Somatic mutations in ALS genes in the motor cortex of sporadic ALS patients

Óscar González-Velasco[1,2*], Rosanna Parlato[2,*], Rüstem Yilmaz[2], Lorena Decker[3], Sonja Menge[3], Axel Freischmidt[3], Xiaoxu Yang[4], Nikshitha Tulasi[2], David Brenner[3], Peter M. Andersen[5], Karin M.E. Forsberg[5] , Johannes C.M. Schlachetzki[6], Benedikt Brors[1], Lena Voith von Voithenberg[1,†], Jochen H. Weishaupt[2,3,†]

[1] *Department of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany*

[2] *Division of Neurodegeneration, Department of Neurology, Mannheim Center for Translational Neurosciences, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*

[3] *Department of Neurology, Ulm University, Ulm, Germany*

[4] *Department of Human Genetics, University of Utah, Salt Lake City, USA*

[5] *Department of Clinical Neurosciences, Neuroscience, Umeå University, Umeå, Sweden*

[6] *Department of Neurosciences, University of California, San Diego, La Jolla, USA*

** These authors contributed equally and share first authorship.*

[†] *Co-senior authors share last authorship.*

Correspondence: lena.voithenberg@dkfz.de, jochen.weishaupt@uni-ulm.de

Abstract

Several lines of evidence suggest that amyotrophic lateral sclerosis (ALS) pathology starts focally and spreads within the central nervous system (CNS). We leveraged DNA and RNA sequencing data from ALS autopsy tissue to explore the presence of somatic mosaic variants in sporadic ALS (sALS) cases. Deep targeted sequencing of known ALS disease genes from motor cortex tissue revealed enrichment of somatic variants in sALS compared to familial ALS (fALS) with a monogenic cause and an increased pathogenicity of mosaic mutations in known ALS mutational hot spots. Experimental validation of the variants identified that the somatic *FUS* variant p.E516X leads to nucleo-cytoplasmic mislocalization and aggregation. Additionally, somatic variant calling from single-cell RNA-sequencing data of sALS tissue showed an accumulation of somatic variants in excitatory neurons. Collectively, using a multi-modal approach, our findings suggest that somatic mutations within the motor cortex, especially in excitatory neurons, may contribute to the development of sALS.

Introduction

Amyotrophic lateral sclerosis (ALS) is an adult-onset neurodegenerative disease characterized clinically by degeneration primarily of motor neurons eventually leading to respiratory failure and death[1,2]. Clinical and pathological evidence suggests focal onset and subsequent systemic spreading of neurodegeneration[3–5]. Neuropathologically, large layer V neurons of the corticospinal tract and motor neurons in the pons, medulla and spinal cord have been hypothesized to be the primary targets of degeneration.

Only about 5-10% of European ALS patients report a positive family history for the disease (fALS)[6], usually with an autosomal dominant mode of inheritance, whereas most ALS patients report a family history that is unremarkable for ALS (sALS). Germline mutations in approximately two dozen genes could be linked to ALS, usually in a heterozygous state, with *C9ORF72*, *SOD1*, *TARDBP*, *FUS*, and *TBK1* being the five most frequently mutated genes in populations of European origin[7,8]. Emerging insight into ALS genetics, together with neuropathological analysis, has fueled hypotheses on the pathomechanisms in the ALS field in recent years, such as protein aggregation and impaired protein quality control or intracellular transport[9].

Rare cases of germline *de novo* mutations, in particular in *FUS* and *SOD1* as a cause of sALS, have been reported[10,11]. However, whereas even in a considerable proportion (~50%) of fALS cases screening for a germline mutation in blood DNA remains inconclusive, conventional genetic testing turns out negative in approximately 90% of the people with sALS[12]. On the other hand, twin studies point to a considerably higher contribution of genetic factors to ALS pathogenesis than explained by the frequency of Mendelian mutations[13]. Clinically a clear distinction between most fALS cases and sALS cases cannot be made with sALS autopsies carrying similar neuronal inclusions as fALS cases[14].

Insights from neuropathology, but also progression of clinical symptoms, suggest that ALS pathology starts focally as a proteinopathy[15,16] and then spreads contiguously within the central nervous system over time[3]. Experimentally, inoculation of aggregate of mutant SOD1 protein purified from the ventral horn spinal cord of a fALS patient with a *SOD1* mutation into the L3-ventral horn spinal cord segment in mice triggers a fulminant spreading ALS disease throughout the neuroaxis[17]. Biological and clinical data supports the notion that a prion-like mechanism with focal initiation may be involved in spreading toxic RNA and protein species including misfolded SOD1, TDP43 and FUS[17,18]. Such a hypothesis entails that

even a small number of pathologically altered cells may be sufficient to initiate the focal development of ALS pathology, with expanding motoneuronal demise, and eventually leading to clinically progressive manifest muscle weakness[19]. Studies in chimeric mice transgenic for fALS *SOD1* mutations suggest a motoneuron-autonomous initiation of ALS[20]. Furthermore, a combination of multiple genomic alterations may increase the risk of ALS development[21]. Considering such a focal onset and subsequent spreading, it is plausible to hypothesize that somatic mosaic mutations could be responsible for a proportion of still genetically unexplained ALS cases.

To test this hypothesis, we here combined whole exome sequencing, deep targeted sequencing of ALS-related genes with up to 2000x coverage, targeted amplicon sequencing, and single cell RNA-sequencing of post-mortem motor cortex tissue followed by proof-of-concept validation of selected mosaic mutations. Our data provide evidence that somatic mutations in selected genes may play a role in ALS pathogenesis, and that within the motor cortex these occur predominantly in the neuron population most vulnerable to ALS.

Results

Somatic SNVs detected in ALS patients by Whole Exome Sequencing (WES)

To detect somatic single nucleotide variants (SNVs) and short insertions/deletions (INDELs) in patients with sALS, we performed a) whole exome sequencing, b) targeted sequencing at a depth of ~2000x, and c) single nucleotide variant calling from single cell RNA-sequencing on human postmortem motor cortex. Whole-exome sequence data were analyzed from 12 individuals: 6 patients diagnosed with sporadic ALS (sALS) and 6 matched control individuals without clinical diagnosis of a neurodegenerative disease (Supplementary Table 1). The average sequencing depth across the samples was ~100x. On average, we detected 4,070 raw SNVs in the controls and 3,400 SNVs in the sALS patients. After stringent filtering for mosaicism, a mean of 290 (σ=130) SNVs were retained in the controls and 240 (σ=70) SNVs in the ALS patients (Supplementary Table 2). Thus, we did not detect any somatic mosaic variants in known ALS genes. We identified SNVs in several other genes with possibly ALS-relevant functions (Supplementary Table 3). For example, we detected SNVs in *RPA2*, which is involved in double-strand break repair and in *NISCH*, which is involved in cytoskeletal organization, although the link between these genes and ALS pathogenesis requires further study.

Targeted deep sequencing of fALS-related genes: Identification of germline mutations in sporadic and familial ALS patients

Considering that we did not detect an increased overall somatic variant burden in sALS by WES, possibly owing to low-coverage sequencing of ALS genes, we designed a panel of genes known for their role in ALS development, for which we performed targeted deep sequencing at a depth of ~2000x (Supplementary Table 4). To detect germline and somatic mutations in sALS brains by deep sequencing, we analyzed motor cortex DNA of 10 individuals diagnosed as sALS patients, 9 familial ALS (fALS) cases, of which 7 carried germline mutations in fALS genes (*SOD1*, *TBK1*, *NEK1*, and *C9ORF72*), and 13 control cases (Supplementary Table 5).

SNVs were detected by combining multiple variant callers and integrating them for the selection of variants of low allele fraction, which relate to variants appearing in a somatic mosaic pattern (Figure S1, Figure S2A). The overall mean number of variants detected in different genomic regions and the target

coverage per gene are shown in the Fig. S2B-D. AlphaMissense was used to assess the pathogenicity and potential functional impact of these mutations[22] and ClinVar to interpret the reported disease associations of these rare variants[23]. Our panel sequencing approach confirmed variants in fALS patients, previously based on blood germline genotyping, for example *NEK1* in patient 13[24].

By using this approach, we firstly focus on the heterozygous germline mutations (Supplementary Table 6). From this analysis, we detected three non-synonymous, heterozygous germline SNVs in two patients diagnosed with sALS. None of these SNVs were detected in any of the controls. One of the SNVs KIF5A p.A268T, detected in patient 20 and reported to have a pathogenic effect at the clinical level, has previously been linked to spastic paraplegia (dbSNP[25]: rs139015012). We observed that the patient did not show any additional somatic SNVs in any of the known ALS-related genes of the targeted panel. Another sample (patient 19) showed a heterozygous mutation in ALS2 p.T1472M (rs201089588), which has been reported previously in two ALS cases and one case of infantile-onset ascending hereditary spastic paralysis[25]. The same patient carried a second heterozygous germline variant in SQSTM1 p.P387L, which was reported as “likely pathogenic” in ClinVar (rs776749939) with studies linking this variant to increased susceptibility to frontotemporal dementia (FTD) and ALS[26,27]. Therefore, we excluded patients 19 and 20 from any quantitative analysis, because they harboured likely pathogenic germline mutations that were not detected by previous examination.

Targeted deep sequencing of fALS-related genes: Somatic mosaic mutational burden in sporadic ALS patients

To identify somatic mosaic mutations by targeted deep sequencing, we specifically selected variants with low variant allele frequency (VAF) in sALS cases. For selecting the allelic frequency (AF) range in which mosaic variants could be identified reliably, we included control and spike-in DNA samples from a patient with the SOD1 p.D90A homozygous germline variant at known concentrations (1.5% and 0.5%). After applying our pipeline for variant calling (Figure S1), we identified somatic variants down to a spike-in variant AF of ~1%. Variants with AF below this cut-off, e.g. at 0.5% could not be detected by any of the variant callers used in their default configurations, although the variant was visible at the expected base exchange frequency in genome browsers. Therefore, we determined a lower threshold for detecting somatic mosaic variants at a variant allele frequency of 1%, and focused the analysis on variants occurring at an allele frequency between 1% and 35%.

We observed the overall relative distribution of SNVs and small insertions and deletions in the different

genomic regions, e.g. exonic, 3'UTR, 5'UTR, intergenic, to be similar between controls, fALS, and sALS patients (Fig. 1A). Interestingly, we observed an increase in the total number of somatic variants in sALS patients (μ=93.7, σ=28.8) compared to controls (μ=45.8, σ=26.6, t-test p-value=0.015); for fALS samples this did not reach statistical significance (μ=56.4, σ=32.9, t-test p-value=0.07) in the selected set of genes of interest (Fig. 1B, Fig. S2B). No significant difference was noted between fALS and controls (p-value=0.59). An analysis of the number of SNVs by AF showed that sALS cases have an overall increase in the number of SNVs at variant AF between 5% and 35% in these genes (Fig. 1C). Whereas the majority of the genes known to play a role in fALS showed an increase in the number of variants in sALS patients versus controls (Fig. 1D), we observed a specifically strong enrichment in the number of SNVs in *VAPB*, *MAPT*, *FUS*, *NEFH*, *CCNF*, *NEK1*, and *TBK1*, for some of which the number of somatic SNVs almost doubled.

Targeted deep sequencing of fALS-related genes: Mutational signatures enriched in sALS patients

Next, we investigated the distribution of base exchanges across the genome and within its trinucleotide context. The distribution of base exchanges observed by deep targeted sequencing in sALS patients and controls showed a similar pattern with most base exchanges occurring as C>A, C>T, T>A, T>C, and T>G (Fig. S3A-B). Base exchange frequencies for C>G were significantly lower for both, controls and sALS samples.

We analyzed single base substitution (SBS) signatures and compared them between the groups of samples. A number of single base substitution signatures can be related to underlying mechanisms for variant accumulation in the genome. We observed an enrichment of the signatures SBS2, SBS85, and SBS89, related to APOBEC and cytidine deaminase, in sALS samples compared to controls (Fig. S3C). Furthermore, the signatures SBS8, SBS10b, and SBS11, although appearing with a small contribution, were only detected in sALS samples. These signatures are associated with unknown aetiology of CC > AA exchanges, polymerase epsilon exonuclease domain mutations, and the exposure to alkylating agents, respectively[28]. We observed an increase in SNVs as part of the signature SBS89 of unknown aetiology for sALS patients as well as a decrease in signatures SBS29 and SBS25 potentially related to tobacco and chemotherapy (Fig. S3D). Furthermore, we observed an overall higher cosine similarity with mismatch repair deficiency signatures for sALS patients compared to controls (Fig. S3E). This supports a potential role of DNA damage repair deficiency in sALS development.

Targeted deep sequencing of fALS-related genes: Somatic mosaic variants detected in sALS patients

We compared the occurrence of somatic variants at an AF between 1% and 35% in genes known to be involved in ALS between fALS, sALS, and control samples and performed validation of the variants by amplicon sequencing (Fig. 2, Table 1). In controls, we detected somatic variants in *FIG4*, *GLE1*, *TUBA4A*, *TBK1*, *SETX*, and *FUS*, as well as an SNV in a 3'UTR of *VAPB*. fALS patients additionally showed variants in *CCNF*. For sALS patients, we observed a significant increase in the patients carrying variants in *FIG4* (9/9), *GLE1* (9/9), and *TBK1* (8/9). Variants in *SETX* were observed in 6 of 9 and variants in *FUS* in 4 of 9 sALS cases. Additionally, we observed exonic non-synonymous variants in *ALS2*, *NEFH*, as well as SNVs in the 5'UTR of *VAPS* in 44% of cases and in 3'UTRs of *NEK1*, *PFN1*, and *TARDBP* unique to individual patients. In comparison, we only detected 2 SNVs in 3'UTRs in *TARDBP* in fALS patients and none in controls, which occur in the general European population at a frequency of 3.2% and 0.13%.

Mutations in *TARDBP* can cause ALS by a loss-of-function principle, accompanied by cytoplasmic aggregation and loss of nuclear TDP-43 at the protein level. To assess the potential functional importance of the 3'UTR variant on *TARDBP* expression, we investigated evolutionary conservation using multiple sequence alignment across a wide range of species (Fig. S4). We focused on the sequence region of the 3'UTR variants in *TARDBP*. Results revealed a 100% sequence conservation across species analyzed in the region harboring the somatic mosaic low AF variant. Evolutionary conservation was less pronounced in the two regions, for which two fALS patients carried heterozygous variants (with 17 or 28 and 22 of 28 species showing variability in each of the regions, respectively). The evolutionary conservation between mosaic sALS and heterozygous fALS 3'UTR variants in *TARDBP* might indicate the functional importance of the 3´UTR region in which the rare low AF variant from the sALS patient is located. It has been reported that TDP-43 is able to autoregulate its expression levels and localization. This autoregulation of TDP43 levels and localization is achieved in part via a negative feedback loop by self-targeting the TDP-43 binding region located in the 3´UTR of its transcript[29–31]. A previous study has reported a similar *TARDBP* 3'UTR variant in a frontotemporal lobar degeneration patient (FTLD) and a frontotemporal lobar degeneration patient with ALS co-morbidity (FTLD-ALS), both with TDP-43 proteinopathy[32]. The *TARDBP* 3'UTR variant was accompagnied by a two-fold increase in total TARDBP mRNA when compared to controls.

Table 1. Somatic mosaic variants detected from targeted deep sequencing, aligned to GRCh37 and selected for validation by amplicon sequencing

Sample number	Sample code	Gene	aa change	chromosomal nucleotide position	validation	MAF	variant type
28	0.148-09	TARDBP	*575C>G	1_11083286_C_G	X	0.018	UTR3
14	O.34-09	TBK1	229-1G>A	12_64858113_G_A	X	0.02631579	splicing
24	15-086	TBK1	Q655K	12_64891431_C_A	X	0.06486486	exonic:nonsynonymous_SNV
27	139TB	TBK1	T657I	12_64891438_C_T	X	0.023809524	exonic:nonsynonymous_SNV
28	0.148-09	FUS	1542-2A>T	16_31202718_A_T	X	0.0709677	splicing
29	11-112	FUS	1542-2A>T	16_31202718_A_T	X	0.0297619	splicing
24	15-086	FUS	1542-2A>T	16_31202718_A_T	X	0.0484581	splicing
28	0.148-09	FUS	E516K	16_31202724_G_A	X	0.0148305	exonic:nonsynonymous_SNV
19	11-112	FUS	E516K	16_31202724_G_A	OX	0.0134409	exonic:nonsynonymous_SNV
24	15-086	FUS	E516K	16_31202724_G_A	X	0.0168421	exonic:nonsynonymous_SNV

Sample number	Sample code	Gene	aa change	chromosomal nucleotide position	validation	MAF	variant type
14	O.34-09	FUS	E516X	16_31202724_G_T	O	0.07083333	exonic:stopgain
19	11-112	FUS	E516G	16_31202725_A_G	O	0.0132626	exonic:nonsynonymous_SNV
28	0.148-09	FUS	E516V	16_31202725_A_T	OX	0.0306748	exonic:nonsynonymous_SNV
28	0.148-09	FUS	R521G	16_31202739_C_G	OX	0.0162272	exonic:nonsynonymous_SNV
24	15-086	ALS2	L878F	2_202593855_G_A	O	0.014	exonic:nonsynonymous_SNV
24	15-086	VAPB	-121G>C	20_56964395_G_C	X	0.07755102	UTR5
19	11-112	VAPB	-117G>C	20_56964399_G_C	X	0.09345794	UTR5
24	15-086	VAPB	-96T>C	20_56964420_T_C	O	0.02813299	UTR5
19	11-112	NEFH	S289A	22_29877116_T_G	O	0.01312336	exonic:nonsynonymous_SNV
14	O.34-09	FIG4	Q155X	6_110053856_C_T	X	0.01803607	exonic:stopgain
30	O.34-09	FIG4	V157L	6_110053862_G_T	X	0.016	exonic:nonsynonymous_SNV

Sample number	Sample code	Gene	aa change	chromosomal nucleotide position	validation	MAF	variant type
28	0.148-09	GLE1	L149V	9_131284959_C_G	OX	0.07377049	exonic:nonsynonymous_SNV
27	139TB	SETX	S2128F	9_135161823_G_A	X	0.01616162	exonic:nonsynonymous_SNV
20	2012-111	SETX	S2128F	9_135161823_G_A	X	0.01616162	exonic:nonsynonymous_SNV
22	16-73	SETX	S2128F	9_135161823_G_A	X	0.01629328	exonic:nonsynonymous_SNV
28	0.148-09	SETX	A2127D	9_135161826_G_T	X	0.01217039	exonic:nonsynonymous_SNV
27	139TB	SETX	L2126I	9_135161830_G_T	X	0.02204409	exonic:nonsynonymous_SNV
14	O.34-09	SETX	L2126I	9_135161830_G_T	O	0.01408451	exonic:nonsynonymous_SNV

O validation by amplicon sequencing (out of three controls), OX partial validation by amplicon sequencing (out of ≥1 control), X no validation by amplicon sequencing.

We observed multiple variants in *FIG4* (Table 1, Fig. 2A-B). Selection based on pathogenicity prediction by AlphaMissense or description of clinical effects in the literature retained one variant that was described to be pathogenic. This missense variant in FIG4 at amino acid position 157 substituting a valine by a leucine (FIG4 p.V157L; NM_014845:exon5:c.G469T) was found in sALS patient 30 with an allele frequency of 1.6%. This position corresponds to the start site of the phosphatase domain of FIG4. A variant at this exact same position but with a valine to methionine substitution (FIG4 p.V157M) was reported in the ClinVar database (rs1455052760), and the phenotype reported is linked to Charcot-Marie-Tooth disease type 4, a neurodegenerative disease that involves also motoneurons.
Furthermore, a non-synonymous pathogenic variant in the *SETX* gene (SETX p.A2127D) was found with an allele frequency of 1.2% in patient 28 (Fig. 2A, C).
One of the sALS patients (patient 27) carried a high impact stop-gain variant in the *TBK1* gene (TBK1 p.Q655X) with an allele frequency of 3.9% (Supplementary Table 6). However, the same TBK1 p.Q655X variant was also detected in a control sample (patient 18) with a slightly lower allele frequency than 1.5% (Fig. 2A, D), suggesting that additional factors contributes to disease development.

We observed an enrichment of *FUS* variants in exon 15 in sALS cases (Fig. 2A, *right*). Exon 15, the extreme C-terminus of the protein, encodes the nuclear localization signal (NLS) and the glycine-rich domain of the protein that enhances the aggregation-prone property of FUS protein[33,34]. It interacts with the nuclear import receptor transportin (karyopherin β2), essential for shuttling FUS between the nucleus and the cytoplasm[35]. Mutations in the NLS domain of FUS lead to cytoplasmic translocation and aggregation of FUS and cause ALS[36–38]. Here, we report six distinct low AF missense somatic mutations within the NLS domain of FUS, one stop-gain (FUS p.E516X), and four non-synonymous SNVs (FUS p.E516K, p.E516G, p.E516V, and p.R521G), all found in sALS patients (Fig. 3A). As a first step, we used AlphaMissense to assess the functional impact of these mutations. All of the variants were flagged as "pathogenic". Only one control reported a missense somatic variant in this very same FUS region (FUS p.G515V), which was the only variant predicted as benign by AlphaMissense. FUS p.R521G missense mutations in FUS have been previously reported in a study of three different fALS pedigrees. The variant led to a cytoplasmic retention of the mutant protein observed both in brain tissue and transfected N2A and SKNAS cells[36]. Additionally, we identified a somatic mutation in three sALS patients reporting an adenine to cytosine substitution in the acceptor splice site of intron 14 (NM_004960:exon15:c.1542-2A>T). This specific splice-site mutation (rs1266751712) was identified as the most probable cause of ALS in a screening of a large French family[39]. Furthermore, the authors showed that the mutant allele

was missing the 40 bp coding sequence of FUS exon 15 as well as the first 203 bp of the 3' untranslated region.

Moreover, our results showed that three sALS patients carried somatic variants of low allele frequency and of functional importance in at least two positions in the same or different ALS-related genes. For patient 28, we detected the variants FUS p.E516V, FUS p.R521G, SETX p.A2127D, as well as mutations in canonical *FUS* and *TARDBP* splice sites. In patient 19, we observed FUS p.E516G and *FUS* splice site variants (c.A1542-2T), and in patient 24 FUS p.E516K and another *FUS* splice site variant (c.A1542-2T) was observed.
Taken together, we identified several pathogenic somatic mutations in ALS-associated genes, suggesting their potential contribution to sALS.

Functional validation of somatic mosaic variants in FUS detected in sALS patients

ALS patients with a mutation in *FUS* show neuronal cytoplasmic aggregation of this protein[36,38]. To functionally validate the predicted variants in *FUS* (FUS E515V, E516X, E516K, E516V) (Fig. 3A), we analyzed the cellular localization of the protein in cultured cells (Fig. 3B). Mutant FUS proteins carrying the 4 variants were cloned and transiently overexpressed in HEK293 cells in comparison to wildtype controls and a known variant (FUS-R495X). Whereas wild type FUS is localized in the nuclei in a well-defined pattern, the previously described Mendelian ALS mutation FUS-R495X showed cytoplasmic aggregation (Fig. 3B, C). The FUS-E516X variant, which we detected in ALS patients in a somatic mosaic pattern, was also localized in aggregated foci in the cytoplasm, losing its nuclear localization (Fig. 3B, C), while this was not observed for FUS-E516K and FUS-G515V under the studied conditions.

Single cell somatic variant calling in ALS and FTLD patient samples

To further understand the cell type-specific distribution of somatic mosaic variants and their frequency in the motor cortex, we performed somatic variant calling in single nuclei RNA-seq (snRNA-seq) data from a public cohort of sporadic and familial ALS (sALS and fALS), sporadic and familial FTLD (sFTLD and fFTLD), and control samples using Scomatic focusing on global patterns of mutations (Fig. 4-5, Supplementary Table 7)[40]. In total, we processed around 850000 cells with their original cell type annotation (number of samples: fALS: 5, fFTLD: 5, control: 15, sALS: 13, sFTLD: 12).

Our results revealed that excitatory neurons were the cell type with the highest mutational burden across all the different patient conditions (Fig. 4A-D), pointing to an inherent vulnerability of these cells to accumulate mutations in translated regions, as observed previously in pyramidal neurons[41] and specifically for Alzheimer's disease-related genes in neurons[42].

We first focused on SNVs within 28 genes included in the deep panel sequencing above. We observed that the mRNA expression levels of these genes were prominently and distinctly expressed in excitatory neurons (Fig. 4E), and in general undetectable in the other cell types, which impacted the discovery of somatic mutations in this specific set of genes. To overcome this limitation, we selected all the variants including the SNVs detected only in individual cell types (genes having no coverage in cells other than excitatory neurons, but otherwise having passed all quality scores). Following this approach, we observed a total of 15 variants in the genes from our panel on excitatory neurons (10 intronic and 5 UTR3), although the comparison between groups did not yield statistical significance difference (sALS vs Controls t-test p-value=0.79, fALS vs Control t-test p-value=0.70). Overall, the genes with the largest average number of variants per patient were *MAPT*, *KIF5A*, *NEK1*, *ALS2,* and *FIG4*, and, apart from *MAPT*.

We also investigated somatic variants in other genes that were detected in excitatory neurons comparing multiple cell types (presence of a variant in a specific cell-type but not in the others). We detected genes with an increased mutational burden in sALS by computing the ratio of the number of variants between sALS samples and controls for each of the genes, selecting genes that showed a log2-difference larger than 2 (19 genes in total) (Fig. 5A, Supplementary Table 7). Additionally, for all of the genes selected they did not show any SNVs in the control samples. These variants were mostly observed in genes with potential functional roles in brain development and neurodegenerative diseases (Table 2). Gene-set enrichment analysis of the top mutated genes with a log2-difference larger than 2 related them with synapse, GABA-ergic synapse, synapse organization and postsynaptic functions (Fig. 5B).

Table 2. Genes for which SNVs were exclusively detected in excitatory neurons of sALS patients from single-cell somatic variant calling analysis.

Gene	number of variants (sALS)	number of variants (control)
PLCB1	8	0
APP	6	0
KCNQ5-IT1	6	0
LRRTM4-AS1	6	0
SNTG1	6	0
CTXN1	5	0
NBEA	5	0
BICD1	4	0
CDH12	4	0
CERS6	4	0
E2F7;NAV3	4	0
GLS	4	0
LRFN5	4	0
MEF2C	4	0
RPE	4	0
TMEM108	4	0
TUSC7;MIR4447	4	0
UNC5D	4	0
ZDHHC11B	4	0

We investigated the cell types of origin in more detail by unsupervised clustering of the cells. Two clusters of subpopulations of excitatory neurons were identified, in which the numbers of SNVs were specifically high (clusters 5 and 9, Fig. 4B-C, 5C). These clusters were mainly composed of sALS cells (Fig. 5D), pointing to a selected high vulnerability of these cells to accumulate SNVs. Whereas cluster 5 (13064 cells) was mainly enriched in variants in coding and non-coding regions (Fig. 4B), cluster 9 (7825 cells) was enriched in non-synonymous variants (Fig. 4C). Compared to the other clusters, cell cluster 5 showed higher expression of genes involved in cell-cell communication and signaling, neuron development, neuron differentiation, and developmental processes. In addition to these cellular processes, cluster 9 was enriched for chemical synaptic transmission. This suggests a phenotypic

similarity between both clusters with cells in cluster 9 additionally showing activity in synaptic transmission. Furthermore, both clusters showed an elevated expression of the transcription factor *CUX2*, a late neuron differentiation and neuronal proliferation marker.

Finally, using all available somatic variants detected in the excitatory neurons, we computed per condition *de novo* mutational signature profiles for the base substitutions and their immediate 5' and 3' sequence context using SigProfileExtractor, and decomposed them into known single base substitution (SBS) signatures. Focusing on signatures resulting from altered DNA damage repair, we observed a trend towards enrichment in these signatures in sALS patients compared to controls (Fig. 5E). SBS14, for example, is related to polymerase ε mutation and DNA mismatch repair and SBS6, SBS15, SBS26, and SBS44 result from defective DNA mismatch repair. SBS26 was exclusively detected in sALS samples, with defective DNA mismatch repair comprising 63.9% of the whole sample signature.

We investigated whether specific mismatch repair mechanisms were affected by SNVs by computing the mutational burden over several DNA mismatch repair gene signatures, including DNA damage repair and cell cycle checkpoints (sALS=1, control=1), direct repair (none), base excision repair (sALS=1), nucleotide excision repair (none), mismatch repair and microsatellite instability (none), homologous recombination (sALS=1), non-homologous end joining (none), and translesion DNA synthesis (sALS=1).

Taken together, our analysis revealed an enrichment of somatic mutations predominantly in excitatory neurons, with an association to DNA damage repair signatures, suggesting that altered DNA repair mechanisms may contribute to sALS disease pathogenesis.

Discussion

Here, we show that sporadic, but not monogenic ALS caused by germline mutations, is linked to somatic mutations in the motor cortex (Fig. 6). We also demonstrate ALS-typical pathology of FUS protein harboring a specific mosaic mutation in sALS patients. Moreover, analysis of single cell transcriptomic data revealed an increased burden of somatic mutations in human motor cortex neurons and specifically in excitatory neurons, the brain cell type suggested to be primarily affected in ALS[43]. The data support a pathogenic role of low-frequency, somatic mutations in sporadic ALS patients, which represent the majority of people affected with ALS.

The possible role that somatic mutations play in the pathogenesis of neurodegenerative diseases has been debated for decades, as germline mutations account only for a minority of cases in most neurodegenerative entities. Whereas Huntington's disease represents an exception in that it is exclusively caused by CAG repeat expansions in *HTT*, only 10-20% of ALS patients and an even smaller proportion of Parkinson's syndromes or Alzheimer's disease cases can be explained by genetic variants that are inherited by classical Mendelian genetics. Neuropathological[3], clinical[19] and biological evidence[18,44] suggests that prion-like mechanisms may be involved. For this to be the case, only a minute "seed" of pathologically altered cells may be sufficient to initiate the focal onset and development of Parkinson, Alzheimer or ALS pathology.

Although it is challenging to provide straightforward evidence for a causative role of somatic mutations in sALS, our results strongly promote the hypothesis that somatic mosaic mutations in a small subset of brain cells could be responsible for a substantial proportion of sALS cases. However, most of these brain mosaic mutations with low allele frequency can be expected to escape detection by testing blood DNA, thus additional evidence for a substantial role of somatic mutations in neurodegenerative diseases is scarce. Low-frequency mosaic mutations in known Alzheimer's disease genes, such as *MAPT*, *PSEN1* or *PSEN2*, have been detected in Alzheimer's disease patient autopsy tissue[45]. Hints to somatic mutations in Parkinson's disease post-mortem tissue are even more limited[46,47]. With regard to motoneuron diseases, using RP-PCR performed on DNA extracted from spinal cord tissue of 44 sALS cases, no C9orf72 HRE was found[48]. Also, a case report described three siblings affected by the upper motor neuron disease HSP owing to somatic mosaicism of a novel SPAST mutation in the mother, who remained asymptomatic[49].

The fact that we observed a similar mosaic variant burden across the whole exome between control and ALS cortex autopsy tissue with high coverage (approx. 300x) could be explained by insufficiently deep coverage to detect these differences. Alternatively, a potential reduction in DNA damage repair capacity

may not be sufficient to induce significant DNA damage signatures across the whole exome in sALS patients. Rare somatic variants stochastically arising in ALS-causing genes could nevertheless be responsible for the initiation of ALS pathology despite a statistically unchanged overall somatic variant burden between patients and controls and analyses in a larger dataset are warranted.

Consequently, we developed a targeted "ultra-deep" comprehensive ALS gene panel test (> 2000x) with unique molecular identifiers for sequencing error correction. Positive "spike-in" controls showed that the protocol was suitable for detecting somatic variants with an allele frequency as low as approximately 1%. Comparison of patient motor cortex tissue from people with sALS, fALS, and without ALS revealed an increased somatic variant burden in sALS genes, but not in fALS when compared to controls. The finding is in agreement with the concept of somatic mutations contributing to non-familial, sporadic ALS. In contrast, a Mendelian mode of inheritance is caused by germline mutations with high penetrance[50] that are often sufficient to instigate ALS and do not depend on genetic mosaic mutations. Moreover, at an allele frequency of 10% or below, the frequency curves of exonic variants in sALS, fALS, and controls are getting closer, possibly because the probability of detecting the SNVs is lower. Alternatively, below a certain threshold, the probability that a given mosaic variant causes ALS declines.

Notably, in several sALS motor cortices, but not in controls, we detected somatic variants in more than one ALS gene, suggesting a possible poly- or oligogenic causation at the mosaic level. Presently, we do not know whether the different variants detected in an individual arose from the same cells or in different cells in a "colony" or even in cell types that could communicate to instigate pathology. It also remains to be shown whether somatic variants can be found in any of the known ALS genes, or whether genetic variation in certain genes is disadvantaged during development and leads to disappearance of the respective cell lineage and absence of these genes from the list of ALS genes with somatic variants. We found that some of the known fALS genes carry a higher somatic variant burden than others but noticeably, mutations in the second most frequently mutated germline fALS gene *SOD1* were not detected at all in our study, supporting the hypothesis that cortical de novo mosaic variants causing sALS is limited to a few genes and not any gene.

The deep panel sequencing revealed several mosaic mutations in the ALS gene *FUS*, but not in control individuals. *FUS* germline mutations can cause familial ALS and *de novo* germline mutations in *FUS* are detected in up to 50% of early-onset (<35 years) ALS patients[10]. Lending further support for pathogenicity, all *FUS* mosaic mutations observed in the motor cortex of sALS patients in this study are in a recognized ALS-linked hotspot for germline mutations, located at the C-terminus of FUS[23]. Respective mutations usually disturb the nuclear localization sequence of the protein, leading to nucleocytoplasmic

redistribution and cytoplasmic aggregation[51], accompanied by nuclear loss-of-function effects[52]. We observed a robust cytoplasmic mislocalization and deposition of the FUS p.E516X mutant protein found in the mosaic state (7% allele frequency in a sALS patient) in cultured cells. This further supports a causative role of this mutation, which would have been classified causal if found as a germline mutation. Not all *FUS* missense mutations showed the same robust change in subcellular distribution in our cell assay within the possible observation time and degree of cell stress applied. The same is sometimes observed with known *FUS* germline mutations with proven pathogenicity[51].

A neuron-autonomous origin of ALS in excitatory cortical neurons has been proposed, and the only approved disease-modifying drug for ALS in Europe (except the recently approved SOD1-targeting tofersen) is riluzole, which inhibits glutamate release. Moreover, integration of GWAS and single cell transcriptomic data showed that ALS-associated genes are predominantly expressed in glutamatergic neurons[53]. We leveraged single cell transcriptome data from ALS patient brain cortex and control individuals to compare the relative frequency of mosaic variants in different cell types. The data confirmed previous resultsthat neurons are indeed more prone to accumulate somatic mutations than other brain cell types[41]. Separate analysis of excitatory and inhibitory neurons revealed an even higher enrichment of mosaics in excitatory neurons, which is in line with the view that ALS is a disease primarily of excitatory neurons. Although the somatic variant calling in single cell data is not sufficiently sensitive to detect a specific enrichment of somatic mutations in ALS genes, we were able to find a gene signature whose overall mutational burden was significantly higher in sporadic ALS patient excitatory neurons when compared to controls in this dataset. Recent analyses suggest that the accumulation of variants in neurons with age and genomic instability as an effect of deficient DNA damage response play a role in ALS disease development[54,55]. Many of the known ALS-related genes impair DNA damage repair directly or indirectly[56]. Thus, an accumulation of variants in DDR-related SBS signatures as observed for excitatory neurons may play a role in the development of sALS in these patients.

Finally, our results may have profound implications for therapeutic design, as a role of genetic mosaics in the brain could render respective patients suitable for gene-specific interventions, such as antisense-oligonucleotides and siRNAs. A necessary prerequisite would be the availability of sensitive diagnostic procedures applicable in living patients. Whereas brain biopsies are still too invasive, liquid biopsies to search for genetic mosaics in CSF samples could lead to the identification of a group of ALS patients amenable for gene-specific therapy. Recently, the first gene-specific ALS therapies for patients with *SOD1* and *FUS* germline mutations have been approved or are in clinical testing[57]. However, our present finding that more than one gene may be mutated in a particular sALS case and that individual sALS patients have

different patterns of gene mosaicism, suggests that to be successful multiple gene intervention will be necessary and tailored to a particular patient (an “ASO/siRNA cocktail”).

In summary, although it remains to be determined whether a general impairment of the DNA repair system causes a ubiquitously higher frequency of somatic variants in sALS or whether the stochastic occurrence of genetic mosaics in ALS-relevant genes leads to neurodegeneration, here we have established a high-coverage ALS gene panel for detection of low-frequency somatic variants in human brain tissue. The data obtained suggest a contribution of rare somatic mutations to sALS, which is further corroborated by the functional validation of a FUS mutation found as a low frequency mosaic in ALS motor cortex tissue. Moreover, single cell transcriptome data demonstrate a high propensity of excitatory neurons to accumulate somatic variants, which could contribute to the susceptibility of cortical motor neurons, in line with the hypothesis of Charcot, Eisen, Braak and others of a focal origin and systemic spreading of ALS pathology (Fig. 6). Most importantly, these tools can be applied to screen additional patients that would benefit from personalized gene-specific therapeutic approaches in sALS.

Materials and Methods

Patient cohort

Fresh frozen autoptic human precentral gyrus (and spinal cord) tissues of donors with ALS and control donors (with similar age and sex distribution) were provided by the ALS Brain Bank at Umeå University in Sweden and the Netherlands Brain Bank. The Swedish ALS patients were diagnosed in accordance with the EFNS Guidelines for the management of ALS[58]. Using genomic DNA, all patients were Sanger sequenced for a panel of known fALS-causing genes. Patients with a pathogenic mutation in one of these genes or patients with a definite family history of ALS within three generations were termed fALS, all others were termed sALS. The autopsies were performed prospectively since 1993 adhering to a specific protocol and the tissue specimens saved in small pieces at -80°C until use. Control samples were collected and stored the same way. A complete list of samples used for whole exome sequencing is included in Supplementary Table 1 and for targeted sequencing in Supplementary Table 5. The library experimental design, including the information about paired motor cortex and spinal cord samples, as well as repeated libraries and spike-in controls, is summarized in Supplementary Tables 4-5.

DNA extraction and library preparation

Genomic DNA was extracted from ca. 20 mg of tissue by the QIAamp DNA Mini kit (QIAGEN, Germantown, USA) according to the manufacturer's recommendations. For whole exome sequencing the DNA was further processed by CeGaT Gmbh (Tübingen, Germany). For targeted sequencing, 50 ng of genomic DNA was used for library preparation and target enrichment by the SureSelect XT HS2 DNA system (Agilent Technologies, Santa Clara, CA, USA). The list of ALS genes targeted in this study is provided in Suppl. Table 4.

Whole exome sequencing

Whole exome sequencing was performed on 12 samples from the motor cortex by CeGaT GmbH (Supplementary Table 1). DNA concentration was quantified using Qubit fluorometer DNA assay (Thermo Fisher Scientific) and samples were RNase treated and provided in EDTA-free EB (Qiagen). After DNA quality control, libraries were prepared by the Twist Human Core Exome Kit and RefSeq and Mitochondrial Panel (Twist Bioscience). After library quality control, the samples were sequenced by the NovaSeq 6000 (Illumina)(reads 2x 100bp, output 35 Gb / sample).

WES data processing and germline variant calling

Paired-end next-generation sequencing FASTQ reads were aligned to human_g1k_v37_decoy reference genome with BWA-MEM (v0.7.17) with -K 100000000 -Y parameters[59], and sorted with Picard (v2.20.7)[60] SortSam. PCR duplicates were then marked with Picard MarkDuplicates. Indel realignment was carried out with GATK2 (v3.8-1)[61] RealignerTargetCreator, and IndelRealigner. Base quality scores were recalibrated with GATK BasaeRecalibrator and output with PrintReads. Germline heterozygous and homozygous SNV and Indels were called with GATK HaplotypeCaller and further genotyped with GenotypeGVCFs with JAVA (jre1.8.0_73). The code and Snakemake wrapper of the pipeline are accessible on GitHub[62].

Mosaic SNV/INDEL detection from WES data: Mosaic single nucleotide variants/mosaic small (typically below 20 bp) INDELs were called by using a combination of four different computational pipelines: Both tissue-shared and tissue-specific variants were called by MosaicHunter (single-mode, v 1.0)[63] with a posterior mosaic probability >0.05[64,65] as well as the single-mode of GATK's (v 4.0.4.0) Mutect2[66] with "PASS" followed by DeepMosaic (v 1.1.1)[67] and MosaicForecast (v 8-13-2019)[68]. Tissue-specific variants were called by the intersection of variants from the paired mode of Mutect2 and Strelka2 (v 2.9.2)[69] (set on "pass" for all variant filter criteria). For the panel of normal samples required for the pipeline of DeepMosaic and MosaicForecast, we employed an in-house panel of similarly sequenced normal tissues (n = 59 blood and brain samples from 30 individuals)[62]. For "tumor-normal" comparisons required by Mutect2 and Strelka2 pipelines, we compared the cortical and spinal cord tissue from each individual. Variants were excluded if they resided in segmental duplication regions as annotated in the UCSC genome browser (UCSC SegDup) or RepeatMasker regions, within a homopolymer or dinucleotide repeat with more than 3 units, or overlapped with annotated germline INDELs from HaplotypeCaller.

Variant annotation and prioritization: ANNOVAR (v2018Apr16)[70] was used to annotate refGene, gnomAD (v2.11)[71], dbSNP (v150)[25], CADD (v1.3)[72], eigen score[73], and fathmm (v2.3)[74] on the variants. The remaining candidates with both gnomAD(All) allelic frequency<0.001 and CADD phred score >10 were considered as initial candidates for the follow-up amplicon sequencing validations.

Design of panel for target amplification

A DNA panel for Agilent SureSelect XT Library Preparation was designed targeting the exonic and adjacent intronic regions of a total of 28 genes covering a genomic region of 1012 kB (Supplementary Table 4). 3358 probes with a total probe size of 107 kbp were designed by using the Agilent SureDesign software covering the coding exons of the genes and extending into the 3' and 5' end by ~50 bases each.

Details about the target ID, intervals, and regions are provided in Supplementary Table 4.

Target enrichment and sequencing library preparation

The Agilent SureSelect XT HS2 DNA Library Preparation and Target Enrichment protocol was followed for target enrichment. All steps were performed as described in the protocol provided by the manufacturer with the following adjustments: 23-50 ng of genomic DNA were used for enzymatic fragmentation for 15 min to obtain a target fragment size of 150-200 bp. During the pre-capture PCR 9-10 amplification rounds (for 50 ng and for 23 ng of input DNA, respectively) were performed. The quality of the libraries was followed by using the D1000 ScreenTape System (TapeStation System, Agilent).

Targeted next-generation DNA sequencing

Sequencing of a pool of 16 samples with a concentration of 10 nM each was performed at the Sequencing Core Facility of the German Cancer Research Center. The samples were sequenced on a HiSeq 4000 instrument (Illumina) by paired-end sequencing of 100 bp in a 4-color patterned flow cell with 235 M reads targeting a coverage of ≥2000x. Per sample ~10-20 Mio reads were obtained.

Targeted DNA sequencing analysis

A schematic overview of the analysis workflow is provided in Figure S1. Reads from individual samples in raw FASTQ files were demultiplexed based on the Agilent indices used. Quality control of sequencing results was performed by using FastQC and MultiQC[75,76]. The molecular barcode sequences were extracted and sequencing adaptors were removed from the raw reads by using AGeNT Trimmer v2.0.3. Reads were then mapped to the reference genome hsGRCh37d5 by using BWA-MEM[59] v0.7.17. The molecular barcode tags were added to the aligned reads in the BAM file by using SAMtools view. Error-corrected consensus reads were generated by using the molecular barcodes and AGeNT LocatIt v2.0.5 in v2Duplex mode with a barcode distance of 0. BAM files were validated by using picard v.2.25.1 ValidateSamFile, read groups were added by SAMtools v1.12 addreplacerg[77], if required, and BAM files were sorted and indexed (SAMtools). Variant calling was performed by multiple variant callers. Octopus v0.7.4[78] was run in the *cancer* setting calling germline and somatic variants with random forest filtering and somatic forest filtering. SAMtools v1.12 mpileup was run with default settings and varscan v2.3.8 mpileup2indel and mpileup2snp were used with a minimum read depth of 8, minimum number of supporting reads of 2, minimum base quality of 15, and minimum variant allele frequency of 0.01.

Variants had to be supported by ≤90% of the reads by one strand to be considered for downstream analysis. GATK v4.2.0.0[61] Mutect2[79] and Mutect2-PoN (GenomicsDBIImport, CreateSomaticPanelOfNormals) were run, the latter with a panel of normals consisting of nine control samples (see Supplementary Table 4). Variant calling by using GATK v4.2.0.0 HaplotypeCaller[80] was performed with the *ploidy* parameter set to 100 to consider all variants as different haplotypes. Strelka v2.9.5[69] was run in default mode and blat v3.6 MosaicHunter[63] was run in the exome and the genome modi. The variants detected in all tools were consolidated and DeepMosaic v1.0.0[67] *draw* and *predict* was used to detect somatic mosaic variants with model type efficientnet-b4.

Variant annotation, filtering and pathogenicity prediction

Variants were annotated using the annovar tool for variant effect prediction, and selected only if having a gnomAD frequency in the general population of < 0.001[70,81].

Predictions for all possible single amino acid substitutions were obtained from AlphaMissense[22] (AlphaMissense_aa_substitutions.tsv.gz file from https://zenodo.org/records/8208688, 2023-09-19). The database of predicted effects was split by gene, and individual coding mutations found on the panel that have passed all filters were interrogated.

Validation and quantification of mosaic mutations with targeted amplicon sequencing

Targeted amplicon sequencing was performed for candidate variants detected from deep panel sequencing as described above, to independently validate the identified mosaic mutations. PCR products were designed with a target length between 170 and 190 bp. PCRs were performed using the GoTAq Colorless Master Mix (Promega, M7832) on the ALS samples and controls as reported previously[62]. The libraries were sequenced on the Illumina NextSeq 550 platform with 151 bp paired-end reads using the NextSeq 500/550 Mid Output kit (300 cycles). The analysis was performed by comparing binomial distributions as described in Chung et al. (2023)[62]. A full validation was achieved, if the difference in the binomial distribution was significant against three control samples, whereas a partial validation was achieved when the difference was significant against at least one of the control samples.

Single cell RNA-Seq analysis and somatic variant calling

We downloaded the raw fastq files from a public single-cell RNA-Seq dataset of sALS, fALS and Controls, and additionally including sFTLD and fFTLD[40] from NCBI. In total we processed around 850.000 cells (n. of samples: c9ALS=5, c9FTLD=5, Control=15, sALS=13, sFTLD=12). Fastq files were processed using

cellranger version 8.0.1, and human genome reference version GRCh38. We then used SComatic for somatic variant calling analysis following the guidelines and scripts provided by the authors[82], annotations of groups of cells for SComatic were defined using the data´s original authors annotation of cell types, namely, astrocytes, microglia, excitatory neurons, inhibitory neurons, oligodendrocytes, and oligodendrocyte precursors (OPCs). Bedtools version 2.29.2, and annovar for variant annotation were used as part of the variant calling pipeline. Variants were filtered by selecting the ones with the PASS flag, as described in the SComatic documentation. Downstream analysis was carried out in R version 4.3.0, using Seurat version 5.1.0 and SeuratObject 5.0.2[83]. We selected the top 2000 most variable genes using FindVariableFeatures, and computed PCA and UMAP with those. Clustering was done as follows: FindNeighbors was computed using the first 25 PCs, and FindClusters with a resolution of 0.8. Finally, information of somatic variants per barcode was appended to the Seurat object metadata. Gene markers for clusters 5 and 9 were selected using FindMarkers specifying as ident.1 the cluster of interest. Gene enrichment and ontology analysis for the selected markers was done using the gprofiler2 package with correction method g_SCS, and using the top 500 markers for the mentioned clusters.
Computation of *de novo* mutational signatures and decomposing the *de novo* signatures into the COSMIC signatures was done using the R package SigProfilerExtractorR version v1.1.16[84].

Plasmids and site-directed mutagenesis

The coding sequence of human wildtype FUS (NM_004960.4) was cloned into BglII and KpnI sites of pCMV-Myc-N (Clontech Laboratories, Mountain View, CA, USA). Whereas FUS R495X was already available[85], other variants were introduced by site directed mutagenesis as described in the manual of the QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, CA, USA). Respective oligonucleotides for FUS mutagenesis are listed in Supplementary Table 8. All sequences were verified by Sanger sequencing.

Transfection of HEK293 cells and immunocytochemistry

HEK293 cells were transfected with plasmid DNA using calcium phosphate precipitation with minor modifications[86]. 24 h post transfection, immunocytochemistry was performed as recently described[85] using primary mouse anti-myc (1:500, Cell Signaling Technology, 2276) and secondary donkey anti-mouse-647 antibodies (1:500, Thermo Fisher Scientific, A-31571). Images were acquired on a Zeiss LSM 980 confocal microscope and analyzed using Image J.

Bibliography

1. Riancho, J., Gonzalo, I., Ruiz-Soto, M. & Berciano, J. Why do motor neurons degenerate? Actualisation in the pathogenesis of amyotrophic lateral sclerosis. *Neurología (English Edition)* **34**, 27–37 (2019).

2. Martin, L. J., Price, A. C., Kaiser, A., Shaikh, A. Y. & Liu, Z. Mechanisms for neuronal degeneration in amyotrophic lateral sclerosis and in models of motor neuron death (Review). *Int J Mol Med* (2000) doi:10.3892/ijmm.5.1.3.

3. Brettschneider, J. *et al.* Stages of pTDP-43 pathology in amyotrophic lateral sclerosis. *Ann Neurol* **74**, 20–38 (2013).

4. Ravits, J., Paul, P. & Jorg, C. Focality of upper and lower motor neuron degeneration at the clinical onset of ALS. *Neurology* **68**, 1571–1575 (2007).

5. Sekiguchi, T. *et al.* Spreading of amyotrophic lateral sclerosis lesions--multifocal hits and local propagation? *Journal of Neurology, Neurosurgery & Psychiatry* **85**, 85–91 (2014).

6. Barberio, J., Lally, C., Kupelian, V., Hardiman, O. & Flanders, W. D. Estimated Familial Amyotrophic Lateral Sclerosis Proportion: A Literature Review and Meta-analysis. *Neurol Genet* **9**, e200109 (2023).

7. Dilliott, A. A. *et al.* Clinical testing panels for ALS: global distribution, consistency, and challenges. *Amyotroph Lateral Scler Frontotemporal Degener* **24**, 420–435 (2023).

8. Akçimen, F. *et al.* Amyotrophic lateral sclerosis: translating genetic discoveries into therapies. *Nat Rev Genet* **24**, 642–658 (2023).

9. Weishaupt, J. H., Hyman, T. & Dikic, I. Common Molecular Pathways in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Trends Mol Med* **22**, 769–783 (2016).

10. Hübers, A. *et al.* De novo FUS mutations are the most frequent genetic cause in early-onset German ALS patients. *Neurobiol Aging* **36**, 3117.e1-3117.e6 (2015).

11. Müller, K. *et al.* De novo mutations in SOD1 are a cause of ALS. *J Neurol Neurosurg Psychiatry* **93**, 201–206 (2022).

12. Ruf, W. P. *et al.* Spectrum and frequency of genetic variants in sporadic amyotrophic lateral sclerosis. *Brain Commun* **5**, fcad152 (2023).

13. Al-Chalabi, A. *et al.* An estimate of amyotrophic lateral sclerosis heritability using twin data. *J Neurol Neurosurg Psychiatry* **81**, 1324–1326 (2010).

14. Turner, M. R. *et al.* Controversies and priorities in amyotrophic lateral sclerosis. *Lancet Neurol* **12**, 310–322 (2013).

15. Strong, M. J., Kesavapany, S. & Pant, H. C. The Pathobiology of Amyotrophic Lateral Sclerosis: A Proteinopathy?: *Journal of Neuropathology and Experimental Neurology* **64**, 649–664 (2005).

16. De Marchi, F. *et al.* Emerging Trends in the Field of Inflammation and Proteinopathy in ALS/FTD Spectrum Disorder. *Biomedicines* **11**, 1599 (2023).

17. Bidhendi, E. E. *et al.* Two superoxide dismutase prion strains transmit amyotrophic lateral sclerosis-like disease. *J Clin Invest* **126**, 2249–2253 (2016).

18. Feiler, M. S. *et al.* TDP-43 is intercellularly transmitted across axon terminals. *J Cell Biol* **211**, 897–911 (2015).

19. Ludolph, A. C. *et al.* Clinical spreading of muscle weakness in amyotrophic lateral sclerosis (ALS): a study in 910 patients. *J Neurol* **271**, 5357–5367 (2024).

20. Clement, A. M. *et al.* Wild-type nonneuronal cells extend survival of SOD1 mutant motor neurons in ALS mice. *Science* **302**, 113–117 (2003).

21. Ganz, J. *et al.* Contrasting somatic mutation patterns in aging human neurons and oligodendrocytes. *Cell* **187**, 1955-1970.e23 (2024).

22. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

23. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46**, D1062–D1067 (2018).

24. Brenner, D. *et al.* NEK1 mutations in familial amyotrophic lateral sclerosis. *Brain* **139**, e28 (2016).

25. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311 (2001).

26. Le Ber, I. *et al.* SQSTM1 mutations in French patients with frontotemporal dementia or frontotemporal dementia with amyotrophic lateral sclerosis. *JAMA Neurol* **70**, 1403–1410 (2013).

27. Van Der Zee, J. *et al.* Rare mutations in SQSTM1 modify susceptibility to frontotemporal lobar degeneration. *Acta Neuropathol* **128**, 397–410 (2014).

28. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

29. Avendaño-Vázquez, S. E. *et al.* Autoregulation of TDP-43 mRNA levels involves interplay between transcription, splicing, and alternative polyA site selection. *Genes Dev* **26**, 1679–1684 (2012).

30. Tollervey, J. R. *et al.* Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat Neurosci* **14**, 452–458 (2011).

31. Polymenidou, M. *et al.* Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat Neurosci* **14**, 459–468 (2011).

32. Gitcho, M. A. *et al.* TARDBP 3'-UTR variant in autopsy-confirmed frontotemporal lobar degeneration with TDP-43 proteinopathy. *Acta Neuropathol* **118**, 633–645 (2009).

33. Rhine, K. *et al.* ALS/FTLD-Linked Mutations in FUS Glycine Residues Cause Accelerated Gelation and Reduced Interactions with Wild-Type FUS. *Mol Cell* **80**, 666-681.e8 (2020).

34. Bonucci, A., Murrali, M. G., Banci, L. & Pierattelli, R. A combined NMR and EPR investigation on the effect of the disordered RGG regions in the structure and the activity of the RRM domain of FUS. *Sci Rep* **10**, 20956 (2020).

35. Zinszner, H., Sok, J., Immanuel, D., Yin, Y. & Ron, D. TLS (FUS) binds RNA in vivo and engages in nucleo-cytoplasmic shuttling. *Journal of Cell Science* **110**, 1741–1750 (1997).

36. Kwiatkowski, T. J. *et al.* Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* **323**, 1205–1208 (2009).

37. Naumann, M. *et al.* Impaired DNA damage response signaling by FUS-NLS mutations leads to neurodegeneration and FUS aggregate formation. *Nat Commun* **9**, 335 (2018).

38. Vance, C. *et al.* Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science* **323**, 1208–1211 (2009).

39. Belzil, V. V. *et al.* Identification of a FUS splicing mutation in a large family with amyotrophic lateral sclerosis. *J Hum Genet* **56**, 247–249 (2011).

40. Pineda, S. S. *et al.* Single-cell dissection of the human motor and prefrontal cortices in ALS and FTLD. *Cell* **187**, 1971-1989.e16 (2024).

41. Miller, M. B. *et al.* Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**, 714–722 (2022).

42. Li, Z. *et al.* Single-neuron whole genome sequencing identifies increased somatic mutation burden in Alzheimer's disease related genes. *Neurobiology of Aging* **123**, 222–232 (2023).

43. Braak, H., Ludolph, A. C., Neumann, M., Ravits, J. & Del Tredici, K. Pathological TDP-43 changes in Betz cells differ from those in bulbar and spinal α-motoneurons in sporadic amyotrophic lateral sclerosis. *Acta Neuropathol* **133**, 79–90 (2017).

44. Kiechle, M. *et al.* In Vivo Protein Complementation Demonstrates Presynaptic α-Synuclein Oligomerization and Age-Dependent Accumulation of 8-16-mer Oligomer Species. *Cell Rep* **29**, 2862-2874.e9 (2019).

45. Sala Frigerio, C. *et al.* On the identification of low allele frequency mosaic mutations in the brains of Alzheimer's disease patients. *Alzheimers Dement* **11**, 1265–1276 (2015).

46. Lobon, I. *et al.* Somatic Mutations Detected in Parkinson Disease Could Affect Genes With a Role in Synaptic and Neuronal Processes. *Front Aging* **3**, 851039 (2022).

47. Proukakis, C. Somatic mutations in neurodegeneration: An update. *Neurobiology of Disease* **144**, 105021 (2020).

48. Nordin, A. *et al.* Extensive size variability of the GGGGCC expansion in C9orf72 in both neuronal and

non-neuronal tissues in 18 patients with ALS or FTD. *Hum Mol Genet* **24**, 3133–3142 (2015).

49. Aulitzky, A. *et al.* A complex form of hereditary spastic paraplegia in three siblings due to somatic mosaicism for a novel SPAST mutation in the mother. *J Neurol Sci* **347**, 352–355 (2014).

50. Andersen, P. M. & Al-Chalabi, A. Clinical genetics of amyotrophic lateral sclerosis: what do we really know? *Nat Rev Neurol* **7**, 603–615 (2011).

51. Dormann, D. *et al.* ALS-associated fused in sarcoma (FUS) mutations disrupt Transportin-mediated nuclear import. *EMBO J* **29**, 2841–2857 (2010).

52. Scekic-Zahirovic, J. *et al.* Toxic gain of function from mutant FUS protein is crucial to trigger cell autonomous motor neuron loss. *EMBO J* **35**, 1077–1097 (2016).

53. 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).

54. Sun, Y., Curle, A. J., Haider, A. M. & Balmus, G. The role of DNA damage response in amyotrophic lateral sclerosis. *Essays in Biochemistry* **64**, 847–861 (2020).

55. Renton, A. E., Chiò, A. & Traynor, B. J. State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci* **17**, 17–23 (2014).

56. Wang, H., Kodavati, M., Britz, G. W. & Hegde, M. L. DNA Damage and Repair Deficiency in ALS/FTD-Associated Neurodegeneration: From Molecular Mechanisms to Therapeutic Implication. *Front. Mol. Neurosci.* **14**, 784361 (2021).

57. Miller, T. M. *et al.* Trial of Antisense Oligonucleotide Tofersen for SOD1 ALS. *N Engl J Med* **387**, 1099–1110 (2022).

58. EFNS Task Force on Diagnosis and Management of Amyotrophic Lateral Sclerosis: *et al.* EFNS guidelines on the clinical management of amyotrophic lateral sclerosis (MALS)--revised report of an EFNS task force. *Eur J Neurol* **19**, 360–375 (2012).

59. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

60. Broad Institute. Picard Toolkit. (2019).

61. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

62. Chung, C. *et al.* Comprehensive multi-omic profiling of somatic mutations in malformations of cortical development. *Nat Genet* **55**, 209–220 (2023).

63. Huang, A. Y. *et al.* MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Research* **45**, e76–e76 (2017).

64. Breuss, M. W. *et al.* Autism risk in offspring can be assessed through quantification of male sperm mosaicism. *Nat Med* **26**, 143–150 (2020).

65. Breuss, M. W. *et al.* Somatic mosaicism reveals clonal distributions of neocortical development. *Nature* **604**, 689–696 (2022).

66. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213–219 (2013).

67. Yang, X. *et al.* Control-independent mosaic single nucleotide variant detection with DeepMosaic. *Nat Biotechnol* **41**, 870–877 (2023).

68. Dou, Y. *et al.* Accurate detection of mosaic variants in sequencing data without matched controls. *Nat Biotechnol* **38**, 314–319 (2020).

69. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018).

70. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164–e164 (2010).

71. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

72. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).

73. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* **48**, 214–220 (2016).

74. Shihab, H. A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation* **34**, 57–65 (2013).

75. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).

76. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

77. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

78. Cooke, D. P., Wedge, D. C. & Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *Nat Biotechnol* **39**, 885–892 (2021).

79. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. Preprint at https://doi.org/10.1101/861054 (2019).

80. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at https://doi.org/10.1101/201178 (2017).

81. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

82. Muyas, F. *et al.* De novo detection of somatic mutations in high-throughput single-cell profiling data sets. *Nat Biotechnol* **42**, 758–767 (2024).

83. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015).

84. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).

85. Freischmidt, A. *et al.* A serum microRNA sequence reveals fragile X protein pathology in amyotrophic lateral sclerosis. *Brain* **144**, 1214–1229 (2021).

86. Jordan, M. & Wurm, F. Transfection of adherent and suspended cells by calcium phosphate. *Methods* **33**, 136–143 (2004).

87. Robinson, H. *Human Brain with Colorful Neurons*. (Zenodo, 2022). doi:10.5281/ZENODO.7058540.

Data and code availability

The raw sequencing data will be made available at the European Genome-Phenome Archive. The code associated with this work will be made available at 10.5281/zenodo.15092262.

Acknowledgements

The authors would like to thank the patients and their families. We acknowledge the support by the Sequencing Core Facility of the German Cancer Research Center and the Omics IT and Data Management Core Facility with special thanks to F. Petermann, M. Vogel, L. Weiser, and G. Warsow. We acknowledge the support of the LIMa Live Cell Imaging Mannheim at Microscopy Core Facility Platform Mannheim (CFPM). We thank K. Hauschulz for suggestions on sequencing library preparation. LV wishes to thank N. Paramasivam and S. Uhrig for discussions on variant calling. The project was in part funded by a grant by the AI Health Innovation Cluster to RP, BB, LV, and JW (AIH23). R. Yilmaz was supported by the Deutsche Forschungsgemeinschaft Walter Benjamin Programme (YI 209/1-1, AOBJ 680080).

Author contributions

OGV analyzed the variants from targeted sequencing and single-cell datasets, prepared graphs, and wrote the manuscript. RP designed the project, prepared patient material for sequencing and performed experimental validation in patient material, prepared graphs, and wrote the manuscript. RY designed, established and performed sequencing validation of variants. AF, LD, and SM performed experimental validation of variants in cultured cells. XY analyzed WES data. NT contributed performing targeted amplicon sequencing. DB, JCMS, and BB supervised aspects of the project. PMA and KMEF prepared and provided patient material. LV designed and supervised the project, established the analysis pipeline for variant calling from targeted sequencing, prepared material for sequencing, prepared figures, and wrote the manuscript. JW designed and supervised the project and wrote the manuscript. All authors have revised or critically reviewed the article.

Declaration of interests

LV is an employee of F. Hoffmann-La Roche.

Figures

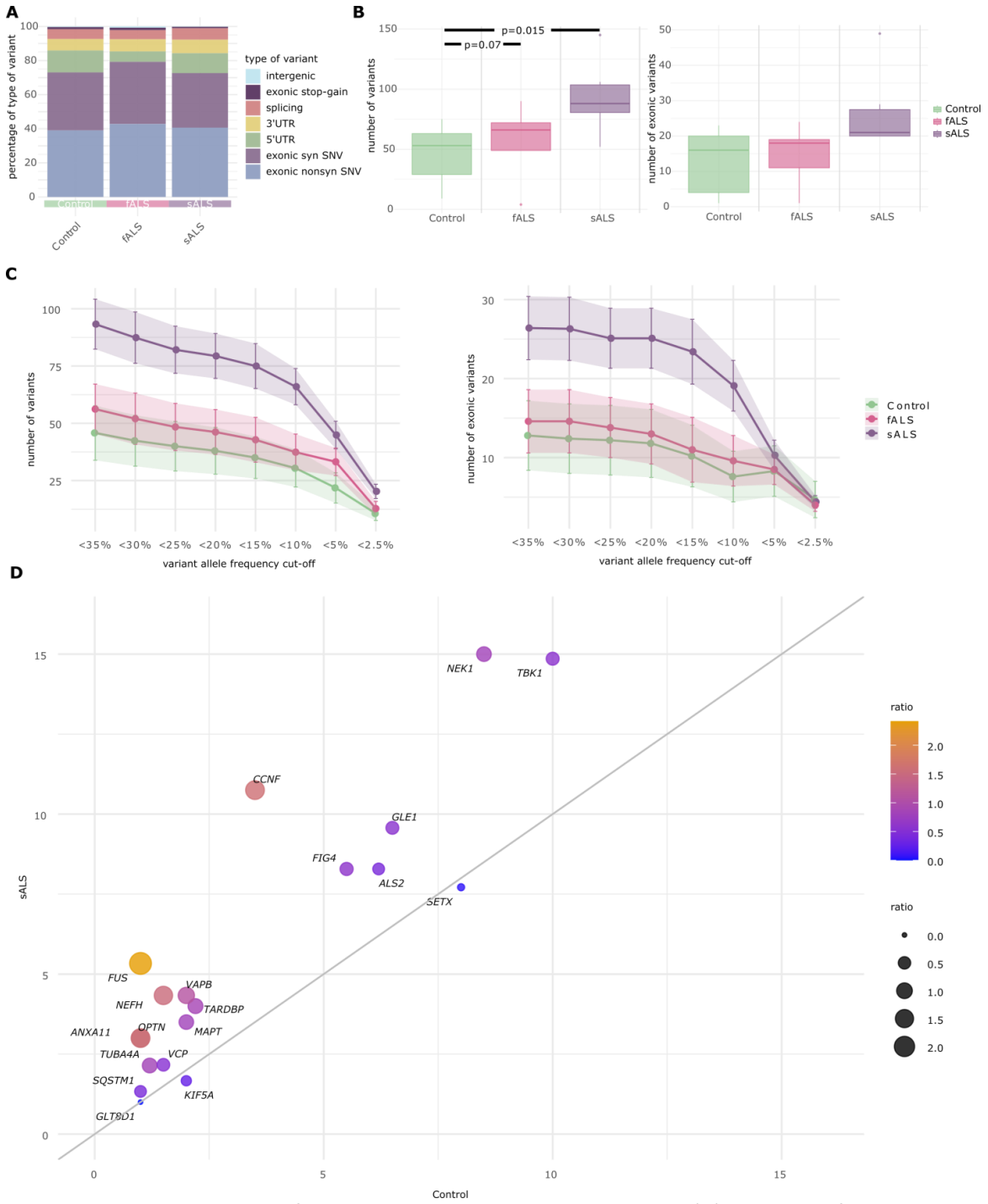

Figure 1. Somatic variant identification by targeted deep sequencing. (A) Number of variants with an allele frequency 35% > AF > 1% (left) and number of exonic SNVs with an allele frequency 35% > AF > 1% (right) detected per sample in all genes of the targeted panel. (B) Frequency of variant types detected per sample group. (C) Number of variants (left) and number of exonic variants (right) detected per

variant allele frequency for the different groups of samples. (D) Frequency of number of variants per sample in sALS versus control in the indicated genes of the targeted panel in the control group versus the sALS group colored by the ratio of the number of variants between sALS and control samples.

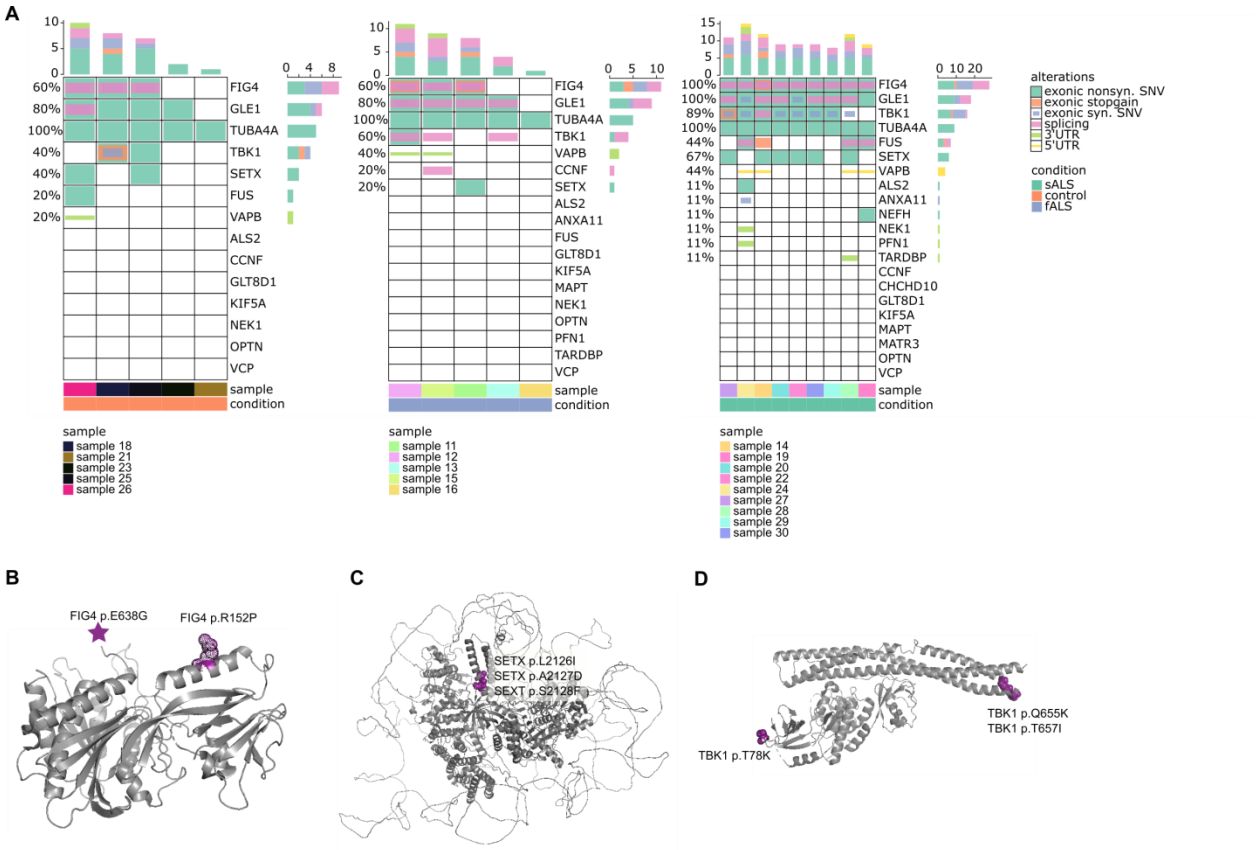

Figure 2. Somatic variants in genes of a targeted panel. (A) Frequency and type of SNVs in fALS genes of interest in control (left), fALS (center), and sALS (right) cases. (B-D) Protein positions of SNVs detected in ALS samples displayed on the protein structures of FIG4 (pdb: 7K1W) (B), SETX (AlphaFold: A2AKX3) (C), and TBK1(pdb: 6NT9) (D).

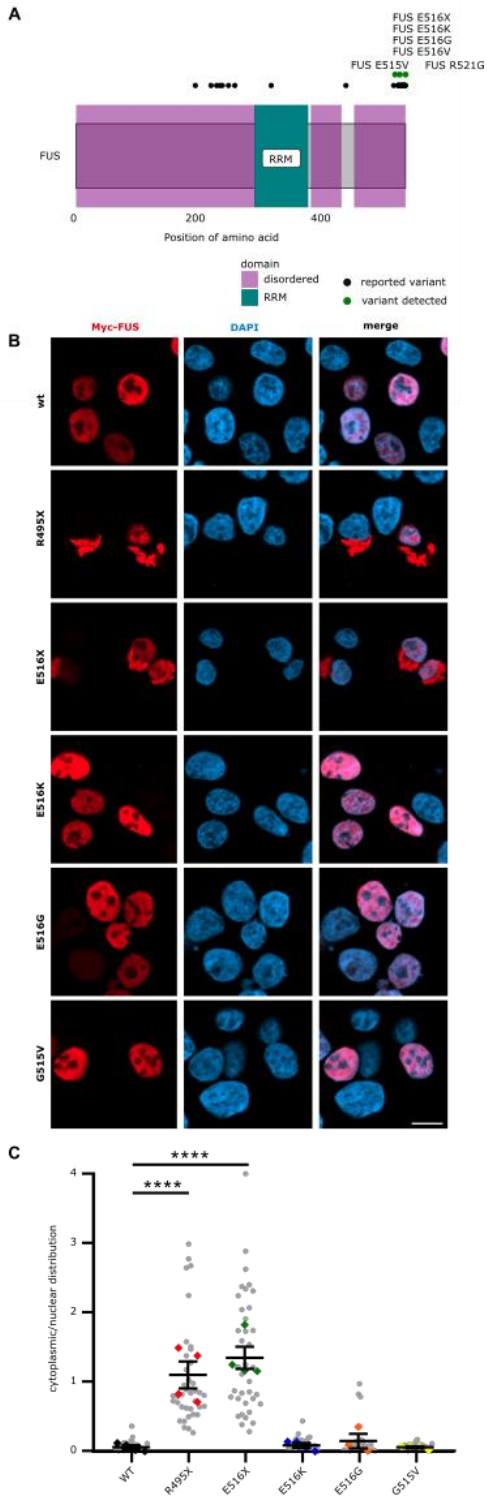

Figure 3. Functional validation of variants in FUS observed as somatic mosaic variants in sALS patients. (A) FUS protein structure with functional domains and known SNVs and SNVs detected in this study. (B) Cellular localization of wild type and mutant FUS in cultured HEK293 cells. Scale bar 10 µm. (C) Ratio of cytoplasmic to nuclear localization of FUS protein for the different FUS variants.

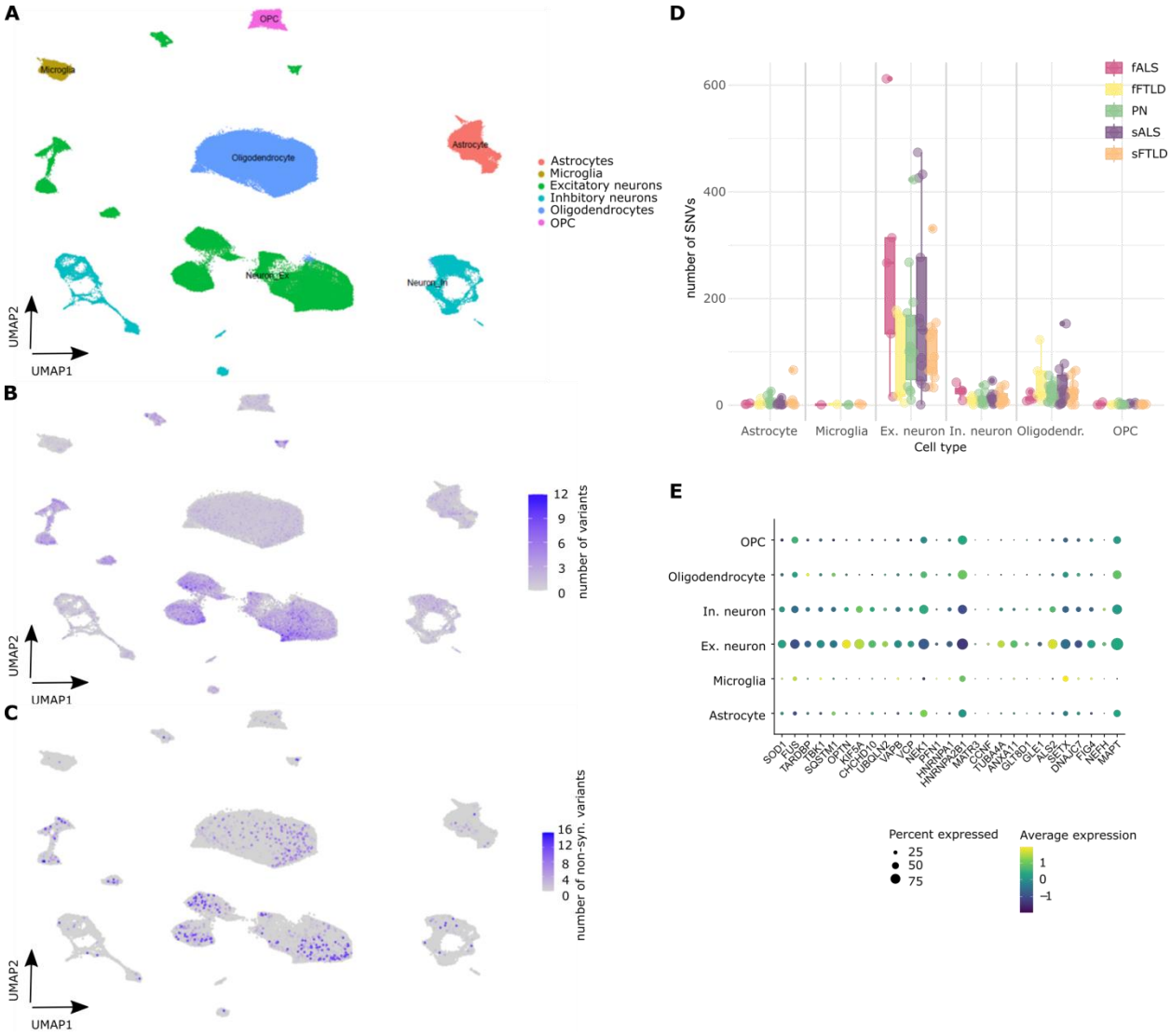

Figure 4. Somatic variant calling from single-cell RNA-sequencing data. (A) Dimension reduction of single-cell RNA expression levels of clusters of cells by Uniform Manifold Approximation and Projection (UMAP) of a dataset with sALS, sFTLD, and annotation of cell types. (B) UMAP of cell populations, color indicating the number of all (coding and non-coding) somatic variants detected by SComatic per cell barcode. (C) UMAP of cell populations, color indicating the number of non-synonymous somatic variants per cell barcode. (D) Bar diagram showing the number of variants per cell type and disease status. (E) Gene expression of the genes included in the targeted deeply sequenced panel.

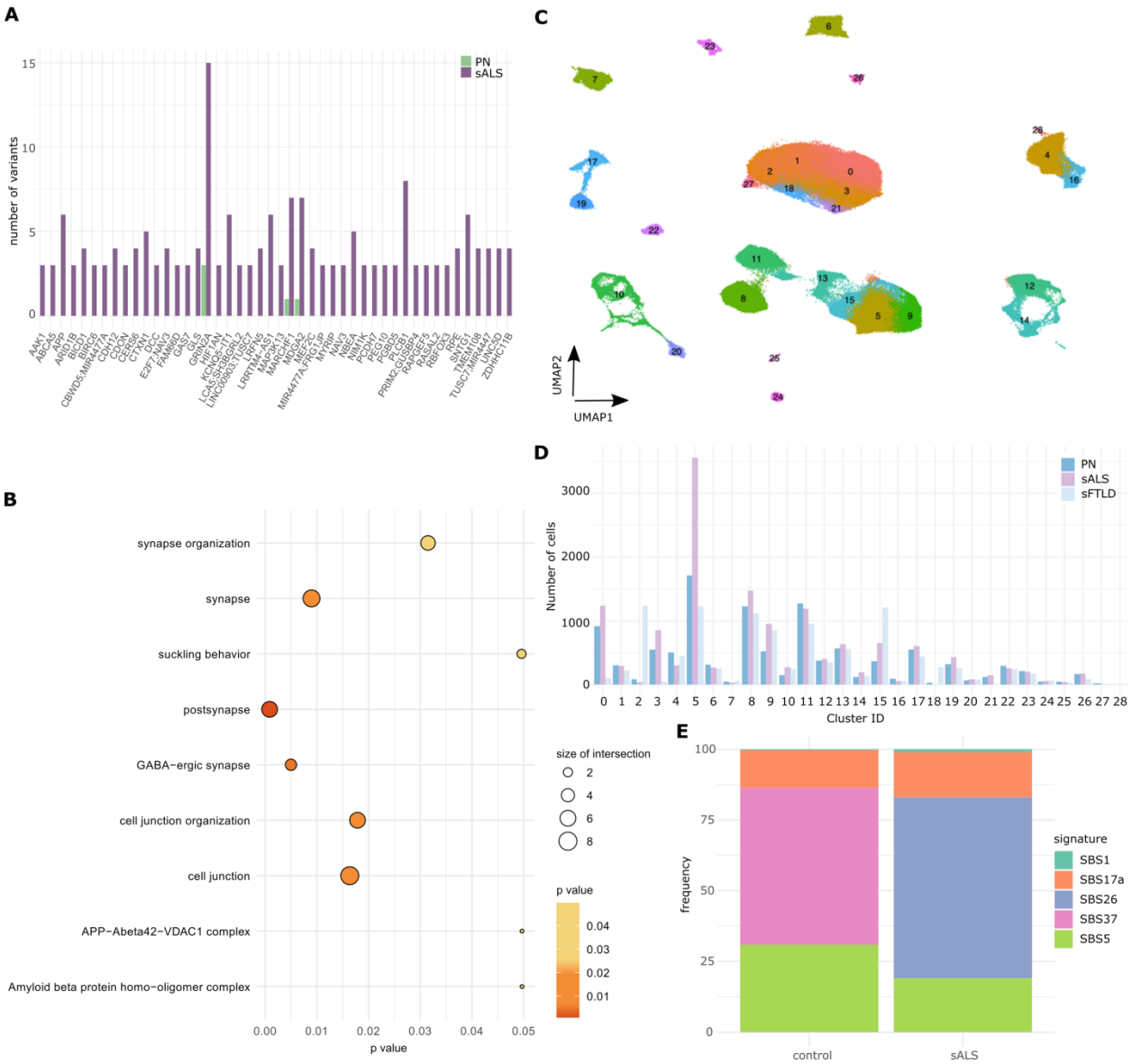

Figure 5. SNVs, signatures, and cell composition in gene expression-based clusters. (A) Number of single nucleotide variants per gene in excitatory neurons for controls and sALS samples. (B) Dimension reduction of single-cell RNA expression levels of clusters of cells by Uniform Manifold Approximation and Projection (UMAP) of a dataset with sALS, sFTLD, and annotation of cell types with unsupervised clustering. (C) Bar diagram displaying the number of cells per cluster and per disease state. (D) Relative fraction of single base substitution signatures in controls in comparison to sALS samples.

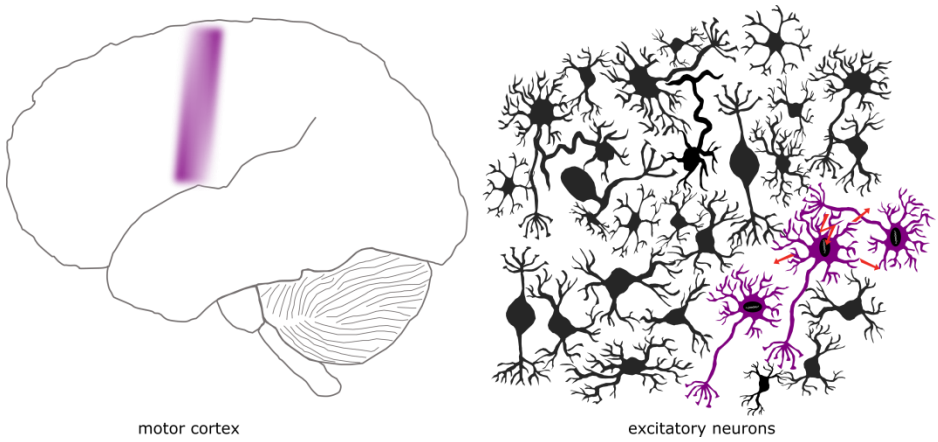

Figure 6. Schematic depiction of a motor cortex region in which an accumulation of somatic variants was observed in excitatory neurons suggesting a focal onset and local spreading of the disease. The motor cortex region of the brain and excitatory neurons are highlighted in violet. Adapted from Robinson[87].