs. $6, n = 9$ vs. $9, n = 12$ vs.12) to evaluate state-of-the-art DTU tools *DEXSeq*, *DRIMSeq*, *RATs*, and *SUPPA2*. They reported that while *SUPPA2* and *RATs* always controlled their FDR, their sensitivity levels remained consistently low across all experiments, hovering around 50%. *DRIMSeq* and *DEXSeq* had considerably higher sensitivity ($\geq 75\%$) while sometimes exceeding their target FDR. Both *DRIMSeq* and *DEXSeq* demonstrated improved FDR control with larger sample sizes, and 12 vs. 12 yielded the most favorable TPRs and FDRs.

Based on these findings, we choose to reproduce the "Swimming Downstream" results obtained with *DEXSeq* and *DRIMSeq* on the $n = 12$ vs. 12 experiment and to evaluate SPIT's performance on the same dataset. We first ran *DEXSeq* and *DRIMSeq* on the released Salmon [63] quantification files by Love *et al*.[64] as outlined in the "Swimming Downstream" workflow. We then applied the stage-wise adjustment tool *stageR* on the preliminary results from both *DEXSeq* and *DRIMSeq* using target OFDR values $0.01, 0.05$, and 0.1. The "Swimming Downstream" evaluation first applied the *DRIMSeq* pre-filters on the simulated counts and defined the set of true positives as the DTU genes and transcripts that pass the *DRIMSeq* filter. To be able to replicate the reported TPR and FDRs and compare results, we applied the same filters and skipped the SPIT pre-filtering process.

Test set with real RNA-Seq reads: GTEx simulation

To simulate each of the 20 GTEx experiments the following steps were executed:

1. Randomly divide the 235 GTEx samples into two sets to create case and control groups, I_{case} and $I_{control}$, comprising of 117 and 118 samples, respectively.
2. Apply the SPIT pre-filter outlined above assuming the randomly assigned I_{case} and $I_{control}$. Note that we skip step 6 of the pre-filtering as it could create an unfair bias in the pre-filtered set of genes towards the DTU genes selected in the next step.
3. We apply the criteria outlined in step 6 of the pre-filtering process to identify genes with consistently dominant isoforms within the $I_{control}$ group. Out of these genes with dominant isoforms, we randomly select 100 to compose our superset of DTU genes, $\mathrm{D} = \{\mathrm{d}_1, \mathrm{d}_2, \dots, \mathrm{d}_{100}\}$.
4. For each splicotype group (subgroup of samples that share the same DTU events) π_s, $s \in \{1, 2, 3, 4, 5\}$, we randomly select 30 DTU genes from set D with replacement to form D_{π_s}. This results in a complex and structured partition within I_{case}, where some DTU genes are shared between the five splicotypes while others are unique to a specific splicotype.
5. For a DTU gene $d_k \in D_{\pi_s}$, let α_k be the dominant isoform of d_k in $I_{control}$ with $\overline{IF} = u$, and β_k be the least dominant isoform in $I_{control}$ with mean $\overline{IF} = v$.
We switch the dominance status of a_k and β_k in I_{case} by allowing $IF_{a_k,i} = v \pm \epsilon$ and $IF_{\beta_k,i} = u \pm \epsilon$ for all $i \in D_{\pi_s}$, where noise parameter $\epsilon = 0.05$.
6. Within all simulated DTU cases, the original transcript counts for a_k and β_k are updated by multiplying the gene counts by $IF_{a_k,i}$ and $IF_{\beta_k,i}$, respectively. The gene counts are updated subsequently as the sum of all updated transcript counts, and IF values are calculated once again with equation (1) so that within each gene IF values add up to 1.

Addressing confounding variables

After completing the preliminary DTU analysis, the main output of the SPIT pipeline is a binary vector v_j for each transcript indicating the presence (1) or absence (0) of a DTU event in each sample in comparison to the control group. Note that v_j carries a 0 for all samples of the control group. Moreover, notice that for the transcripts that SPIT reports as significant DTU events, the v_j vector represents a partitioning of all samples, case and control, into two groups with relatively high and low IF_j values.

In the presence of a confounding effect, this partition of the high and low IF_j values can also be achieved via the confounding variable if included in the experimental design. Based on this assumption, SPIT filters out the DTU events with potential confounding effects using a random-forest-based method.

Given a set of covariates $X = \{x_1, x_2, \dots, x_k\}$, we define a matrix X_j for every candidate DTU transcript j such that $\mathrm{X}_{j_i} = [v_{j_i}, x_{1_i}, x_{2_i}, \dots, x_{k_i}]$ for any sample i in either group. We also define a vector y_j based on the IF_j values such that $y_{j_i} = IF_{j_i}$ in the same sample order as in X_j.

We then fit a random forest regressor[65, 66] $\phi_j(\mathrm{X}_j) \rightarrow y_j$ on each candidate DTU transcript. The same number of samples as in the input matrix is bootstrapped for the construction of each tree with maximum tree depth 1, and we minimize the L_2 loss on the mean IF_j in terminal nodes. Notice that with tree-depth 1, our goal is not to precisely predict IF_{j_i} for samples as much as it is to assess which covariates might be contributing into observable variance in IF_j values. We require at least n_{small} number of samples to split the root node. An illustrative case of detecting a confounding effect can be seen in the random forest depicted in Figure 1.d. Building on the modeled demonstration in Figure 1, assume that a candidate DTU event was detected for the subgroup in Case-Complex samples. Supposing one covariate (age) was provided as input, the random forest attempts to regress IF_j based on $\mathrm{X}_{j_i} = [v_{j_i}, \mathrm{age}_i]$. On the upper panel, the first tree T_1 finds the expected effectiveness of vector v_j in separating low IF_j values, as it was primarily inferred based on IF_j. A similar effective partition cannot be achieved with the provided covariate age in tree T_2.

On the lower panel, however, a partition by age in T'_2 demonstrates that age works as well as v_j in T_1, which implies that the identified DTU event cannot be confidently distinguished from a possible confounding effect of the covariate.

With the objective of estimating the importance of each covariate as well as v_j in the partitioning of high vs. low IF_j samples, we conduct a permutation importance test[66, 67] on each random forest ϕ_j. The permutation importance test is based on the coefficient of determination R_j^2 of ϕ_j, which is a score of how well IF_j is predicted in tree leaf nodes.

Let ϕ_j have L leaf nodes $\lambda_1, \dots, \lambda_l, \dots, \lambda_L$ with IF_j means $\overline{IF_{j_{\lambda_l}}}$. Then,

$$R_j^2 = 1 - {}^{u_j}/_{v_j}, \text{ where}$$

$$u_j = \sum_{l=1}^{L} \sum_{\forall i \in \lambda_l} \left(IF_{j_i} - \overline{IF_{j_{\lambda_l}}} \right)^2, \text{ and}$$

$$v_j = \sum_I \left(IF_{j_i} - \overline{IF_j} \right)^2.$$

Once the R_j^2 of ϕ_j is calculated on $\phi_j(\mathrm{X}_j) \rightarrow y_j$, one of the covariate columns of the X_j matrix is randomly permuted to form $\mathrm{X}_j^{\zeta_{k,\rho}}$, where $\zeta_{k,\rho}$ denotes a random permutation $\rho \in \mathrm{P}$ of the covariate x_k column. $R_j^{2^{\zeta_{x_k,\rho}}}$ is then calculated on $\phi_j\left(\mathrm{X}_j^{\zeta_{x_k,\rho}}\right) \rightarrow y_j$. The importance of covariate x_k is then defined as the decrease in score:

(4)

$$\gamma_{x_k} = R_j^2 - R_j^{2^{\zeta_{x_k,\rho}}}.$$[66]

Although the significance criteria can be changed by the user, in the default settings of SPIT a candidate transcript is only labeled as DTU with the following condition:

(5)

$$Q_1 \bigcup_{\mathrm{P}} \gamma_{v_j,\rho} \; > \; \max \bigcup_{X} Q_3 \bigcup_{\mathrm{P}} \gamma_{x_k,\rho} \, ,$$

where Q_1 and Q_3 refer to the 1st and 3rd quartiles of the permutation importance scores, respectively. The number of permutations for the permutation importance test is a user-set parameter and defaults to 100.

Parameter-fitting

SPIT has two main hyperparameters that affect its behavior: bandwidth (h) for KDE-fitting, and κ for p-value thresholding. The choice of bandwidth (h) directly determines the level of smoothing in the KDE function, with larger values of h leading to oversmoothed and smaller values leading to undersmoothed IF distributions.[68] In contrast to the conventional interpretation of an optimal bandwidth, selecting an optimal bandwidth for SPIT does not require achieving the highest possible accuracy in representing the underlying IF histograms. This is due to the fact that overdispersion in RNA-Seq data can lead to overly erratic histograms, which may be identified as multimodal by traditional approaches. Rather, selecting high values of h allows us to reduce the risk of false discoveries by "oversmoothing" the input IF distributions and only detecting only the most significant partitions in the data.

Similar to the choice of bandwidth, the optimal κ value also depends on the level of dispersion present in the input dataset. Smaller values of κ lead to more stringent behavior by setting smaller p-value thresholds for detecting DTU events. To estimate the optimal values of h and κ for each dataset, SPIT implements a parameter-fitting process similar to cross-validation. This involves creating a set of experiments by introducing simulated DTU events into the input control group, following the same approach as used in the GTEx test experiments. Then, different combinations of h and κ values are evaluated based on their accuracy.

Given the set of case samples I_{case} and the set of control samples $I_{control}$, we define a number (n_e) of experiments, $T = \{t_1, t_2, \dots, t_{n_e}\}$. To simulate each of the parameter-fitting experiments:

1. Randomly divide $I_{control}$ into two sets of equal size to create the simulation case and control groups, I^S_{case} and $I^S_{control}$, respectively.
2. Apply the SPIT pre-filter outlined above assuming the randomly assigned I^S_{case} and $I^S_{control}$. As with the GTEx test experiments, we skip step 6 of the pre-filtering process.
3. We repeat the steps 3-5 of the GTEx test experiment simulation on I^S_{case} and $I^S_{control}$, where the number of splicotypes introduced into I^S_{case} is a user-set parameter (n_g, defaults to 5). For simple genetic disorders and experiments with small sample sizes, n_g can be set to 1 as a complex partition within the case group is either not

expected or cannot be detected. The noise parameter ϵ can also be set by the user, and defaults to 0.05 as in the GTEx simulation.

In order to estimate the optimal values of h and κ (i.e. h^* and κ^*) out of all combinations within user-set search ranges (with defaults $0.02 \leq h \leq 0.20$ and $\kappa \in \{0.1, 0.2, \ldots, 1\}$), we employ a leave-one-out cross-validation (LOOCV) approach on the simulated set of experiments, T. For each step s in $n_e$number of iterations:

1. Let $T_{(s)} = T \setminus t_s$. We run SPIT on $T_{(s)}$ with all $(h_i, \kappa_j) \mid h_i \in \{0.02, 0.03, \ldots, 0.20\}, \kappa_j \in \{0.1, 0.2, \ldots, 1\}$ to yield estimated F-scores, F_{h_i,κ_j}.
2. Select (h_s^*, κ_s^*) such that $F_{h_s^*,\kappa_s^*} = \max \bigcup_{I,J} F_{h_i,\kappa_j}$.
3. Run SPIT on t_s with (h_s^*, κ_s^*) to get F_s.

After n_e iterations, we obtain a set of optimal hyperparameters and their corresponding F-scores: $\{(h_1^*, \kappa_1^*), (h_2^*, \kappa_2^*), \ldots, (h_{n_e}^*, \kappa_{n_e}^*)\}$ and $\{F_1, F_2, \ldots, F_{n_e}\}$. We select the hyperparameter values with the highest consensus among the iterations as our estimated optimal values (h^*, κ^*). The average F-score $\overline{(F)}$ across all iterations can be interpreted as the overall F-score of the SPIT pipeline on the provided dataset, which can help determine if SPIT is an appropriate analysis tool for the dataset. In general, larger sample sizes of the control group ($n \geq 16$) are expected to improve accuracy of SPIT test as the U-statistic is nearly normal with $n = 8$ vs. 8.[69] Consequently, the parameter-fitting experiments are expected to reveal the best results with control group sizes ≥ 32.

For the parameter-fitting experiments in this work, we used the default search ranges with $n_e = 10$ and $n_g = 5$. (h^*, κ^*) were estimated as $(0.09, 0.6)$ for the GTEx simulation experiments, and $(0.06, 0.4)$ for the Lieber brain samples. Final $\bar{F}$ across 10 experiments were $0.911$and 0.930, respectively.

SPIT's parameter-fitting process can be time-consuming and computationally intensive, and it is an optional step. Running 10 experiments ($n\_e = 10$) on a typical personal laptop can take up to 24 hours, however, multithreading is available through GNU parallel.[70] Without parameter-fitting, the default values of (h, κ) are set to the estimated optimal (h^*, κ^*) for the GTEx dataset, $(0.09, 0.6)$.

Removing outlier effects and tie-correction

Assume that a global minimum was detected in the IF distribution of case samples in order to partition subgroups for an arbitrary transcript, and the left and right tails of the case and control groups were determined as l_{case}, r_{case}, $l_{control}$, and $r_{control}$.
We define a parameter n_{small}, which defines the minimum size for subgroups that can be confidently detected and interpreted in the given dataset. If either or both of the sizes of $l_{control}$ and $r_{control}$ are smaller, they can be expanded to the right and to the left, respectively, until each contains at least n_{small} samples for comparison. Unlike the tails of the control group,

l_{case} and r_{case} represent meaningful stratifications within the case group that may have biological implications. Therefore, the group sizes of both l_{case} and r_{case} need to be at least n_{small}. Otherwise, the stratification is considered unreliable due to potential influence of outliers. In such cases a Mann-Whitney U test is conducted between the entire groups of I_{case} and $I_{control}$.

Additionally, in order to reduce the impact of insignificant differences between IF values in the Mann-Whitney U test, SPIT rounds all IF values to two decimal points.[71] The normal approximation for the U-statistic[69] is used for tie-correction for group sizes larger than 8. Although SPIT works well with smaller sample sizes ($n \geq 12$) for simple genetic architectures, it requires $n \geq 24$ samples for each group for the normal approximation to be reliable in SPIT-Test module. Exact U-statistic p-values are computed for group sizes smaller than 8 when there are no ties.

Data availability

The “Swimming Downstream” dataset is uploaded to Zenodo by Love *et al.*:
Quantification files: https://zenodo.org/record/1291522
Scripts and simulation data: https://zenodo.org/record/1410443

All 20 of the GTEx simulation experiments are uploaded to Zenodo:
https://zenodo.org/record/8128846

Quantification files and phenotype information for the GTEx samples used in the detection of tissue-dependent DTU events are uploaded to Zenodo: https://zenodo.org/record/8128945

The RNA-Seq data used in the schizophrenia analysis are made available by the Lieber Institute for Brain Development at http://eqtl.brainseq.org/phase2/.

Acknowledgements

This work was supported in part by the U.S. National Institutes of Health under grants R01-MH123567 and R01-HG006677.

References

1. Ezkurdia, I. et al. Most highly expressed protein-coding genes have a single dominant isoform. *J Proteome Res* **14**, 1880-1887 (2015).
2. Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C. & Huang, T.H. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* **24**, 167-177 (2008).
3. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476 (2008).
4. Salomonis, N. et al. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. *Proc Natl Acad Sci U S A* **107**, 10514-10519 (2010).
5. de Morrée, A. et al. Self-regulated alternative splicing at the AHNAK locus. *Faseb j* **26**, 93-103 (2012).
6. Kellermayer, D., Smith, J.E., 3rd & Granzier, H. Novex-3, the tiny titin of muscle. *Biophys Rev* **9**, 201-206 (2017).
7. Vitting-Seerup, K. & Sandelin, A. The Landscape of Isoform Switches in Human Cancers. *Mol Cancer Res* **15**, 1206-1220 (2017).
8. Gupta, M.P. Factors controlling cardiac myosin-isoform shift during hypertrophy and heart failure. *J Mol Cell Cardiol* **43**, 388-403 (2007).
9. Gandal, M.J. et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat8127 (2018).
10. Costa, V., Aprile, M., Esposito, R. & Ciccodicola, A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *European Journal of Human Genetics* **21**, 134-142 (2013).
11. Arnedo, J. et al. Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. *Am J Psychiatry* **172**, 139-153 (2015).
12. Liu, Z. et al. Resolving heterogeneity in schizophrenia through a novel systems approach to brain structure: individualized structural covariance network analysis. *Molecular Psychiatry* **26**, 7719-7731 (2021).
13. Tsuang, M.T., Lyons, M.J. & Faraone, S.V. Heterogeneity of Schizophrenia: Conceptual Models and Analytic Strategies. *The British Journal of Psychiatry* **156**, 17-26 (1990).
14. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
15. Marshall, C.R. et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics* **49**, 27-35 (2017).
16. Singh, T., Neale, B.M. & Daly, M.J. Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. *medRxiv*, 2020.2009.2018.20192815 (2020).
17. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91 (2013).
18. Wray, N.R. et al. Research review: Polygenic methods and their application to psychiatric traits. *J Child Psychol Psychiatry* **55**, 1068-1087 (2014).

19. Murray, R. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics* **27**, 832-837 (1956).
20. Emanuel, P. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **33**, 1065-1076 (1962).
21. Bernard, S.W. Density estimation for statistics and data analysis, Vol. 26. (CRC press, 1986).
22. Hawinkel, S., Rayner, J.C.W., Bijnens, L. & Thas, O. Sequence count data are poorly fit by the negative binomial distribution. *PLoS One* **15**, e0224909 (2020).
23. Li, J. & Tibshirani, R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* **22**, 519-536 (2013).
24. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
25. Westfall, P.H. & Young, S.S. Resampling-based multiple testing: Examples and methods for p-value adjustment, Vol. 279. (John Wiley & Sons, 1993).
26. Varabyou, A., Salzberg, S.L. & Pertea, M. Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments. *Genome Res* **31**, 301-308 (2020).
27. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017 (2012).
28. Nowicka, M. & Robinson, M.D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Res* **5**, 1356 (2016).
29. Van den Berge, K., Soneson, C., Robinson, M.D. & Clement, L. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biology* **18**, 151 (2017).
30. Love, M.I., Soneson, C. & Patro, R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Res* **7**, 952 (2018).
31. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).
32. Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* **46**, 582-592 (2018).
33. Varabyou, A. et al. CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. *bioRxiv*, 2022.2012.2021.521274 (2022).
34. Hopitzan, A.A., Baines, A.J., Ludosky, M.-A., Recouvreur, M. & Kordeli, E. Ankyrin-G in skeletal muscle: Tissue-specific alternative splicing contributes to the complexity of the sarcolemmal cytoskeleton. *Experimental Cell Research* **309**, 86-98 (2005).
35. Hakim, N.H.A., Kounishi, T., Alam, A.H.M.K., Tsukahara, T. & Suzuki, H. Alternative splicing of Mef2c promoted by Fox-1 during neural differentiation in P19 cells. *Genes to Cells* **15**, 255-267 (2010).
36. Sielski, N.L., Ihnatovych, I., Hagen, J.J. & Hofmann, W.A. Tissue specific expression of myosin IC isoforms. *BMC Cell Biol* **15**, 8 (2014).

37. Cook, A.W., Gough, R.E. & Toseland, C.P. Nuclear myosins – roles for molecular transporters and anchors. *Journal of Cell Science* **133** (2020).
38. Collado-Torres, L. et al. Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron* **103**, 203-216.e208 (2019).
39. Gallego Romero, I., Pai, A.A., Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biology* **12**, 42 (2014).
40. Bhalala, O.G., Nath, A.P., Inouye, M. & Sibley, C.R. Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue. *PLoS Genet* **14**, e1007607 (2018).
41. Germain-Lee, E.L., Obie, C. & Valle, D. NVL: A New Member of the AAA Family of ATPases Localized to the Nucleus. *Genomics* **44**, 22-34 (1997).
42. Nagahama, M. et al. NVL2 is a nucleolar AAA-ATPase that interacts with ribosomal protein L5 through its nucleolar localization sequence. *Mol Biol Cell* **15**, 5712-5723 (2004).
43. Jiao, L. et al. Ribosome biogenesis in disease: new players and therapeutic targets. *Signal Transduction and Targeted Therapy* **8**, 15 (2023).
44. Stein, K.C., Morales-Polanco, F., van der Lienden, J., Rainbolt, T.K. & Frydman, J. Ageing exacerbates ribosome pausing to disrupt cotranslational proteostasis. *Nature* **601**, 637-642 (2022).
45. Flach, J. et al. Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. *Nature* **512**, 198-202 (2014).
46. Ding, Q., Markesbery, W.R., Chen, Q., Li, F. & Keller, J.N. Ribosome Dysfunction Is an Early Event in Alzheimer's Disease. *The Journal of Neuroscience* **25**, 9171-9175 (2005).
47. Ding, Q. et al. Increased 5S rRNA Oxidation in Alzheimer's Disease. *Journal of Alzheimer's Disease* **29**, 201-209 (2012).
48. Healy-Stoffel, M., Ahmad, S.O., Stanford, J.A. & Levant, B. Altered nucleolar morphology in substantia nigra dopamine neurons following 6-hydroxydopamine lesion in rats. *Neuroscience Letters* **546**, 26-30 (2013).
49. Yang, R. et al. La-related protein 4 binds poly(A), interacts with the poly(A)-binding protein MLLE domain via a variant PAM2w motif, and can promote mRNA stability. *Mol Cell Biol* **31**, 542-556 (2011).
50. Glatt, S.J. et al. Similarities and differences in peripheral blood gene-expression signatures of individuals with schizophrenia and their first-degree biological relatives. *Am J Med Genet B Neuropsychiatr Genet* **156b**, 869-887 (2011).
51. Devireddy, L.R., Hart, D.O., Goetz, D.H. & Green, M.R. A mammalian siderophore synthesized by an enzyme with a bacterial homolog involved in enterobactin production. *Cell* **141**, 1006-1017 (2010).
52. Zughaier, S.M., Stauffer, B.B. & McCarty, N.A. Inflammation and ER stress downregulate BDH2 expression and dysregulate intracellular iron in macrophages. *J Immunol Res* **2014**, 140728 (2014).
53. Vidal, R., Caballero, B., Couve, A. & Hetz, C. Converging pathways in the occurrence of endoplasmic reticulum (ER) stress in Huntington's disease. *Curr Mol Med* **11**, 1-12 (2011).

54. Matus, S., Glimcher, L.H. & Hetz, C. Protein folding stress in neurodegenerative diseases: a glimpse into the ER. *Curr Opin Cell Biol* **23**, 239-252 (2011).
55. Bai, B. et al. Deep Multilayer Brain Proteomics Identifies Molecular Networks in Alzheimer's Disease Progression. *Neuron* **105**, 975-991.e977 (2020).
56. Chi, B. et al. The neurodegenerative diseases ALS and SMA are linked at the molecular level via the ASC-1 complex. *Nucleic Acids Res* **46**, 11939-11951 (2018).
57. Knierim, E. et al. Mutations in Subunits of the Activating Signal Cointegrator 1 Complex Are Associated with Prenatal Spinal Muscular Atrophy and Congenital Bone Fractures. *Am J Hum Genet* **98**, 473-489 (2016).
58. Oliveira, J., Martins, M., Pinto Leite, R., Sousa, M. & Santos, R. The new neuromuscular disease related with defects in the ASC-1 complex: report of a second case confirms ASCC1 involvement. *Clin Genet* **92**, 434-439 (2017).
59. Shima, A., Matsuoka, H., Yamaoka, A. & Michihara, A. Transcription of CLDND1 in human brain endothelial cells is regulated by the myeloid zinc finger 1. *Clin Exp Pharmacol Physiol* **48**, 260-269 (2021).
60. Patel, H., Dobson, R.J.B. & Newhouse, S.J. A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data. *J Alzheimers Dis* **68**, 1635-1656 (2019).
61. Neff, R.A. et al. Molecular subtyping of Alzheimer&#x2019;s disease using RNA sequencing data reveals novel mechanisms and targets. *Science Advances* **7**, eabb5398 (2021).
62. Soneson, C., Matthes, K.L., Nowicka, M., Law, C.W. & Robinson, M.D. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol* **17**, 12 (2016).
63. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417-419 (2017).
64. Love, M.I., Edn. 1.0 (Zenodo, 2018). https://doi.org/10.5281/zenodo.1291522
65. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).
66. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
67. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340-1347 (2010).
68. Jones, M.C., Marron, J.S. & Sheather, S.J. A Brief Survey of Bandwidth Selection for Density Estimation. *Journal of the American Statistical Association* **91**, 401-407 (1996).
69. Mann, H.B. & Whitney, D.R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18**, 50-60, 11 (1947).
70. Tange, O. GNU Parallel 2018. (Ole Tange, 2018).
71. Neuhäuser, M. & Ruxton, G.D. Round Your Numbers in Rank Tests: Exact and Asymptotic Inference and Ties. *Behavioral Ecology and Sociobiology* **64**, 297-303 (2009).