Analysis of RNA stability in ALS patients

January 2021

Wesley de Nooijer (4272382)
Utrecht University and UMC Utrecht
Supervisors: dr. Kevin Kenna and Yan Wang
Second reviewer: dr. Onur Basak

Abstract

Studies of the pathological features of Amyotrophic Lateral Sclerosis (ALS) implicate anomalous RNA misprocessing with the disease. Here, we investigate motor cortex RNA stability and its genetic underpinning in a cohort of ALS patients. RNA stability captures information about RNA misprocessing and is estimated using total RNA sequencing data. Analyses of RNA stability estimates show that outliers occur disproportionally in neuronal pathways relevant to ALS such as synaptic vesicle recycling and neuron projection regeneration. The genetics underlying RNA stability are studied firstly by relating the most common mutation underlying ALS, C9orf72 expansions, to RNA stability. We find that C9orf72 positive samples generally have lower RNA stabilities. Next, evolutionary scores as well as scores for impact on RNA-binding affinity are calculated for genetic variants. However, relating these scores with RNA stabilities did not yield any significant results. Overall, we demonstrate the importance of using RNA stability for studying ALS and recommend several improvements to the methodology, including the incorporation of micro-RNAs and transcript features into the statistical models, to capitalize on its potential for further discoveries in ALS and other phenotypes.

Layman's summary

Amyotrophic Lateral Sclerosis (ALS) is a disease that is characterized by the degradation of neurons. One possible cause of this degradation is mistakes during the processing of RNA transcripts. In this report, we investigated whether there was a link between misprocessed RNA and ALS by estimating the stability of RNA transcripts. Using this approach, we show several links between RNA stability and biological mechanisms relevant to ALS. Overall, we demonstrate that there is potential for future discoveries in ALS by using RNA stability estimates and making several adjustments to our methods.

Introduction

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease characterized by progressive degeneration of motor neurons in the brain and spinal cord (Rowland and Shneider, 2001). Symptoms include muscle atrophy, spasticity, and dysphagia, with patients typically dying from neuromuscular respiratory failure 2-5 years after symptoms onset. ALS is the third most common neurodegenerative disease in humans, with a cumulative lifetime risk of ~1 in 300-400 (Brown and Al-Chalabi, 2017; Johnston et al., 2006). ALS can affect people of all ages, although the age of onset peaks at 65 years in Europe (Logroscino et al., 2010).

Heritability analyses have indicated that individual risk for ALS includes a substantial genetic component of ~50-60% (Al-Chalabi et al., 2010; Ryan et al., 2019). For 10-15% of ALS patients, genetic risk factors have now been identified and this has led to clinical trials involving gene therapy that show some clinical promise (Miller et al., 2020). The identification of additional genetic risk factors could help further elucidate the mechanisms causing ALS and provide new opportunities for treatment and genetic counseling.

Genetics of ALS

Around 10% of ALS patients have a positive family history of ALS (Talbott et al., 2016). Major progress has been made in discovering risk genes for familial ALS (fALS), with four genes (C9orf72, TARDBP, SOD1, FUS) accounting for over two-thirds of all familial cases (Chia et al.,

2018; Hardiman et al., 2017). The remaining 90% of ALS cases have no known affected family members and are classified as sporadic ALS (sALS). Despite the substantial implied genetic component for individual ALS risk, only about 15% of sALS cases can be explained by risk genes identified thus far (Chia et al., 2018). The most common cause of ALS is a C9orf72 repeat expansion that is found in at least 8% of sALS cases and over 40% of familial ALS cases. (Donnelly et al., 2013; Majounie et al., 2012).

Genome-wide association studies (GWAS) of sALS suggest that rare variants underpin the genetic architecture of the disease (Hardiman et al., 2017; Van Rheenen et al., 2016). The genetics of ALS are complex as oligogenic inheritance, pleiotropy, and gene-environment interactions obscure links between variants and phenotype.

Nevertheless, over 40 genes have now been associated with ALS (Gregory et al., 2020). Studying the putative function of these genes may give insights to the mechanisms underlying the disease. Studies following this approach identified three major types of upstream mechanisms that converge to the pathway of neuron degeneration characteristic to the disease. These processes are 1) impaired protein homeostasis 2) RNA misprocessing, and 3) disruptions in axonal transport (Barmada, 2015; Ghasemi and Brown, 2018; Masrori and Damme, 2020). The focus of this research will be on the second.

RNA processing

Between DNA transcription and fulfilling its final function, RNA is subject to several regulatory processes, including pre-mRNA processing, nuclear export, localization, and decay (Figure 1; Alberts et al., 2014). RNA-binding proteins (RBPs) and micro-RNAs (miRNAs) that interact with the transcript control many of these processes. Loss of function of the FUS and TDP-43 RBPs have often been associated with ALS (Da Cruz and Cleveland, 2011; Ishigaki and Sobue, 2018; Paez-Colasante et al., 2015). Specifically, FUS and TDP-43 regulate the transcription, splicing and transport of many mRNAs. (Buratti, 2008; Ishigaki and Sobue, 2018). Collectively, these observations underscore the importance of RNA misprocessing in ALS. The aggregate effects of RNA misprocessing can be studied using a proxy metric: the stability of RNA transcripts. Indeed, Alkallas et al. (2017) showed that the number of RBP and miRNA binding sites in 3' UTRs in a regulatory model consisting of the RBFOX1 and ZFP36 RBP families together with four miRNAs was significantly predictive of brain mRNA stability.

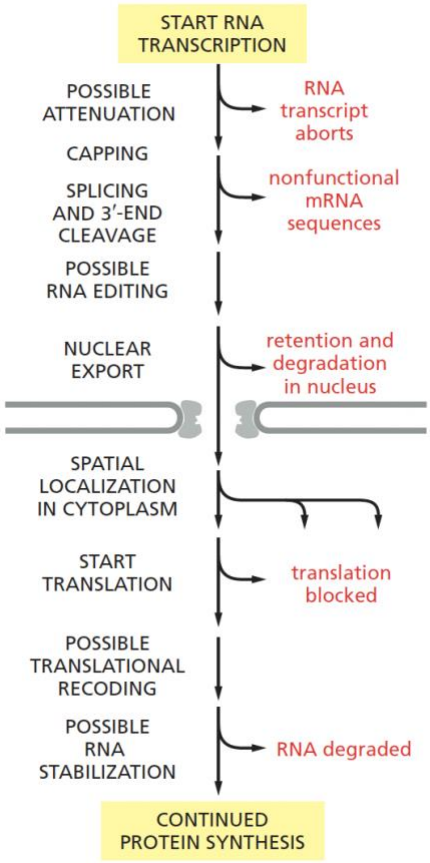

Figure 1: Post-transcriptional controls on gene regulation. Taken from Alberts et al. (2014; Figure 7-54, pg. 413).

RNA stability

The stability of an RNA transcript is a key determinant of its abundance. Generally, exonic read abundance is assumed to be a proxy purely for transcriptional activity. However, this is a simplification that fails to account for differences in stability between transcripts (Furlan et al., 2021). For instance, lowly expressed genes with highly stable transcripts may have the same exonic read abundance as highly expressed genes with low stability transcripts. While the consequences of post-transcriptional regulation are captured by exonic read abundance, it is hard to distinguish it from transcriptional regulation. Gaidatzis et al. (2015) showed that measuring changes in both exonic and intronic read abundance enables the separation of transcriptional and post-transcriptional regulation. This approach seems especially suited for the brain, where intronic reads are relatively abundant, unlike in other tissues (Ameur et al., 2011). However, one caveat of this approach is that it assumes that intronic read abundance is solely influenced by transcription rate, thereby neglecting the RNA processing rate. Nevertheless, transcription rate and splicing rate seem strongly linked, possibly through co-transcriptional splicing (Ameur et al., 2011).

The stability of a transcript is primarily determined by its sequence. Transcript features such as transcript length, number of miRNA targets, and UTR-lengths are negatively correlated with RNA stability (Duan et al., 2013). GC-content is weakly positively correlated with stability. Additionally, the sequence will affect the ability of RBPs and miRNAs to bind, and thereby impact processes such as localization and degradation.

Project outline

In this project we aim to answer the following research questions:

- Can total RNA sequencing be used to identify anomalous RNA stability in ALS patients?
- Can individual estimates of RNA stability be related to individual whole genome sequencing data to identify relevant variants for follow up gene discovery analyses?

To answer these questions, we first estimate RNA stability from total RNAseq. Subsequently, we investigate the RNA stability outliers and test for enrichment in genes or pathways. Lastly, we investigate the genetic underpinning of RNA stability by pairing the estimates with whole genome sequencing (WGS) data.

Methods

Data description

RNAseq and WGS data used in this study were from New York Genome Center (NYGC, https://www.nygenome.org/als-consortium/) and Answer ALS (AALS, https://www.answerals.org/). The NYGC dataset comprises 178 ALS patients and 62 controls, of which 29 patients had other neurological disorders. The data were derived from post-mortem motor cortex tissue samples. The AALS dataset contains 121 patients and 24 controls motor neurons samples derived from induced pluripotent stem cells. WGS data was available only for the NYGC dataset. Given the difference in the type of samples between the two datasets, each of the following processing and analyses steps were performed separately for the two datasets.

Estimating RNA stability

RNA stability was estimated by leveraging both intronic and exonic read counts derived from RNAseq data. Exonic and intronic segment coordinates were extracted from ENSEMBL annotation file release 99 (ENSEMBL, 2021). HTSeq-count (Anders et al., 2015) was then used to count the number of reads that were mapped to exonic and intronic segments of the human reference genome, respectively.

The exonic and intronic read counts form the input for Rembrandts (REMoving Bias from Rna-seq ANalysis of Differential Transcript Stability; Alkallas et al., 2017), a software package for estimating RNA stability. Rembrandts is based on a simple model for gene expression consisting of RNA transcription, processing, and decay. Each of these processes is assumed to be at steady state, which makes it possible to separate transcriptional and post-transcriptional processes as described below.

The Rembrandts RNA stability estimates are made with the assumption that intronic read abundances solely reflect transcriptional controls on gene expression. Exonic read abundances, on the other hand, reflect both transcriptional and post-transcriptional processes. By subtracting the logarithm of the fold-change of intronic reads (Δintron) from the logarithm of the fold-change in exonic reads (Δexon), the post-transcriptional component can be isolated. The aggregate effects of post-transcriptional regulation manifest in the degradation rate. Δexon – Δintron is therefore a proxy for mRNA stability: Δstability. It is important to note that Δstability is only a relative measure of changes in stability between different steady state conditions.

Extending this approach for estimating RNA stability, Alkallas et al. (2017) show that using Δexon – Δintron gives a highly biased measure of Δstability. The bias arises due to the limited availability of RNA processing machinery. If there are more unprocessed pre-mRNA molecules present than can be handled at a given moment, then the intronic read abundance will be relatively higher. This gives a higher weight to transcriptional processes (Δintron) over post-transcriptional (Δexon- Δintron), leading to an overestimation of stability. This bias is regressed out in Rembrandts. The final measure for RNA stability is given by Δstability = Δexon – Δintron - bias. Larger differences between Δexon and Δintron indicate that there was a likely large post-transcriptional component in the differential gene regulation.

To calculate RNA stability, generally sufficient read coverage across the samples in a gene is needed. Rembrandts estimates a minimum read cutoff threshold that nearly maximizes the correlation between Δexon and Δintron. This correlation is present because both metrics are a function of changes in transcriptional rate. Maximizing this correlation minimizes the effects of noise due to low read coverage. If the read coverage is below this threshold, no stability value is calculated for that gene.

The stability estimates made by Rembrandts were subsequently corrected for confounders. BRETIGEA (McKenzie et al., 2018) was used to correct for heterogeneity of brain cell types in the samples. BRETIGEA uses a validated set of brain cell type-specific marker genes to estimate the relative proportion of six major brain cell types, i.e., astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and oligodendrocyte precursor cells. For the NYGC dataset, extensive metadata were available and corrections for sequencing platform, sex, and RNA integrity number (RIN) were made. The AALS data was also corrected for sex.

WGS data

WGS data from NYGC were mapped to genome hg38 using Burrows-Wheeler alignment (BWA; Li and Durbin, 2009) following the protocols laid out by Regier et al. (2018). For variant calling, a joint callset was generated using GATK v4.1.0 (Auwera and O'Connor, 2020).

RNA stability in ALS patients

We performed several statistical tests to investigate anomalous RNA stability in ALS patients. In these analyses we used measured stability values to identify outlier genes or pathways.

The first analysis was aimed at identifying outlier genes. Measured stability values were compared to simulated values using a permutation approach. This analysis works by calculating a likelihood of the stability values occurring in a gene. This likelihood is calculated by converting stability values to z-scores and subsequently calculating the probability of a z-score of that magnitude occurring in a normal distribution. The resulting p-value of each gene is then multiplied to calculate a final likelihood for the entire gene (Figure 2a). The likelihoods of the observed stability values are subsequently compared to those of 1 million simulated likelihoods. The simulated likelihoods are calculated by taking one p-value from each sample (Figure 2b). If the likelihood score for a gene is more extreme than many of the 1 million simulations, it may indicate that this gene is characterized by anomalous RNA stability.

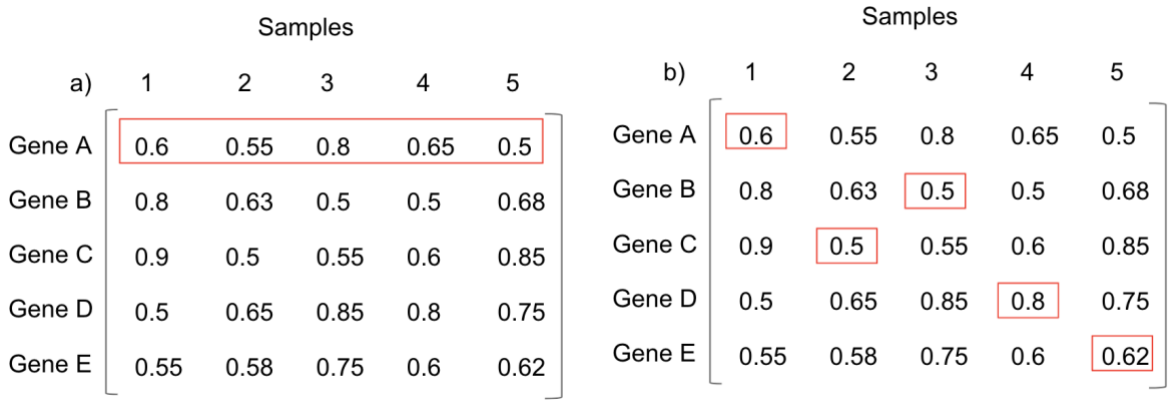

Figure 2: Visualization of the permutation test methodology: a) A likelihood score is calculated for each gene by taking the product of all values of that gene, b) simulated likelihoods are calculated by taking one value from each sample and calculating the product of those values.

Next, we used the p-values resulting from the permutation approach to identify pathways enriched in RNA stability outliers. To do this, we performed a gene set analysis using gene ontology (GO) terms. Gene ontology terms represent biologically relevant groups of genes. The GO-terms were taken from the Molecular Signature Database (MSigDb; Liberzon et al., 2015; Subramanian et al., 2005). The gene set analysis involves correlating gene set membership (true/false) with the p-value of each gene. This enables us to identify pathways influenced by anomalous RNA stability.

To further investigate the role of RNA stability in ALS, we performed two statistical tests linking stability values and case-control status. These two tests are a logistic regression and an odds ratio test. The former tests for general association between RNA stability values and case-control status. The latter test is focused on outliers specifically and investigates whether the number of outliers among all samples in a gene is related to case-control status. The chosen threshold for outlier stability is 1.96 standard deviations from the mean, which corresponds to 5% of the area of a normal distribution. The odds ratio test compares the fraction of samples that are outliers among cases with the fraction of outliers of controls. If outliers occur disproportionally in either group, the gene may be predictive of case-control status.

We also performed replication analyses for each test. Such analyses can be performed to combine the results of multiple statistical tests with the same null hypothesis. In this way, we can combine the results of the statistical tests performed on the two different datasets for all genes that occur in both datasets. An important caveat here is that the data are derived from different sources (post-mortem samples and iPSC). However, results with a strong effect may still replicate in both datasets. The replication analyses were done using Fisher's method, which involves taking the sum of the log of the p-values as denoted in Equation 1, with k = 2014, the number of overlapping genes.

$$\sum_{i=1}^{k} -2 * \log\left(p_i\right) \tag{1}$$

The genetics of RNA stability

To investigate the link between genetic variants and RNA stability, we first performed a logistic regression to test for association between the most common cause of ALS: C9orf72 repeat expansions, and RNA stability values.

Next, we used two different metrics that aggregate information about the downstream effects of the variants. If links between a metric and RNA stability are found, individual variants can then be investigated further. The two metrics are Combined Annotation Dependent Depletion (CADD) and DeepClip scores. CADD scores inform about the relative pathogenicity of a variant (Kircher et al., 2014), while DeepClip scores (Grønning et al., 2020) represent the relative effect a variant has on the binding of a specific RNA-binding protein. The genetic variants in the dataset were annotated using reference genome BSgenome.Hsapiens.UCSC.hg38.

CADD scores are calculated by support vector machines that were trained on evolutionarily conserved and simulated single-nucleotide variants. The higher the CADD score, the more likely the variant is simulated, and therefore the more likely to have deleterious effects.

DeepClip is a deep learning-based tool that predict scores of RNA-binding proteins binding to a specific sequence. DeepClip is trained to work on relatively short sequences (maximum 75 basepairs), and we therefore cut the genome into 75 basepair long contigs surrounding a mutation. The effects of long-range interactions are thus not captured. DeepClip scores were calculated for nine RBPs: AGO1-4, ALKBH5, CAPRIN1, ELAVL1, FUS, PUM2, TDP-43, hnRNPC, and RBFOX1. Loss of function of two of these RBPs has often been associated with ALS: FUS and TDP-43 (Da Cruz and Cleveland, 2011; Ishigaki and Sobue, 2018; Paez-Colasante et al., 2015). The other seven RBPs serve as controls, potentially enabling the isolation of ALS-specific effects. For each specific RBP, the parameters of the deep learning model were taken from the DeepClip database. DeepClip scores inform about the binding affinity of a specific RBP to a particular sequence. By calculating the difference between DeepClip score of the reference sequence and that of the sequence containing the variant (ΔDC_score), we can estimate the effects of that variant on binding. Given the importance of RBPs in post-transcriptional regulatory processes, mutations that disrupt such processes may lead to large ΔDC_scores. These ΔDC_scores may therefore have predictive power regarding RNA stability.

We also used CADD scores and ΔDC_scores of variants on the 3’ UTR and 5’ UTRs. These two regions of the mRNA are heavily involved in many post-transcriptional processes such as localization and repression (Araujo et al., 2012; Mayr, 2019). Focusing on genetic variants in these regions may therefore provide a clearer signal of mutations that disrupt post-transcriptional processes. Lastly, we performed odds ratio tests to test whether outlier stabilities are more likely if a sample contains variants with high CADD or ΔDC_scores.

Results

RNA stability estimates in ALS patients

RNA stability estimates made by Rembrandts are variance stabilized after which the average value of all samples is subtracted. This yields RNA stability values centered around zero. Figure 3 depicts the resulting RNA stability values for five example genes. Positive values imply that the sample had a relatively high transcript stability for that specific gene, as compared to

other samples. Values departed from zero indicate post-transcriptional regulation was likely more important.

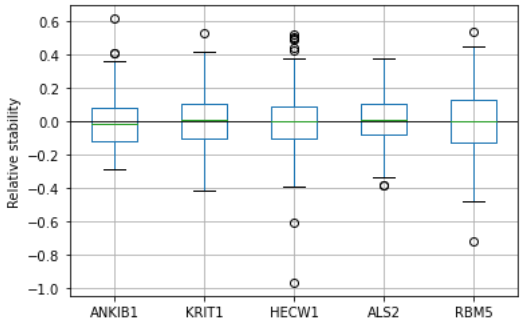

Figure 3 Box-and-whisker plot of RNA stability values for five example genes to illustrate Rembrandts output visually. Positive values imply transcripts were relatively stable. The further values depart from zero, the more likely post-transcriptional regulation was important.

2061 genes passed the read count cut-off threshold for the NYGC dataset, and 11441 for the AALS dataset. The large difference in the number of genes passing the threshold is due to a large difference in read coverage. The read coverage of the AALS dataset is generally much higher.

The permutation test compared the stability values obtained by Rembrandts with permuted values to identify outlier genes. The results of these tests indicate that stability outliers did not occur disproportionally in any gene (Figure 4a and 4b). After false discovery correction, no p-values were smaller than 1.

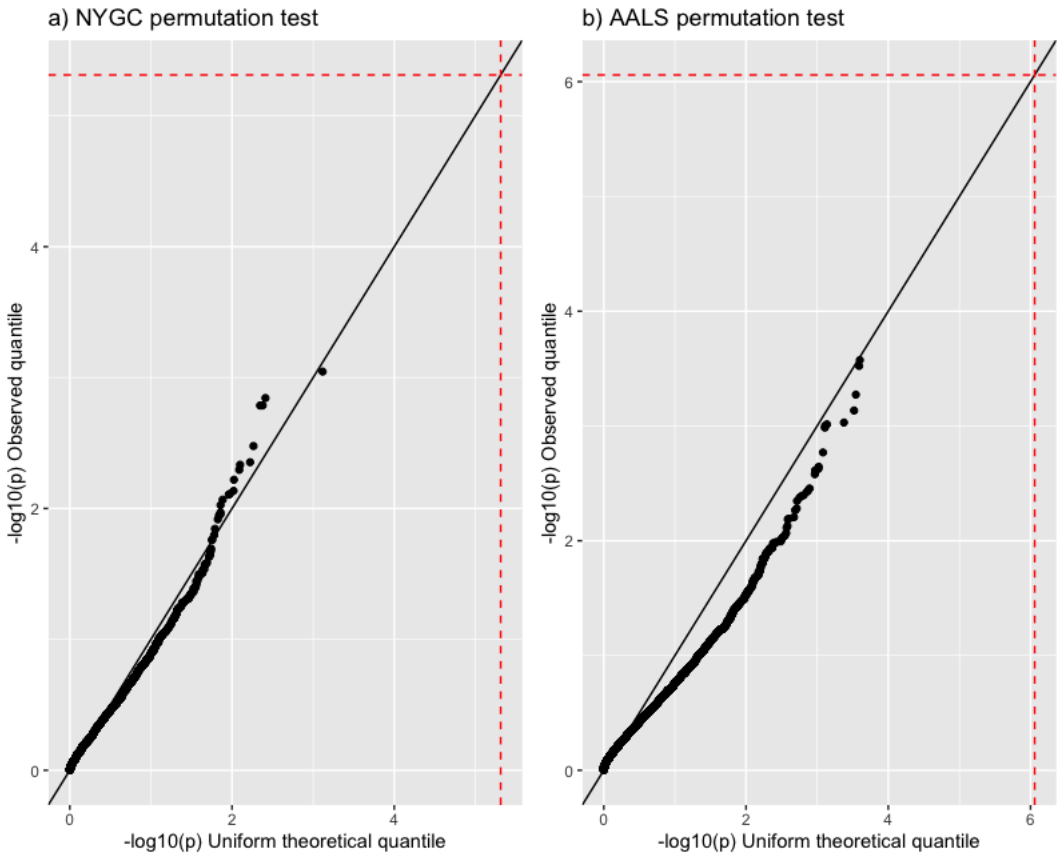

Figure 4: Distribution of permutation test -log10 gene p-value for the a) NYGC and b) AALS datasets. Red dashed line denotes threshold of p < 0.01 after FDR-correction. No p-values were significant after FDR-correction

The next analysis was the gene set analysis, and it was aimed at identifying pathways that were enriched in stability outliers. These tests yielded a significant result at $p < 0.05$ for each dataset. Namely, regulation of synaptic vesicle recycling ($p = 0.02$) was significant for the NYGC dataset and neuron projection regeneration for the AALS dataset ($p = 0.042$) at $p < 0.05$.

Figure 5 depicts the range of p-values of the gene set and the background. The top three pathways for each dataset in terms of lowest p-values are listed in Table 1. Additionally, the combined p-value, based on the Fisher's exact test, is listed in the table. Both pathways have a combined p-value of 0.19. One caveat regarding these results is that the difference in source from which the samples were derived, i.e., post-mortem brain tissue and iPSCs, the replication test invalidates the assumption of a shared null hypothesis.

Table 1: Top pathways in terms of significance from the GSEA analysis.

Pathway	NYGC p-value	AALS p-value	Combined p-value
Neuron projection regeneration	1	0.042	0.19
Postexertional Malaise	NA	0.089	NA
Negative regulation of activin receptor signaling pathway	NA	0.10	NA
Regulation of synaptic vesicle recycling	0.020	1	0.19
Spinal rigidity	0.18	1	1
Limb girdle muscle weakness	0.59	1	1

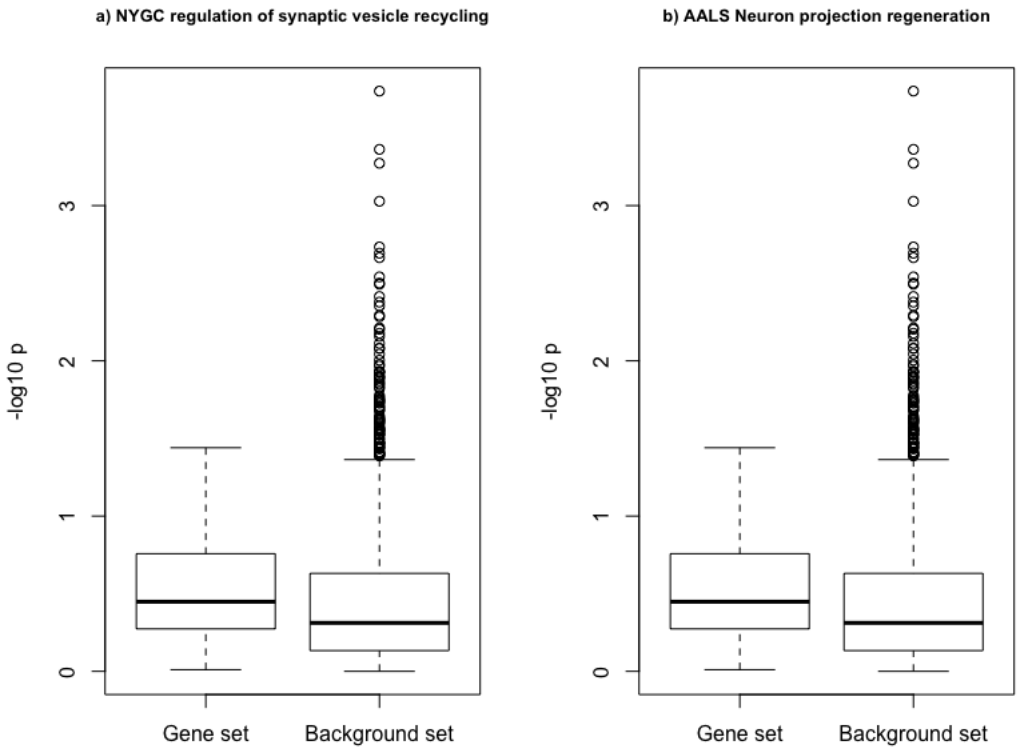

Figure 5: Box-and-whisker plot of gene set and background p-values for the a) regulation of synaptic vesicle recycling pathway in the NYGC dataset and b) the neuron projection regeneration pathway in the AALS dataset.

Investigating the relationship between RNA stability and ALS further, we find that individual gene RNA stabilities are generally not strong predictors of case-control status in our datasets. After false discovery rate (FDR) correction, the GDA gene ($p = 0.036$), is significant at a $p < 0.05$ level for the NYGC dataset, and no genes are significant for AALS. The logistic regression for GDA is plotted in Figure 6. Combining the adjusted p-values of the two datasets did not yield significant results.

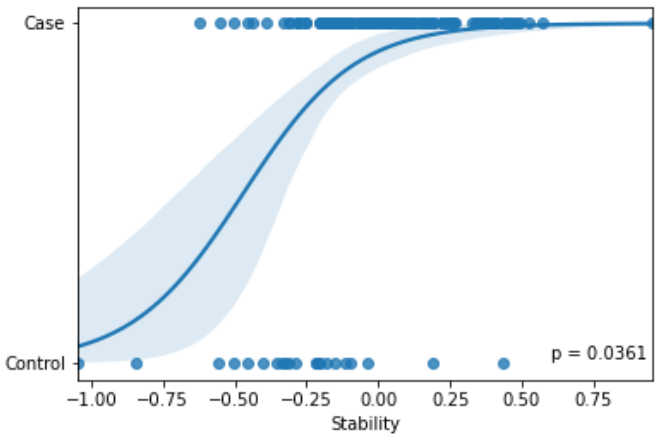

Figure 6: Logistic regression of stability and case-control status for GDA, the gene with the lowest p-value. The blue line denotes the best-fit curve for predicting case-control status based on stability values. Shading denotes the 95% confidence interval.

Next, the odds ratio tested whether RNA stability outliers occurred disproportionally among cases or controls. The results indicate that the number of outliers in a gene is not significantly different between cases or controls. This is visualized in Figure 7, where p-values are plotted against the odds ratio. The distribution of values is skewed towards positive odds ratios (>1) which indicates that outliers were generally more likely to occur in cases than controls.

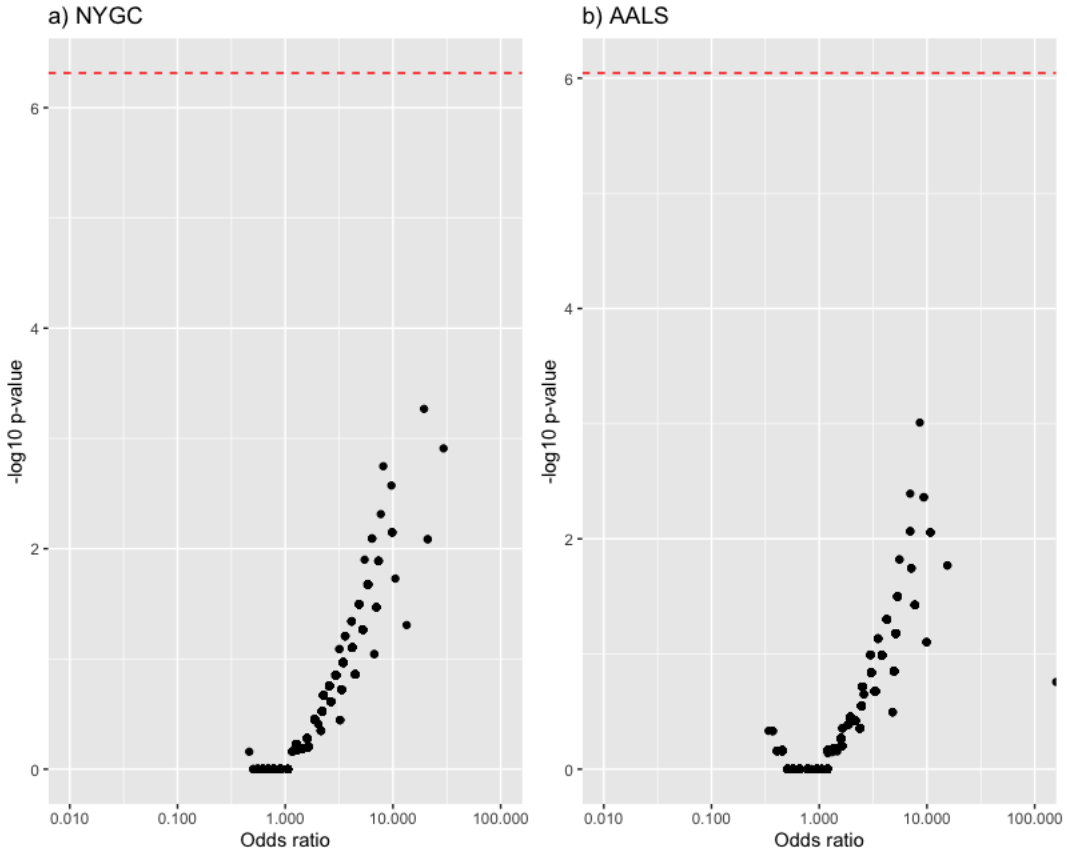

Figure 7: Odds ratios versus FDR-corrected p-values for the a) NYGC and b) AALS datasets. Red dashed line denotes threshold of p < 0.01 after FDR-correction.

Linking genetic variants and RNA stability

Firstly, as a common cause of ALS, it is worthwhile to investigate the correlation between C9orf72 repeat expansions and RNA stability estimates. Figure 8 indicates that samples with the C9orf72 expansion generally have lower RNA stabilities (R=0.18, p < 2.2e-16).

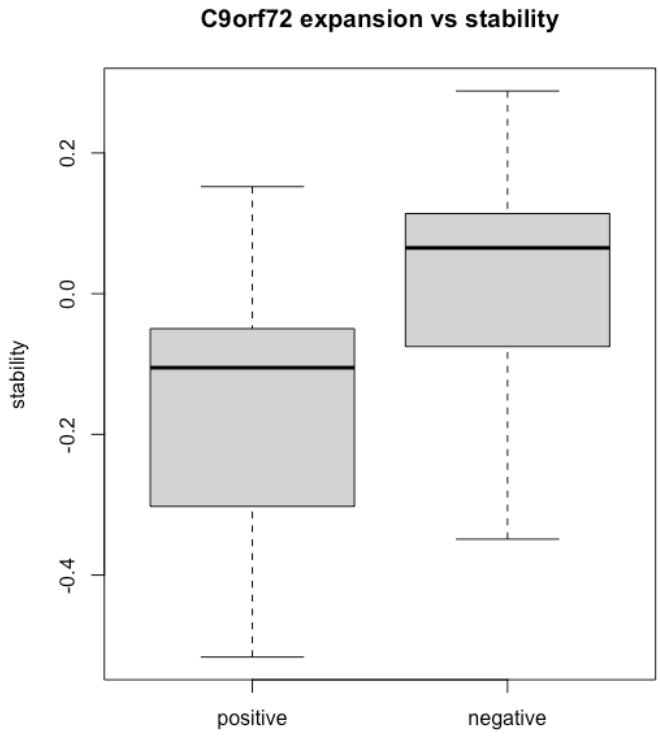

Figure 8: Samples with C9orf72 expansion generally have lower RNA stabilities ($p < 2.2e-16$). Box-and-whisker plot of RNA stabilities for samples with (positive) and without (negative) C9orf72 expansions.

Next, we related CADD scores and RNA stability through a gene burden test counting the number of mutations with a score of >15 per sample. This threshold was chosen because missense variants have a median CADD score of around 15 (Kircher et al., 2014). The test was performed in several different configurations: all mutations, 3' UTR mutations only, and 5' UTR mutations only. However, no significant links were found between CADD and RNA stability for any genes after FDR-correction.

To test for a relationship between ΔDC_scores and RNA stability, we used several summary statistics made up out of the scores of all samples in a gene. The first of these metrics was a gene burden test consisting of the number of mutations per sample that were among the top 10% most impacting for RBP binding. We then performed linear regressions to test how predictive this summary statistic was for RNA stability. The FGF12 and ABLIM3 genes were the only genes for which multiple RBPs returned p-values under 1 after FDR-correction (Table 2). Their combined p-values were 0.98 and 0.89, respectively. The lowest p-value, 0.04, is observed for the gene burden test of top 10% mutations for the ELAVL1 RBP predicting the RNA stability of ABLIM3. Largely, the number of top mutations in a sample is not a great predictor of RNA stability for any genes.

Table 2: P-values of linear models consisting of the number of mutations in a sample that are among the top 10% most impacting for RBP binding and RNA stability. The most significant result for each RBP was listed in this table.

Genes	AGO1-4	ALKBH5	CAPRIN1	ELAVL1	FUS	PUM2	TDP43	hnRNPC	RBFOX1	Combined
FGF12	0.12	1	1	0.28	0.53	1	1	1	1	0.98
ABLIM3	1	0.09	0.9	0.04	1	1	1	1	1	0.89
EPAS1	1	0.17	1	1	1	1	1	1	1	1

SCAF8	1	1	0.22	1	1	1	1	1	1	1
EPAS1	1	0.18	1	1	1	1	1	1	1	1
RAB3C	1	1	1	1	1	0.06	1	1	1	1
FAM219A	1	1	1	1	1	1	1	1	0.09	1

We also tested whether significance would increase if the sum of ΔDC_scores or the most impacting mutation were related to RNA stability. Additionally, we performed an ordinal regression by converting the stability values into three discrete categories consisting of z-scores smaller than -1, -1 to 1, and larger than 1. None of these changes to the method lead to increases in significance. Overall, these results indicate that CADD scores and ΔDC_scores generally lack predictive power for RNA stability in the current set-up.

Discussion

Using total RNAseq data to estimate RNA stability provided us with a novel perspective for studying ALS. The data indicated clear RNA stability outliers that could be studied further using statistical tests to identify links with ALS. These tests, namely the permutation test, logistic regression, odds ratio test, and gene set analysis, showed that while individual genes could generally not be linked to ALS based on stability outliers, several pathways could.
The general lack of correlation between gene RNA stability and ALS, even for known ALS genes, may signify that anomalous stability in single genes is not a prominent feature of ALS. However, it may also be explained by the relatively small size of the datasets. Given the size of the datasets (260 and 145 samples, respectively) and the complexity of the ALS phenotype, it is plausible that the study is underpowered with regards to the effect sizes of individual genes. The specific method of estimating RNA stability employed by Rembrandts may be another explanation for the lack of correlation. Indeed, Furlan et al. (2020) developed INSPEcT, a software tool to estimate RNA stability from total RNAseq, which used a different method for estimating RNA stability. This method was based on a global power law relationship between pre-mRNA and mature mRNA. Using INSPEcT, they did find a signature of brain genes involved in post-transcriptional regulation. Notably, they found that these genes were enriched in FUS and TARDBP binding motifs, for which loss of function has often been implicated with ALS (Da Cruz and Cleveland, 2011; Ishigaki and Sobue, 2018; Paez-Colasante et al., 2015).

A quantitative comparison of RNA stability estimations made by Rembrandts, INSPEcT and EISA (Gaidatzis et al., 2015) that was performed in the same study may explain these contrasting results. Comparing estimated RNA stabilities to empirically measured RNA degradation showed Rembrandts and EISA returning stability changes opposite to those measured. Conversely, INSPEcT did corroborate the experimental measurements. These contrasting results can be explained by the different assumptions underlying the RNA stability estimates. Degradation was artificially suppressed in the experiment providing the RNAseq data. The drug responsible for this suppressed degradation is thought to also have impacted the RNA processing machinery negatively. Given that Rembrandts assumes changes in processing rates opposite to changes in synthesis rates, these processing rates were now overestimated. By extension, the degradation rate is then overestimated and stability, which is inversely proportional to degradation, underestimated. INSPEcT does not distinguish between processing and synthesis rates, and is therefore not susceptible to this particular kind of error. Collectively, these findings show that the stability estimates made by

Rembrandts can be invalidated by factors affecting RNA processing machinery, such as genetic variants that affect splicing or mRNA export.

Performing the analyses described in this report using INSPEcT could help strengthen conclusions or help identify the limitations of Rembrandts in our study. Moreover, unlike Rembrandts, INSPEcT provides estimates of absolute stability values and these could be used to compare and rank genes which opens up new possibilities for analyses. Finally, the results could be corroborated or challenged through replicating the analysis using larger datasets, which is especially relevant here given the relatively small number of samples.

The gene set analyses indicated that RNA stability outliers may occur disproportionally in certain neuronal pathways. None of the individual genes in these pathways had a particularly high significance, but their aggregate effect is significantly different than the background set. These pathways, regulation of synaptic vesicle recycling and neuron projection regeneration pathways, are highly relevant to ALS. Highlighting the crucial role of vesicle function in ALS are findings that vesicle associated membrane protein B (VAPB; no stability estimate available) mRNA levels are suppressed in the spinal cords of ALS patients (Anagnostou et al., 2010) and that a missense mutation in this gene causes ALS (Nishimura et al., 2004). Impaired regeneration of neurons has also been implicated with ALS (Kang et al., 2013). Important to note is that the causation of the relationship between RNA stability outliers and these pathways is unclear. It is both possible that anomalous stabilities of the transcripts in this pathway cause ALS or that the neuron degeneration characteristic to ALS affects the stability of transcripts.

Furthermore, we studied whether RNA stability estimates could be related to WGS data. The finding that samples with C9orf72 expansions generally have lower RNA stability estimates highlights the potential RNA stability has for new discoveries in ALS. It is unclear how C9orf72 expansion cause neurodegeneration, with both loss-of-function and gain-of-function mechanisms being proposed (Rohrer et al., 2015). Research into the consequences of lower RNA stabilities for C9orf72 may help shed light on this issue.

Next, no significant links were established between CADD scores or ΔDC_scores and RNA stability. This lack of correlation between variants and RNA stability may be attributed simply to the limited size of the dataset, but also to the fact that RNA stability is affected by many factors, including environment, transcript attributes such as GC content and transcript length, miRNAs and RBPs. CADD scores inform primarily about the deleteriousness of a variant and its effects do not necessarily manifest in RNA stability. Similarly, disrupted RBP binding, to the extent that it is captured by ΔDC_scores, is only one factor affecting RNA stability. Extending the model to include both genetic and non-genetic factors may aid in establishing relationships between variants and RNA stability.

Additionally, ΔDC_scores may not accurately represent disrupted RBP binding due to several caveats in their calculation. First, DeepClip takes sequences as its only input, and therefore does not incorporate any structural information. However, structural changes in mRNA can affect RBP binding (Soemedi et al., 2017). Second, due to the sequence length limitation of DeepClip (max 75 bp), the long-range effects of variants are not captured in the ΔDC_scores. Additionally, computation and time constraints as well as the limited availability of RBP

models greatly limited the number of RBPs for which we could obtain results. Ideally, simulations would be run for the RBFOX1 and ZFP36 (no model available) RBP families, which were found to be significantly predictive of brain mRNA stability together with four other miRNAs by Alkallas et al. (2017). Notably, our results do not show RBFOX1 ΔDC_scores to be significantly predictive of mRNA stability. The incorporation of the ZFP36 RBP family as well as the miRNAs into the regulatory model should shed more light on if this failure to replicate is due to our specific method for using variants that affect RBP binding to predict RNA stability. Replication testing using different software tools for estimating RNA stability, such as INSPEcT, could provide great information into the reliability of the Rembrandts stability estimates.

The statistical power of models relating variants to RNA stability could be increased by incorporating the quantified effects of disrupted miRNA binding, for example by recording the number of matches of seed sequences in the 3' UTR (Alkallas et al., 2017). The incorporation of known stability confounders such as GC-content and transcript length (Duan et al., 2013) can also help to reduce noise. If relevant loci for ALS patient RNA stability are identified they can be replicated in larger whole genome sequencing datasets, including a larger dataset collected by Project MinE (Project MinE ALS Sequencing Consortium, 2018s) comprising over 10000 ALS patients and controls.

Overall, the strengths of this research include devising methodologies for the study of RNA stability in ALS patients and studying the genetic basis of RNA stability. These analyses investigating the genetics of RNA stability could have resulted in the identification of relevant loci for ALS, as well as other phenotypes affected by disrupted RNA stability. The methodologies described in this report may lead to such discoveries if they are improved in several aspects. This is highlighted by the GSEA and C9orf72 expansion analyses results which indicate that RNA stability may capture some of the signal for ALS. The results of this study do come with several caveats due to limited data availability and possible flaws in the RNA stability estimates. Additionally, the statistical models linking genetic variants and RNA stability were limited in scope, and the ΔDC_score estimates had several flaws, as described above. Ultimately, our findings highlight the importance of RNA stability in ALS and the many possible improvements in the methods applied here for studying it, the potential for new discoveries using RNA stability remains promising.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. Molecular Biology of the Cell, 4th ed. Garland Science.

Al-Chalabi, A., Fang, F., Hanby, M.F., Leigh, P.N., Shaw, C.E., Ye, W., Rijsdijk, F., 2010. An estimate of amyotrophic lateral sclerosis heritability using twin data. J. Neurol. Neurosurg. Psychiatry 81, 1324–1326. https://doi.org/10.1136/jnnp.2010.207464

Alkallas, R., Fish, L., Goodarzi, H., Najafabadi, H.S., 2017. Inference of RNA decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer's disease. Nat. Commun. 8, 909. https://doi.org/10.1038/s41467-017-00867-z

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., Feuk, L., 2011. Total RNA sequencing reveals nascent transcription and widespread co-

transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. 18, 1435–1440. https://doi.org/10.1038/nsmb.2143

Anagnostou, G., Akbar, M.T., Paul, P., Angelinetta, C., Steiner, T.J., de Belleroche, J., 2010. Vesicle associated membrane protein B (VAPB) is decreased in ALS spinal cord. Neurobiol. Aging 31, 969–985. https://doi.org/10.1016/j.neurobiolaging.2008.07.005

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169. https://doi.org/10.1093/bioinformatics/btu638

Araujo, P.R., Yoon, K., Ko, D., Smith, A.D., Qiao, M., Suresh, U., Burns, S.C., Penalva, L.O.F., 2012. Before It Gets Started: Regulating Translation at the 5′ UTR. Comp. Funct. Genomics 2012, e475731. https://doi.org/10.1155/2012/475731

Auwera, G.A.V. der, O'Connor, B.D., 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, Inc.

Barmada, S.J., 2015. Linking RNA Dysfunction and Neurodegeneration in Amyotrophic Lateral Sclerosis. Neurotherapeutics 12, 340–351. https://doi.org/10.1007/s13311-015-0340-3

Brown, R.H., Al-Chalabi, A., 2017. Amyotrophic Lateral Sclerosis. N. Engl. J. Med. 377, 162–172. https://doi.org/10.1056/NEJMra1603471

Buratti, E., 2008. Multiple roles of TDP-43 in gene expression, splicing regulation, and human disease. Front. Biosci. 13, 867. https://doi.org/10.2741/2727

Chia, R., Chiò, A., Traynor, B.J., 2018. Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. Lancet Neurol. 17, 94–102. https://doi.org/10.1016/S1474-4422(17)30401-5

Da Cruz, S., Cleveland, D.W., 2011. Understanding the role of TDP-43 and FUS/TLS in ALS and beyond. Curr. Opin. Neurobiol., Neurobiology of disease 21, 904–919. https://doi.org/10.1016/j.conb.2011.05.029

Donnelly, C.J., Zhang, P.-W., Pham, J.T., Haeusler, A.R., Mistry, N.A., Vidensky, S., Daley, E.L., Poth, E.M., Hoover, B., Fines, D.M., Maragakis, N., Tienari, P.J., Petrucelli, L., Traynor, B.J., Wang, J., Rigo, F., Bennett, C.F., Blackshaw, S., Sattler, R., Rothstein, J.D., 2013. RNA Toxicity from the ALS/FTD C9ORF72 Expansion Is Mitigated by Antisense Intervention. Neuron 80, 415–428. https://doi.org/10.1016/j.neuron.2013.10.015

Duan, J., Shi, J., Ge, X., Dölken, L., Moy, W., He, D., Shi, S., Sanders, A.R., Ross, J., Gejman, P.V., 2013. Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines. Sci. Rep. 3, 1318. https://doi.org/10.1038/srep01318

Furlan, M., de Pretis, S., Pelizzola, M., 2021. Dynamics of transcriptional and post-transcriptional regulation. Brief. Bioinform. 22. https://doi.org/10.1093/bib/bbaa389

Furlan, M., Galeota, E., Gaudio, N.D., Dassi, E., Caselle, M., de Pretis, S., Pelizzola, M., 2020. Genome-wide dynamics of RNA synthesis, processing, and degradation without RNA metabolic labeling. Genome Res. 30, 1492–1507. https://doi.org/10.1101/gr.260984.120

Gaidatzis, D., Burger, L., Florescu, M., Stadler, M.B., 2015. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nat. Biotechnol. 33, 722–729. https://doi.org/10.1038/nbt.3269

Ghasemi, M., Brown, R.H., 2018. Genetics of Amyotrophic Lateral Sclerosis. Cold Spring Harb. Perspect. Med. 8. https://doi.org/10.1101/cshperspect.a024125

Gregory, J.M., Fagegaltier, D., Phatnani, H., Harms, M.B., 2020. Genetics of Amyotrophic Lateral Sclerosis. Curr. Genet. Med. Rep. 8, 121–131. https://doi.org/10.1007/s40142-020-00194-8

Grønning, A.G.B., Doktor, T.K., Larsen, S.J., Petersen, U.S.S., Holm, L.L., Bruun, G.H., Hansen, M.B., Hartung, A.-M., Baumbach, J., Andresen, B.S., 2020. DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. Nucleic Acids Res. gkaa530. https://doi.org/10.1093/nar/gkaa530

Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E.M., Logroscino, G., Robberecht, W., Shaw, P.J., Simmons, Z., van den Berg, L.H., 2017. Amyotrophic lateral sclerosis. Nat. Rev. Dis. Primer 3, 1–19. https://doi.org/10.1038/nrdp.2017.71

Hedlund, E., Karlsson, M., Osborn, T., Ludwig, W., Isacson, O., 2010. Global gene expression profiling of somatic motor neuron populations with different vulnerability identify molecules and pathways of degeneration and protection. Brain 133, 2313–2330. https://doi.org/10.1093/brain/awq167

Ishigaki, S., Sobue, G., 2018. Importance of Functional Loss of FUS in FTLD/ALS. Front. Mol. Biosci. 5, 44. https://doi.org/10.3389/fmolb.2018.00044

Johnston, C.A., Stanton, B.R., Turner, M.R., Gray, R., Blunt, A.H.-M., Butt, D., Ampong, M.-A., Shaw, C.E., Leigh, P.N., Al-Chalabi, A., 2006. Amyotrophic lateral sclerosis in an urban setting. J. Neurol. 253, 1642–1643. https://doi.org/10.1007/s00415-006-0195-y

Kang, S.H., Li, Y., Fukaya, M., Lorenzini, I., Cleveland, D.W., Ostrow, L.W., Rothstein, J.D., Bergles, D.E., 2013. Degeneration and impaired regeneration of gray matter oligodendrocytes in amyotrophic lateral sclerosis. Nat. Neurosci. 16, 571–579. https://doi.org/10.1038/nn.3357

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., Shendure, J., 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315. https://doi.org/10.1038/ng.2892

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P., 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 1, 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Logroscino, G., Traynor, B.J., Hardiman, O., Chiò, A., Mitchell, D., Swingler, R.J., Millul, A., Benn, E., Beghi, E., Eurals, F., 2010. Incidence of amyotrophic lateral sclerosis in Europe. J. Neurol. Neurosurg. Psychiatry 81, 385–390. https://doi.org/10.1136/jnnp.2009.183525

Majounie, E., Renton, A.E., Mok, K., Dopper, E.G., Waite, A., Rollinson, S., Chiò, A., Restagno, G., Nicolaou, N., Simon-Sanchez, J., van Swieten, J.C., Abramzon, Y., Johnson, J.O., Sendtner, M., Pamphlett, R., Orrell, R.W., Mead, S., Sidle, K.C., Houlden, H., Rohrer, J.D., Morrison, K.E., Pall, H., Talbot, K., Ansorge, O., Hernandez, D.G., Arepalli, S., Sabatelli, M., Mora, G., Corbo, M., Giannini, F., Calvo, A., Englund, E., Borghero, G., Floris, G.L., Remes, A.M., Laaksovirta, H., McCluskey, L., Trojanowski, J.Q., Van Deerlin, V.M., Schellenberg, G.D., Nalls, M.A., Drory, V.E., Lu, C.-S., Yeh, T.-H., Ishiura, H., Takahashi, Y., Tsuji, S., Le Ber, I., Brice, A., Drepper, C., Williams, N., Kirby, J., Shaw, P., Hardy, J., Tienari, P.J., Heutink, P., Morris, H.R., Pickering-Brown, S., Traynor, B.J., 2012. Frequency of the C9orf72 hexanucleotide repeat expansion in patients with

amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. Lancet Neurol. 11, 323–330. https://doi.org/10.1016/S1474-4422(12)70043-1

Masrori, P., Damme, P.V., 2020. Amyotrophic lateral sclerosis: a clinical review. Eur. J. Neurol. 27, 1918–1929. https://doi.org/10.1111/ene.14393

Mayr, C., 2019. What Are 3' UTRs Doing? Cold Spring Harb. Perspect. Biol. 11, a034728. https://doi.org/10.1101/cshperspect.a034728

McKenzie, A.T., Wang, M., Hauberg, M.E., Fullard, J.F., Kozlenkov, A., Keenan, A., Hurd, Y.L., Dracheva, S., Casaccia, P., Roussos, P., Zhang, B., 2018. Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. Sci. Rep. 8, 8868. https://doi.org/10.1038/s41598-018-27293-5

Miller, T., Cudkowicz, M., Shaw, P.J., Andersen, P.M., Atassi, N., Bucelli, R.C., Genge, A., Glass, J., Ladha, S., Ludolph, A.L., Maragakis, N.J., McDermott, C.J., Pestronk, A., Ravits, J., Salachas, F., Trudell, R., Van Damme, P., Zinman, L., Bennett, C.F., Lane, R., Sandrock, A., Runz, H., Graham, D., Houshyar, H., McCampbell, A., Nestorov, I., Chang, I., McNeill, M., Fanning, L., Fradette, S., Ferguson, T.A., 2020. Phase 1–2 Trial of Antisense Oligonucleotide Tofersen for SOD1 ALS. N. Engl. J. Med. 383, 109–119. https://doi.org/10.1056/NEJMoa2003715

Nishimura, A.L., Mitne-Neto, M., Silva, H.C.A., Richieri-Costa, A., Middleton, S., Cascio, D., Kok, F., Oliveira, J.R.M., Gillingwater, T., Webb, J., Skehel, P., Zatz, M., 2004. A Mutation in the Vesicle-Trafficking Protein VAPB Causes Late-Onset Spinal Muscular Atrophy and Amyotrophic Lateral Sclerosis. Am. J. Hum. Genet. 75, 822–831. https://doi.org/10.1086/425287

Paez-Colasante, X., Figueroa-Romero, C., Sakowski, S.A., Goutman, S.A., Feldman, E.L., 2015. Amyotrophic lateral sclerosis: mechanisms and therapeutics in the epigenomic era. Nat. Rev. Neurol. 11, 266–279. https://doi.org/10.1038/nrneurol.2015.57

Project, M.A.S.C., van Rheenen, W., Pulit, S.L., Dekker, A.M., Al Khleifat, A., Brands, W.J., Iacoangeli, A., Kenna, K.P., Kavak, E., Kooyman, M., McLaughlin, R.L., Middelkoop, B., Moisse, M., Schellevis, R.D., Shatunov, A., Sproviero, W., Tazelaar, G.H.P., van der Spek, R.A.A., van Doormaal, P.T.C., van Eijk, K.R., van Vugt, J., Basak, A.N., Blair, I.P., Glass, J.D., Hardiman, O., Hide, W., Landers, J.E., Mora, J.S., Morrison, K.E., Newhouse, S., Robberecht, W., Shaw, C.E., Shaw, P.J., van Damme, P., van Es, M.A., Wray, N.R., Al-Chalabi, A., van Den Berg, L.H., Veldink, J.H., 2018. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. Eur. J. Hum. Genet. 26, 1537–1546.

Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.-J., Kher, M., Banks, E., Ames, D.C., English, A.C., Li, H., Xing, J., Zhang, Y., Matise, T., Abecasis, G.R., Salerno, W., Zody, M.C., Neale, B.M., Hall, I.M., 2018. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun. 9, 4038. https://doi.org/10.1038/s41467-018-06159-4

Rohrer, J.D., Isaacs, A.M., Mizielinska, S., Mead, S., Lashley, T., Wray, S., Sidle, K., Fratta, P., Orrell, R.W., Hardy, J., Holton, J., Revesz, T., Rossor, M.N., Warren, J.D., 2015. C9orf72 expansions in frontotemporal dementia and amyotrophic lateral sclerosis. Lancet Neurol. 14, 291–301. https://doi.org/10.1016/S1474-4422(14)70233-9

Rowland, L.P., Shneider, N.A., 2001. Amyotrophic Lateral Sclerosis. N. Engl. J. Med. 344, 1688–1700. https://doi.org/10.1056/NEJM200105313442207

Ryan, M., Heverin, M., Pender, N., McLaughlin, R., Hardiman, O., 2019. Heritability of ALS: A Population-based study over 24 years. (S54.001). Neurology 92.

Soemedi, R., Cygan, K.J., Rhine, C.L., Glidden, D.T., Taggart, A.J., Lin, C.-L., Fredericks, A.M., Fairbrother, W.G., 2017. The effects of structure on pre-mRNA processing and stability. Methods, Structural biology of the spliceosome 125, 36–44. https://doi.org/10.1016/j.ymeth.2017.06.001

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Talbott, E.O., Malek, A.M., Lacomis, D., 2016. Chapter 13 - The epidemiology of amyotrophic lateral sclerosis, in: Aminoff, M.J., Boller, F., Swaab, D.F. (Eds.), Handbook of Clinical Neurology, Neuroepidemiology. Elsevier, pp. 225–238. https://doi.org/10.1016/B978-0-12-802973-2.00013-6

Van Rheenen, W., Shatunov, A., Dekker, A.M., McLaughlin, R.L., Diekstra, F.P., Pulit, S.L., Van Der Spek, R.A., Võsa, U., De Jong, S., Robinson, M.R., 2016. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat. Genet. 48, 1043–1048.