Genomewide paired DNA-RNAseq analyses to discover intronic splice mutation hotspots in neurological disorders

Yan Wang[1], Charlotte van Dijk[1], Ilia Timpanaro[1], Paul Hop[2], Brendan Kenna[2], Maarten Kooyman[2], Eleonora Aronica[3], Jeroen Pasterkamp[1], Leonard van den Berg[2], Johnathan Cooper-Knock[4], Project MinE ALS sequencing consortium, NYGC ALS consortium, Jan Veldink[2], Kevin Kenna[1*]

1. Department of Translational Neuroscience, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht 3584 CG, The Netherlands
2. Department of Neurology, UMC Utrecht Brain Center, University Mediale Center Utrecht, Utrecht 3584 CX, The Netherlands
3. Department of (Neuro)Pathology, Amsterdam UMC, University of Amsterdam, Amsterdam, 1105 AZ Amsterdam, The Netherlands
4. Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield S10 2HQ, UK.
5. Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, New York 10032

Abstract

Paired DNA-RNA sequencing analyses enhance diagnostic yield in rare disorders by revealing pathogenic splice mutations that cannot be identified using DNAseq based methods alone. However, such analyses are typically restricted to known disease gene panels and focus on causal variants close to intron-exon boundaries. Here, we developed *SpliPath* (Splicing Pathology), a computational framework to conduct genomewide paired DNA-RNAseq analyses and empower discovery of splice mutation hotspots deeper into intronic sequences. We demonstrate the utility of SpliPath through analyses of CNS tissues in patients afflicted by amyotrophic lateral sclerosis (ALS). Using SpliPath we *1*) re-discover known ALS mutations, *2*) discover and experimentally verify a first of kind intronic mutation hotspot within a known ALS gene (*KIF5A*), *3*) nominate 17 candidate mutation hotspots beyond known ALS genes. Nominated mutation hotspots were supported by enrichment analyses of independent genome sequencing data and provide new evidence implicating *ELP3* and tRNA processing dysfunction in ALS.

Introduction

In theory, paired DNA-RNA sequencing offers a simple solution for the challenge of identifying rare genetic variants that cause pathogenic mRNA splicing defects. Moreover, in practice, there has been substantial success in deploying this approach for targeted investigations of known disease gene panels. Perhaps most famously, pioneering work in rare Mendelian muscle disorders demonstrated that incorporating the direct detection of aberrant mRNA splicing events in RNA-seq of muscle biopsies, improved the diagnostic yield of genetic testing by an impressive 35% [1]. However, concealed within the apparent simplicity of this approach are methodological challenges that limit the scope and scale with which traditional paired DNA-RNAseq analysis pipelines are applied. The first key challenge is the influence of technical and biological artefacts within short read RNA-seq data, which is especially challenging in complex tissues with high isoform diversity such the brain. The second key challenge is the ablility to link genetic variants to specific aberrant splicing events using traditional bioinformatic predictions of canonical splice site disruption or more modern deep learning based classifiers. These challenges might be manageable when investigations focus only on relatively small gene panels in small sample sizes, however they would compound quickly if investigations were to be extended to systematic genomewide analyses of larger cohorts. Moreover, even within the scope of known disease gene panels, analyses performed better for the detection of splice altering DNA variants close to known intron-exon boundaries. Deeper intronic splice variants are recognized as major contributors to disease phenotypes but still remain substantially more difficult to discover [2,3].

The potential gains of applying paired DNA-RNAseq splicing analyses at genomewide scale in large cohorts include the opportunity to nominate undiscovered mutation hotspots and novel pathogenic splicing phenotypes for disorders with unexplained missing heritability. This could be particularly impactful for disorders that exhibit high genetic heterogeneity and an architecture that is dominated by rare variants with incomplete penetrance. For these disorders, GWAS approaches to disease gene discovery are limited by a lack of statistical power to evaluate rare disease variants. Alternative rare variant burden testing (RVB) strategies can be very effective [4–6] but rely heavily on the appropriate selection of "qualifying variants" prior to testing. Optimizing this selection process can be challenging, but is critical as the inclusion of benign non-pathogenic will dilute the differences between cases and controls. This dilution effect masks RVB associations from *bona fide* disease variants, causing a loss of sensitivity for disease gene discovery. Conversely, excessive stringency in selecting qualifying variants will also attenuate power for disease gene discovery. This can occur where pathogenic DNA variants are not readily identifiable as being disruptive, such as is typically the case for deeper intronic splice variants. Genomewide paired DNA-RNAseq analyses could theoretically contribute to resolving these limitations when mRNA splicing defects are involved. For example, they could be used to nominate possible locations of relevant splice regulatory elements for validation in independent genetics data. We recently encountered an example to support this rationale in the context of whole exome sequencing (WXS) analyses of amyotrophic lateral sclerosis (ALS)[6]. In our study in 2018, we observed that 9 patients harbored one of 8 different rare variants, all of which were predicted to cause exon 27 mis-splicing of *KIF5A*. These variants also associated with a striking endophenotype characterized by early onset and long survival. Remarkably, there was no evidence of disease association for splice variants detected across other exons of KIF5A, most of which were not supported as *bona fide* splice mutations in subsequent bioinformatic analyses. There was also no evidence for signatures of disease association across any other genetic variants beyond exon 27. However, our analysis could also only evaluate putative splice variants in the immediate vicinity of the exon 27 junctions, leaving open the question as to whether additional patients are impacted by splice mutations in deeper intronic elements.

Here, we aim to evaluate the potential of integrative analysis of genetic data and paired mRNA splicing data to discover novel pathogenic events in genomewide analyses. We emphasized the discovery of deeper intronic splice altering variants, with the aim of identifying anchor points within introns that could be tested for additional deleterious variants in DNAseq only analyses. We focus on a cohort of 306 ALS patients and 86 controls with paired whole genome sequencing (WGS) and RNA-seq of multiple CNS regions (motor cortex, frontal cortex, cervical spinal cord, lumbar spinal cord). Notably, our study is specifically geared towards ultra-rare genetic variants, here the numbers of observed minor allele carriers

are too low to be assessed by the splicing quantitative trait loci (sQTL) frameworks deployed in GWAS and common variant studies. We identify numerous key challenges when extending paired DNA-RNAseq analyses from a known disease gene panel to a genomewide investigation. We highlight the risks for erroneous conclusions and provide QC, analysis and visualization solutions in our computational framework, SpliPath (Splicing Pathology), to mitigate key challenges. We also conduct follow up genetic analyses in an independent whole genome sequencing (WGS) dataset (6,652 ALS and 2,472 controls) to further evaluate putative intronic mutation hotspots and novel candidate genes for ALS.

Results

Unimodal RNA/DNA analyses show high numbers of candidate mRNA splicing defects for both ALS patients and neurotypical controls

The ALS transcriptomics dataset from New York Genome Center (NYGC) included a total of 1,103 RNA-seq profiles, each representing 1 of 4 brain or spinal cord regions within a cohort of 392 unique individuals (306 cases, 86 controls). Using RegTools [7] and our splice junction classification pipeline (Methods), we detected a total of 418,524 unique candidate novel splice junctions (Fig. 1). These events spanned 14,110 genes, and 41% of events were detected in just 1 of the 392 individuals. We observed no significant difference in the frequency of novel splice junctions in patients vs controls for any of the 4 tissues (frontal cortex: p=0.41; motor cortex: p=0.85; cervical cord: p=0.78; lumbar cord: p=0.72; T test), with each patient harboring a mean of 7,472 novel junctions and each control harboring a mean of 7,575 novel junctions (Fig. 1c). We also observed no significant case – control differences in the frequencies of more specific subcategories of novel splice events, namely the use of a novel acceptor site, the use of a novel donor site, novel exon skipping or the inclusion of novel (cryptic) exons (Fig. S2).

SpliceAI[8] and dbscSNV[9] were used to predict splice disrupting DNA variants within whole genome sequencing (WGS) data for the same 392 individuals. Each patient harbored a mean of 4,185 splice altering variants (SpliceAI score >= 0.2 or dbscSNV score >= 0.7, applying thresholds suggested previously [6]) in protein coding genes. Again, we observed no evidence for any gross systematic excess of candidate splice disrupting DNA variants in patients when compared to the baseline rate observed in controls (Fig. 1d).

Novel pathogenic splice variants of *KIF5A* and *TBK1* are readily discoverable through simple integrative analyses that are restricted to known disease genes

To discover which of the observed aberrant mRNA splicing events might be attributable to rare ALS-causing DNA variants, we next putatively assigned novel junctions inferred from the RNA-seq analyses to candidate splice disrupting DNA variants inferred from the DNAseq analyses. Specifically, we searched for instances where an observed novel splicing event could be linked to a candidate DNA variant within the same exon-intron-exon cluster in the same individual (Fig. S1). By integrating RNA and DNA sequencing data for 17 known ALS genes (Fig. 2a), two novel junction – variant pairs were identified that exhibited both moderate PSI and high splice-altering prediction score. The first variant was in a consensus splice site region for the *TBK1* gene (12:64496407G>C, dbscSNV score = 1.0, SpliceAI score = 1.0, n=1 patient, Fig. 2b). The second variant was 58bp away from the intron/exon boundary of exon 27 in *KIF5A* (12:57582544A>C, SpliceAI score = 0.64, n=1 patient, Fig. 2d). Using SVM-BPfinder [10] and BPP [11], we predicted that the *KIF5A* variant corresponded to a branchpoint disruption. Follow up analyses of the acceptor – donor site reconfigurations predictions made by SpliceAI revealed that for both variants, SpliceAI correctly predicted the exact *TBK1* exon 16 skipping and *KIF5A* exon 27 skipping that we observed in RNA-seq data (*TBK1* exon 16 skipping PSI = 0.33; *KIF5A* exon 27 skipping PSI = 0.36 in motor cortex sample). Combined these findings strongly suggested that the exon skipping events were driven by the two variants, respectively.

A key objective of our study was to evaluate whether candidate intronic variants identified by integrated DNA-RNAseq analyses, could serve as “anchor points” to localize hidden intronic splice elements that are disrupted in additional patient. Since the availability of paired DNA-RNAseq is limited, we also wanted to explore whether we could further evaluate candidate anchor points as *bona fide* mutation hotspots using DNAseq data alone. Accordingly, we screened a 10bp flanking region around the identified *KIF5A* branch point variant for additional variants in an independent DNAseq only dataset of 6,652 ALS patients and 2,472 controls included in Project MinE [12]. We discovered two patients that carried ultra-rare variants (n=1 carrier per variant; observed variants: 12:57582544A>C, 12:57582544A>G), both of which were predicted by SpliceAI to cause the same exon 27 skipping event observed for the branchpoint variant. To further test whether these variants indeed induced the predicted exon-skipping effect, we designed a mini-gene reporter assay based on a construct used in a recent massively parallel splice reporter assay (Fig. 2f) [13].

These constructs included an mScarlet reporter gene where the coding sequence was disrupted by the insertion of exon 27 of *KIF5A* plus the entirety of the flanking introns. Under wildtype conditions *KIF5A*-exon 27 will be retained after splicing and block the production of functional mScarlet protein, however ALS patient variants that induce skipping of exon 27 will attenuate this blocking effect and rescue fluorescent signals from the mScarlet protein (which are quantified relative to a control eGFP reporter). Our assay revealed no differences in exon skipping signals between cells expressing wildtype *KIF5A* sequence insertions and negative control DNA variants. However, exon skipping signals were significantly higher for positive control variants that we identified at the intron-exon boundaries in previous work [6], and for newly discovered intronic variants that we report in this study (Fig. 2e). Combined these observations reveal a key intronic branchpoint that was invisible to previous assessments of the *KIF5A* gene [6], and which can be evaluated in future genetic testing of ALS patients without the need for parallel RNA-seq analyses.

Enrichment for *bona fide* splice altering variants in genomewide analyses requires more rigorous integrative analyses of paired DNA-RNAseq data

Genomewide integration of paired WGS-RNAseq data of the 306 patients using the same workflow as deployed for the known ALS gene panel revealed 2,983,284 novel junctions putatively assigned to 1,863,099 rare DNA variants (Fig. S1a-d). We observed no significant global correlations between the strength of RNA and DNA based predictors of splice disruption across the novel junction – variant pairs (Fig. 3, S3-S5). To restrict the results to the highest confidence novel junction – variant pairs, we introduced a minimum PSI threshold that would robustly recover novel splice junctions induced by high impact rare variants such as the true positive events detected in our known genes analysis or the Cumming et al paper [1]. We also applied the same minimal SpliceAI / dbscSNV score thresholds as used in our unimodal analyses above. Filtering novel splice junctions based on PSI criteria will be sub-optimal in the event that aberrant transcripts are rapidly cleared by non-sense mediated decay (NMD) [14]. We therefore also computed z scores from corrected gene expression levels to identify expression outliers and possible instance of NMD (Fig. S6). Novel junction – variant pairings supported by at least one RNA-seq read and evidence for NMD in the same individual (z score≤-3), were retained for further analysis. Following these filters, there were on average 9 novel introns per patient in motor cortex samples, which included 4 novel donor junctions for a known exon, 3 novel acceptor junctions for a known exon and 2 exon skipping events. We have also inferred 2 cryptic exons per patient motor cortex sample. Notably these cryptic exons occur at much higher PSI in patient tissues than the cryptic exons introduced by TDP-43 pathology [15]. Our study was not designed to detect new TDP-43 dependent cryptic exons as bulk tissue RNAseq does not provide the sensitivity needed to do this accurately. For reliable splicing evidence, we required that each novel junction – variant pair passed the thresholds in at least two tissues' samples of an individual. As a result, 2,294 junction variant pairings involving 2063 (0.07%) of the starting set of novel junctions and 1790 (0.10%) of the starting set of rare variants were retained for further analysis. The genome-wide integrative analysis was also applied to WGS-RNAseq data of 86 non-neurological controls and resulted in 519 pairs that included 469 (0.06%) of the starting set of novel junctions and 429 (0.07%) of the starting set of rare variants.

For the novel junction – DNA variant pairs passing all of the above thresholds, we next compared computational predictions of the precise alterations in donor/acceptor splice site usage made using SpliceAI, with specific splice alterations observed empirically in the RNA-seq data. We observed a high number of discordant DNA-RNA based inferences, and accordingly classified each novel junction – variant pairing as being either a "full match", "partial match" or "mismatch" (Fig. S8). We observed that the SpliceAI scores for DNA variants in "mismatch" pairings were significantly lower than those of variants within the "full match" and "partial match" groups ($P < 2.2e-16$, T test). Moreover, the distances of the mismatch variants to any known annotated splice site were significantly longer than those of full and partial match variant sets ($P < 2.2e-16$, T test). These data support that the full and partial match variant pairs were likely of higher confidence. After excluding mismatch events, the set of nominated novel junction – variant pairings included 719 observations of 475 unique variants within the patient group and 153 observations of 133 unique variants within the control group (Table S3, S4). Most of these variants occurred within the splicing consensus regions (51%), but a large fraction occurred in deeper intronic

regions (39%, 0.3% within predicted branch points). Of the remaining variants, 8% occurred within protein coding sequences and 3% occurred within UTRs (Fig. 3f). The variants were linked to 793 observed novel junctions in the patient group and 163 observed novel junctions in the control group, of which 41% were classified as novel donor junctions, 32% were classified as novel acceptor junctions, and 27% were classified as exon skipping events. From these junctions, we inferred 47 putative cryptic exon inclusion events (Fig. 3e). We observed no significant GO term enrichments when comparing the genes nominated through analyses of ALS patients with a reference geneset prioritized through analyses of neurotypical controls (Table. S2). We also did not observe any over-representation for genes that were recently reported to exhibit differential splicing under TDP-43 loss of function conditions [15,16].

Secondary analysis of nominated splice mutation hotspots in independent whole genome sequencing data implicates *ELP3* and tRNA processing genes in ALS

We next investigated if independent ALS whole genome sequencing data could be used to corroborate candidate mutation hotspots nominated by our genomewide paired DNA-RNAseq analyses. As per our previous evaluation of the *KIF5A* intronic mutation hotspot, we defined candidate mutation hotspots as ± 10bp windows around each of the 475 candidate variants nominated through genomewide analyses of ALS patients. We also defined a set of "control regions" as ±10bp windows around each of the 133 potential splice-altering variants that were identified through genomewide analyses of controls. These regions were screened for additional ultra-rare variants in the Project MinE WGS dataset. We observed a tentative excess of ultra-rare variants in patients when all candidate hotspots were considered jointly (RVB test, Odds Ratio = 1.2994, P = 0.0459). To validate this finding, a follow-up permutation test (N = 1000) was conducted on joint ±10bp windows surrounding 475 randomly selected splice-altering variants. The permutation test yielded a P = 0.042, confirming the excess of ultra-rare variants in the mutation hotspots. We did not detect an excess of ALS patient variants when analyzing the 133 control regions (Odds Ratio = 0.9458, P = 0.6781). The size of the project MinE dataset limited statistical power to conduct more localized rare variant association testing analyses at the level of individual hotspots. We therefore next nominated 18 mutation hotspots across 18 ALS candidate genes where at least two ALS patients were observed to carry additional ultra-rare variants within the restrictive ±10 bp hotspot windows (Table 1). Among the nominated mutation hotspots, 11 occurred in consensus splice site regions, 7 occurred in deeper intronic regions. Notably, the intronic *KIF5A* mutation hotspot was still recovered, validating that our approach facilitated the recovery of *bona fide* disease genes even when applied in an unbiased genomewide investigation.

Competitive gene set enrichment analyses of the 18 nominated candidate genes with respect to the 14,110 genes in which any novel junction was detected revealed 6 over-represented pathways (p.adj < 0.05, Table 1). Strikingly 3 of the top enriched pathways related to tRNA processing (p.adj=1.45e-3), with the association being driven by rare aberrant splicing events in the genes *ELP3, RTCB* and *TRDMT1*. Of these genes *ELP3*, has been implicated in ALS before [17], but its relevance to ALS has remained controversial as no causal variant has been definitively fine mapped. In the paired DNA-RNAseq dataset, we observed 2 patients carrying a splice-altering variant (chr8:28110250G>A) in the nominated mutation hotspot. RNAseq data for these patients revealed a novel cryptic exon inclusion event that precisely corresponded to the consequences predicted for the assigned DNA variant using SpliceAI (chr8:28110085-28110251, Fig. 4). *In silico* translation of the altered mRNA sequence, indicated that the cryptic exon may lead to a truncated amino acid sequence and disruption of a SAM functional domain that has been reported as essential for a proposed as key for neuroprotective effects of wildtype *ELP3* [18]. Notably we observed no evidence for NMD in the carriers of *ELP3* variants. Finally, as part of an effort to define the potential boundaries of hidden intronic splice elements detected in our analyses, we conducted saturated *in silico* mutagenesis evaluations of 5 candidate ALS hotspots located >100bp from known intron-exon junctions (Fig. 4,S9). In the case of *ELP3* the distribution of SpliceAI prediction scores exhibited clear peaks that corresponded perfectly to the boundaries of the observed cryptic exon (Fig. 4). This implied the potential for a second mutation hotspot at the 3' end of the cryptic exon, however analyses of this locus in Project MinE did not reveal candidate variants in this region. For *RTCB* and *TRDMT1* we observed exon skipping events which

were linked to variants in splice consensus regions. Of the remaining candidate genes, *IQCE* was notable as being the only candidate to have also been associated with ALS relevant cryptic exon inclusion events in recent studies of TDP-43 loss of function models [15]. The gene has also been reported as causal for ciliopathy phenotypes reminiscent of those associated with known ALS genes *NEK1* and *CFAP410* [5,19].

SpliPath to facilitate discovery of pathogenic splice-altering variants in paired DNA-RNAseq analyses

The workflow described here can naturally be reused to conduct paired DNA-RNAseq analyses in the context of any disease phenotype and tissue. To support future use-cases, we therefore implemented the entirety of our workflow in a dedicated "SpliPath" R package (https://github.com/KennaLab/SpliPath)(Fig. S1,S10). In addition to functions to conduct all data processing steps, this package includes an R-shiny based visualization tool that enables easy browsing of the data on a local machine, or hosting of SpliPath results on interactive web interfaces (example browser: https://yanwang271.shinyapps.io/splipath_browser/)

Discussion

In this work, we demonstrated the feasibility of using genomewide paired DNA-RNAseq analyses to identify rare pathogenic splice altering mutations for a complex neurological disorder that exhibits high locus and allelic heterogeneity. This extends the scope of such analyses beyond their more conventional use in evaluating known disease gene panels. In particular, we demonstrated that the approach can be used to nominate hidden functional domains within non-coding intronic sequences that represent hotspots for disease mutations. We demonstrated that while paired DNA-RNAseq analyses of modest cohort sizes were required to nominate such hotspots initially, thereafter the detection of additional pathogenic variants could be achieved for at least some cases using WGS data alone. This has relevance for disorders such as ALS where key disease tissues are not accessible for transcriptomics analyses in living patients.

Our work revealed several key challenges to implementing this approach. By genome-wide screening of novel splice junctions in brain and spinal cord samples, we found thousands of novel candidate splice junctions per individual in both ALS and controls. We observed similar frequencies of novel donor sites, novel acceptor sites, novel exon skipping events and cryptic exon inclusion between ALS and control groups. Our work indicated that the vast majority of these events could not be corroborated as likely consequences of rare *cis*-acting DNA variants. Likewise, our screening of DNA variants predicted to be splice disrupting using SpliceAI, revealed millions of candidate events per individual. Again, the vast majority of these could not be linked to reliable evidence for mRNA mis-splicing or NMD in matched transcriptomics data. As well as highlighting the limitations of unimodal analyses, our study also demonstrated that simplistic integrative analyses that work well for small panels of known disease genes, breakdown when applied in a genomewide setting.

To mitigate these challenges and extend paired DNA-RNAseq analysis to genomewide identification of pathogenic mutation hotspots, we developed SpliPath, an R package to perform splice junction annotation, pairing of DNA variants and candidate splice junctions, nomination of splice-altering mutation hotspots, and interactive data visualization. Notably this framework is entirely distinct from the splicing quantitative trait loci (sQTL) used to associate common variants with splicing changes [20], as such methods cannot be applied to discover ultra-rare pathogenic variants in rare genetic diseases.

Using SpliPath, we identified two novel ALS variants that respectively drive *TBK1* exon 16 skipping and *KIF5A* exon 27 skipping. The novel variant identified in *TBK1* impacted a consensus splice site and would be readily discoverable through typical DNAseq only analyses. At the protein level, the *TBK1* exon 16 skipping event is expected to result in deletion of functional coiled-coiled domain 2, and is in-keeping with proposed loss of function models reported in previous studies [21,22]. The novel *KIF5A* variant impacted a putative branchpoint 58bp upstream of exon 27. We validated the exon-skipping potential of all SNVs at

this position using a minigene reporter assay, substantiating that this nucleotide is key in regulating exon 27 inclusion. We have previously demonstrated the selective pathogenicity of exon 27 skipping in prior work [6], but the intronic variants we describe here were not recognizable in prior investigations and this study represents the first discovery of an intronic mutation hotspot in ALS. The putative *KIF5A* branchpoint and *TBK1* splice site can be included in future genetic testing of ALS patients and our findings again emphasize the importance of (deep) intronic variants in disease pathogenicity studies.

By applying SpliPath genomewide, we prioritized 793 (0.03%) novel junctions and 719 (0.04%) splice-altering DNA variants from a starting dataset of 2.98 million novel junctions and 1.86 million DNA variants. We observed that this refined search space could be used to establish candidate mutation hotspots that exhibited an enrichment for rare ALS patient variants in independent WGS data and that nominated genes were enriched for biological processes that included tRNA processing, protein transport along microtubule, centriolar satellite, and spindle assembly. As discussed more extensively below, the nomination of *ELP3* and tRNA processing pathways through these analyses is especially notable. However, other nominated candidates will also be of interest for future research, most notably genes that have either been associated with TDP-43 dependent cryptic splicing events in ALS disease models (*IQCE*) or previously established ALS pathways such as microtubule transport (*SSX2IP*). Finally, it is worth highlighting that the *KIF5A* intronic hotspot identified during our known disease gene analyses remained discoverable amongst the 18 candidates nominated by our genomewide analyses. Collectively these observations provide proof of principle that demonstrates the validity of our methodological approach.

Multiple studies outside ALS research have already implicated disruptions of tRNA processing as primary disease mechanisms in phenotypes involving motor neuron degeneration. In Charcot-Marie-Tooth disease type 2 (CMT2D), mutations of 6 different tRNA synthetase genes cause a peripheral neuropathy that appears to be driven by aberrant sequestration of $tRNA^{Gly}$ [23]. Notably, compensation for this sequestration through overexpression of $tRNA^{Gly}$ provides phenotypic rescue in both Drosophila and mouse models of the disease. Similarly, distinct gene mutations impacting the tRNA splicing endonuclease complex have been associated with a severe motor phenotype in mice that was characterized by a progressive loss of spinal motor neurons and aberrant processing of tyrosine pre-tRNA [24]. Within ALS, tRNA disruptions are a less central hypothesis but have been implicated through studies of the genes *ELP3* [17,18] and *ANG* [25]. *ELP3*, which mediates tRNA acetyltransferase functions of the Elongator complex, was implicated as a potential ALS gene through a microsatellite based genomewide association study and parallel Drosophila mutagenesis screen [17]. The work suggested that risk alleles within the microsatellite study associated with reduced *ELP3* expression, and that this may compromise a neuroprotective function performed by *ELP3*. Subsequent work supported tRNA dependent neuroprotective effects for *ELP3* in zebrafish and murine models of multiple ALS genotypes, and also reported reductions of *ELP3* expression in the motor cortices of ALS patients [18]. However, causal risk alleles have never definitively been established for *ELP3* and thus the genetic link has yet been finalized. Accordingly, our study provides new evidence towards resolving controversies over the contributions of *ELP3* to genetic risk in ALS. Here, the nominated *ELP3* variant was linked to a cryptic exon inclusion event in multiple independent patients, and is predicted to cause deletion of an S-adenosyl-methionine (SAM) domain that is believed to form the elongator tRNA modification active site [26]. Beside these tRNA disruptions related to ALS, inhibited $tRNA^{Phe}$ function was recently suggested to contribute to *C9orf72* pathology[27]. Amongst the other tRNA processing genes nominated in this study, *TRDMT1* (previously named *DNMT2*) methylates tRNA and protects tRNA from stress induced cleavage by factors such as angiogenin (encoded by the gene *ANG*). *ANG* has also been implicated as a genetic risk factor in ALS and recent work hints that measures of its tRNA cleavage activity could provide a prognostic biomarker for disease progression in both animal models and ALS patients [25,28,29]. Finally *RTCB*, the third gene nominated within the tRNA processing pathway, has also been reported to protect against stress induced tRNA cleavage by angiogenin [30]. Intriguingly, *RTCB*, also directly modulates the abundance of *XBP1* protein-coding formation [31], and deletion of *XBP1* has been shown to enhance clearance of mutant *SOD1* aggregates in mouse models [32]. Collectively, our findings therefore provide new insights to corroborate the contributions of *ELP3* and aberrant tRNA processing in ALS genetic risk.

One limitation of our study is that in aiming to focus on the highest confidence splice altering genetic variants, true pathogenic mutation hotspots may have been missed. The sensitivity of the analysis was also constrained by the complexity, noise and lack of single cell resolution inherent to short read RNAseq profiling of bulk tissues. Furthermore, the number of samples within the paired DNA-RNAseq study cohort was modest compared to the scale of corresponding DNAseq only studies in ALS. Future datasets that increase the quality of the transcriptomics data through single cell resolution, long read sequencing or larger sample numbers will create new opportunities to discover missing hotspots and/or characterize identified hotspots more extensively. Large DNAseq only WGS datasets will also provide greater power to validate *bona fide* disease mutation hotspots. Of note for ALS, single cell analyses may also create the opportunity to explore TDP-43 dependent aberrant splicing events that are subject to genetic modifiers, such as the recently discovered *UNC13A* cryptic exon inclusion events [15,16]. Finally, our study incorporated popular bioinformatic tools for the prediction of DNA variant effects on splicing, but such tools are also under active development and future versions will undoubtedly also allow for additional performance gains in future research.

In conclusion, our work demonstrates the feasibility of using genomewide analyses of paired DNA-RNAseq data to guide the discovery of causal splice altering mutations and novel intronic mutation hotspots in a neurological disorder. Within ALS, our work implicates a novel intronic mutation hotspot of *KIF5A* for future genetic testing and indicates that rare splice altering variants of *ELP3* and other tRNA processing genes constitute genetic risk factors in ALS. We also provide the SpliPath R package to facilitate reuse of our workflow in future research and to support interactive data access and visualization of interpretable results.

Methods

RNAseq and DNAseq datasets

All paired DNAseq-RNAseq data were provided by the New York Genome Center (NYGC) ALS sequencing consortium (https://www.nygenome.org/als-consortium/). These data included full WGS for each donor, plus RNAseq of up to 4 brain and spinal cord regions (motor cortex, frontal cortex, cervical spinal cord, lumbar spinal cord). The cohort included 306 ALS / motor neuron disease (MND) spectrum patients as well as 86 neurotypical controls. Follow-up genetic analyses of novel candidate ALS mutation hotspots were conducted using an independent genetic dataset from the Project MinE consortium (data-freeze 2) [33]. The Project MinE data included WGS for 6,652 ALS patients and 2,472 controls.

Alignment and pre-processing of RNAseq and DNAseq datasets

All RNA-seq data was aligned to the GRCh38 reference genome using STAR v2.7.3a [34]. All transcript annotations were based on Ensembl version 98. To enhance capture of novel junctions in downstream analyses, STAR was used in two-pass mode (whereby novel junctions identified during first pass mapping are used to generate a refined splice junction annotation file for second pass mapping). All WGS data from the NYGC was aligned to the hg38 reference genome using BWA [35]. Variant calling and individual genotyping were performed in accordance with best practices outlined for the Genome Analysis Toolkit (GATK v4.1.0) [36]. Processing of the WGS dataset from project MinE is described van Rheenen et al [12]. Sample duplication events between project MinE and the NYGC dataset were identified by using KING to analyze genetic relatedness [37]. These relatedness analyses were conducted using a quality controlled set of common variant markers that were extracted from VCF files using plink v1.9 [38]. Quality control filters for the common variant markers included filtering for a set of biallelic LD pruned single nucleotide variants that exhibited a MAF>0.01, genotype call rate>0.9 and no evidence for deviation from Hardy Weinberg equilibrium (P<0.001).

Novel splice junction identification

RegTools v0.5.2 [7] was used to define candidate splice junctions based on the detection of split reads within the aligned RNAseq bam files. Split reads were only considered in the event that they mapped to the reference genome with a minimum overlap of 6bp at both ends. Candidate splice junctions were only considered during downstream analyses if supported by a minimum of 5 split reads in at least one sample. Observed splice sites were classified as being “known” or “novel” by comparison to reference splice junction coordinates retrieved from Ensembl 98 and the Snaptron database [39] Novel candidate splice junctions were classified as being either novel exon skipping junction, novel donor junction, or novel acceptor junction (Fig. S1d). For designation as a novel exon skipping, we required that the data supported skipping of annotated exons without creating novel donor/acceptor sites. For designation as a novel donor/acceptor junction, we required that an observed novel donor/acceptor site was paired with a known annotated splice site. Where a novel acceptor junction and a novel donor junction were observed within the same intron, we inferred putative cryptic exon inclusion. More complex splice aberrations involving multiple genes or multiple novel splice sites were regarded as low confidence events and excluded from our analyses in this study. Candidate junctions occurring within ENCODE blacklist regions were also excluded [40].

Prediction of splice-altering consequences for DNA variants

We annotated all observed DNA variants with predictions scores from dbscSNV [9] and SpliceAI[8]. The dbscSNV score was defined as the maximum of an AdaBoost and a random forest prediction score. With SpliceAI, we scored the predicted effect of every observed minor allele on every potential splice site located within 500bp of the variant position. The final SpliceAI score was taken as the maximum delta score predicted for any given donor gain, donor loss, acceptor gain, and acceptor loss event. DNA variants were only considered as potentially splice altering in the event of a dbscSNV score >= 0.7 or a SpliceAI score >= 0.2.

Nomination of splice altering DNA variants and candidate ALS mutation hotspots

For each novel splice junction identified through RNA-seq analyses, the genomes of the corresponding tissue donors were screened for DNA variants that could feasibly represent causal *cis*-acting splice variants (detailed Fig. S1). For any given novel splice junction, we restricted our DNA variant search space to include the entirety of the disrupted intron plus both flanking exons. All novel junction events were also required to exhibit either a tissue PSI ≥ 0.2, a supporting tissue read count ≥ 20 or an NMD z-score ≤ -3 (see below). We also required that observed novel junctions were reproducible across RNAseq profiles of at least 2 tissue types from the same donor. For the assignment of novel junctions to DNA variants predicted to be splice altering using SpliceAI, we required "matching" of the exact novel junction events detected from RNAseq analyses with the expected novel junction events predicted from SpliceAI analyses (detailed Fig. S7). dbscSNV does not make specific event predictions and only annotates DNA variants within highly restricted splice consensus regions (−3 to +8 bp at the donor site and −12 to +2 bp at the acceptor site), accordingly no stringent event matching criteria were imposed. Candidate mutation hotspots nominated for secondary analyses in project MinE WGS included ±10bp flanking windows around ultra-rare splice altering variants (MAF < 0.01% in gnomAD v3.1.2[41]). Follow up analyses of nominated mutation hotspots in project MinE WGS data also focused exclusively on ultra-rare variants (singleton events with MAF < 0.01% in gnomAD v3.1.2).

Nomination of candidate NMD events

To identify instances where DNA variants induce splice altering events that are rapidly degraded through NMD, we generated gene expression z scores for each gene in each tissue in each donor. These z-scores were derived from read count data prepared using htseq-count [42] . For each tissue from each donor, the expression of protein-coding genes were normalized to number of transcripts per million mapped reads (TPM). Adjusted TPMs were generated by using linear regression to correct for sequencing platform, RIN, sex, percentage European ancestry as well as 30 PEER factors computed using the PEER R package [43]. The residuals from these regressions at each gene were then transformed to z scores. Candidate NMD events were required to exhibit a z score ≤ -3.

Association testing

RVB test were performed using the Rare Variant Analysis Toolkit (RVAT) (https://github.com/KennaLab/rvat, manuscript in preparation). Firth penalized logisitic regression was used to model case-control status with respect to rare variant burden while correcting for sex, sequencing platform and the first 4 principal components derived from PCA of common variant profiles. To confirm the specificity of association at the nominated mutation hotspots, permutation tests (N=1,000 permutations) were performed on ±10bp windows around randomly selected 475 splice-altering rare variants genomewide. Geneset enrichment analyses were performed using logistic regression and ontology terms downloaded from MSigDB C5 [44]. For the geneset enrichment analyses of the 18 nominated mutation hotspots, the background geneset was defined as the 14,110 genes where at least one novel junction event could be detected in the NYGC RNAseq data. The p values for all geneset enrichment analyses were adjusted for multiple testing using *Benjamini–Hochberg* false discovery rate method.

SpliPath

The SpliPath R package was developed to perform the analyses described in this study including splice junction annotation, integrative RNA splice junction and DNA variants analysis and the mutation hotspots nomination as described in the Methods. For reviewing the nominated mutation hotspots with splicing evidence, or visually nominating the mutation hotspots, a Shiny based interactive data browser was implemented within SpliPath R package (example browser: https://yanwang271.shinyapps.io/splipath_browser/). With SpliPath data browser (Fig. S10), users can find all the possible splice-altering variants in genes of interest, view novel junctions possibly caused by the variants, and putatively link variants with novel junctions by viewing junction's expression in variants carriers compared to that in other individuals. SpliPath R package can be installed from https://github.com/KennaLab/SpliPath. The *make_sashimi_like_plot* and *make_gene_wise_plot* function in *SpliPath* were adapted from *make_differential_splicing_plot* function in LeafCutter [20]. The analyses in this study were performed using R 3.6.

KIF5A minigene reporter assay

The *KIF5A* mini-gene reporter construct (Fig. S11) was based on the design described by Cheung et al [13].Vector assembly was performed using the Gibson Assembly kit from NEB (Ca. No. E5510S, New England BioLabs)[45] and primers 1-12 (Table S5). Sequences were obtained from the plasmids pGEMT-PT2A-GFP-Tdtomato-iRFP670 for the bacteria backbone [46] and an in-house plasmid containing the EF1alpha-mScarlet and polyA tail sequences. The *KIF5A* minigene sequence was extracted from genomic DNA derived from SH-SY5Y cells. ALS patient and control mutations (Table S1) were introduced using the site directed mutagenesis kit from NEB and primers 13-21 (Table S5) (Ca. No. E0554S, New England BioLabs) [47]. Construct sequences were validated by Sanger sequencing. SH-SY5Y cells were cultured in a 24 well plate at a density of 100.000 cells and 1 ml of medium (DMEM with GlutaMAX, 10% FBS and 0.5% Penicillin-Streptomycin). Cells were transfected with 100 ng of plasmid using the Lipofectamine Transfection Reagent (Ca. No. L3000001, ThermoFisher) as per manufactur instructions. After 48 hours of incubation, cells were harvested and stained with DAPI to select live cells. Fluorescence was measured on a BD FACSCanto II Flow Cytometer. Gates were set to select living and single cells using untransfected cells. A compensation protocol was set up using cells transfected with in house plasmids containing eGFP or mScarlet only. The data was analyzed using the FlowJo software. mScarlet signals were obtained from eGFP positive cells to ensure transfection. The exon skipping signal was defined as the median mScarlet signal divided over the median mScarlet signal of the wildtype plasmid. One-way ANOVA followed by T tests (adjusted for multiple testing using *Benjamini–Hochberg* false discovery rate method) were performed to compare the exon skipping signals of wildtype insertions to that of negative control, positive control, and target variants insertions.

Acknowledgements

We are grateful to the patients with ALS and their families allowing donation of tissue for research. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 772376 - EScORIAL. The collaboration project is co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. K.K. is supported by grants from the Dutch Research Council (grant no. ZonMW-VIDI 91719350) and the ALS Foundation Netherlands. K.K., E.A., J.V. are supported by ALS Stichting (grant " ALS Tissue Bank – NL "). All NYGC ALS Consortium activities are supported by the ALS Association (ALSA, 19-SI-459) and the TOW Foundation.

Consortia

MinE ALS sequencing consortium

Philip van Damme, Philippe Corcia, Philippe Couratier, Patrick Vourc'h, Orla Hardiman, Russell McLaughin, Marc Gotkine, Vivian Drory, Nicola Ticozzi, Vincenzo Silani, Jan H. Veldink, Leonard H. van den Berg, Mamede de Carvalho, Jesus S. Mora Pardina, Monica Povedano, Peter Andersen, Markus Weber, Nazli A. Başak, Ammar Al-Chalabi, Chris Shaw, Pamela J. Shaw, Karen E. Morrison, John E. Landers, Jonathan D. Glass, Clifton L. Dalgard

NYGC ALS consortium

Hemali Phatnani, Justin Kwan, Dhruv Sareen, James R. Broach, Zachary Simmons, Ximena Arcila-Londono, Edward B. Lee, Vivianna M. Van Deerlin, Neil A. Shneider, Ernest Fraenkel, Lyle W. Ostrow, Frank Baas, Noah Zaitlen, James D. Berry, Andrea Malaspina, Pietro Fratta, Gregory A. Cox, Leslie M. Thompson, Steve Finkbeiner, Efthimios Dardiotis, Timothy M. Miller, Siddharthan Chandran, Suvankar Pal, Eran Hornstein, Daniel J. MacGowan, Terry Heiman-Patterson, Molly G. Hammell, Nikolaos. A. Patsopoulos, Oleg Butovsky, Joshua Dubnau, Avindra Nath, Robert Bowser, Matthew Harms, Eleonora Aronica, Mary Poss, Jennifer Phillips-Cremins, John Crary, Nazem Atassi, Dale J. Lange, Darius J. Adams, Leonidas Stefanis, Marc Gotkine, Robert H. Baloh, Suma Babu, Towfique Raj, Sabrina Paganoni, Ophir Shalem, Colin Smith, Bin Zhang, Brent Harris, Iris Broce, Vivian Drory, John Ravits, Corey McMillan, Vilas Menon, Lani Wu, Steven Altschuler,

Yossef Lerner, Rita Sattler, Kendall Van Keuren-Jensen, Orit Rozenblatt-Rosen, Kerstin Lindblad-Toh, Katharine Nicholson, Peter Gregersen, Jeong-Ho Lee M.D, Maze Therapeutics, Bristol-Myers Squibb, Sulev Koks, Stephen Muljo, Bryan J. Traynor, Pfizer, Regeneron, Insitro

Figures

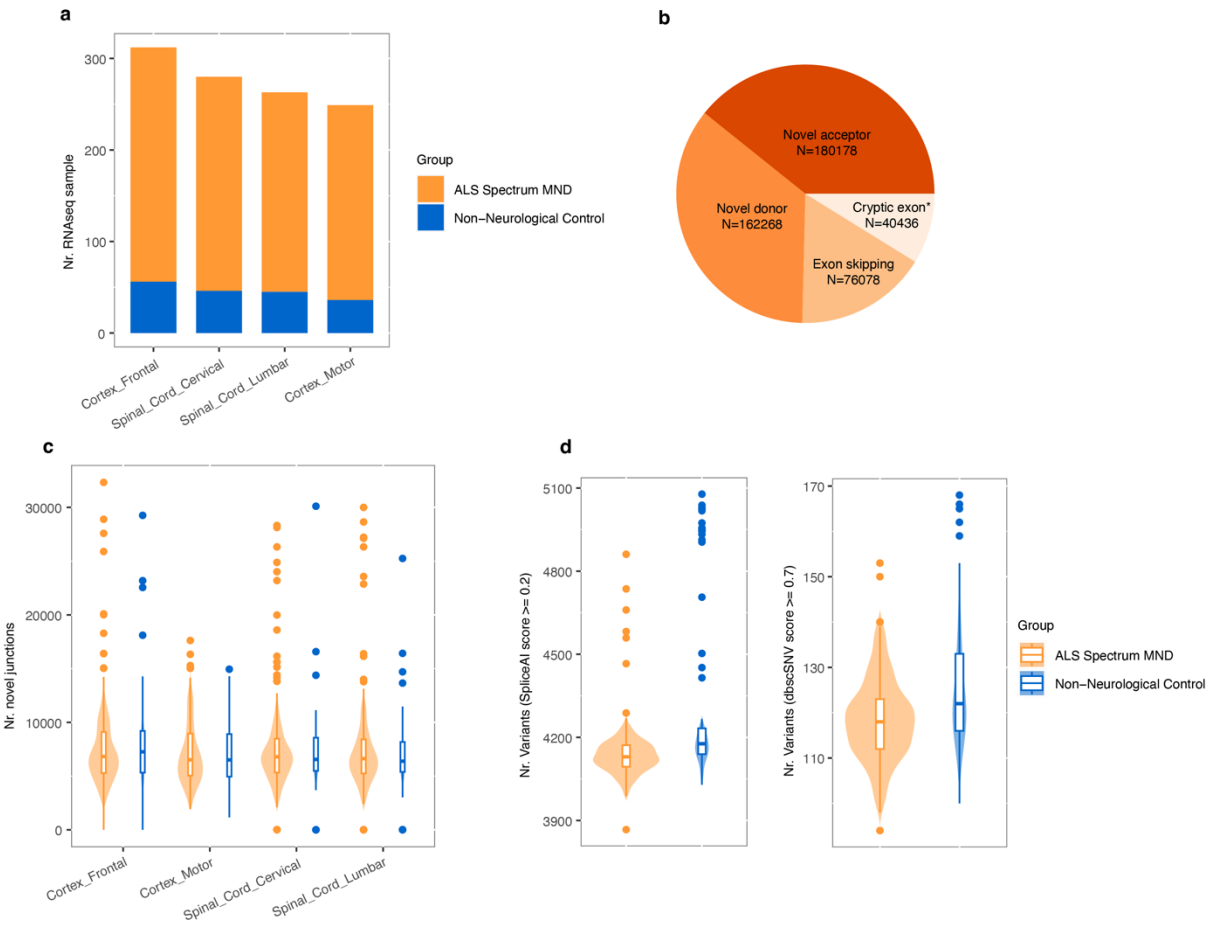

Figure 1. Thousands of novel RNA splice junctions and millions of splice-altering DNA variants are nominated through unimodal analyses in both ALS patients and neurotypical controls.
a) We investigated 1,103 RNAseq samples of brain and spinal cord from 392 individuals in NYGC ALS cohort. b) Four types of novel splicing events were observed, with exon skipping and cryptic exon inclusion events being the rarer than the use of novel donor/acceptor sites. *Cryptic exon inclusion events were putatively inferred from observation of a novel acceptor junction and a novel donor junction within the same intron. c) Number of novel splice junctions (supported by at least 5 split-reads in a sample) inferred from RNAseq of x4 tissue types in patients and controls. d) Number of splice altering variants predicted by SpliceAI and dbscSNV for patients and controls. SpliceAI applies to both exonic and intronic genetic variants[8]; while dbscSNV applies to single nucleotide variants (SNVs) within splicing consensus regions: −3 to +8 bp at the donor site and −12 to +2 bp at the acceptor site [9].

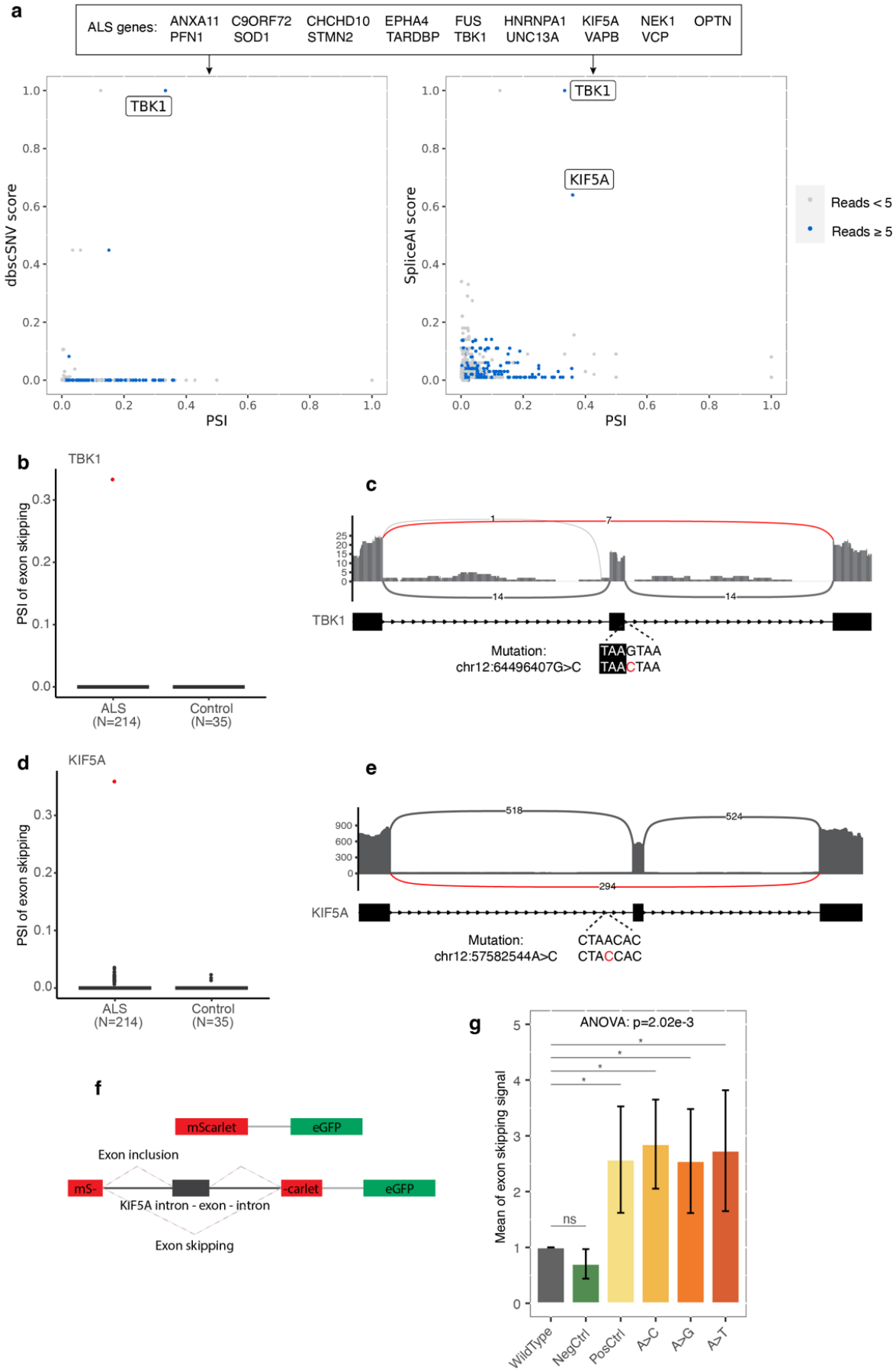
a
ALS genes: ANXA11 PFN1 C9ORF72 SOD1 CHCHD10 STMN2 EPHA4 TARDBP FUS TBK1 HNRNPA1 UNC13A KIF5A VAPB NEK1 VCP OPTN
TBK1
dbscSNV score
PSI
TBK1
KIF5A
SpliceAI score
PSI
Reads < 5
Reads ≥ 5
b
TBK1
PSI of exon skipping
ALS (N=214)
Control (N=35)
c
TBK1
Mutation: chr12:64496407G>C
TAAGTAA
TAACTAA
d
KIF5A
PSI of exon skipping
ALS (N=214)
Control (N=35)
e
KIF5A
Mutation: chr12:57582544A>C
CTAACAC
CTACCAC
f
mScarlet
eGFP
Exon inclusion
mS-
-carlet
eGFP
KIF5A intron - exon - intron
Exon skipping
g
ANOVA: p=2.02e-3
Mean of exon skipping signal
ns
WildType
NegCtrl
PosCtrl
A>C
A>G
A>T

Figure 2. Integrative analysis of novel splice junctions and DNA variants reveal pathogenic splice-altering variants.

a) Integrating novel splice junctions with DNA variants from same individuals in 17 known ALS genes in motor cortex samples. Each dot is a novel junction – variant pair in a patient's motor cortex RNAseq profile, where the variant was putatively linked to the novel junction based on their positions in gene (See Fig. S1). The x axis is PSI of novel junction and y axes are SpliceAI score and dbscSNV scores for a variant. Two novel junction – variant pairs in *TBK1* and *KIF5A* of two patients that exhibit moderate PSI signals and high splice-altering scores are highlighted. The novel exon skipping events in the two patients' RNAseq profile and the linked splice-altering variants are shown in c) and e). The box plots b) and d) show the PSI of the *TBK1* and *KIF5A* exon skipping in all individuals, and the red dots refer to those of the two variants carriers. f) Design of mini-gene reporter to validate the exon skipping effects of *KIF5A* variants discovered in DNAseq only data. The plasmid contained an eGFP and a split mScarlet in which the *KIF5A* intron 26 - exon 27 - intron 27 sequences with(out) variants were inserted (See design of whole plasmid in Fig. S11) g), Bar plot of exon skipping signals of mini-gene reporter assays. The negative control variant was from gnomAD database with allele count = 5 in non-neurological disease group, and the positive control variant was a splice consensus region variant that we reported previously[6](Table S1). Quantification of exon skipping was based on normalized median mScarlet signal over 3 technical replicates and the experiment was repeated four times. There were no differences of exon skipping signals between cells with wildtype *KIF5A* sequence insertions and negative control sequences insertions ("ns"), while exon skipping signals of cells with positive control and target sequences insertions were significantly higher than that of cells with wildtype *KIF5A* sequence insertions ("*", $P.adj < 0.05$) (T test, FDR adjusted).

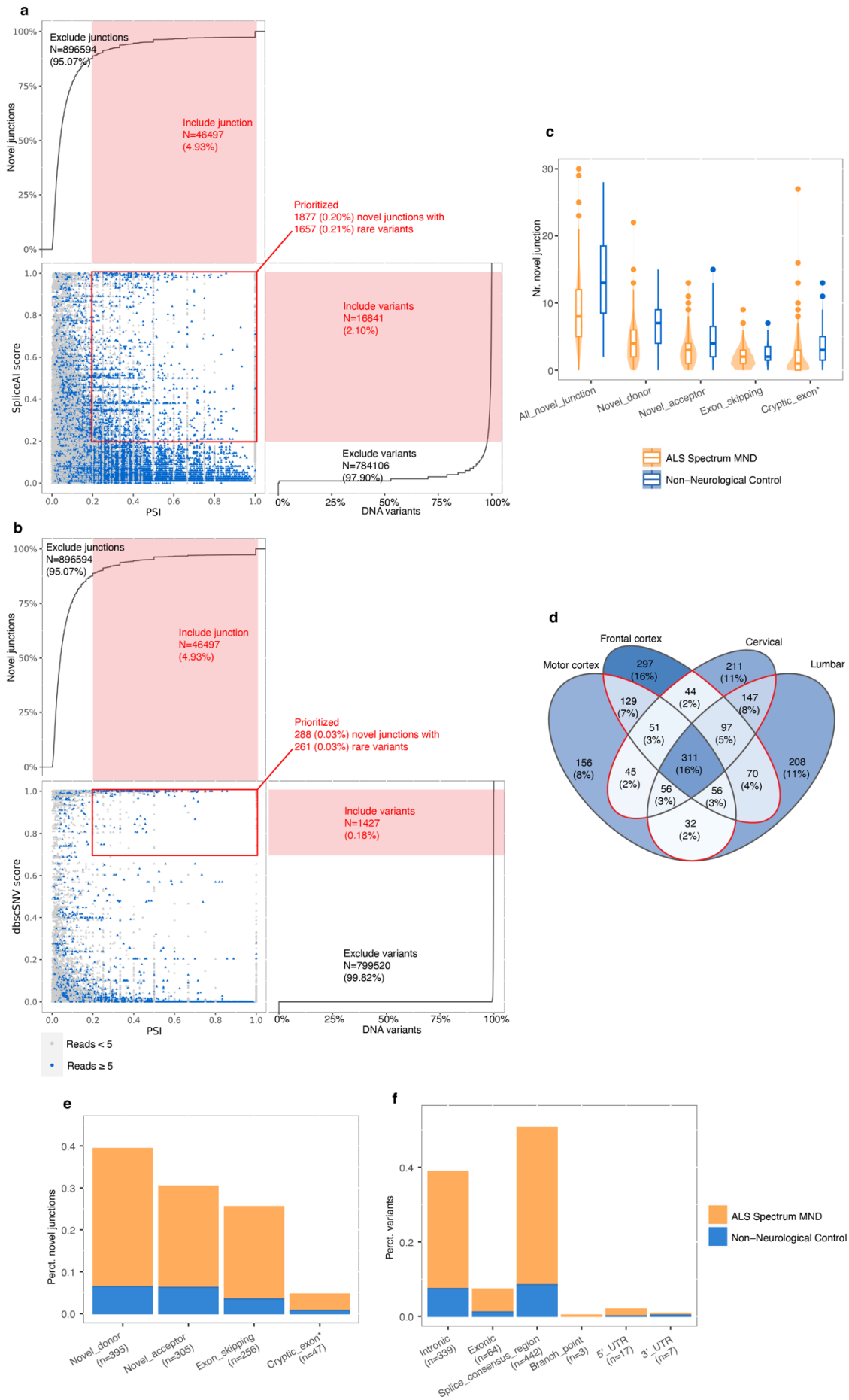
a
Exclude junctions
N=896594
(95.07%)
Include junction
N=46497
(4.93%)
Novel junctions
Prioritized
1877 (0.20%) novel junctions with
1657 (0.21%) rare variants
Include variants
N=16841
(2.10%)
Exclude variants
N=784106
(97.90%)
SpliceAI score
PSI
DNA variants
b
Exclude junctions
N=896594
(95.07%)
Include junction
N=46497
(4.93%)
Novel junctions
Prioritized
288 (0.03%) novel junctions with
261 (0.03%) rare variants
Include variants
N=1427
(0.18%)
Exclude variants
N=799520
(99.82%)
dbscSNV score
PSI
DNA variants
Reads < 5
Reads ≥ 5
c
Nr. novel junction
All_novel_junction
Novel_donor
Novel_acceptor
Exon_skipping
Cryptic_exon*
ALS Spectrum MND
Non−Neurological Control
d
Frontal cortex
Cervical
Motor cortex
Lumbar
297 (16%)
211 (11%)
44 (2%)
129 (7%)
147 (8%)
51 (3%)
97 (5%)
311 (16%)
156 (8%)
208 (11%)
45 (2%)
70 (4%)
56 (3%)
56 (3%)
32 (2%)
e
Perct. novel junctions
Novel_donor (n=395)
Novel_acceptor (n=305)
Exon_skipping (n=256)
Cryptic_exon* (n=47)
f
Perct. variants
Intronic (n=339)
Exonic (n=64)
Splice_consensus_region (n=442)
Branch_point (n=3)
5' UTR (n=17)
3' UTR (n=7)
ALS Spectrum MND
Non−Neurological Control

Figure 3. Genome-wide integrative analysis of variants and novel splice junction yields 872 putative splice-altering variants for further investigation.

a), b) Genomewide integrative analysis of novel splice junctions with candidate splice altering DNA variants from 306 patients. Each dot is a variant – novel junction pair for a single patient, where the variant was putatively assigned to the novel junction based on genomic proximity (Fig. S1, pairings that included common DNA variants with MAF>0.05 were excluded). The x axis depicts PSI for a novel junction in the motor cortex dataset, while y axes depicts SpliceAI and dbscSNV scores for the variant. The results from the other tissues types provided in Fig. S3-S5. The red rectangles define the thresholds of novel junction PSI, SpliceAI score, and dbscSNV score. The cumulative density plots show the distribution of PSI scores for all novel junctions in the dataset, and the distribution of splice-altering scores for all DNA variants in the dataset. Filtering on these PSI thresholds as well as a minimal read count ≥ 5 reduced the search space to 0.20% of the novel junctions and 0.21% of the variants when SpliceAI was used to annotate splice-altering variants, and 0.03% of the novel junctions and 0.03% of the variants when dbscSNV was used to annotate splice-altering variants. c) The number of novel junctions remaining per individual in case and control motor cortex samples after applying the thresholds. d) The concordance of variant – novel junction pairs that passed the thresholds across multiple tissue types from the same donor. The numbers in the Venn plot show the number of genes that harbor these novel junction – variant pairs. For reliable splicing evidence. The pairs that passed the thresholds in at least two tissues were prioritized for further analysis (within red border lines). We also excluded novel junction – variant pairs that failed stringent SpliceAI prediction vs observed junction matching criteria (Fig. S7, S8). The retained set of higher confidence events involved e) 956 novel junctions and f) 872 splice-altering variants.

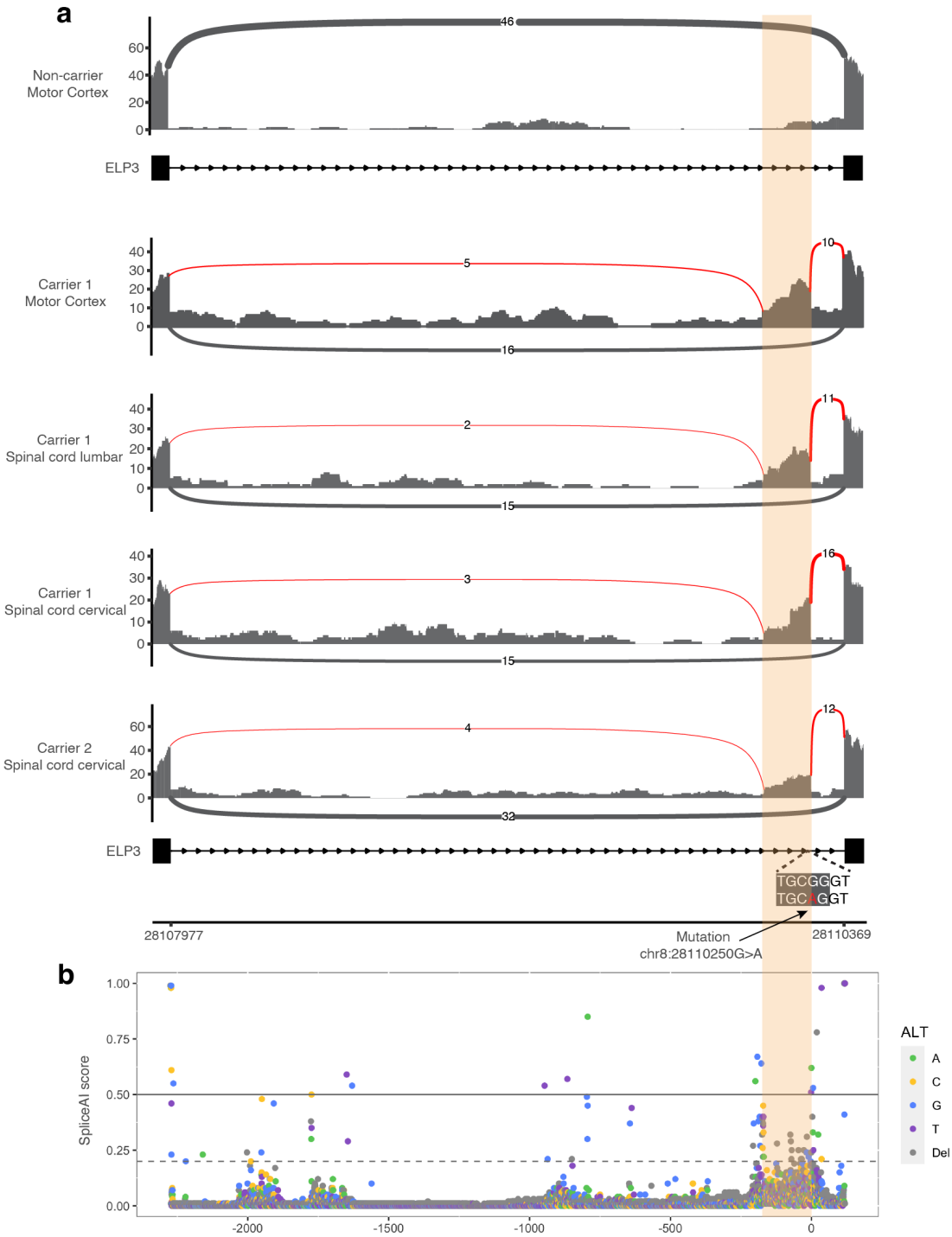

Figure 4. *ELP3* cryptic exon inclusion events caused by a recurrent intronic splice-altering variant in ALS patients.

a) *ELP3* cryptic exons (marked in orange) were observed consistently across tissue types within both chr8:28110250G>A carriers. These events were absent in patients not carrying chr8:28110250G>A. SpliceAI predicted the exact cryptic exon boundaries as were observed. b) *In silico* mutagenesis of the *ELP3* intron 6. X axis is the relative genomic position to chr8:28110250G>A. Y axis is the SpliceAI score of the mutations. The dots were colored by the mutated alternative allele. As expected, SpliceAI scores of mutations at the boundaries of two known exons were nearly 1 (dots stacked upon one another). SpliceAI predicted multiple mutations of moderate splice-altering scores around the boundaries of the cryptic exons.

Table 1. Candidate pathogenic mutation hotspots identified from independent Project MinE WGS dataset (hg19).

Gene	Hotspot (hg19)	Nr. Ultra rare variant (Nr. Patients)	Position
IQCE	7:2632751-2632771	4(4)	splicing consensus region*
INCENP	11:61914120-61914140	3(3)	splicing consensus region
SNX22	15:64445212-64445232	3(3)	intronic
SSX2IP	1:85155719-85155739	3(3)	intronic
CLMN	14:95677207-95677227	2(2)	splicing consensus region
COA6	1:234509965-234509985	2(2)	splicing consensus region
CRB2	9:126134611-126134631	2(2)	splicing consensus region
ELP3	8:27967757-27967777	2(2)	intronic
KIF5A	12:57976317-57976337	2(2)	intronic
MROH7	1:55119746-55119766	2(2)	intronic
NEK6	9:127089714-127089734	2(2)	splicing consensus region
PCSK7	11:117101666-117101686	2(2)	intronic
PLEKHA8	7:30092314-30092334	2(2)	splicing consensus region
RTCB	22:32789869-32789889	2(2)	splicing consensus region
TEP1	14:20845642-20845662	2(2)	splicing consensus region
TRDMT1	10:17216535-17216555	2(2)	splicing consensus region
USP40	2:234460158-234460178	2(2)	splicing consensus region
ZNF720	16:31730362-31730382	2(2)	intronic

Splicing consensus regions: −3 to +8 bp at the donor site and −12 to +2 bp at the acceptor site[9].

Table 2. Competitive gene set enrichment analysis of 18 candidate ALS genes

GO annotation	P.adj	Coefficient	Candidate genes
tRNA processing	1.45e-3	3.3627	ELP3, RTCB, TRDMT1
tRNA metabolic process	4.50e-3	3.1078	ELP3, RTCB, TRDMT1
Protein transport along microtubule	4.72e-3	3.3983	KIF5A, SSX2IP
tRNA modification	7.70e-3	3.2145	ELP3, TRDMT1
Centriolar satellite	1.96e-2	2.9781	NEK6, SSX2IP
Spindle assembly	2.28e-2	2.9355	INCENP, NEK6

The significant pathways in which at least 2 of the 18 nominated genes were involved.

Reference

1. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing Genotype-Tissue Expression Consortium. *Sci Transl Med* **5209**, 1–12 (2016).
2. Vaz-Drago, R., Custódio, N. & Carmo-Fonseca, M. Deep intronic mutations and human disease. *Hum Genet* **136**, 1093–1111 (2017).
3. Qian, X. *et al.* Identification of Deep-Intronic Splice Mutations in a Large Cohort of Patients With Inherited Retinal Diseases. *Front Genet* **12**, (2021).
4. Smith, B. N. *et al.* Exome-wide Rare Variant Analysis Identifies TUBA4A Mutations Associated with Familial ALS. *Neuron* **84**, 324–331 (2014).
5. Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat Genet* **48**, 1037–1042 (2016).
6. Nicolas, A. *et al.* Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* **97**, 1268-1283.e6 (2018).
7. Feng, Y.-Y. *et al.* RegTools: Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. *bioRxiv* 436634 (2018) doi:10.1101/436634.
8. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
9. Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res* **42**, 13534–13544 (2014).
10. Corvelo, A., Hallegger, M., Smith, C. W. J. & Eyras, E. Genome-Wide Association between Branch Point Properties and Alternative Splicing. *PLoS Comput Biol* **6**, e1001016 (2010).
11. Zhang, Q. *et al.* BPP: a sequence-based algorithm for branch point prediction. *Bioinformatics* **33**, 3166–3172 (2017).
12. van Rheenen, W. *et al.* Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat Genet* **53**, 1636–1648 (2021).
13. Cheung, R. *et al.* A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Mol Cell* **73**, 183-194.e8 (2019).
14. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat Rev Genet* **176**, 139–148 (2016).
15. Brown, A. L. *et al.* TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature* **603**, 131–137 (2022).
16. Ma, X. R. *et al.* TDP-43 represses cryptic exon inclusion in the FTD–ALS gene UNC13A. *Nature* **603**, 124–130 (2022).
17. Simpson, C. L. *et al.* Variants of the elongator protein 3 (ELP3) gene are associated with motor neuron degeneration. *Hum Mol Genet* **18**, 472–481 (2009).
18. Bento-Abreu, A. *et al.* Elongator subunit 3 (ELP3) modifies ALS through tRNA modification. *Hum Mol Genet* **27**, 1276–1289 (2018).
19. Van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet* **48**, 1043–1048 (2016).
20. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* **50**, 151–158 (2018).

21. Schmidt, A. *et al.* TBK1 and TNFRSF13B mutations and an autoinflammatory disease in a child with lethal COVID-19. *NPJ Genom Med* **6**, 55 (2021).
22. Chia, R., Chiò, A. & Traynor, B. J. Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. *Lancet Neurol* **17**, 94–102 (2018).
23. Zuko, A. *et al. tRNA overexpression rescues peripheral neuropathy caused by mutations in tRNA synthetase*. https://www.science.org.
24. Hanada, T. *et al.* CLP1 links tRNA metabolism to progressive motor-neuron loss. *Nature* **495**, 474–480 (2013).
25. Greenway, M. J. *et al.* ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat Genet* **38**, 411–413 (2006).
26. Glatt, S. *et al.* Structural basis for tRNA modification by Elp3 from Dehalococcoides mccartyi. *Nat Struct Mol Biol* **23**, 794–802 (2016).
27. Malnar, M. *et al.* Phenylalanine-tRNA aminoacylation is compromised by ALS/FTD-associated C9orf72 C4G2 repeat RNA. (2022) doi:10.21203/rs.3.rs-2166624/v1.
28. Schaefer, M. *et al.* RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev* **24**, 1590–1595 (2010).
29. Hogg, M. C. *et al.* 5'ValCAC tRNA fragment generated as part of a protective angiogenin response provides prognostic value in amyotrophic lateral sclerosis. *Brain Commun* **2**, (2020).
30. Akiyama, Y. *et al.* RTCB Complex Regulates Stress-Induced tRNA Cleavage. *Int J Mol Sci* **23**, (2022).
31. Liu, X. *et al.* A Functional Non-coding RNA Is Produced from xbp-1 mRNA. *Neuron* **107**, 854-863.e6 (2020).
32. Hetz, C. *et al.* XBP-1 deficiency in the nervous system protects against amyotrophic lateral sclerosis by increasing autophagy. *Genes Dev* **23**, 2294–2306 (2009).
33. van der Spek, R. A. A. *et al.* The project MinE databrowser: bringing large-scale whole-genome sequencing in ALS to researchers and the public. *Amyotroph Lateral Scler Frontotemporal Degener* **20**, 432–440 (2019).
34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017) doi:10.1101/201178.
37. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
38. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).
39. Wilks, C., Gaddipati, P., Nellore, A. & Langmead, B. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics* **34**, 114–116 (2018).
40. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
41. Chen, S. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. doi:10.1101/2022.03.20.485034.

42. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
43. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500–507 (2012).
44. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).
45. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343–345 (2009).
46. Liu, Z. *et al.* Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. *Sci Rep* **7**, 2193 (2017).
47. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences* **82**, 488–492 (1985).