THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Exploring Blood Collection Techniques in ALS through Transcriptomics

BIOX7002: Research Project A

Summer Semester 2022/23

Institute for Molecular Bioscience

Supervisor: Dr Fleur Garton

Co-supervisor: Dr Anna Freydenzon

Jayden Beckwith

46267539

Abstract

RNA-sequencing (RNA-seq) is a critical tool to identify novel molecular mechanisms involved in motor neurone disease (MND). An overlooked element of RNA-seq is the way samples are collected and processed prior to sequencing. A technique to economically extract RNA from blood in EDTA tubes (hereafter Ethylenediamine (eRNA)) has been recently optimised in our laboratory. It remains unknown if the resulting transcriptome (using RNA-seq) is comparable to the gold standard PAXgene™ blood collection. Here, we designed an RNA-seq pipeline to compare 6 samples with a matched eRNA and PAXgene™ collection. Read counts, differential gene expression and a GO analysis was carried out. Additionally, with 5 MND cases and 1 control, a preliminary pipeline to examine rare aberrant splice-isoforms was established to examine molecular mechanisms involved in MND that could be further explored in larger sample sizes.

Supporting eRNA as a viable alternative to PAXgene™ collections, there was no significant difference between the average read count of RNA. Differential expression (DE) analyses did reveal subtle differences between both collection techniques. DE identified several protein coding ribosomal genes (*RPL/RPS*) that were differentially expressed (upregulated in eRNA and downregulated in PAXgene™ collection types), which was subsequently detected in GO enrichment analysis. When looking for aberrant splice events, more frequent events in the MND PAXgene™ collections with similar (but reduced magnitude) events found in the matched eRNA. These were in genes involved in neurodegenerative pathways (e.g., *UBC, TOR1AIP1, CR1 and HLA-DRB5*) and may indicate reduced power in eRNA. These results suggest fewer ribosomal genes in the PAXgene™ collection method and reduced variance compared to eRNA.

Overall, our results detect high quality RNA-seq data from both eRNA and PAXgene™ collections. We do find variation in the transcriptome due to the collection method, to support the use of a single collection method within an experiment. Refinement of the eRNA collection method (i.e., to reduce ribosomal transcriptome), larger sample sizes and bioinformatic simulations (i.e., test the expected stochasticity based on the correlation between collection types) could provide further advice on the impact of eRNA vs. PAXgene™ in detecting true biological effects. The economical savings of eRNA at the sample banking stage compared to PAXgene™ method are substantial and could have diverse applications across the wider scientific community. This preliminary investigation provides important insights into the usefulness of this method and detects some novel aberrant splice events that could be further examined in the pathogenesis of MND.

Introduction

Amyotrophic lateral sclerosis

Motor neurone disease (MND) describes a group of neurodegenerative disorders characterised by progressive loss of motor neurones, responsible for controlling voluntary muscle movement. The most common form of MND is amyotrophic lateral sclerosis (ALS). ALS accounts for ~80-90% of MND cases, with death from respiratory failure usually just 2-5 years after onset[22]. The incidence of ALS is relatively rare with just ~2-3 per 100,000 people per year across different populations, and while no clear geographical pattern is indicated, aging is a significant risk factor[1,22]. Onset of disease is typical after the fifth decade and an ALS diagnosis is indicated by signs of degeneration of both the upper and lower motor neurones. Symptoms can include initial twitches, progressive muscle atrophy, weakness in the limbs, slurred speech, difficulty swallowing and fatigue[2,28], however, it is worth noting that there is a high degree of variability in presentation as well as the spread and rate of disease progression. Riluzole, the only prescribed therapy for ALS in Australia, provides a modest (estimated 2-3 month) extension to life expectancy. Since its approval for use over 25 years ago, limited progress has been made. As such, there remains a significant need to heighten research efforts into the underlying mechanisms to identify novel therapeutic targets.

Genetic architecture and mechanisms of disease

ALS is a heterogenous disease both in its presentation and genetic architecture. Most ALS cases are considered complex with heritability estimates (the genetic contribution to disease liability) to be ~40%[30]. A proportion of cases, ~10-15% can be identified to have a mendelian genetic architecture as they carry a pathogenic variant in 1 in 30 genes and typically have a first degree relative with ALS and/or frontal temporal dementia (FTD). The three most commonly identified mutated genes to cause ALS are TAR DNA-bind protein (*TARDBP*), hexanucleotide expansions in chromosome 9 open reading frame 72 (*C9orf72*) and superoxide dismutase (*SOD1*)[28]. Investigating the mechanisms of these genes have implicated several pathways such as neuroinflammation, disturbed RNA metabolism, proteostasis failure, mitochondrial dysfunction, excitotoxicity, cytoskeletal disturbances and axonal transport defects, impaired DNA repair, oligodendrocyte dysfunction and nucleocytoplasmic transport deficits[3,17]. Indeed, some genetic cases – in particular, the identification of mutations in *TARDBP*, have provided crucial insight into the common pathogenic themes of ALS[27].

Cytoplasmic aggregation and nuclear depletion of TAR DNA-bind protein 43 (TDP-43, encoded by *TARDBP*) has been identified within 97% of ALS cases, reinforcing this mechanism as a central focus of disease pathogenesis[5]. TDP-43 is a highly conserved and ubiquitously expressed RNA-binding protein that is primarily localised to the nucleus, linking both sporadic and familial forms of the disease[24]. It contains two RNA-recognition motifs, a nuclear localisation and nuclear export sequence, allowing shuttling between the cytoplasm and nucleus within neurons[7]. The exact molecular mechanism that leads to TDP-43 aggregate formation causing neurodegeneration remains elusive. However, TDP-43 plays a key regulatory role in RNA metabolism, including acting as a splicing repressor[17]. This major regulatory role is critical for the repression of cryptic exon (CE) inclusion, which are typically excluded from mature mRNAs. Unlike normal exons, CEs are concentrated within introns and are a biproduct of aberrant RNA metabolism leading to the production of abnormal isoforms. Previous studies have strongly suggested that when TDP-43 is depleted in nuclear cells, CEs are spliced into mature mRNAs leading to the formation of premature stop codons by nonsense-mediated decay, inducing frameshift mutations and transcript degradation that drives disease progression[5]. Truncated forms of TDP-43 have been found in ALS aggregates, predominately in the cortex but also in the spinal cord of patients (Figure 1). Although TDP-43 mislocalisation is a common histopathological signature of ALS, it is not unique to patients with only *TARDBP* mutations but is also present in patients with *C9orf72* expansions and *TBK1* mutations in sporadic ALS[7,22]. This is suggestive that there may be other aberrant events that occur at other genetic loci, causing TDP-43 proteinopathies that promote ALS pathogenesis.

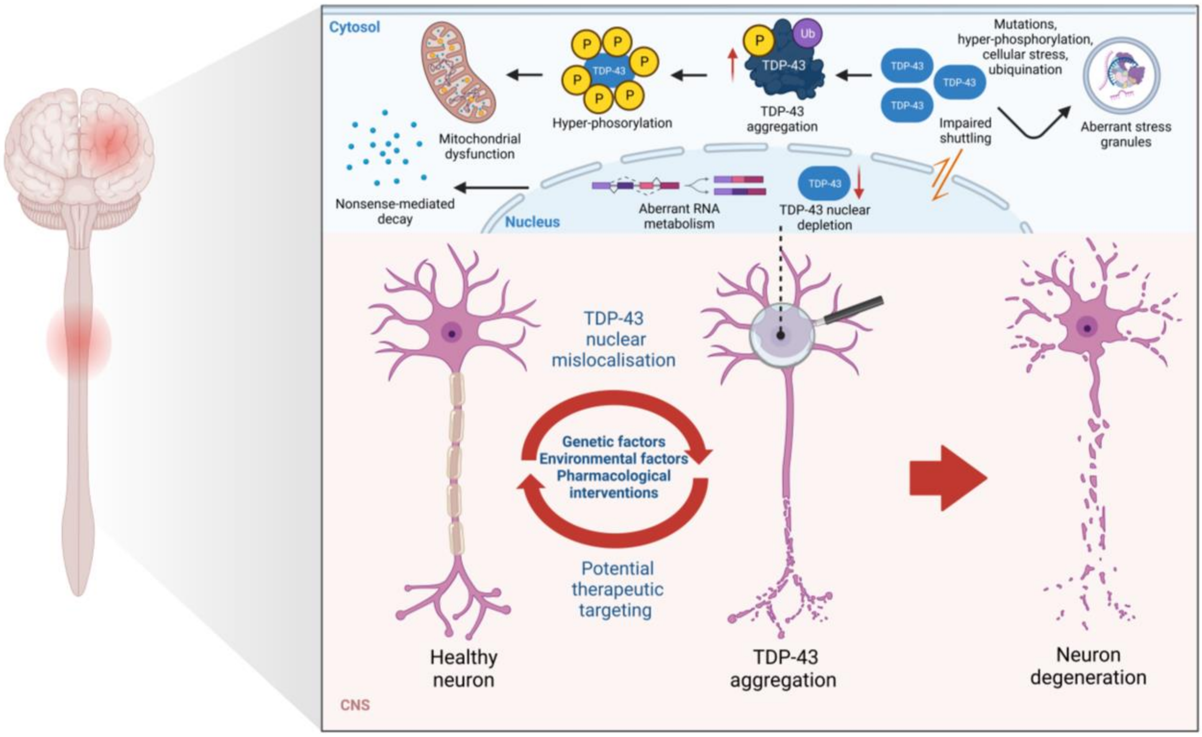


Figure 1. TDP-43 mislocalisation with nuclear depletion and cytoplasmic aggregation, inducing motor neuron deterioration via multiple downstream pathways (e.g., formation of aberrant stress granules and mitochondrial

dysfunction) in the central nervous system of ALS patients. Although TDP-43 has been identified within 97% of ALS cases, it is not well understood how genetic and environmental factors contribute to TDP-43 mislocalisation. This has limited the development of therapeutics to treat ALS to date due to its heterogeneous nature.

ALS genome-wide association studies have identified a significant risk locus within *UNC13A,* but the mechanism has not been known. Recently, *UNC13A* has been identified as a target of TDP-43 as it participates in vesicle maturation and neurotransmitter release as a critical synaptic gene[13]. Loss of *UNC13A* activity via TDP-43 nuclear depletion have demonstrated to compromise neuronal function. As such, it has recently been proposed that ALS risk SNPs in the *UNC13A* locus may modify disease risk through novel CEs that promote nonsense-mediated decay of *UNC13A*. This has been evidenced by CE transcripts found in high levels within patient neurons[17]. Previous studies have also shown that *UNC13A* is one of the genes with the most significant levels of alternative splicing in neurons with nuclear TDP-43 depletion, specifically between exons 20 and 21 (Hg38; Chr19: 17642414 – 17642541)[17]. Subsequently, this has made *UNC13A* a genetic locus of interest that could provide a potential avenue for therapeutic targeting. However, it is likely other CE sites have yet to be discovered that may account for patient symptoms and disease variability.

A transcriptome-focused approach to understand ALS

An approach to understanding ALS and identifying potential therapeutic targets, is to study the changes in gene expression and exon splicing through RNA sequencing (RNA-seq). Disruption of RNA homeostasis has been consistently associated with models of ALS and more recently, been linked as a mechanism of disease risk[10,28]. As yet, no large RNA-seq study has been carried out in an ALS cohort. This method is increasing in popularity due to its high-throughput nature and ability to detect *de novo* aberrant splicing – not possible with earlier transcriptome approaches such as microarray. RNA-seq pipelines have provided a streamline method to detect these events. Typically, an RNA-seq pipeline involves using quality controlled sequenced reads and aligning to a reference genome for downstream analysis. Differential gene expression (DGE) analysis is one of the most common downstream analyses of an RNA-seq pipeline. It allows to detect differentially expressed genes across two or more set conditions[18]. In combination with DGE analysis, there is a myriad of different splicing tools that allow for detection of rare splice sites that may play a role in the mechanism of pathological disease. A recent innovative tool called FRASER (Find RAre Splicing Events in RNA-seq) has captured research interest for detecting pathological splicing. It is an R/Bioconductor machine learning algorithm that provides a count-based statistical test for aberrant splicing detection in RNA-seq, while controlling for confounding variables[19]. Unlike other splicing detection methods, there are three features that render FRASER unique: (1) it considers non-split reads overlapping splice sites, allow for the detection of intron retention events; (2) automatically controlling for confounding effects; and (3) assesses statistical significance using a count distribution[19]. Ultimately, these components are vital in an RNA-seq pipeline to explain molecular events that may promote ALS pathogenesis.

A review of collection methods

To obtain quality RNA-seq data, samples must be collected in a manner that preserves the RNA and then allow for the enrichment or depletion of the RNA for a particular species. Peripheral blood is an ideal sample due to its non-invasive nature of collection (large samples can be analysed) and its ability to be a proxy for other tissues[21]. The general workflow for obtaining RNA includes: (1) Collection: a small amount of blood is collected from the individual using an anticoagulant clotting agent; (2) Isolation: RNA in the blood sample is isolated using a variety of methods, such as column-based purification or centrifugation; (3) Conversion to cDNA: RNA is converted to cDNA and then amplified using polymerase chain reaction (PCR); (4) Library preparation that can remove the rRNA (95% of the total cellular RNA, to avoid it consuming the bulk of the RNA-seq reads) before input for high-throughput sequencing[23].

Ethylenediamine RNA (eRNA) and PAXgene™ are two alternative methods for the isolation and preservation of RNA from biological samples. Both methods have been utilised in research and have advantages and limitations depending on the specific needs of the study. The eRNA method involves the isolation of RNA from cells using a silica-based matrix in an EDTA tube. The matrix is treated with a detergent solution to lyse the cells and release the RNA, which is then bound to the matrix. The RNA is then eluted from the matrix and can be used for downstream applications such as RNA-seq or RT-qPCR. An advantage of the eRNA method is the low cost (tubes and labour) without any loss in purity (minimal RNA degradation and contamination). It is also relatively fast and easy to perform, making it a popular choice for many research applications. However, the eRNA method is not widely used (it is a refined technique by the University of Queensland) and needs further validation. The PAXgene™ method is the most widely used. It preserves the RNA using a proprietary stabilisation reagent (within the blood collection tube) to form a complex with the RNA, protecting it from degradation. The stabilised sample can then be stored at room temperature for several weeks or shipped to a different location without degrading the RNA. One of the main advantages of the PAXgene™ method is its ability to preserve RNA for extended periods of time and in a wide range of temperatures, making it suitable for shipping and storing samples. However, this method is more labour intensive, time consuming and the tube is currently ~8x more expensive than the eRNA method. Furthermore, the PAXgene™ method is not suitable for all downstream applications, as the stabilisation reagent may interfere with some RNA-based assays. Even though both sampling systems have the same purpose, the techniques can result in gene expression profiles that differ between the systems[23]. However, there may be a superior technique to enhance the quantification of RNA-seq data in ALS samples.

This investigation aims to examine RNA-seq from different collection techniques (eRNA and PAXgene™) in ALS/control samples. It is hypothesised that each collection method will enable RNA to be profiled with comparable ranked transcriptome profiles between matched samples. Given PAXgene™ proprietary blend of reagents and reputation as the gold standard collection technique, this may prove to be superior in the interrogation of the biologically relevant transcriptome (despite being more expensive and laborious in nature). We will use gold standard pipelines to process the RNA data and carry out differential expression analyses as well as a recently developed splice-detection bioinformatic technique, to explore isoform detection including novel splice isoforms. Validation of collection techniques typically involve less than five isoforms and so the investigation proposed is novel[14]. The design will allow us to determine if there are any critical differences in the transcriptome between two collection techniques and/or identify if there is a preferred method to investigate the molecular mechanisms of ALS. Ultimately, this will inform the methods to be used in a much larger RNA-seq study on ALS and controls.

Materials and Methods

Dataset, quality control & pre-processing

The Human Studies Unit (HSU) at the Institute for Molecular Bioscience performed the initial steps in sample collection, RNA isolation and sequencing. Briefly, peripheral blood was collected from ALS patients (n = 6) and controls (n = 6) at the University of Queensland in the Otto Hirschfeld Building (OTTO). Samples were collected sequentially (from a single draw) in an EDTA (eRNA) and a PAXgene™ tube. On the same day, these were walked to HSU (building next door) for immediate processing. PAXgene™ tubes were stored at room temperature for 24 hours before transferring to -20°C freezers. Extraction was carried out with PAXgene™ Blood RNA kit as per the manufacturer's instructions. **UQ METHOD REDACTED.** RNA integrity number (RIN) was measured (all samples > 7.9 using Agilent Technologies Bioanalyzer Eukaryote Total RNA Nano Quantitation) and prepared for sequencing both using the Illumina stranded total RNA Ribozero Plus library. The final dataset comprised of 12 RNA-seq samples, N = 6 eRNA patient collections and N = 6 matched PAXgene™ collection. This consisted of 5 cases and 1 control (3 males and 3 females) with a mean age of 58.6

years. Adhering to human research ethics, all 6 subjects were appropriately consented and deidentified. Samples were sequenced (in the same lane) with the NextSeq 150 cycle high output kit (2x 75bp) as reverse stranded. An RNA-seq pipeline was developed to analyse the reads from the samples to investigate the impact of collection between biological replicates (Figure 2).

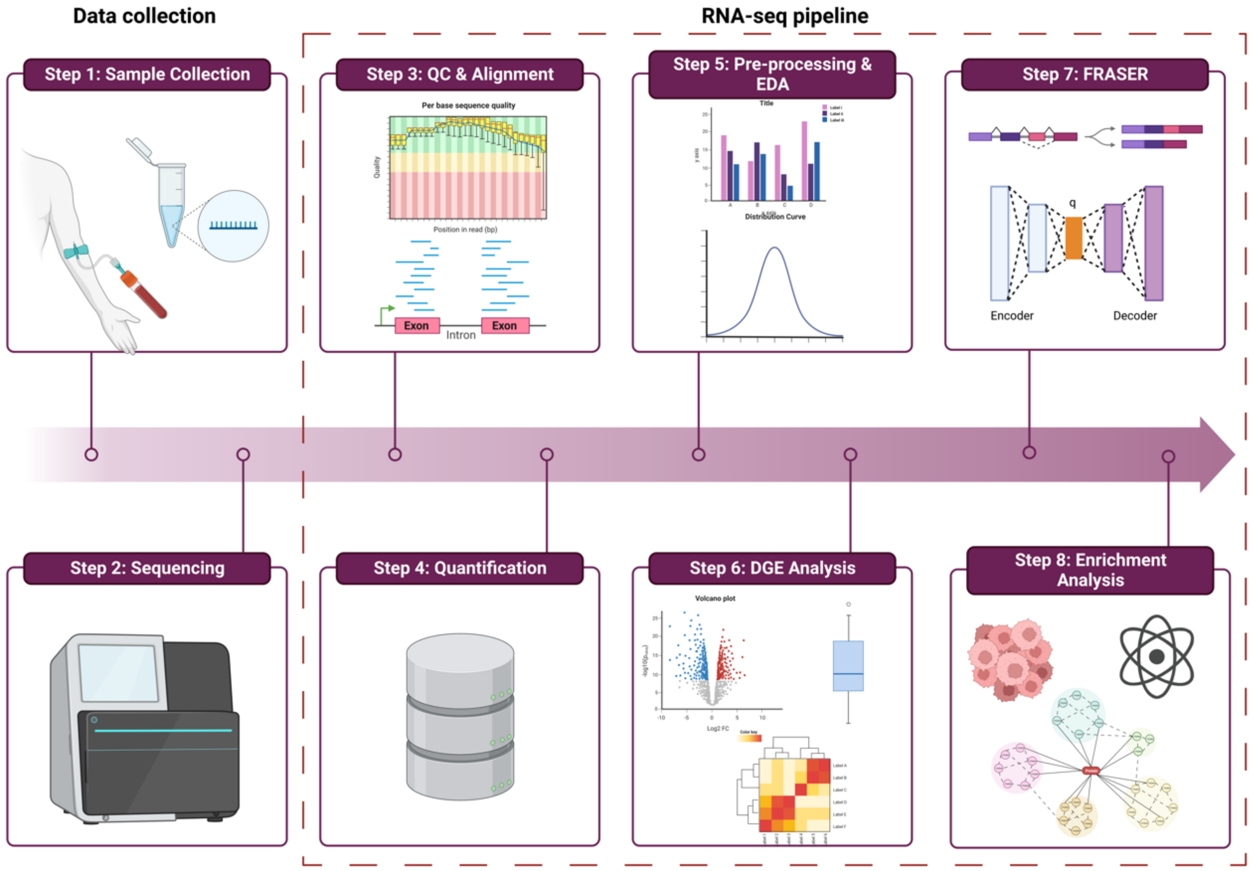


Figure 2. Overview of the project workflow and analysis pipeline. The data was collected from blood samples of 5 ALS patients and 1 control (Step 1) before RNA was isolated and sequenced (Step 2). This was coordinated and carried out by the Human Studies Unit at IMB. The RNA-seq pipeline (Steps 3-8) was developed and carried out for this project. This involved quality controlled FASTQ reads that were mapped to Hg19 reference genome and then used for DGE and FRASER analyses, followed by enrichment analysis to highlight classes of genes that were over-represented between collection types.

Short paired-end reads were obtained in FASTQ format and underwent quality control with FastQC (35-73bp; 45-47% GC) and were aligned with STAR RNA-aligner to the human annotation reference, GENECODE v39 (Hg19). The output BAM files were sorted using Rsamtools (v2.12.0) and were counted using Rsubread's (v2.10.5) *FeatureCounts* as reverse-stranded at the gene level. Data pre-processing consisted of filtering out low read counts across samples to enhance relative power for downstream analyses. Genes were then normalised, and log2-transformed using *SummarizedExperiment* from DeSeq2 (v1.3.6) Bioconductor package for principal component analysis (PCA) and downstream exploratory analyses. Exploratory analysis involved using normalised gene sum and mean counts across collection types to: (1) visualise the distribution of counts across samples; (2) understand which genes overlapped and the composition between eRNA and PAXgene™ collection techniques; (3) genes that have the highest abundance of reads mapped between techniques; and (4) check the correlation between samples based on covariate effects. Statistically significant differences between collection types were also performed using two-tailed Mann-Whitney U test and Student's t-test, defined at p-value (*p*) < 0.05.

Differential gene expression analysis

DGE analysis was performed with three software packages in R/Bioconductor, including DeSeq2 (v1.3.6), Limma/Voom (v3.52.4) and Dream, a part of the edgeR (v3.38.4) library. All metadata was factored accordingly, to ensure no misreading from the software packages. The first design experiment used DeSeq2 and investigated the effect of collection type (with PAXgene™ set as the reference), accounting for the study ID pairing for each patient. The second design consisted of investigating the effect of collection type, accounting for gender. This was performed to see if there was any sex effect confounding in the sample. DeSeq2 uses raw counts as input and normalises data via the median of ratios, accounting for sequencing depth and RNA composition. Differentially expressed genes were annotated using the biomaRt (v2.52.0) package via Ensembl IDs. Additionally, random effect of study identification for patients was also tested on collection type to check if there were any confounding effects between the two variables, using Limma and Dream. Limma was used to normalise counts after log2 transformation and normalisation. The random effect was modelled using *duplicatedCorrelation* and the variancePartition (v1.26.0) library (where eRNA was set as the reference due to limited functionality). The Dream library allowed for differential expression testing by using linear mixed models for repeated measures, by increasing power and decreasing false positives. This permitted modelling of the variance explained by study identification of patients as a random effect.

FRASER analysis

FRASER (v1.8.1) consisted of three steps: (1) counting; (2) modelling; and (3) detection. It utilised RSubread's *FeatureCounts* to obtain counts of splice junctions and sites through in-built functions, *CountRNAdata* and *CalculatePSIvalues*. The quantification of alternative splicing is based on the percent-spliced-in (PSI, Ψ), which captures local information related to splicing of each exon[28]. The Ψ metric is defined as the number of reads supporting each exon inclusion divide by the number of reads that span the exon. However, FRASER uses an intron-centric approach adapted from Pervouchine *et al.* (2013), where each intron is defined by the combination of a 5'-splice site (D, donor) and 3'-splice sites (A, acceptor). Denoted Ψ5 and Ψ3, respectively, as:

$$(1)\quad \Psi_5(D,A) = \frac{n(D,A)}{\Sigma_{A'} n(D,A')} \qquad (2)\quad \Psi_3(D,A) = \frac{n(D,A)}{\Sigma_{D'} n(D',A)'}$$

Where n(D, A) is the number of split reads spanning the intron of interest between the donor (D), and the acceptor (A), with the summation of the denominators are computed over the acceptors that are spliced with the donor (equation 1); and all donors that are spliced with the acceptor of interest (equation 2). $\Psi_5(D,A)$ and $\Psi_3(D,A)$ are represented as a conditional probability of splicing from D to A, or the percentage of transcripts spliced D to A, respectively to the number of transcripts at the D site[20]. Additionally, to detect partial or full intron retention, FRASER incorporates a splicing efficiency metric that measure the percent of intron retained (θ). FRASER does not distinguish between θ_5 and θ_3, quantifying intron retention as a combined variable, θ. It is defined by Pervouchine *et al.* (2013), specifically:

$$(3)\quad \theta_5 = \frac{\Sigma_{A'} n(D,A')}{n(D) + \Sigma_{A'} n(D,A')} \qquad (4)\quad \theta_3 = \frac{\Sigma_{D'} n(D',A)}{n(A) + \Sigma_{D'} n(D',A)}$$

Similarly, n(D) is the number of non-split reads that span the exon-intron boundary of the donor (D), (equation 3) and n(A) is the same, but for the acceptor (A) (equation 4). Raw intron-centred Ψ correlation heatmaps were developed for all splicing metrics based on gender, study id and collection technique to observe any covariate interactions. FRASER controls for confounders by optimising simulated outliers through training a denoising autoencoder, consisting of an encoder and decoder matrix. The fitted encoding dimensions were estimated using FRASER's hyperparameter tuning tool, *optimHyperParams* (Figure 3).

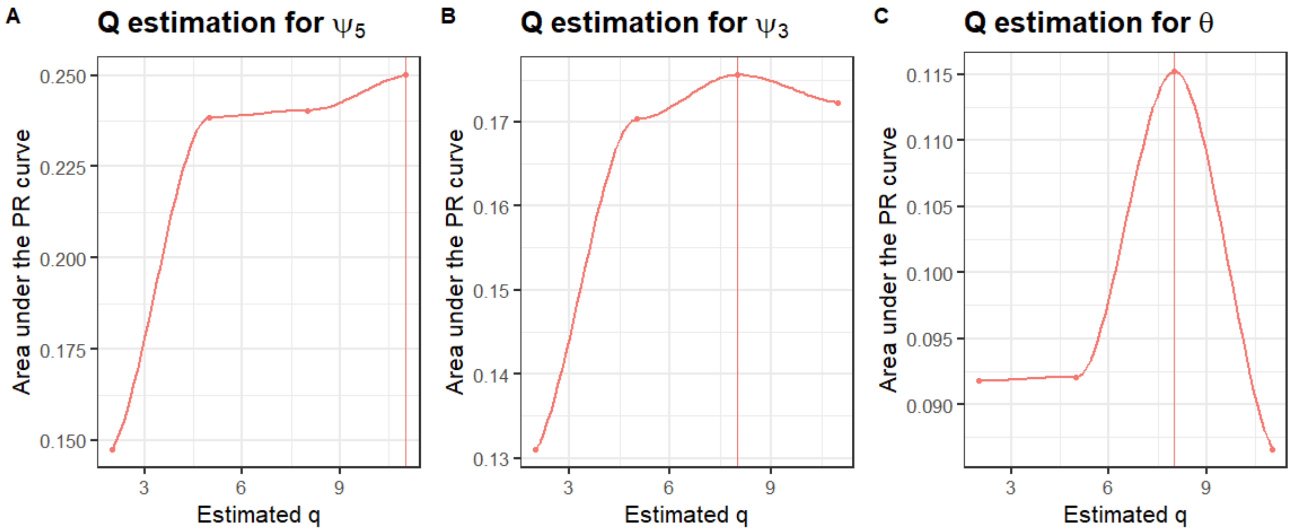


Figure 3. Hyperparameter tuning of splicing metrics for FRASER autoencoder. **A**. showed ψ_5 optimised at an estimated latent space dimension of 11 against the area under the PR curve; **B.** explained ψ_3 and **C.** showed θ optimised at a latent space dimension of 8.

It works by artificially injecting outliers into the data and then comparing the AUC-PR (area under the precision recall curve) of recalling these outliers for different values of q (dimension of latent space between the encoder and decoder matrices). A hybrid approach (*PCA-BB-decoder*) was utilised, which used PCA to estimate the encoder matrix and then using a beta-binomial loss function to optimise the weights in the decoder matrix of the autoencoder. Since FRASER offers other implementations for analysis, the hybrid approach was selected based on the faster processing speed over other methods. Additionally, the Hg38 UCSC genome was used for intron annotation corresponding to HGNC symbols of splice sites and junctions. As recommended by the FRASER documentation, aberrant splicing detection parameters ($\Delta\Psi > 0.3$ and adjusted $p < 0.05$) were used in conjunction with multiple-testing correction across all samples by the Benjamini–Yekutieli false discovery rate method.

Enrichment analysis

To consider if certain parts of the transcriptome were impacted by the collection method, any gene that was identified as differentially expressed and were unique to each isolation type underwent further GO enrichment analysis. Enrichment analysis consisted of identifying the biological processes, molecular pathways and cellular components of genes that were selected. The analyses were performed using clusterProfiler (v3.0.4) library.

Results

Quantitation, exploratory differential and principal component analysis

Quantitation of reads showed no significant difference between average read count of collection types (p = 0.15). Initial quality control correlations showed all matched samples correlated (range R^2 = 0.73-0.93, $p < 2.2 \times 10^{-16}$) between collection types. Reads were quantified in terms of expression counts and revealed a dataset that consisted of 61,533 genes. The dataset was subset on collection type and a mean count threshold >= 10 was applied across each gene of both data frames for exploratory analyses (where 10 is typically the staple of pre-processing for RNA-seq pipeline read filtering). Further filtering genes via collection type revealed 16,222 genes corresponding to eRNA and 17,570 genes in PAXgene™ samples. Restricting isoforms revealed 29,577 unique transcript and genes between both collection data frames. It was observed that there was no significant difference between collection methods based on expression counts (PAXgene™, N = 15,111 and N = 14,466, eRNA) (p = 0.48).

It was observed that both collection techniques consisted of 14,051 unique overlapping genes, with 1,060 isolated to PAXgene™ and 415 to eRNA (Figure 4A). Between collection techniques, 37 different gene and transcript types (e.g., protein coding, miRNA, pseudogenes) were identified. It was examined that 84.6% (N = 12,234) of gene composition was comprised of protein coding genes in eRNA and 81.9% (N = 12,372) in PAXgene™ (Figure 4B) samples. Additionally, we also measured a sub-category of "non-protein coding" genes. This consisted of 18.1% (N = 2,738) in PAXgene™ and 15.4% (N = 2,231) in genes eRNA samples. Of the top five protein coding genes, two overlapped with the highest number of expression counts. In eRNA samples these were *NAMPT, BINP3L, TENT5C, SLC25A37,* and *ACTB* and in PAXgene™ samples these were *ACTB, HLA-B, CSF3R, SORL1* and *NAMPT.* Interestingly, ALS-related genes from literature such as *UNC13A, TARDBP, SOD1* and *C9orf72* all had lower counts than expected in case samples based on gene length, with no significant difference between counts based on collection types (p = 0.35).

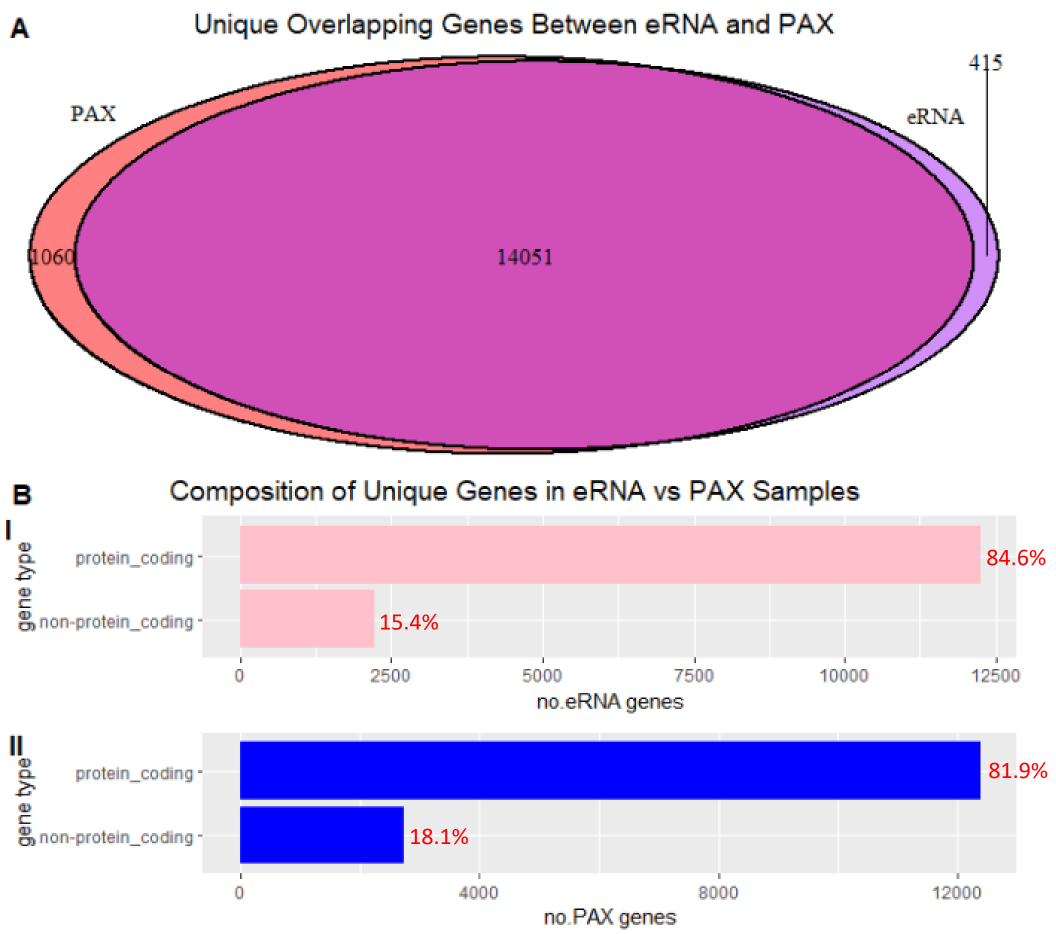


Figure 4. **A.** Venn diagram with the number of unique overlapping genes between PAXgene™ and eRNA collection techniques**. B.** Breakdown between i. eRNA and ii. PAXgene™ bar plots showed that there was no significant difference between the two collection types based on gene composition (*p* = 0.57).

Filtered counts based on a threshold >= 10 (sum of gene count) from the original data frame (N = 27,050), underwent log2 transformation and normalisation. Preliminary visual inspection of the density distribution of counts (Supplementary Figure 1) and dimensionality reduction was performed. Additionally, a normalised Pearson correlation matrix was developed between collection types based on normalised expression counts and showed a high positive correlation (all samples R > 0.89) (Supplementary Figure 2). PCA revealed a definitive separation between PC1 and PC2 based on collection types, where PC1 captured 57% of the total variance explained by collection type (as well as study individuals); while PC2 captured 18% of the total variance explained separating patients by sex (Figure 5). Ultimately, it was observed that larger variation was more apparent in eRNA.

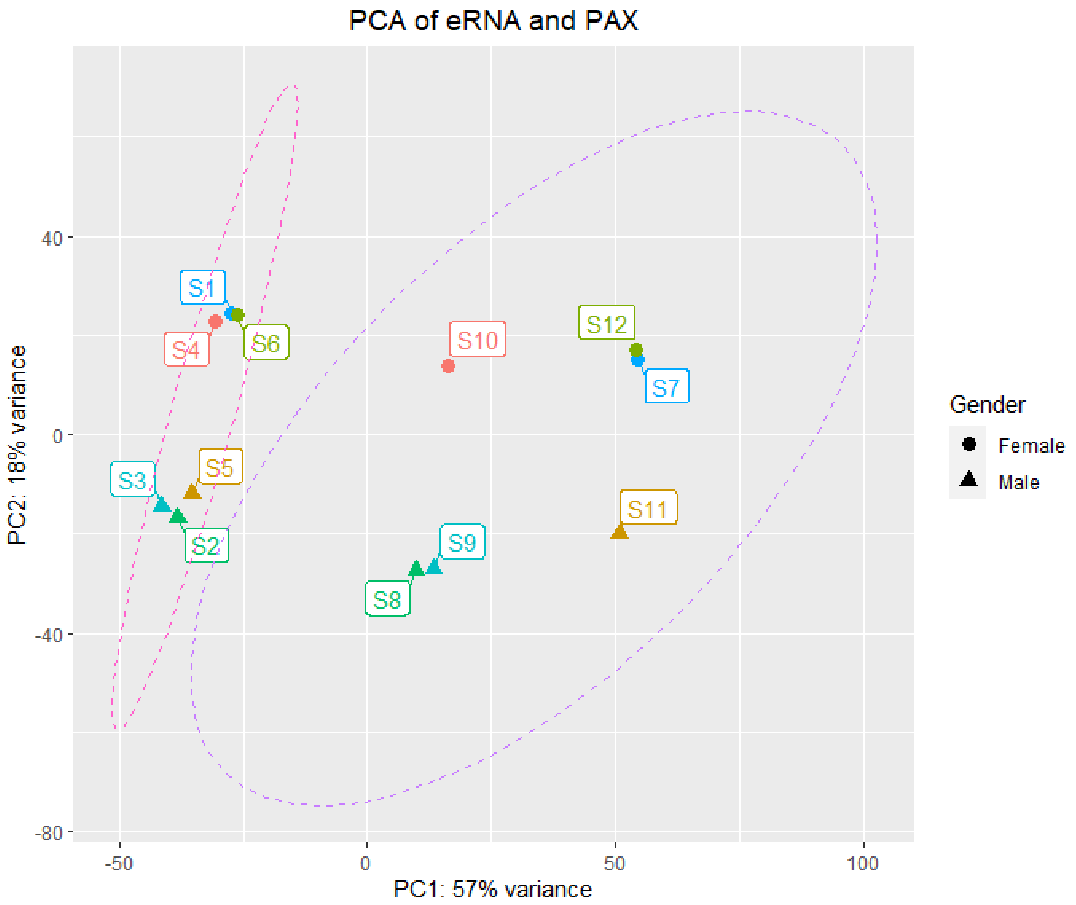


Figure 5. Principal component analysis of PAXgene™ (S1-S6) and eRNA (S7-S12) collection types, coloured by patient identification. PC1 captures 57% variance total variance explained in type (as well as individuals) and PC2 explains 18% of total variance explained separating gender. The variation explained is mostly in the eRNA samples, whereas PCA doesn't particularly capture variation of expression in PAXgene™ samples.

To compare each collection method transcriptome profile, we ranked each gene based on its variation between samples after normalisation (accounting for read/coverage), and log2 transformation. Looking at the top 1000 most variable genes, there was a weak but significant correlation between top 1000 ranks of PAXgene™ and eRNA (non-parametric Kendall's rank correlation coefficient test, p = 3.5 x 10^{-6}, R = 0.098). This result was reinforced by observing a positively correlated distribution of the top 1000 most heritable genes from an external dataset (CAGE eQTLs). This list was based on gene expression heritability estimates from RNA-seq data from 5000 blood samples (*Consortium for the Architecture of Gene Expression*)[16]. The distribution observed reinforced the weak correlation between the two collection types (Figure 6) with a high left skewed peak tail. Out of the top 1000 most heritable genes, it was detected that 232 genes exhibited R > 0.8. This indicated that the relationship between the two collection types is unlikely due to chance.

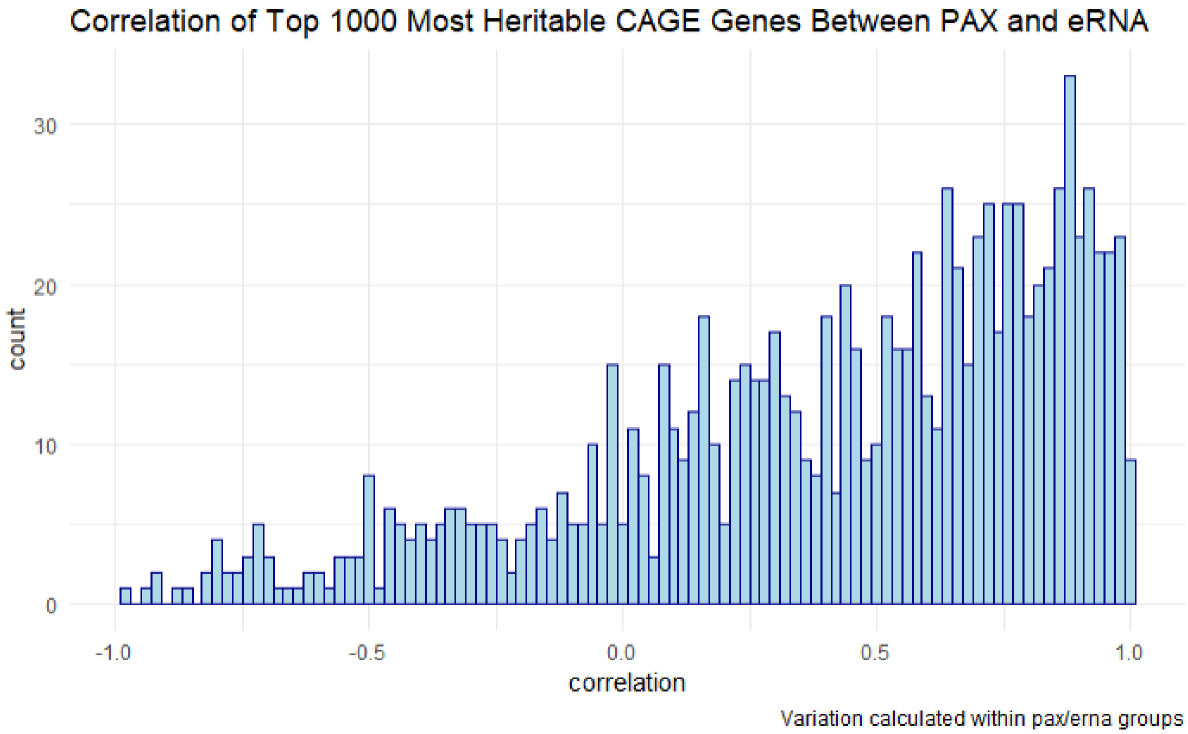


Figure 6. Correlation distribution of the top 1000 most heritable CAGE eQTLs identified from Lloyd-Jones *et al*. (2017) in our eRNA and PAXgene™ samples. This showed a left skewed distribution between collection types of the most heritable genes.

Differential gene expression analysis

Differential gene expression was carried out across three Bioconductor software packages. DeSeq2 was used to examine the impact of the collection method on gene expression accounting for each matched pair. Using a log2 fold change = +/-1 and adjusted p < 0.05, 2604 upregulated and 2904 downregulated protein coding genes were identified between the two collection types. It was evident clustering analysis of the top differentially expressed (DE) genes (i.e., N = 3,000 and N = 50) appeared to show consistent patterns with collection method (Figure 7A and B respectively). The top 50 DE genes had a higher z-score in the eRNA collection to indicate a consistent difference in isoform capture with a noticeable number of ribosomal protein L and S genes (*RPL/RPS*).

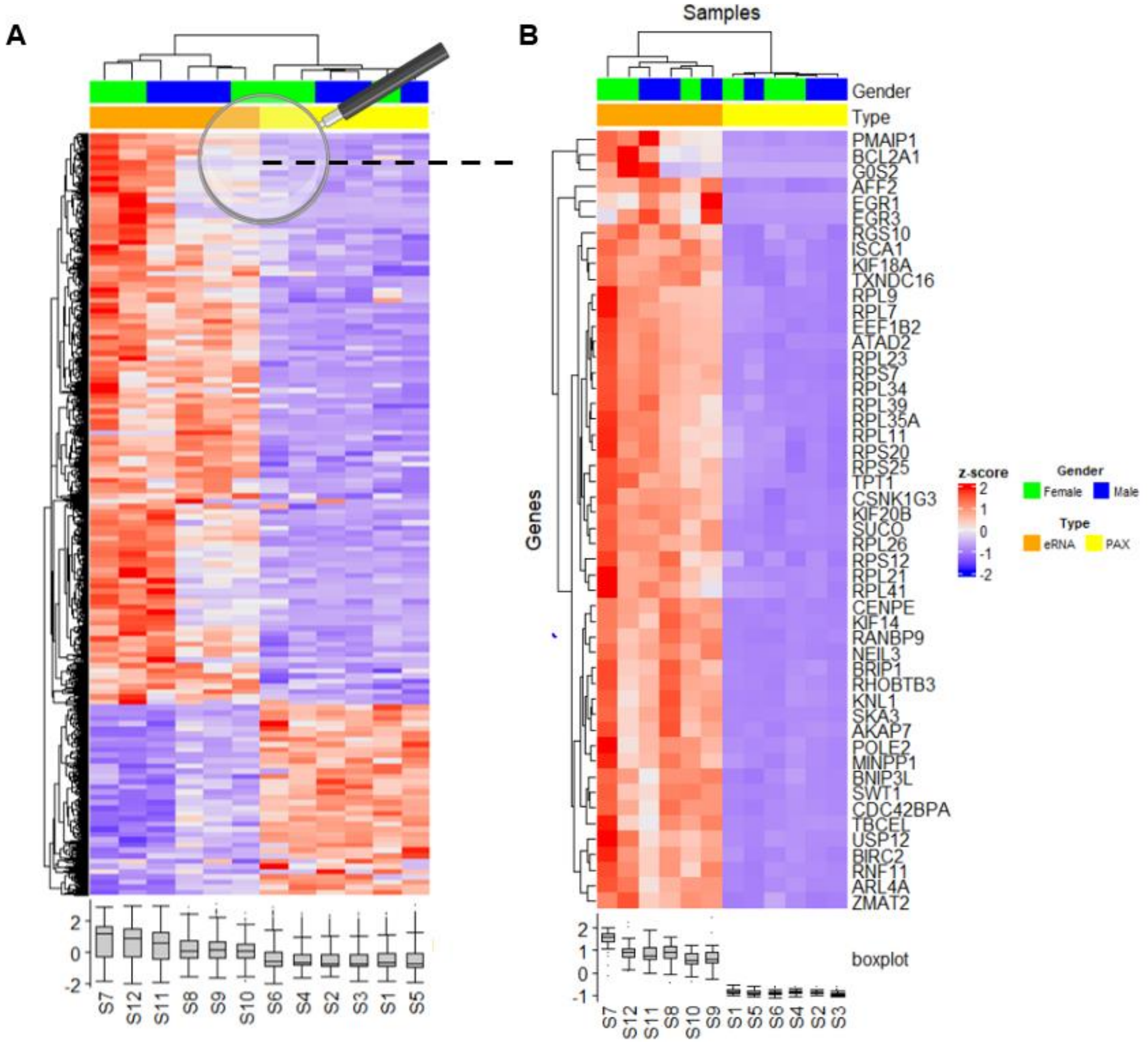


Figure 7. A. Top 3000 differentially expressed genes based on collection type and accounting for patient identification pairing, with **B.** enhanced visual zoom on the top 50 differentially expressed genes based on z-score changes across samples. Particularly, sample 7 exhibited a median with the greatest level of differential expression based on visual inspection of boxplots.

Volcano plot visualisation indicated inflation based on the magnitude of the log-fold change of eRNA (Supplementary Figure 3). Adding sex into the model (to check for sex effects differentially impacting the collection methods) (Supplementary Figure 4) had little change on the result (N = 149 additional DE genes). Next, we modelled a random effect of study identification for samples to check if there were any confounding effects between the two collection methods, in Limma and Dream. This substantially reduced genomic inflation in the volcano plot (Supplementary Figure 5) between the two variables to indicate collection type accounts for a large proportion (but not all) the variance explained (Figure 8). Other contribution factors being Study ID and the Residuals between both variables.

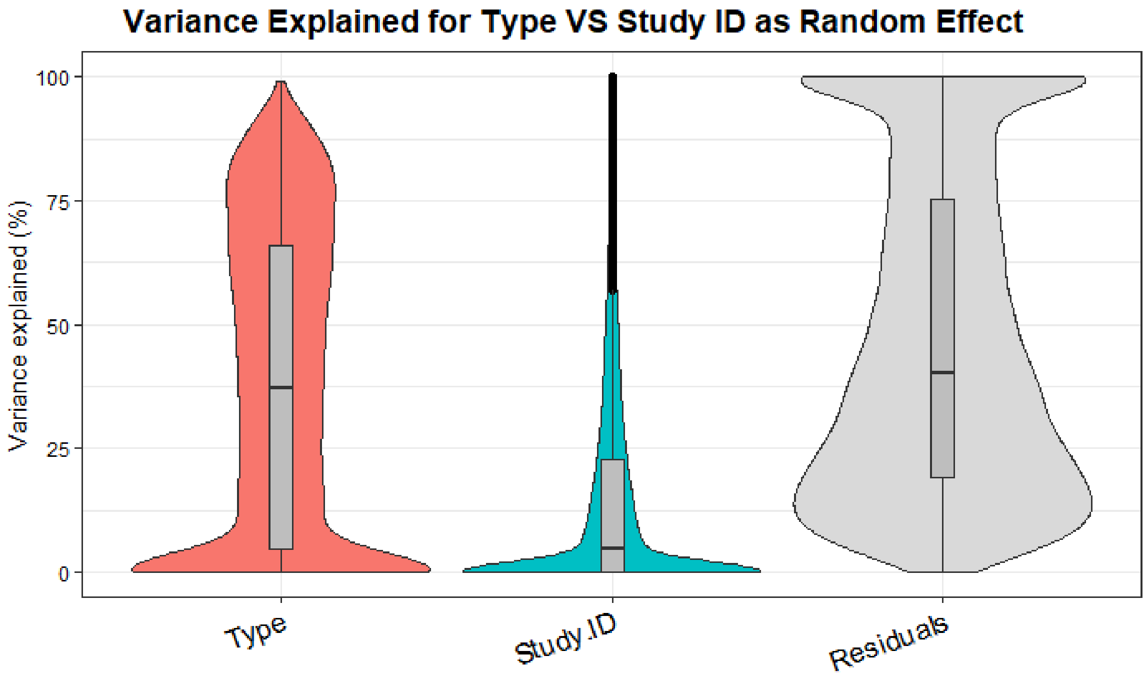


Figure 8. Variance explained for isolation type verses study identification of patients as a random effect; displaying a large proportion of the variance explained by isolation type, based on the residuals between the two variables.

Detection of aberrant splicing

To examine changes within genes, at an isoform level, we used FRASER. A total of 750,965 junctions and 127,406 splice sites were counted in 12 samples. Filtering (based on mean expression as recommended by FRASER documentation) revealed 53,388 junctions and 102,975 splice sites for analysis. The PAXgene™ collection type had the higher count post-filtering, which revealed 29,012 junctions and 56,746 splice sites, in comparison to the eRNA samples that showed 24,376 junctions and 46,229 splice sites. The average number of splice sites and junctions between collection types did not differ (p = 0.35). FRASER covariates (gender, study identification) and collection type were clustered after data pre-processing (expression filtering, log-transformation, and normalisation) revealed a higher positive Pearson correlation (R = 0.3-0.5) present in intron retention (θ) sites, which were more pronounced in controls (Supplementary Figure 8). A total of 153 aberrant splicing events (PAXgene™ n = 86, eRNA n = 67) were observed between RNA collection types based on splicing metrics. There were no clear differences between collection methods based on 5' and intron retention splicing ($p_{\Psi 5}$ = 0.67, p_{θ} = 0.95) with a trend for increased 3' splicing in PAXgene™ samples ($P_{\Psi 3} < 0.05$). Additionally, intron retention splicing events (θ) comprised 62% of all aberrant events observed across samples, against Ψ_3 and Ψ_5 splicing events ($p < 0.05$) (Table 1).

Table 1. *Number of Genes with Aberrant Splicing Events in PAXgene™ and eRNA Samples*

	Collection	PAXgene™						eRNA					
Splice metric	Sample	1	2	3	4	5*	6	7	8	9	10	11*	12
5’ splicing(Ψ_5)		3	1	2	1	2	8	1	2	4	2	3	2
3’ splicing (Ψ_3)		4	1	3	2	8	4	0	0	2	3	0	0
Intron retention (θ)		3	9	14	15	3	3	4	1	14	11	12	6
Total		10	11	19	18	13	15	5	3	20	16	15	8

Note: Aberrant splice events are filtered based adjusted $p < 0.05$ and $\Delta\Psi > 0.3$, as recommended by FRASER documentation. Counts also contain isoforms of detected genes. * represented control samples.

Within samples, it was observed that sample 6 had the highest statistically significant aberrant splicing event in *UBC* (adjusted p = 1.4 x 10^{-38}), with four other statistically significant genes meeting the threshold; *TOR1AIP1* (adjusted p = 5.9 x 10^{-18})*, MAX* (adjusted p = 2.4 x 10^{-6}) *and HLA-DRB5* (adjusted p = 1.72 x 10^{-2}) (Figure 9A). When comparing these splicing events to its matched eRNA equivalent (sample 12), it was evident that there was a difference in magnitude between observed genes. The *UBC* gene (adjusted p = 5.9 x 10^{-18}) met the threshold in eRNA, but the other identified statistically significant genes did not meet the aberrant threshold that were present in the PAXgene™ sample (Figure 9B).

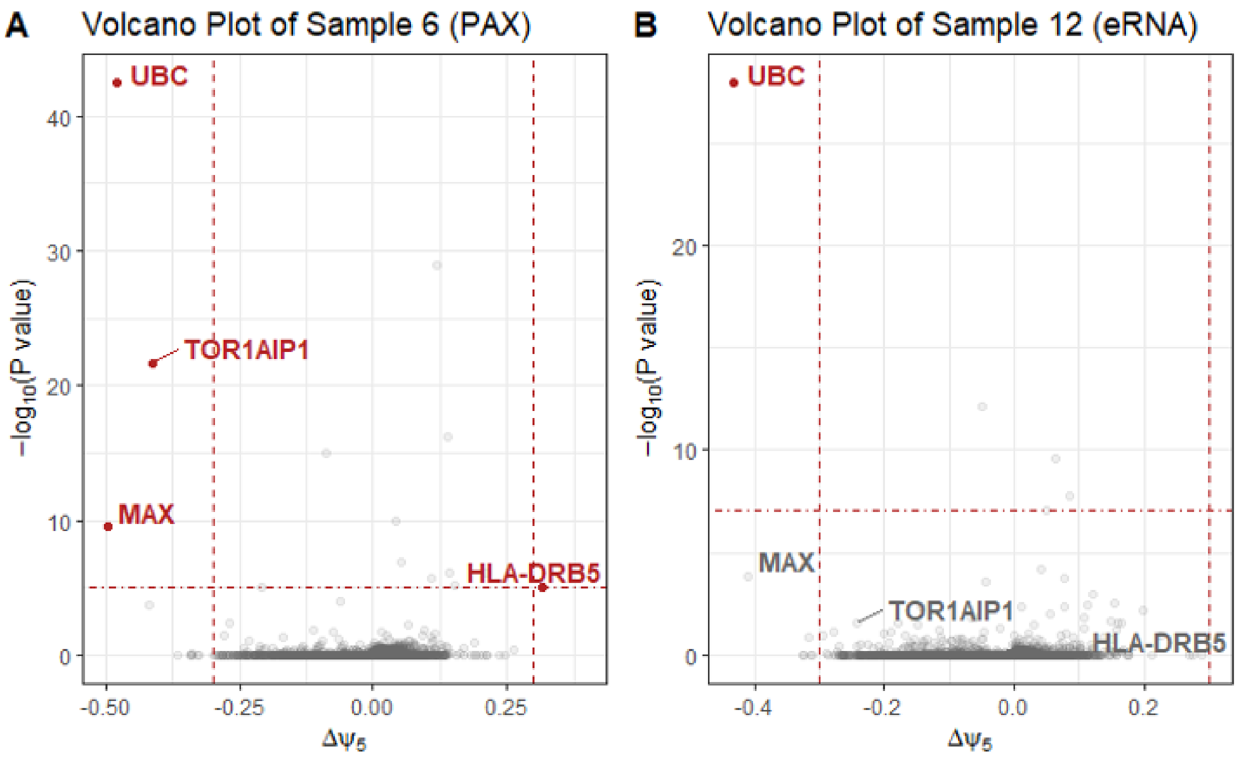


Figure 9. Aggregated volcano plots (the dotted red lines represent the aberrant splice thresholds based on adjusted $p <$ 0.05 and $\Delta\Psi > 0.3$) of statistically significant Ψ_5 aberrant splicing of sample 6 (PAXgene™) verses sample 12 (eRNA) from the same ALS patient. **A.** showed a higher statistically significant magnitude and $\Delta\Psi_5$ in the PAXgene™ case sample 6, in comparison to sample 12 in **B**. for observed aberrant gene loci.

GO enrichment analysis revealed differential transcripts between collection methods

GO enrichment analysis consisted of further exploring categories of differentially expressed genes specific to collection methods. Molecular function and cellular components were part of the GO analysis, however here we limit this report to biological processes (BP). GO-BP of the top 100 DE genes found ten enriched BP, four that were directly relevant to ribosomal RNA (Figure 10). An independent enrichment analysis on the upregulated and downregulated genes (PAXgene™ as the reference) indicated that upregulation genes were associated with cytoplasmic translation and ribosomal processes, while downregulated genes were quantified primarily in protein peptidyl-isomerization processes (Supplementary Figure 6 and 7).

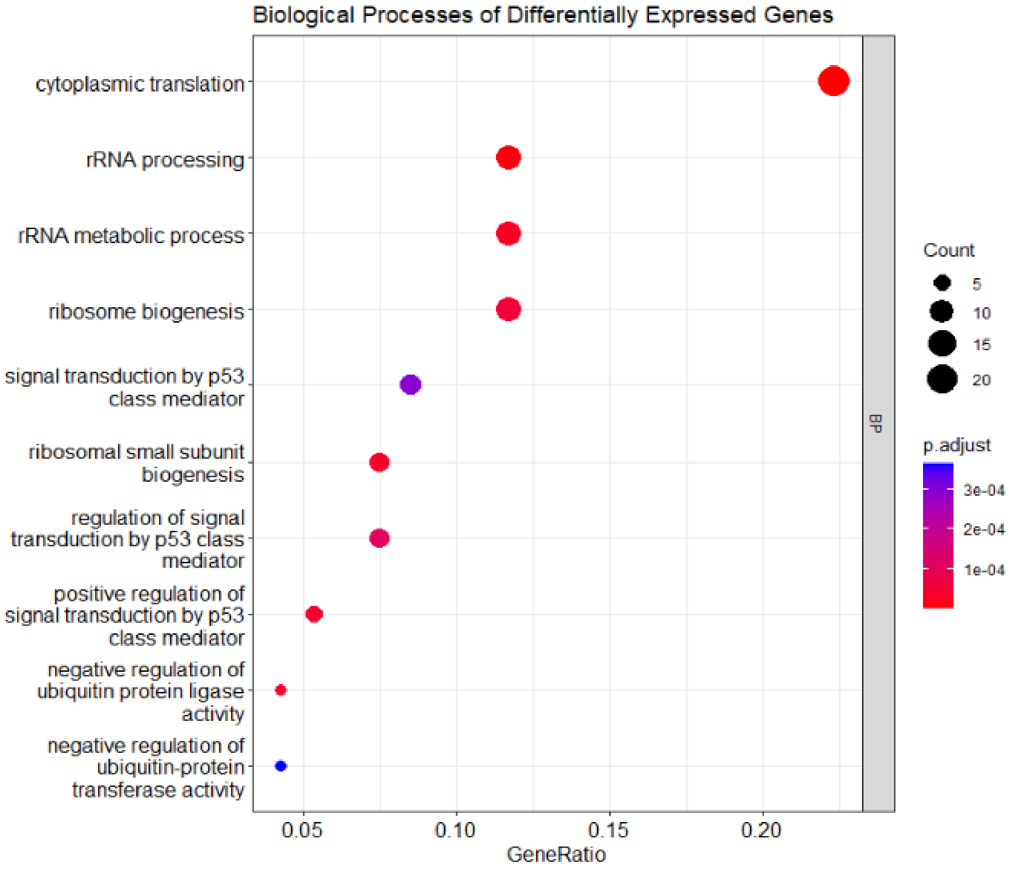


Figure 10. GO enrichment analysis of the top 10 biological processes from the top 100 differentially expressed genes revealed a large proportion of genes related to ribosomal processes, which have been linked to upregulation in eRNA and downregulated in PAXgene™ isolated samples (adjusted $p < 0.05$).

Discussion

The aim of this investigation was to compare the transcriptomes of RNA collected from blood in EDTA (eRNA) and PAXgene™ tubes. The study was designed to validate if a quicker and more efficient protocol (eRNA) was comparable to the gold standard PAXgene™ RNA collection method, and whether it could be taken forward for a large RNA-seq ALS study. Issues unique to measuring RNA-expression include maintaining stability of transcripts and stopping unintended gene induction after blood collection (*ex-vivo*). Typical validation studies of RNA collection methods are limited to a few transcripts via real-time PCR and so here we develop a comprehensive transcriptome-wide comparison at both a gene and isoform-level.

Accordingly to our hypothesis, we expected that each collection method would enable RNA to be profiled with comparable ranked transcriptome profiles between matched samples. Given the PAXgene™ proprietary blend of reagents and reputation as the gold standard collection technique, we also thought this would be superior in the interrogation of the biologically relevant transcriptome (despite being more expensive and laborious in nature). Surprisingly, we only found subtle differences between collection techniques. Analysis of the RNA-seq data demonstrated high-quality sequencing data that was produced using the eRNA method and was comparable to gold standard method. Unexpectedly, the raw number of reads was higher in eRNA compared to PAXgene™ collections. Although there was no significant difference between the average raw read count of both methods, this provides sufficient evidence that the eRNA method can quantify reads comparable to the gold standard technique more efficiently and at ~8x less of collection price. However, previous studies have suggested that there is a rapid decay of free RNA in EDTA-anticoagulated blood, that represents a substantial source of error in RNA-based diagnostics[4,29].

We visualised that the variance between samples were higher in eRNA samples via PCA. Comparatively, PAXgene™ tubes showed less variation between samples and accounted for greater consistency based on expression counts. Through DGE (study identification paired with collection type) and GO enrichment analyses, we assessed differentially expressed protein coding genes that were clustered together. Specifically, genes that were identified to be upregulated were enriched in pathways associated with cytoplasmic translation and rRNA processing, while downregulated genes were associated with sensory perception of taste and embedded in protein peptidyl-isomerisation processes. Notably, *RPL/RPS* were identified as a significant group of protein coding ribosomal genes in the top 50 DE genes during clustering analysis. This group of specific ribosomal genes were observed to be upregulated in eRNA and downregulated in PAXgene™ samples. Given the same library preparation was used (a kit that depletes rRNA) it is possible that these differences could have arisen at the collection stage. We speculate that the time-delay in adding *RNALater* to the EDTA tube (1-2hrs) and/or the proprietary reagent/kit in PAXgene™ (both minimising the *ex vivo* effects known to occur) could have influenced the rRNA in the sample[25]. Without access to further detail on the PAXgene™ tubes, collecting blood directly into *RNALater* may be worth investigating in future studies to minimise rRNA differences.

Potential technical effects in the eRNA were evident by high transcriptomic inflation in the DeSeq2 volcano plots between study identification and collection type paired testing. We ruled out sex-effects by adding this into the model. This was not as evident using other DEG software (Limma/Voom) and could be a result of DeSeq2's approach of controlling for multiple-corrections via Benjamini-Hochberg's procedure, since it does not guarantee error control and can result inflated type I error[11]. This also may be related to the collection procedure and difference in transcripts (i.e., enriched rRNA in eRNA). To combat the inflation, we used a random effect model to increase power and decrease false positives. This substantially reduced transcriptomic inflation in the volcano plot. We note it also constricted the magnitude of significance in DE genes. Controlling for the technical effects of collection (by using a random effect model) may hinder the ability to detect biological effects if both collection methods were used in a larger study.

Given the differences described, we performed ranked correlation analyses between the top 1000 most variable and top 1000 heritable genes. This revealed a peak correlation of ~0.9 but also a long tail (negative correlation) via histogram visualisation for both lists. While this suggested over a fifth of genes are highly correlated, it remains unknown what a typical correlation would be using the same method in a small sample size. Therefore, for further investigation it would be ideal to perform a second experiment of matched collections and/or stochastic simulation of correlations between heritable genes of collection types. This will allow us to see if RNA-seq from PAXgene™ and eRNA do correlate as expected and no different to what could be expected from any biological replicate.

A novel aspect of this study was the use of FRASER to detect evidence of rare aberrant splicing events. Most samples in this study were MND cases and it is expected that splicing is altered in disease. Comparing the ability of each collection method to detect these events, did not find any significant differences in total events. Interestingly, we did not identify any aberrant splicing associated within pathological genes from literature (*TARDP, C9orf72, SOD1 or UNC13A*). It is worth noting that our sample size falls below the recommendation of FRASER documentation. This may impact FRASER’s machine learning approach by small sample size parameters, since the vignette recommended to use more than 20+ samples for optimal results.

Nonetheless, examining novel splice events in PAXgene™ and testing to see if they were replicated in eRNA collection revealed some interesting findings. Overall, it suggested that the power to detect these events may be depleted in eRNA. as the pattern was similar but just fewer events were detected. More specifically, some of the aberrantly spliced genes have been identified in other neurological disease (Alzheimer’s disease (AD), motor neuropathy, and familial adult-onset spinal muscular atrophy (FAMSA)). These were all detected in case samples and included: *CR1* (complement cascade in AD and FTD)*, WARS1* (distal hereditary motor neuropathy)*, HLA-DRB5* (neuroinflammation and AD)*, SIRPB1* (sporadic AD), *TOR1AIP1* (muscular dystrophy and FASMA) and the most statistically significant gene locus, *UBC*[8,9,12,15,26]. *UBC* is known to play a vital role in maintaining ubiquitin homeostasis. Since neurons have a unique dependence on ubiquitination as a critical component for cellular function, 5’ aberrant splicing could implicate imbalanced homeostasis that promotes ubiquitin-positive inclusions and protein aggregation as a mechanism in ALS pathogenesis[6]. Larger samples and or confirmatory isoform analyses are still needed to confirm this.

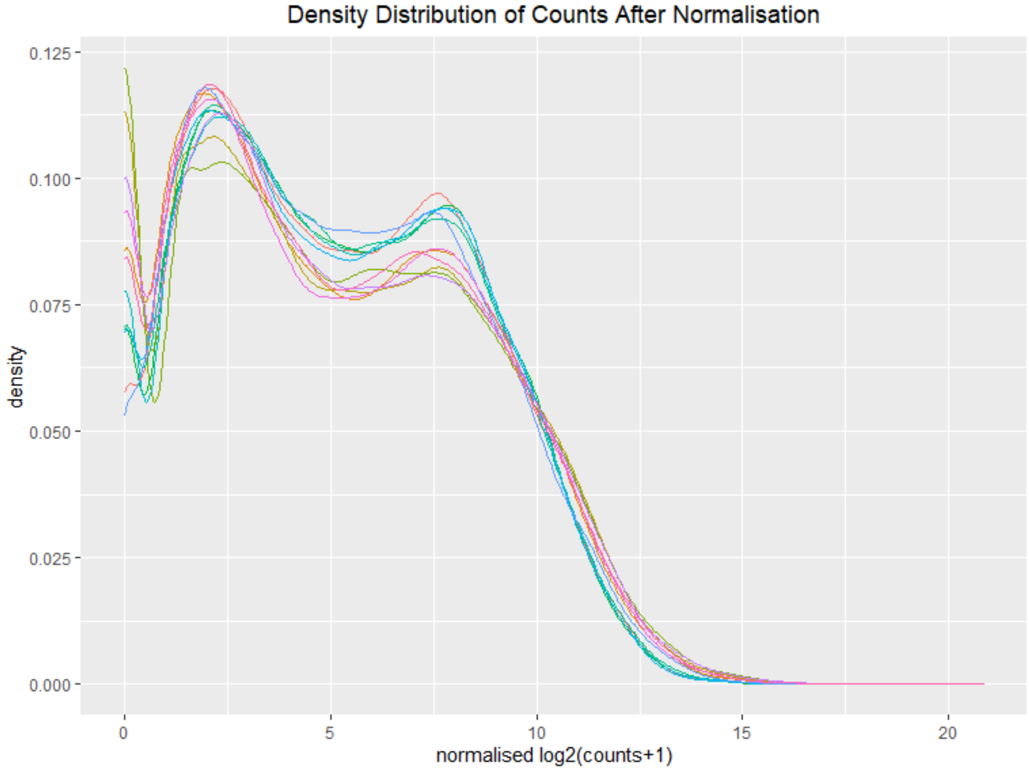
Our results find eRNA collections can produce high quality transcriptome results, that are comparable to the gold standard PAXgene™ collection. The eRNA produced similar numbers of reads and gene counts, and the distribution correlation of the top 1000 (most variable and most heritable) ranked genes both peaked at 0.8-0.9. Differential gene expression did reveal differences between the two collection methods. This could hinder the ability to detect biological effects if both methods were used. More specifically, the higher presence of rRNA transcripts/biological processes in the eRNA collections support refinement of the technique to further optimise its use for transcriptome analysis. Without this refinement, the power to detect rare events or small effects i.e., rare aberrant splicing might be hindered in eRNA compared to PAXgene™ collection. As such, due to the interest in splice events, a large ALS transcriptome study is likely best placed using PAXgene™ collections only.

For other samples, we note that the eRNA collection method is a viable alternative to the gold standard PAXgene™. The significant cost-saving nature of being able to “bank” an eRNA collection may outweigh the potential loss in power that our initial analyses show. Further sample analysis is warranted, particularly a set of collection replicates to gauge what optimal correlation can be expected. The preliminary analyses are promising for this technique and could have a significant impact for the wider scientific community.

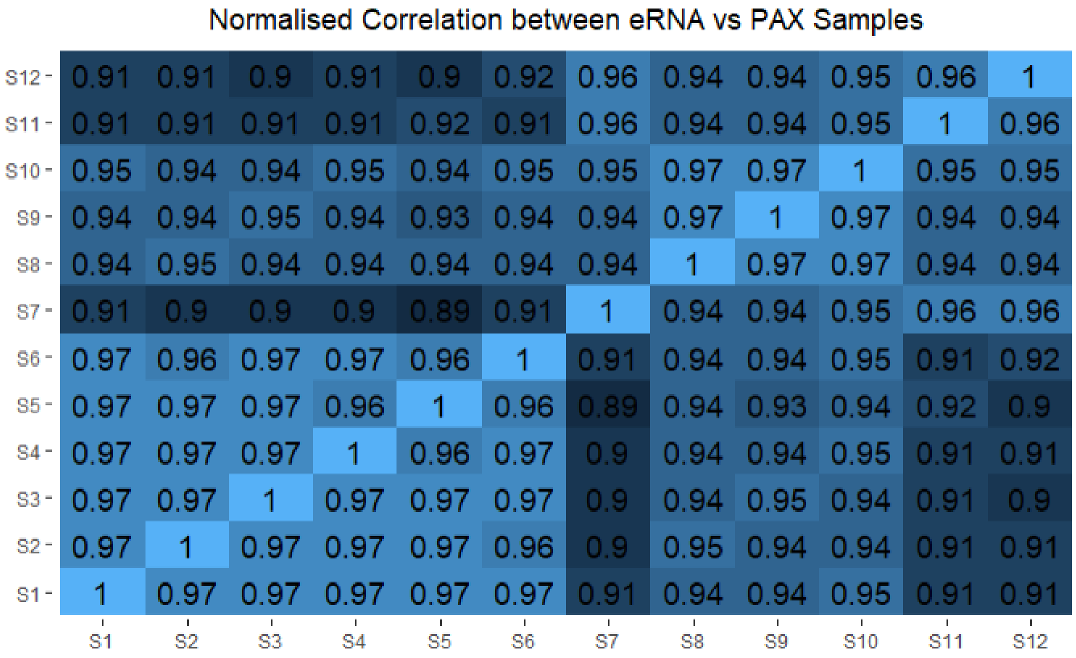
Conclusion

This preliminary investigation showed high quality RNA-seq data from both eRNA and PAXgene™ collections. Further refinement of the eRNA method by incorporating bioinformatic simulations should provide further guidance on the impact of eRNA vs. PAXgene™ in detecting true biological effects. The economical savings of eRNA compared to PAXgene™ method are significant and could have various applications across the wider scientific community. Overall, our results provide promising insights into the practicality of this method and detects some novel aberrant splice events that could be further examined in the pathogenesis of MND.

Supplementary Materials

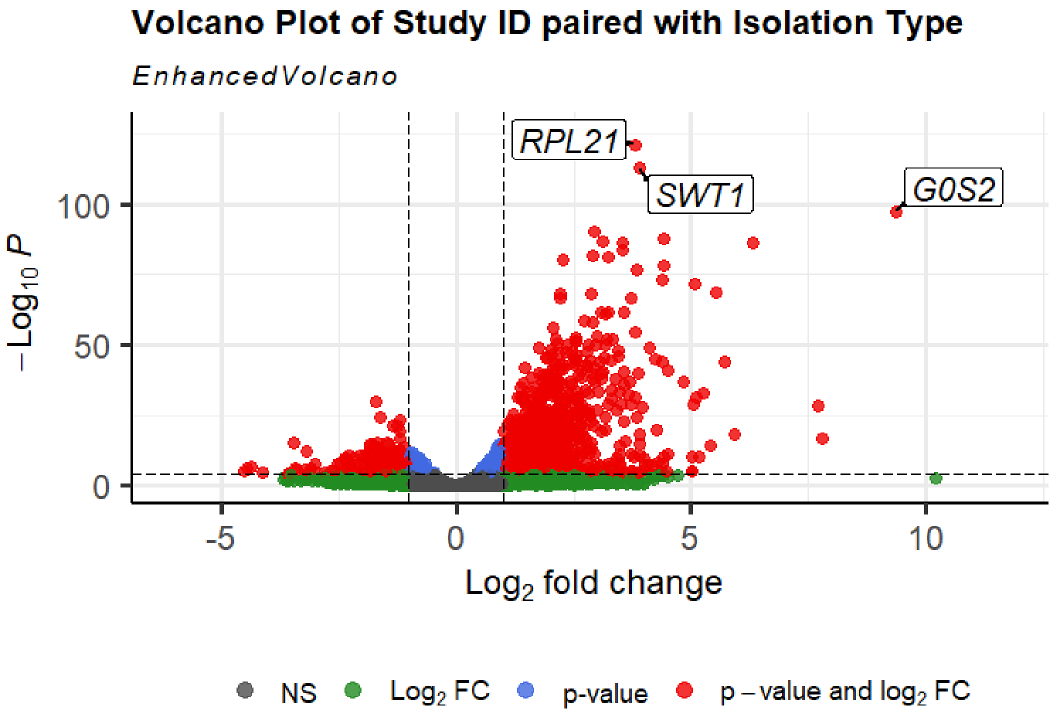


Supplementary Figure 1. Distribution of normalised log2 transformed counts of samples across PAXgene™ and eRNA collection types.

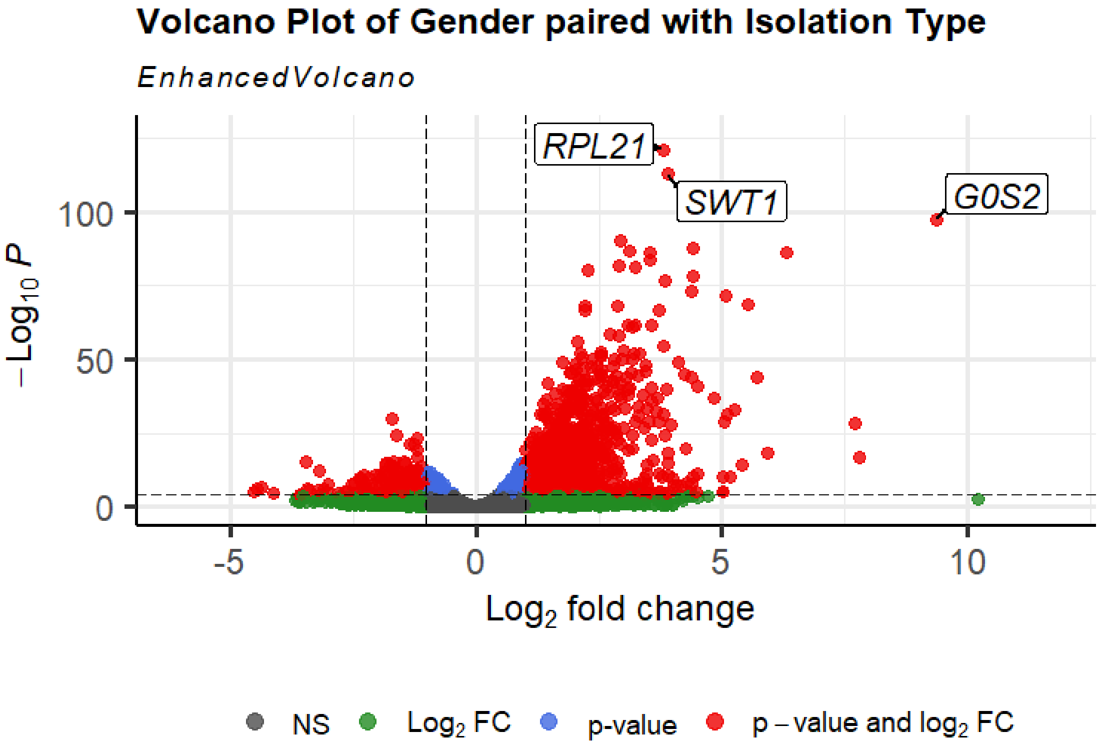


	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
S12	0.91	0.91	0.9	0.91	0.9	0.92	0.96	0.94	0.94	0.95	0.96	1
S11	0.91	0.91	0.91	0.91	0.92	0.91	0.96	0.94	0.94	0.95	1	0.96
S10	0.95	0.94	0.94	0.95	0.94	0.95	0.95	0.97	0.97	1	0.95	0.95
S9	0.94	0.94	0.95	0.94	0.93	0.94	0.94	0.97	1	0.97	0.94	0.94
S8	0.94	0.95	0.94	0.94	0.94	0.94	0.94	1	0.97	0.97	0.94	0.94
S7	0.91	0.9	0.9	0.9	0.89	0.91	1	0.94	0.94	0.95	0.96	0.96
S6	0.97	0.96	0.97	0.97	0.96	1	0.91	0.94	0.94	0.95	0.91	0.92
S5	0.97	0.97	0.97	0.96	1	0.96	0.89	0.94	0.93	0.94	0.92	0.9
S4	0.97	0.97	0.97	1	0.96	0.97	0.9	0.94	0.94	0.95	0.91	0.91
S3	0.97	0.97	1	0.97	0.97	0.97	0.9	0.94	0.95	0.94	0.91	0.9
S2	0.97	1	0.97	0.97	0.97	0.96	0.9	0.95	0.94	0.94	0.91	0.91
S1	1	0.97	0.97	0.97	0.97	0.97	0.91	0.94	0.94	0.95	0.91	0.91

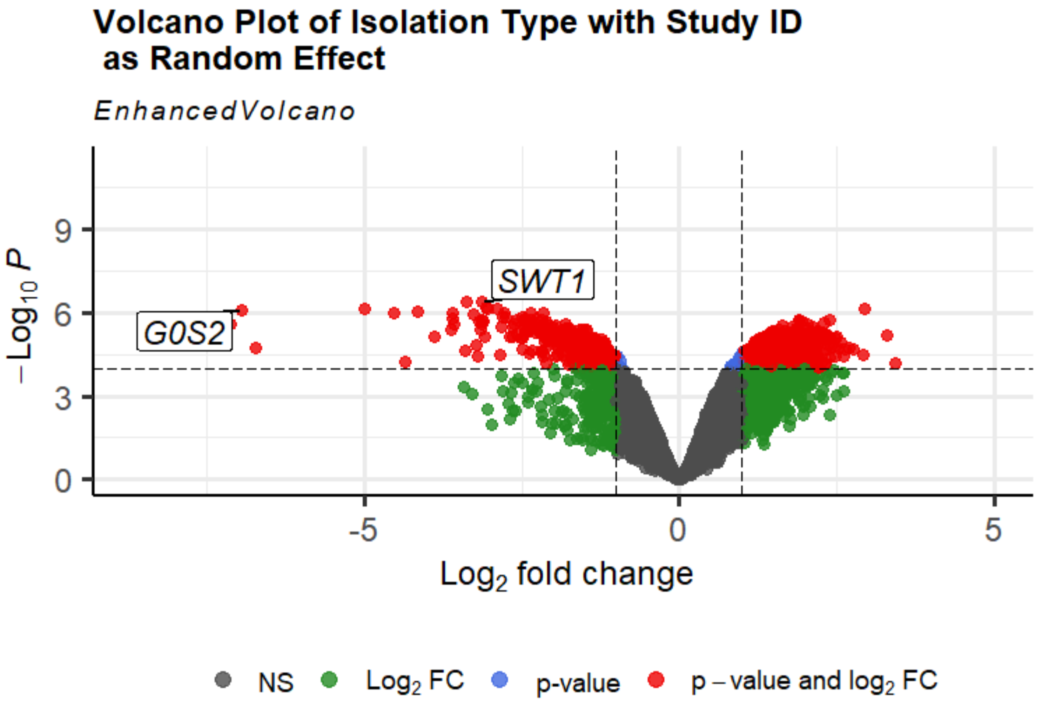
Supplementary Figure 2. Pearson correlation matrix across all samples (S1-S6) of PAXgene™ and eRNA (S7-S12) collection types.



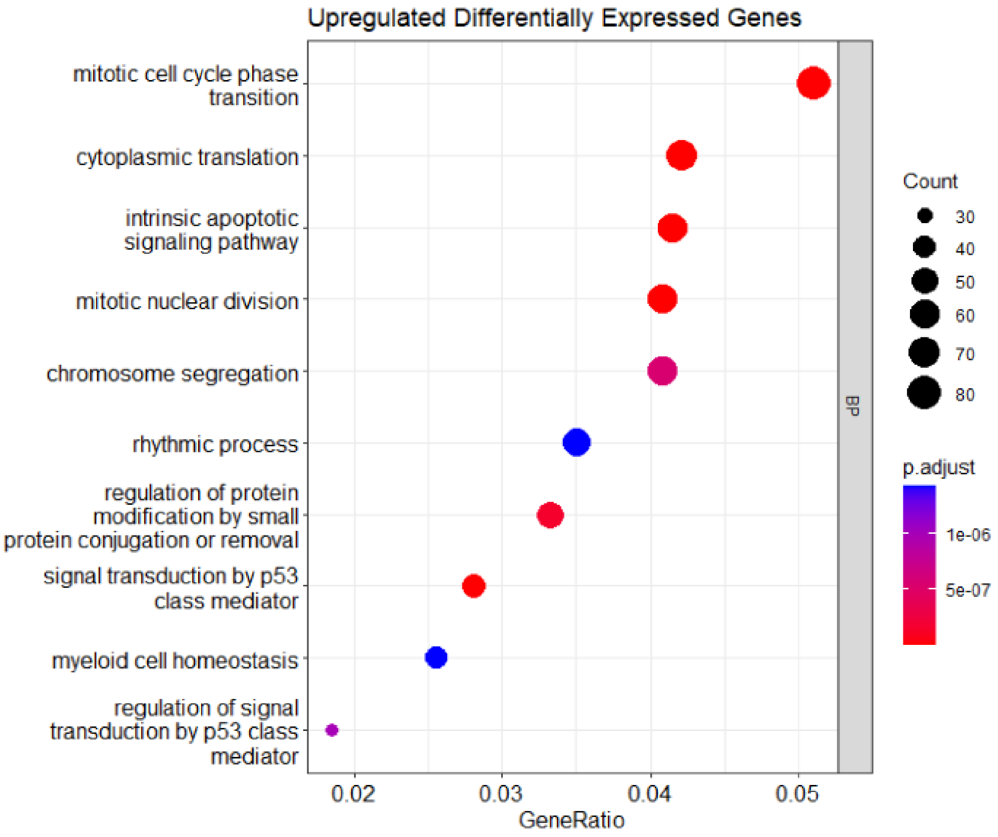
Supplementary Figure 3. Volcano plot of pair testing of collection type and patient identification in DeSeq2, showing high inflation in eRNA magnitude (PAXgene™ set as reference).



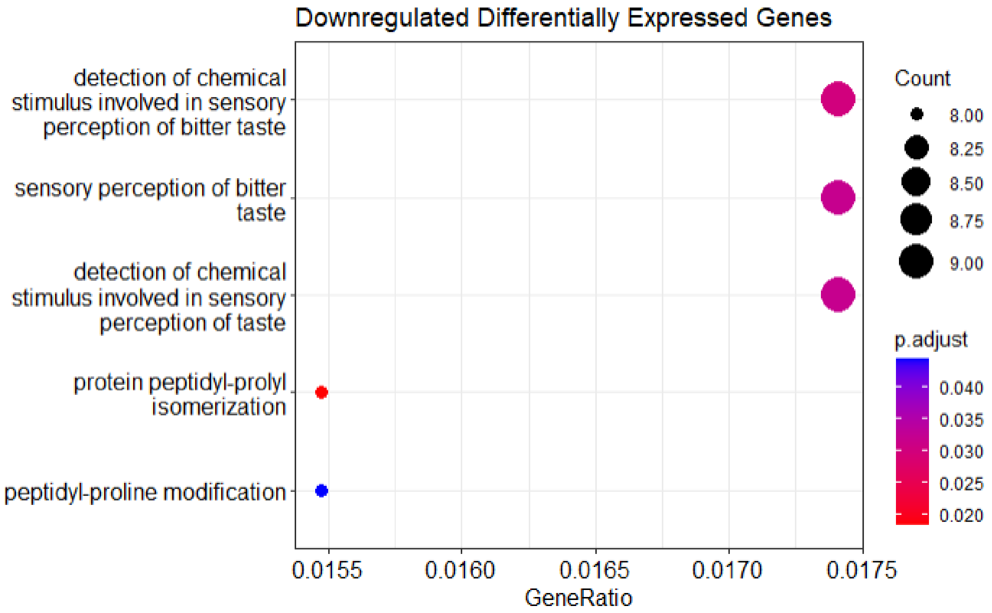
Supplementary Figure 4. Volcano plot of pair testing of collection type and gender in DeSeq2, showing high inflation in eRNA magnitude with little gender effect in comparison to supplementary figure 3 (PAXgene™ set as reference).



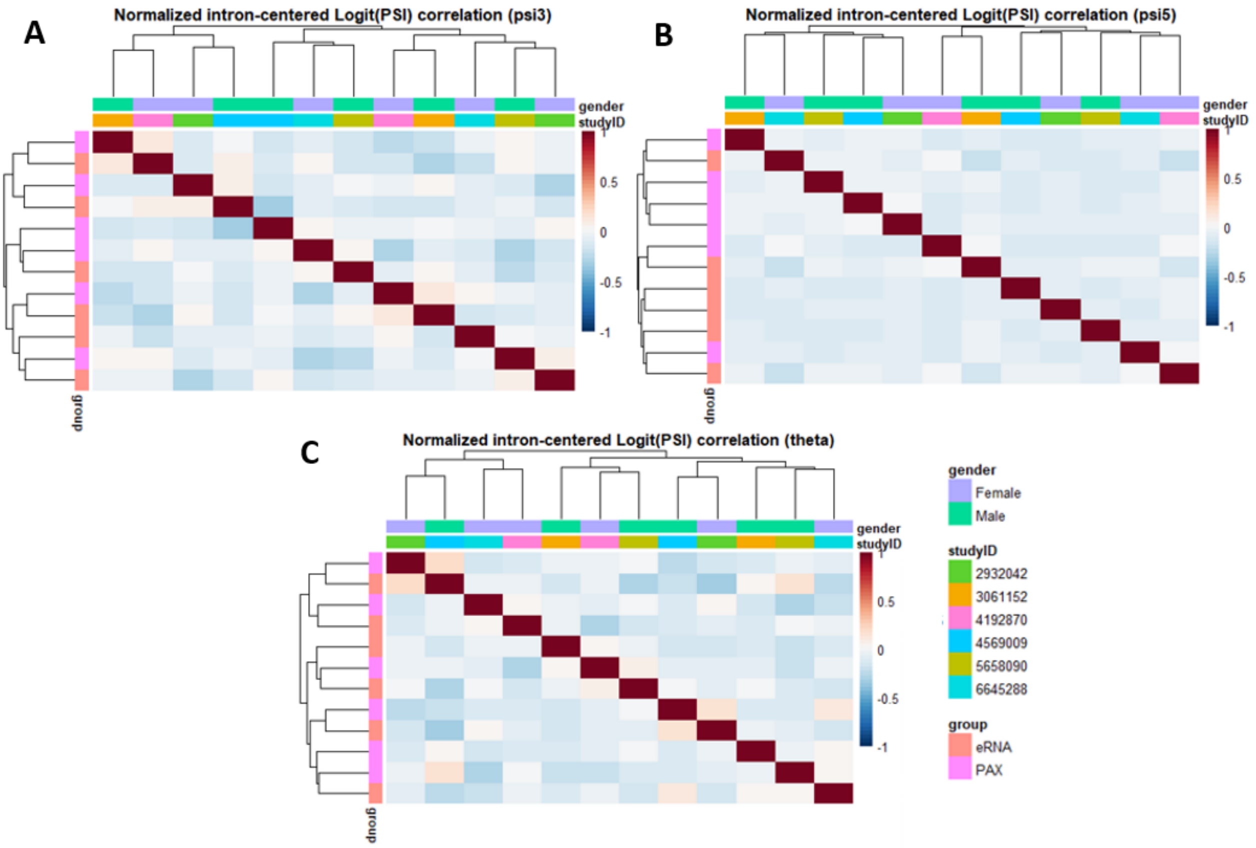
Supplementary Figure 5. Volcano plot of pair testing of collection type and study identification set as a random effect in Limma in comparison to supplementary figure 3 (eRNA set as reference, since can't set reference point in Limma).



Supplementary Figure 6. Top 10 biological Processes of upregulated differentially expressed genes.



Supplementary Figure 7. Top 10 biological processes of downregulated differentially expressed genes.



Supplementary Figure 8. Covariate clustering between FRASER splicing metrics based on gender, isolation type and study identification. 8A showed higher correlation of control cases, particularly in eRNA. 8B showed little changes in correlation of covariates across both isolation types. 8C showed higher positive correlation in cases of eRNA.

References

1. Alonso, A., Logroscino, G., Jick, S. S., & Hernan, M. A. (2009). Incidence and lifetime risk of motor neuron disease in the United Kingdom: a population-based study. *Eur J Neurol*, *16*(6), 745-751. DOI: 10.1111/j.1468-1331.2009.02586.x.
2. Arnold, E. S., Ling, S. C., Huelga, S. C., Lagier-Tourenne, C., Polymenidou, M., Ditsworth, D., Kordasiewicz, H. B., McAlonis-Downes, M., Platoshyn, O., Parone, P. A., Da Cruz, S., Clutario, K. M., Swing, D., Tessarollo, L., Marsala, M., Shaw, C. E., Yeo, G. W., & Cleveland, D. W. (2013). ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proc Natl Acad Sci U S A*, *110*(8), E736-745. DOI: 10.1073/pnas.1222809110.
3. Benson, B. C., Shaw, P. J., Azzouz, M., Highley, J. R., & Hautbergue, G. M. (2021). Proteinopathies as Hallmarks of Impaired Gene Expression, Proteostasis and Mitochondrial Function in Amyotrophic Lateral Sclerosis. *Front Neurosci*, *15*, 783624. DOI: 10.3389/fnins.2021.783624.
4. Bonner, K., Siemieniuk, R. A., Boozary, A., Roberts, T., Fajardo, E., & Cohn, J. (2014). Expanding access to HIV viral load testing: a systematic review of RNA stability in EDTA tubes and PPT beyond current time and temperature thresholds. *PLoS One*, *9*(12), e113813. DOI: 10.1371/journal.pone.0113813.
5. Brown, A. L., Wilkins, O. G., Keuss, M. J., Hill, S. E., Zanovello, M., Lee, W. C., Bampton, A., Lee, F. C. Y., Masino, L., Qi, Y. A., Bryce-Smith, S., Gatt, A., Hallegger, M., Fagegaltier, D., Phatnani, H., Consortium, N. A., Newcombe, J., Gustavsson, E. K., Seddighi, S., . . . Fratta, P. (2022). TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature*, *603*(7899), 131-137. DOI: 10.1038/s41586-022-04436-3.
6. Chisholm, C. G., Lum, J. S., Farrawell, N. E., & Yerbury, J. J. (2022). Ubiquitin homeostasis disruption, a common cause of proteostasis collapse in amyotrophic lateral sclerosis? *Neural Regen Res*, *17*(10), 2218-2220. DOI: 10.4103/1673-5374.335786.
7. Chou, C. C., Zhang, Y., Umoh, M. E., Vaughan, S. W., Lorenzini, I., Liu, F., Sayegh, M., Donlin-Asp, P. G., Chen, Y. H., Duong, D. M., Seyfried, N. T., Powers, M. A., Kukar, T., Hales, C. M., Gearing, M., Cairns, N. J., Boylan, K. B., Dickson, D. W., Rademakers, R., . . . Rossoll, W. (2018). TDP-43 pathology disrupts nuclear pore complexes and nucleocytoplasmic transport in ALS/FTD. *Nat Neurosci*, *21*(2), 228-239. DOI: 10.1038/s41593-017-0047-3.
8. Cossins, J., Webster, R., Maxwell, S., Rodriguez Cruz, P. M., Knight, R., Llewelyn, J. G., Shin, J. Y., Palace, J., & Beeson, D. (2020). Congenital myasthenic syndrome due to a TOR1AIP1 mutation: a new disease pathway for impaired synaptic transmission. *Brain Commun*, *2*(2), fcaa174. DOI: 10.1093/braincomms/fcaa174.
9. Crehan, H., Holton, P., Wray, S., Pocock, J., Guerreiro, R., & Hardy, J. (2012). Complement receptor 1 (CR1) and Alzheimer's disease. *Immunobiology*, *217*(2), 244-250. DOI: 10.1016/j.imbio.2011.07.017.
10. Donnelly, C. J., Grima, J. C., & Sattler, R. (2014). Aberrant RNA homeostasis in amyotrophic lateral sclerosis: potential for new therapeutic targets? *Neurodegener Dis Manag*, *4*(6), 417-437. DOI: 10.2217/nmt.14.36.
11. Ebrahimpoor, M., & Goeman, J. J. (2021). Inflated false discovery rate due to volcano plots: problem and solutions. *Brief Bioinform*, *22*(5). DOI: 10.1093/bib/bbab053.
12. Foveau, B., Correia, A. S., Hebert, S. S., Rainone, S., Potvin, O., Kergoat, M. J., Belleville, S., Duchesne, S., LeBlanc, A. C., & the, C.-Q. C. f. t. e. i. o. A. s. d.-Q. (2019). Stem Cell-Derived Neurons as Cellular Models of Sporadic Alzheimer's Disease. *J Alzheimers Dis*, *67*(3), 893-910. DOI: 10.3233/JAD-180833.
13. Halim, D., & Gao, F. B. (2022). RNA targets of TDP-43: Which one is more important in neurodegeneration? *Transl Neurodegener*, *11*(1), 12. DOI: 10.1186/s40035-021-00268-9.
14. Hieronymus, K., Dorschner, B., Schulze, F., Vora, N. L., Parker, J. S., Winkler, J. L., Rosen-Wolff, A., & Winkler, S. (2021). Validation of reference genes for whole blood gene expression analysis in cord blood of preterm and full-term neonates and peripheral blood of healthy adults. *BMC Genomics*, *22*(1), 489. DOI: 10.1186/s12864-021-07801-0.
15. Hooli, B., & Tanzi, R. E. (2016). Chapter 34 - The Genetic Basis of Alzheimer's Disease: Findings From Genome-Wide Studies. In T. Lehner, B. L. Miller, & M. W. State (Eds.), *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry* (pp. 547-571). Academic Press. DOI: 10.1016/B978-0-12-800105-9.00034-2.
16. Lloyd-Jones, L. R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., Gibson, G., Spector, T., Montgomery, G., Esko, T., Visscher, P. M., & Powell, J. E. (2017). The Genetic Architecture of Gene Expression in Peripheral Blood. *Am J Hum Genet*, *100*(2), 371. DOI: 10.1016/j.ajhg.2017.01.026.

17. Ma, X. R., Prudencio, M., Koike, Y., Vatsavayai, S. C., Kim, G., Harbinski, F., Briner, A., Rodriguez, C. M., Guo, C., Akiyama, T., Schmidt, H. B., Cummings, B. B., Wyatt, D. W., Kurylo, K., Miller, G., Mekhoubad, S., Sallee, N., Mekonnen, G., Ganser, L., . . . Gitler, A. D. (2022). TDP-43 represses cryptic exon inclusion in the FTD-ALS gene UNC13A. *Nature*, *603*(7899), 124-130. DOI: 10.1038/s41586-022-04424-7.
18. McDermaid, A., Monier, B., Zhao, J., Liu, B., & Ma, Q. (2019). Interpretation of differential gene expression results of RNA-seq data: review and integration. *Brief Bioinform*, *20*(6), 2044-2054. DOI: 10.1093/bib/bby067.
19. Mertes, C., Scheller, I. F., Yepez, V. A., Celik, M. H., Liang, Y., Kremer, L. S., Gusic, M., Prokisch, H., & Gagneur, J. (2022). Author Correction: Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat Commun*, *13*(1), 3474. DOI: 10.1038/s41467-022-31242-2.
20. Pervouchine, D. D., Knowles, D. G., & Guigo, R. (2013). Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics*, *29*(2), 273-274. DOI: 10.1093/bioinformatics/bts678.
21. Qi, T., Wu, Y., Zeng, J., Zhang, F., Xue, A., Jiang, L., Zhu, Z., Kemper, K., Yengo, L., Zheng, Z., e, Q. C., Marioni, R. E., Montgomery, G. W., Deary, I. J., Wray, N. R., Visscher, P. M., McRae, A. F., & Yang, J. (2018). Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun*, *9*(1), 2282. DOI: 10.1038/s41467-018-04558-1.
22. Rojas, P., Ramirez, A. I., Fernandez-Albarral, J. A., Lopez-Cuenca, I., Salobrar-Garcia, E., Cadena, M., Elvira-Hurtado, L., Salazar, J. J., de Hoz, R., & Ramirez, J. M. (2020). Amyotrophic Lateral Sclerosis: A Neurodegenerative Motor Neuron Disease With Ocular Involvement. *Front Neurosci*, *14*, 566858. DOI: 10.3389/fnins.2020.566858.
23. Skogholt, A. H., Ryeng, E., Erlandsen, S. E., Skorpen, F., Schonberg, S. A., & Saetrom, P. (2017). Gene expression differences between PAXgene and Tempus blood RNA tubes are highly reproducible between independent samples and biobanks. *BMC Res Notes*, *10*(1), 136. DOI: 10.1186/s13104-017-2455-6.
24. Suk, T. R., & Rousseaux, M. W. C. (2020). The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. *Mol Neurodegener*, *15*(1), 45. DOI: 10.1186/s13024-020-00397-1.
25. Tang, R., She, Q., Lu, Y., Yin, R., Zhu, P., Zhu, L., Zhou, M., & Zheng, C. (2019). Quality Control of RNA Extracted from PAXgene Blood RNA Tubes After Different Storage Periods. *Biopreserv Biobank*, *17*(5), 477-482. DOI: 10.1089/bio.2019.0029.
26. Tsai, P. C., Soong, B. W., Mademan, I., Huang, Y. H., Liu, C. R., Hsiao, C. T., Wu, H. T., Liu, T. T., Liu, Y. T., Tseng, Y. T., Lin, K. P., Yang, U. C., Chung, K. W., Choi, B. O., Nicholson, G. A., Kennerson, M. L., Chan, C. C., De Jonghe, P., Cheng, T. H., . . . Lee, Y. C. (2017). A recurrent WARS mutation is a novel cause of autosomal dominant distal hereditary motor neuropathy. *Brain*, *140*(5), 1252-1266. DOI: 10.1093/brain/awx058.
27. Van Deerlin, V. M., Leverenz, J. B., Bekris, L. M., Bird, T. D., Yuan, W., Elman, L. B., Clay, D., Wood, E. M., Chen-Plotkin, A. S., Martinez-Lage, M., Steinbart, E., McCluskey, L., Grossman, M., Neumann, M., Wu, I. L., Yang, W. S., Kalb, R., Galasko, D. R., Montine, T. J., . . . Yu, C. E. (2008). TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *Lancet Neurol*, *7*(5), 409-416. DOI: 10.1016/S1474-4422(08)70071-1.
28. Williams, T. L. (2013). Motor neurone disease: diagnostic pitfalls. *Clin Med (Lond)*, *13*(1), 97-100. DOI: 10.7861/clinmedicine.13-1-97.
29. Wilson, C., Dias, N. W., Pancini, S., Mercadante, V., & Biase, F. H. (2022). Delayed processing of blood samples impairs the accuracy of mRNA-based biomarkers. *Sci Rep*, *12*(1), 8196. DOI: 10.1038/s41598-022-12178-5.
30. Wingo, T. S., Cutler, D. J., Yarab, N., Kelly, C. M., & Glass, J. D. (2011). The heritability of amyotrophic lateral sclerosis in a clinically ascertained United States research registry. *PLoS One*, *6*(11), e27985. DOI: 10.1371/journal.pone.0027985.