# Unraveling temporal dynamics of the post-mortem transcriptome in amyotrophic lateral sclerosis

Ting Shen, PhD[1], Barbara E. Spencer, PhD[1], Pavel P. Kuksa, PhD[2], Vivianna M. Van Deerlin, MD, PhD[3], Hemali Phatnani, PhD[4,5], NYGC ALS Consortium, Edward B. Lee, MD, PhD[3] and Corey T. McMillan, PhD[1]

[1] Penn Frontotemporal Degeneration Center, Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[2] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[3] Center for Neurodegenerative Disease Research, Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[4] Center for Genomics of Neurogenerative Disease, New York Genome Center, New York, NY, USA

[5] Department of Neurology, Columbia University Irving Medical Center, New York, NY, USA

*Correspondence:

Corey T. McMillan, PhD

Penn Frontotemporal Degeneration Center, Department of Neurology

Perelman School of Medicine, University of Pennsylvania

3700 Hamilton Walk, Richards 606B

Philadelphia, PA 19104

mcmillac@pennmedicine.upenn.edu

## Abstract

Human tissue transcriptomics are crucial for understanding neurodegeneration but limited by cross-sectional post-mortem sampling, which represents end-stage disease. Subtype and Stage Inference (SuStaIn) modeling addresses these limitations by inferring temporal gene expression dynamics while also identifying potential transcriptomic subtypes. Applied to bulk RNA-seq data from post-mortem lumbar spinal cord in amyotrophic lateral sclerosis (ALS), SuStaIn unraveled that more advanced transcriptomic stages were associated with higher microglia and reduced neuron proportions, which mapped onto two ALS subtypes: Immune/Apoptosis/Proteostasis subtype with early immune/apoptotic/proteostatic dysregulation, worse prognosis and higher microglia proportions; Synapse/RNA-Metabolism subtype with early synaptic/RNA-processing deficits, lower male prevalence and neuron loss. Lumbar patterns demonstrated high concordance with cervical patterns and a strong correlation in staging across regional tissues. These findings revealed subtype-specific mechanisms underlying ALS heterogeneities, prioritized key genes driving subtyping/staging as potential therapeutic targets. More broadly, we established a framework to decode temporal dynamics from traditionally constrained post-mortem transcriptomic studies.

## Introduction

Human post-mortem tissue studies available through autopsy provide critically important sources of material for understanding the molecular and genomic features of neurodegenerative disease. Particularly, in amyotrophic lateral sclerosis (ALS), multiple molecular mechanisms have been reported to contribute to disease risk and prognosis, including disturbances in RNA metabolism and protein homeostasis, malfunctioning glial cells, axonal disorganization and disrupted transport, mitochondrial dysfunction, endoplasmic reticulum stress, apoptosis and cell survival, as well as exhibiting with a prion-like propagation pattern[1,2]. Specifically, the transcriptomic landscape of ALS has nominated genes associated with many biological pathways including cryptic exon splicing[3], chromatin accessibility[4], immune and inflammatory[5,6], autophagy and endocytosis[7], neurological processes such as synaptic transmission and neurotransmitter secretion[6]. Studies also highlight cell-specific features including upregulated genes in microglia and astrocytes and downregulated genes in neurons[8,9]. Moreover, specific autosomal dominant genetic mutations may engage distinct transcriptomic features and molecular pathways, suggesting heterogeneous transcriptomic landscapes across ALS[10,11].

To capture potential driving mechanisms of ALS heterogeneity, distinct subtypes have been identified utilizing diverse resources, including neuroimaging[12,13], neuropathology[14], and transcriptomics[15]. These findings suggest potential subtype-specific mechanisms that may underlie disease occurrence and progression. However, disease progression is inherently dynamic, a major challenge of autopsy-studies is that human brain tissue is typically end-stage disease following expiration after a long disease course and is by nature cross-sectional preventing an understanding of the temporal or longitudinal dynamics of gene expression in post-mortem human tissue. Recent analytic strategies of cross-sectional data have aimed to reveal disease heterogeneity and address the lack of longitudinal data computationally but have typically applied either subtype-only[13,16,17] or stage-only[18-20] models. While subtype-only models capture phenotypic heterogeneity, they fail to account for temporal heterogeneity. Conversely, stage-only models describe the dynamic trajectory of disease progression but do not adequately address its heterogeneity. The Subtype and Stage Inference (SuStaIn)

algorithm has been employed to identify distinct subtypes and extract their progression patterns simultaneously[21]. This model has already been successfully applied to identify data-driven neuropathology-based[14] and neuroimaging-based[22] subtypes within ALS-FTD spectrum, allowing for a more nuanced analysis of spatial and temporal disease progression.

Here, we sought to leverage bulk RNA sequencing (RNA-seq) data from human post-mortem spinal cord tissue to unravel the temporal dynamics of transcriptomic signatures in ALS, disentangling the disease heterogeneities by transcriptomic subtyping and staging. We specifically focus on bulk RNA-seq to enable a comprehensive assessment of proportional changes in cell-types across disease stages. With this unsupervised machine learning tool, we were able to identify two transcriptomic subtypes with distinct temporal transcriptomic trajectories, which reflect the sequential involvement of different molecular pathways during the neurodegenerative process in ALS.

## Results

### RNA-seq data processing, module construction, and cell deconvolution

In this study, we employed transcriptomic datasets from both lumbar and cervical spinal cord regions, utilizing the lumbar dataset as the discovery cohort for SuStaIn model training and the cervical dataset as an independent validation cohort to confirm the model's generalizability. We report an initial analysis of bulk RNA-seq of lumbar spinal cord from 151 ALS, 21 ALS-FTD and 45 neurologically normal controls (Supplementary Table 1), which was adjusted for RNA integrity number (RIN) and contributing sites (Supplementary Fig.1). We then used weighted gene co-expression network analysis (WGCNA) to derive 22 transcriptomic modules (Supplementary Data 1, Supplementary Fig.2). Of these, 13 modules showed significantly differential expression in the ALS/ALS-FTD group compared to controls, with 6 modules being down-regulated and 7 modules up-regulated (Fig.1a,b). Functional enrichment analyses were conducted separately for each module (Supplementary Data 2), then we were able to map 11 out of these 13 significant modules to identifiable functional biodomains (Fig.1c, Supplementary Fig.2e), which are discussed when relevant below. The modules eigengenes (MEs) of these 11 modules were z-scored, adjusting for covariates including age at death and

sex, with non-neurologic controls serving as the reference group. For modules exhibiting downregulation, the z-scores were inverted (multiplied by -1) to ensure that all biomarkers considered for the SuStaIn model are indicative of upregulation in disease. Subsequently, these adjusted z-scored MEs were then used as the empirically-derived feature set of input data for the SuStaIn model.

Additionally, the bulk RNA-seq data were deconvolved using Scaden to estimate cell-type proportions[23], serving as indicators of neurodegeneration and neuroinflammation (see Methods: Cell-type deconvolution). In the bulk RNA-seq datasets from both lumbar and cervical spinal cord, microglia proportions were higher, while oligodendrocyte proportions were lower in individuals with ALS/ALS-FTD relative to controls (Fig.1d,e and Supplementary Fig.3).

**Novel transcriptomic subtypes of ALS lumbar spinal cord**

The SuStaIn algorithm was applied to the expression files of the 11 empirically-derived modules, resulting in the identification of two transcriptomic subtypes that exhibit distinct transcriptomic dynamic patterns (Supplementary Fig.4a). The trajectory of gene expression changes for each subtype were constructed (Fig.2a). The z-score ranging from 1 to 3, indicating the degree of gene expression level from mild to moderate to severe. The most noticeable differences between the two subtypes were their distinct transcriptomic dynamic trajectories, specifically the order in which different modules were involved throughout the progression.

The first identified subtype, that we labelled “Immune/Apoptosis/Proteostasis subtype”, initially had specific early evidence of altered gene expression in the “Immune System”, “Apoptosis”, “Proteostasis”, and “Neuron Generation, Myelination, Gliogenesis” biodomains. This subtype subsequently showed abnormality in the “Mitochondrion” biodomain at SuStaIn stage 5, followed by involvement of the “Cytoskeleton” biodomain at SuStaIn stage 6, the “Neuron Generation, Synapse” biodomain at SuStaIn stage 7, and the “RNA Metabolism” biodomain at SuStaIn stage 9-11. The second identified subtype that we labelled “Synapse/RNA-Metabolism subtype” initially involved biodomains of “Synapse”, “RNA Metabolism”, and “Neuron Generation, Myelination, Gliogenesis” at SuStaIn stage 1. This subtype later showed abnormalities in the “Immune System” biodomain, progressing at a

slower rate compared to the Immune/Apoptosis/Proteostasis subtype. It subsequently showed involvement of the “Apoptosis” biodomain at SuStaIn stage 7, the “Cytoskeleton” biodomain at SuStaIn stage 11, and the “Mitochondrion” and “Proteostasis” biodomains at SuStaIn stage 13. In addition to these two subtypes with gene expression changes, individuals assigned to SuStaIn stage 0 were labeled as “normal-appearing group”, which showed no detectable gene expression changes. The predominant modules for each transcriptomic subtype exhibited more pronounced abnormal expression levels relative to the other transcriptomic subtype and the normal-appearing group (Fig.2b).

**Characteristics of subtypes**

The distribution of the 140 individuals with ALS/ALS-FTD distributed across subtypes included 75 (53.6%) categorized as Immune/Apoptosis/Proteostasis subtype, 37 (26.4%) as Synapse/RNA-Metabolism subtype, and 28 (20.0%) as normal-appearing group (Table 1). The detailed characteristics of subtypes are summarized in Table 1 and Fig.2c,d. Relative to the normal-appearing group, the Immune/Apoptosis/Proteostasis subtype had a shorter disease duration, indicative of worse prognosis, while the Synapse/RNA-Metabolism subtype exhibited an earlier age at death and lower proportion of males. Both subtypes showed a trend toward a higher frequency of repeat expansions in *C9orf72* compared to the normal-appearing group. However, while the Synapse/RNA-Metabolism subtype exhibited a trend toward a lower frequency of bulbar-onset compared to the Immune/Apoptosis/Proteostasis, these two subtype groups did not statistically differ from each other on any of the evaluated clinical factors. Cell-type analysis revealed that the two transcriptomic subtypes showed lower neuron proportions and higher microglia proportions than the normal-appearing group (Fig.2e). While not significant, the Immune/Apoptosis/Proteostasis subtype exhibited higher trends in microglia and neuron proportions compared to the Synapse/RNA-Metabolism subtype, consistent with their assignment to the involved biodomains.

**Transcriptomic dynamics of subtypes**

The transcriptomic dynamic trajectories of the two transcriptomic subtypes were constructed, each showing a distinct sequence of module involvement throughout the progression. Each

individual was assigned a SuStaIn stage within their transcriptomic subtype. The SuStaIn stage showed a significant correlation with the z-scored MEs of each module (Supplementary Fig.5), indicating their relevance in the transcriptomic progression. This SuStaIn stage effectively reflected the severity of gene expression dysregulation. We investigated the relationship between SuStaIn stages and clinical profiles. Only in the Synapse/RNA-Metabolism subtype, the SuStaIn stage was negatively correlated with the age at symptom onset (Fig.3a). We also compared the magnitude of SuStaIn stages across various subgroups (Fig.3b). Females exhibited higher SuStaIn stages compared to males. Individuals carrying pathogenic variants in *C9orf72* had significantly higher SuStaIn stages, compared to non-carrier. Moreover, the Synapse/RNA-Metabolism subtype showed higher SuStaIn stages than the Immune/Apoptosis/Proteostasis subtype. Fig.3c illustrates the distribution of individuals assigned to each SuStaIn stage and their corresponding cell type proportions, indicating that as SuStaIn stages advance, neuronal proportions decrease while microglial proportions increase. Further correlation analysis between SuStaIn stage and estimated cellular proportions revealed that within each subtype, the stage was significantly negatively correlated with the proportions of neurons and positively correlated with the proportions of microglia and endothelial cells (Fig.3d).

**Evaluation of Lumbar-trained SuStaIn model in cervical spinal cord**

The 11 modules constructed in samples from lumbar spinal cord, which were used to train the SuStaIn model, were highly conserved in samples from the cervical spinal cord (Fig.4a). Thus, we projected the Lumbar-constructed modules to the cervical spinal cord dataset, and generated new MEs for these modules in cervical spinal cord. Except for the darkgreen module, which was not a driving factor for subtyping, the other modules exhibited similar expression patterns between the two regions (Fig.1a,b and Fig.4b). The black, cyan, magenta, midnightblue, salmon, and tan modules were significantly up-regulated in individuals with ALS/ALS-FTD, while the grey60, and royalblue modules were significantly down-regulated compared to controls.

To evaluate the consistency of transcriptomic subtype assignment, we applied the Lumbar-trained SuStaIn model to the cervical dataset in order to predict the transcriptomic subtypes

and stages within the cervical spinal cord. Of the 130 individuals with ALS/ALS-FTD who also had samples from cervical spinal cord, 93 (71.5%) remained consistent with their lumbar subtype assignments, were regarded as “subtype consistent” individuals, while the remaining 37 cases (28.5%) were considered as “subtype inconsistent” due to differing subtype assignments (Fig.4c). The probability of subtype assignment in lumbar dataset was significantly lower in the “subtype inconsistent” group relative to the “subtype consistent” (Fig.4d). Individuals with a higher probability of being classified into their subtype in the lumbar dataset were more likely to show a stable subtype in testing samples. Conversely, those with inconsistent subtype assignments had less certainty in cross-dataset reclassification. Moreover, the lumbar SuStaIn stage was strongly correlated with the cervical SuStaIn stage in the “subtype consistent” group (Fig.4e).

**Evaluation of SuStaIn model Compared to subtype-only model**

To assess the ability of different models to capture ALS transcriptomic subtypes, we compared the SuStaIn model with a subtype-only model, that does not account for temporal dynamics, generated using non-negative matrix factorization (NMF) clustering. We employed a non-smooth NMF (nsNMF) model on the same preprocessed bulk RNA-seq dataset from the lumbar spinal cord (see Methods: RNA-seq data preprocessing), which comprised a variance stabilizing transform (VST) transformed gene expression matrix with 19,614 genes across 171 samples, to identify subgroups of ALS. Since the SuStaIn model identified two transcriptomic subtypes, a rank of 2 was chosen for the nsNMF model.

Principal component analysis (PCA) successfully demonstrated that two NMF subtypes form distinguishable clusters, suggesting distinct gene expression profiles associated with each NMF model-derived subtype (Fig.5a). To select representative genes for each subtype, genes were ranked by their contributions to the identified subtypes. We selected the top 2,000 genes for each subtype, and found 947 common genes were identified by calculating the intersection of the top gene lists from both subtypes, and 1,053 unique genes for each subtype (Fig.5b). Subsequently, we conducted functional enrichment analysis on the unique and common genes to explore the biological annotations of the two nsNMF-derived subtypes (Fig.5c). Both subtypes involved a majority of the biodomains; however, one subtype showed unique

involvement in the “Immune System” and “Oxidative Stress” biodomains, which we referred to as the “Immune subtype”, while the other subtype had unique involvement in the “Synapse”, “Cognition and Behavior”, “Ion Transport and Homeostasis” biodomains, which we referred to as the “Synapse subtype”. As the name suggests, the NMF-derived subtypes partially capture similar information to the SuStaIn model. However, because the NMF model is cross-sectional, it primarily reflects individuals who already exhibit involvement in multiple biodomains, making it challenging to capture early-stage ALS features, which the SuStaIn model is designed to achieve. We observed an approximate ratio of 1.1:1 between the Immune subtype and the Synapse subtype among individuals with ALS/ALS-FTD. The first principle component (PC1) from PCA differed significantly between the two NMF subtypes, and the second principle component (PC2) was significantly correlated with disease duration (Fig.5d). The Immune subtype also exhibited shorter disease duration compared to the Synapse subtype, mirroring a similar trend observed in the SuStaIn-derived subtypes (Fig.5e). Moreover, the Immune subtype exhibited significantly lower neuron proportion, and higher proportions of microglia and astrocyte (Fig.5e).

**Potential driver genes of subtyping and staging**

We have identified core modules and their corresponding hub genes that show strong correlations with both module eigengenes and disease subtypes, suggesting their potential pivotal for ALS subtyping. For the Immune/Apoptosis/Proteostasis subtype, the core modules include magenta, midnightblue, salmon, and cyan, while for the Synapse/RNA-Metabolism subtype, the core modules are royalblue, black, and lightyellow. Each of these modules yielded a set of hub genes, which are likely drivers of subtyping (Supplementary Fig.6a,b and Supplementary Data 3). We then additionally identified hub genes within core modules that showed strong correlations with SuStaIn stages (Supplementary Fig.6c, Supplementary Data 3), suggesting their potential as critical drivers of transcriptomic staging. For hub genes of each core module, we observed that the hub genes associated with subtyping largely overlapped with those associated with staging (Fig.6a).

To explore whether the hub genes within core modules are specifically expressed in different cell types, we conducted an overlap analysis with cell type-specific marker genes (Fig.6b). We

observed that the hub genes within the magenta and salmon modules, which are associated with the “Immune system”, were predominantly enriched in microglia. This further supports that these modules might play a role in mediating immune responses within the central nervous system. In contrast, the royalblue module that associated with “Synapse” and “Cognition and Behavior”, showed a stronger association with other cell types, such as neurons, astrocytes, and oligodendrocytes, suggesting a different functional profile for this module. Other core modules associated with "RNA metabolism" and "Proteostasis" biodomains did not show specific enrichment in these neural cell types, which might suggest broader systemic involvement beyond the central nervous system.

We further performed differential expression analyses on the lumbar RNA-seq data, identifying a total of 5,429 genes that were significantly up- or down-regulated in individuals with ALS/ALS-FTD relative to controls (Fig.7a and Supplementary Data 4). These differentially expressed genes (DEGs) were mapped onto the core modules (Fig.7b). Notably, all hub genes within the magenta module were DEGs, while the other core modules contained hub genes, which were potential drivers of subtyping and staging, were not detected in traditional differential expression analyses.

## Discussion

In this study, we employed a transcriptomic data-driven SuStaIn model to unravel temporal dynamics of distinct transcriptomic subtypes in individuals with ALS. By analyzing cross-sectional post-mortem bulk RNA-seq data from spinal cord, we identified two distinct patterns of transcriptomic change that predominate in specific molecular pathways, which included an Immune/Apoptosis/Proteostasis subtype, and a Synapse/RNA-Metabolism subtype. These data-driven transcriptomic subtypes exhibited variations in clinical and genetic characteristics, as well as neuron and microglia proportions. Moreover, the data-driven SuStaIn stages constructed the temporal transcriptomic trajectories of each subtype, which aligned with cell-type proportional loss of neurons and gains of microglia that we interpret as increased microglial activation. Together, our findings provided new insights into the heterogeneity in the spatiotemporal transcriptomic signatures of ALS, and underscored the potential utility for

identifying therapeutic targets and facilitating patient stratification in precision medicine. More broadly, this approach opens a new computational strategy for understanding temporal dynamics of the transcriptome in post-mortem human tissues.

Extensive research efforts to uncover the molecular underpinnings of ALS have implicated a number of molecular mechanisms contributing to disease heterogeneity[1,24]. The co-expressed modules we constructed for transcriptomic subtyping and staging, directly captured the majority of previously proposed molecular pathways. Each module can be considered representative of a specific or a group of relevant molecular pathways. Our findings are consistent with previous reports that show upregulation of genes linked to inflammation and immune response, apoptosis, oxidative damage, RNA Metabolism, and mitochondrial function, alongside the downregulation of genes involved in synaptic transmission, myelination, and cytoskeletal organization[6,25-27]. The involvement of different molecular pathways is highly interconnected and create a compounding effect that ultimately drive neurodegeneration in ALS. The activation sequence of these pathways can vary, with each potentially acting as a trigger, modulator, or consequence of the disease process. These pathways interact in dynamic ways, forming a complex network that drives disease progression. For example, inflammation and immune responses may be secondary to protein aggregation released by damaged neurons, mitochondrial dysfunction or other stresses[28], while it might also reported to be among the earliest pathways activated, triggering a cascade of processes includes oxidative stress, mitochondrial dysfunction, and ultimately neuronal apoptosis[29]. Therefore, the predominant molecular pathways affected may vary among patients, necessitating precise patient stratification for the delivery of tailored therapies.

Supporting evidence has demonstrated that ALS exhibits significant clinical, genetic and pathological heterogeneities, with the distinct predominant molecular mechanisms mentioned above potentially forming the basis of these heterogeneities. To gain a deeper understanding of disease heterogeneity, recent emphasis has been placed on defining subtypes and elaborating distinct characteristics to capture potential driving mechanisms of neurodegeneration[13,15-17,30]. Molecular subtypes have been identified in individuals with ALS, each subtype was defined by predominate involvement of specific molecular pathways[15,30]. We identified two data-driven transcriptomic subtypes, the Immune/Apoptosis/Proteostasis

subtype and the Synapse/RNA-Metabolism subtype, exhibited abnormal gene expression in shared and distinct molecular pathways. The two subtypes were characterized by their distinctive molecular pathways as their names suggested, with the Immune/Apoptosis/Proteostasis subtype exhibiting early involvement of the “Immune system”, “Apoptosis”, and “Proteostasis” related pathways, while the Synapse/RNA-Metabolism subtype exhibited predominate abnormalities in “Synapse” and “RNA Metabolism” related pathways. In addition, both subtypes exhibited transcriptomic change in several shared biodomains such as “Neuron Generation”, “Myelination”, and “Gliogenesis”, all of which are part of the same module. The "Gliogenesis" biodomain, categorized under neural systems, includes microglia-related terms that are also associated with the immune system. This explains why abnormalities in this module were present in the early stages of both subtypes. Subsequently, transcriptomic changes progressively involved other molecular pathways, including those related to the “Mitochondrion” and “Cytoskeleton”. The most noticeable differences between the two subtypes were observed in the pathways that initially involved during the early SuStaIn stages, and the sequences in which different pathways were affected throughout the progression trajectories. These findings suggested potential subtype-specific mechanisms driving disease occurrence and progression. As previously stated, the differences in the basis of pathogenic mechanisms among individuals of different subtypes may lie in the distinct initially affected molecular pathways that act as triggers[28,29].

Our SuStaIn findings were partly consistent with previous studies that identified an ALS-Glia subtype characterized by significant enrichment for immunological activities, pro-neuroinflammatory microglia state and neural cell death, which had a worse prognosis[15]. Similarly, the Immune/Apoptosis/Proteostasis subtype we identified also exhibited a shorter disease duration. Accordingly, individuals exhibiting predominant upregulation of inflammatory genes and activated microglia are more likely to experience accelerated progression of disease pathology. Eshima and colleagues also identified an ALS-Ox subtype characterized by oxidative stress and altered synaptic signaling, and an ALS-TD subtype defined by transcriptional dysregulation[15]. They suggested that the ALS-TD and ALS-Ox subtypes shared some overlapping molecular mechanisms, as both were associated with synapse-related pathways and post-translational modification. In our study, the Synapse/RNA-Metabolism

subtype might provide a more comprehensive summary of these two subtypes. As we expected, the transcriptomic subtypes exhibited increased neuronal loss and microglial activation, consistent with the hallmark upregulated neuroinflammation observed in the ALS-Glia subtype[15]. To further assess the SuStaIn model, we conducted a parallel subtype-only analysis on the same RNA-seq dataset in our study, identifying an NMF-derived Immune subtype and a Synapse subtype. This observation provided additional converging evidence that an immune-predominant ALS subtype may experience poorer prognosis outcomes, consistent with previous observations linking heightened immune activation and neuroinflammatory states to accelerated ALS pathology[31].

While these subtypes were broadly similar to the Immune/Apoptosis/Proteostasis and Synapse/RNA-Metabolism subtypes identified by the SuStaIn model, there were notable differences. The subtype-only models captured only cross-sectional characteristics, and did not account for the temporal trajectory of gene expression dynamics during disease progression. The SuStaIn model offered a dynamic view of transcriptomic progression of distinct subtypes that the subtype-only model lacks. This allowed us to observe that neuronal loss and microglial activation may intensify with the progression through successive SuStaIn stages. Moreover, the Immune/Apoptosis/Proteostasis subtype showed a trend towards decreased neuron proportion and increased microglia proportion compared to the Synapse/RNA-Metabolism subtype. Although this trend was not statistically significant, it may be attributed to some patients having progressed to middle or later stages of the disease, where multiple relevant molecular pathways have already been affected. Overall, this observation highlights the potential for transcriptomic subtypes to reflect different stages of disease progression and the importance of considering the temporal dimension in ALS research.

The SuStaIn algorithm provided strong subtyping and staging capabilities[32,33]. In addition to identifying distinct subtypes, we further assessed the robustness of our SuStaIn model by evaluating its performance across different spinal cord regions. The transcriptomic profiles from the lumbar region closely mirrored those from the cervical spinal cord, which enabled us to utilize the cervical samples as a secondary test dataset. Our analysis revealed a 71.5% consistency in subtype assignment when comparing results between lumbar and cervical

regions. However, some patients may exhibit transcriptomic features that classify one region (lumbar or cervical) into a transcriptomic subtype, while the other region appear normal and unable to be categorized into a subtype yet. As the disease progresses, the region initially classified as normal-appearing might evolve to match the subtype of the more affected region. To be precise, 10.8% of patients showed clear inconsistencies in subtype assignments between regions, while the remaining 17.7% had one region assigned to a subtype and the other classified as normal-appearing. For these cases, whether the regions would become consistent with further disease progression remains unclear at the current time point. In general, the transcriptomic subtype of ALS was thought to be consistent across different regions of the spinal cord, suggesting that the SuStaIn model performs reliably in distinguishing subtypes.

Distinct transcriptomic subtypes in ALS were characterized by specific core modules, within which we identified hub genes that may act as critical drivers of both subtyping and staging in the disease (Supplementary Data 3). For the Immune/Apoptosis/Proteostasis subtype, core modules highlighted several hub genes associated with immune responses and microglial activation (*TYROBP*, *GRN*, *TREM2*, *STAT6*, *FCGRT*, etc.), as well as genes linked to apoptosis (*TNFRSF1B*, *CASP4*, *CDKN1A*, etc.), along with *GPX1* gene involved in responding to oxidative stress. Additionally, this subtype featured a “Proteostasis” module with hub genes involved in protein degradation (*PSME1*, *PSMA7*, etc.) and some genes encoding ribosomal proteins (*RPS2*, *RPL36*, *RPL8*, etc.) that contribute to protein synthesis. All of these genes were upregulated in individuals assigned to this subtype. For the Synapse/RNA-Metabolism subtype, we identified hub genes that associated with synaptic plasticity and signaling (*CLSTN1*, *SCAMP5*, *IQCJ-SCHIP1*, *RASGRF1*, etc.), neurodevelopment (*ZNF536*, etc.), myelination and oligodendrocyte survival (*MAG*, *TUBB4A*, *ABCA2*, *S1PR5* etc.), and genes that play a role in regulating the cytoskeleton such as *MAST3*. We further performed a conventional differential expression analyses that is commonly used to identify transcriptomic features; however, it may overlook genes of physiological or pathological importance that are not statistically prominent. Some of the hub genes within core modules identified by the SuStaIn model, did not survive statistical significance tests in differential expression analyses, yet were still captured by our model, highlighting the ability of SuStaIn to uncover additional

layers of meaningful information. In general, identifying hub genes specific to each subtype provides insights into distinct molecular mechanisms underlying each subtype's pathology, potentially guiding targeted therapeutic strategies and improving our understanding of disease heterogeneity.

Several limitations should also be considered in future work. Firstly, the lack of clinical profiles from the NYGC ALS Consortium, including motor assessments and neuropsychological testing, limits the ability to characterize the clinical features of distinct subtypes. Future investigations require larger cohorts with comprehensive clinical assessments for further validation. Secondly, the SuStaIn model's assumption of a monotonic progression trajectory may not fully capture the complexity of transcriptomic changes throughout the disease course. This assumption might overlook the possible fluctuations in transcriptomic dynamics that might occur due to various factors, including treatment effects. Thus, this remains a noteworthy limitation that lies in the potential incomplete representation of disease progression trajectory. Future studies may develop models that can accommodate the possible fluctuations or reversals in biomarker changes, providing a more nuanced understanding of disease trajectories. Moreover, bulk RNA-seq data might be affected by cell type heterogeneity, however, it is critical to note that understanding heterogeneity in individuals with ALS necessitates a substantial dataset, ideally exceeding 100 samples, to effectively train the data-driven SuStaIn model. This requirement is currently cost-prohibitive when using snRNA-seq. Future studies could consider employing snRNA-seq data when a sufficiently large sample size can be acquired. Additionally, a common criticism of post-mortem molecular studies is that observed patterns may reflect consequential rather than causal associations. We identified key modules driving the subtyping and staging process, with some hub genes within these modules standing out as potential contributors. These key genes further require validation through forthcoming animal experiments to assess their functions and impact, which may offer promising therapeutic targets, paving the way for future experimental validations and the development of targeted interventions. Furthermore, transcriptomic features in brain regions such as the frontal, temporal, and motor areas may exhibit different patterns, which may lead to the identification of distinct subtypes and subtype assignment may differ from that of the spinal cord. Future studies could consider the potential regional specificity by using data

from other brain regions to examine the concurrent changes of different molecular pathways in different brain regions along the same timeline.

In conclusion, we used bulk RNA-seq data from human post-mortem tissue to unravel the temporal transcriptomic signatures within individuals with ALS, disentangling the disease heterogeneities by transcriptomic subtyping and staging. We identified one subtype characterized by predominance in immune system and apoptosis-related pathways, and another primarily associated with synaptic pathways, with subsequent involvement of other molecular pathways as the disease progresses. Identification of subtype-specific transcriptomic signatures supporting the existence of transcriptomic drivers of disease heterogeneity, and also enable the prediction of the dynamic progression across distinct subtypes. This knowledge might pave the way for precisely tailored strategies in diagnosis, management, and potential treatments.

## Methods

### Study Oversight

The New York Genome Center (NYGC) Amyotrophic Lateral Sclerosis (ALS) Consortium samples presented in this work were acquired through various institutional review board (IRB) protocols from member sites and the Target ALS postmortem tissue core and transferred to the NYGC in accordance with all applicable foreign, domestic, federal, state and local laws and regulations for processing, sequencing and analysis. The Biomedical Research Alliance of New York (BRANY) IRB serves as the central ethics oversight body for the NYGC ALS Consortium. Ethical approval was given and is effective.

### NYGC ALS Consortium cohort

We downloaded postmortem bulk RNA-seq data of the spinal cord from the NYGC ALS Consortium, contributed by eight medical centers. The dataset consists of 366 RNA-seq samples, comprising 216 samples from cervical and 202 from lumbar regions. The dataset included 151 individuals with ALS, 21 with ALS-FTD, and 45 non-neurological controls. A total of 150 individuals have RNA-seq data from both the lumbar and cervical spinal cords, including

113 with ALS, 17 with ALS-FTD, and 20 non-neurological controls. Clinical variables including age at symptom onset, disease status, age at death, onset site, sex, and contributing site, genetic information including the *C9orf72* repeat expansions, as well as sequencing variables including RIN, sequencing preparation method and sequencing platform were extracted from the database. Demographic and technical information for donors and samples was summarized in Supplementary Table 1. Individuals with ALS-FTD had shorter disease duration (from symptom onset to death), higher percentage of bulbar onset (50.0%), and higher frequencies of *C9orf72* repeat expansion compared to those with ALS.

**RNA-seq data preprocessing**

The Consortium's bulk RNA-seq data processing was previously described[8,34,35]. RNA was extracted from flash-frozen postmortem tissue, RIN was assessed using the Bioanalyzer, RNA-seq libraries were prepared from 500 ng of total RNA, and pooled libraries (average insert size, 375 bp) passing the quality criteria were sequenced on either an Illumina HiSeq 2500 (125-bp paired-end) or an Illumina NovaSeq (100-bp paired-end). For each sample (GRCh38/hg38 genome-mapped paired-end RNA-seq reads), gene expression was quantified using the "featureCounts" program (v2.0.2)[36]. Due to the large impact of RIN on expression, only samples with RIN ≥ 5 were included in this study. The expression for each gene was calculated as the sum of expression levels of individual exons. The raw expression for each exon was computed as the number of the paired reads assigned/mapped to the exon (only non-duplicated read pairs were counted towards the expression value). Multi-mapped reads were counted towards expression using weighted/fractional counts (1/x, where x is the number of genomic locations the read has been mapped to). The set of genes quantified for each sample was derived from the GENCODE v37 hg38 annotation using the Genotype Tissue Expression (GTEx) gene model pipeline (https://github.com/broadinstitute/gtex-pipeline). A total of 59,782 genes were quantified for each sample. Infrequently expressed genes were filtered using a threshold of >= 10 counts in at least 75% of samples[37]. We employed t-distributed stochastic neighbor embedding (t-SNE) to project this high-dimensional gene expression data into a two-dimensional space, facilitating the detection of potential isolated clusters within the data, potentially influenced by the contributing sites and RIN

(Supplementary Fig.1a,b). Normalized counts of the remained genes were VST transformed using DESeq2 package, and adjusted for contributing site and RIN using limma::removeBatchEffect from Limma package (Supplementary Fig.1c,d). This procedure was performed in lumbar and cervical spinal cord separately.

**Differential expression analyses**

The preprocessed RNA-seq data from the lumbar spinal cord was employed to conduct differential expression analyses using the DESeq2 package. We utilized a significance threshold of adjusted p-value < 0.05 to identify genes that were significantly differentially expressed between the individuals with ALS/ALS-FTD and controls (Supplementary Data 4).

**Cell-type deconvolution**

Cell-type deconvolution was conducted with Scaden, a deep neural network algorithm for cell type deconvolution of bulk RNA-seq samples[23]. Scaden model is trained on simulated bulk RNA-seq samples generated from snRNA-seq data. The cell fraction predictions can be compared to ground-truth cell composition, and is able to precisely and robustly deconvolve bulk RNA-seq data into cell types. We utilized two publicly available snRNA-seq datasets as training sets and established two models. One dataset comprises 24,190 cells from the most anterior part of the cervical spinal cord of one donor, which was downloaded from the Human Brain Cell Atlas (https://data.humancellatlas.org/hca-bio-networks/nervous-system/atlases/brain-v1-0)[38]. Another scRNA-seq dataset consisted of 55,289 cells from 7 donors of the lumbar spinal cord, downloaded from the Gene Expression Omnibus (GSE190442) using the count matrices and cell type labels provided[39]. Estimated proportions of each cell type were generated from two Scaden models separately, resulting in two different sets of predictions per sample. We reported the deconvolution estimates for neuron, astrocyte, microglia, oligodendrocyte, endothelial cell and ependymal cell in both lumbar and cervical spinal cord. Predictions for each estimated cell type proportions were highly correlated between these two snRNA-seq references (Supplementary Fig.2b).

**Weighted correlation network analysis**

The preprocessed gene expression data in lumbar spinal cord was used to establish co-expression networks using the WGCNA package[40]. A soft thresholding power was selected by pickSoftThreshold function to build a scale-free network with a scale-free topology fit index > 0.8. Based on this, we construct a topology overlap matrix (TOM), which was used as input for hierarchical clustering analysis. We used a dissimilarity threshold of 0.25 to merge the similar modules, and identified 22 modules for the lumbar spinal cord dataset (Supplementary Data 1). The module eigengene is the first principal component of a module, which can be considered as a representation of the gene expression profiles in a module. Of these, 13 modules showed significantly differential expression in the ALS/ALS-FTD group compared to controls, with 6 modules being down-regulated and 7 modules up-regulated (Fig.1a,b).

Module membership (MM) was calculated as the correlation between an individual gene and the ME of a module. Genes with high MM are strong representatives of the overall expression profile within their respective modules. Gene significance (GS) represents the correlation between an individual and group features. To identify key genes within core modules that may drive subtyping and staging, we calculated the GS using features such as subtype assignment and stage. Hub genes within core modules in current study were identified using a screening criterion of GS > 0.2 and MM > 0.8. We then applied the top 1,000 genes enriched in each cell type[41], and identified the overlapped genes between cell type specific genes and hub genes within each module.

We used the “modulePreservation” function of WGCNA package to evaluate whether the modules established in the lumbar spinal cord are preserved in the cervical spinal cord dataset. The module preservation statistics, Zsummary for each module was calculated via permutation testing (1,000 times), with Zsummary > 10 indicates high quality/preservation, 2 < Zsummary < 10 indicates moderate preservation, and Zsummary < 2 were considered non-preserved[42]. We found that the Zsummary of all 22 modules were higher than 10, then we projected the Lumbar-established modules in the cervical spinal cord dataset and calculated cervical module eigengenes.

**Functional enrichment analyses**

Genes with similar expression patterns were clustered into a module. They can be involved in same biological pathways. To test whether genes within each module map to relevant biological pathways, functional enrichment analyses were conducted separately for each module using gprofiler2 package in R[43]. It helps to identify GO terms that contain a statistically greater number of genes in common with the genes of each module than expected by chance. The g:SCS (Set Counts and Sizes) method was applied to compute multiple testing corrections for enrichment p-values[44]. Terms with g:SCS corrected p-value < 0.05 were reported as significant. We focused on "biological process" subontology of Gene Ontology (GO) (GO:BP) terms. To biologically annotate each module, we selected relevant biodomains that characterize the ALS/FTD endophenotypes[24,45]. Following the strategy of a previous study[46], we ignored those GO terms that were too general to be mapped into specific endophenotypic space, or mapped to high-order positions within the ontology. We primarily focused on GO terms that could be categorized into these relevant biodomains. The significant GO terms identified for each module were assigned to specific biodomains, thus each module is represented by specific biological pathways, as indicated by the top significant terms, with the majority of the terms within this module predominantly associated with specific biodomains.

By doing so, the identified GO terms were then assigned to biological domains, allowing 11 of the significant modules to be biologically annotated for specific molecular pathways (Fig.1g and Supplementary Data 2). The grey60 and royalblue modules were primarily enriched in the "Neuron Generation" biodomain. Additionally, the grey60 module was linked to the "Myelination" and "Gliogenesis" biodomains, while the royalblue module was enriched in the "Synapse" biodomain. The darkgreen module was enriched in the "Cytoskeleton" biodomain, and darkred module was primarily associated with "Mitochondrion" biodomain. The cyan module was involved in the "Proteostasis" biodomain, while black and lightyellow modules were linked to the "RNA Metabolism" biodomain. The tan module was involved in multiple biodomains including "Mitochondrion", "Proteostasis" and "RNA Metabolism". The magenta, midnightblue and salmon modules were primarily enriched in the "Immune System", with the salmon module also involved in "Apoptosis" biodomain.

**SuStaIn modelling of lumbar spinal cord dataset**

We expressed the MEs of each module as a z-score relative to the non-neurologic control group, while adjusting for age at death and sex. Because the z-scores of down-regulated modules became negative, so for simplicity we took the negative value of their z-scores. The SuStaIn algorithm combines clustering and disease progression modelling to identify subtypes with distinct patterns of transcriptomic progression. Using a 10-fold cross-validation, we selected the optimal number of subtypes based on the out-of-sample log-likelihood and cross-validation information criterion to better balance the model complexity with accuracy (Supplementary Fig.3b). The transcriptomic progression of each subtype was described as a linear z-score model based on an event-based model, in which each module follows a piecewise linear trajectory over a common timeframe[21]. The transcriptomic progression was described as a sequence of events, with each event corresponding to a module transitioning from a normal to abnormal expression levels. For the inclusion of z-score events, we utilized z-score waypoints of 1, 2, and 3 for each module, while excluding those z-score events where less than 10 subjects surpassed that z-score waypoint. The final point of this progression trajectory is marked by the maximum z-score, designated as 2, 3, or 5 depending on the maximum z-score event being 1, 2 or 3, respectively[21]. Each subtype initially showed evidence of altered gene expression in specific molecular pathways, followed by the sequential involvement in other pathways, occurring in varying orders across different subtypes. Each stage corresponded to a module reaching a new expression level, representing the dynamics of specific molecular pathways. For each subject, the SuStaIn model calculated a probability value to each subtype and stage. The assignment strength was determined by their maximum probability of belonging to one of the subtypes, thereby enabling their assignment to a specific subtype and stage within the trajectory of this subtype.

**Lumbar-trained SuStaIn model to cervical spinal cord dataset**

Bulk RNA-seq data from cervical spinal cord was preprocessed using same procedure as lumbar spinal cord, and withheld from the SuStaIn model calculations and then used in a secondary analysis to determine whether the transcriptomic subtypes and progression patterns identified in the lumbar spinal cord could be similarly detected in cervical spinal cord.

The Lumbar-established modules were highly preserved and showed similar expression patterns in cervical spinal cord. We projected this Lumbar-established module set to cervical spinal cord dataset, and calculated the cervical MEs. The Lumbar-trained SuStaIn model fitted to the cervical spinal cord dataset was used to infer transcriptomic subtype and stage in cervical spinal cord. Subtype stability was determined as the proportion of individuals who were either assigned to the same subtype or progressed from normal-appearing group to a SuStaIn subtype in another region.

**Non-negative matrix factorization (NMF) clustering**

To evaluate the performance of the SuStaIn model, we conducted a comparative analysis with a subtype-only model generated through unsupervised clustering using[15,47]. We utilized preprocessed RNA-seq data from the lumbar spinal cord for the NMF clustering analysis. The analysis was performed using the NMF package[48], with the following parameters: a factorization rank of 2, 200 iterations, and the NMF method set to nsNMF.

To ensure robust subtype assignment, 11 rounds of NMF clustering were performed. For each individual, the NMF subtype with a simple majority was selected. The basis matrix, representing the gene contributions to each component, was averaged across the 11 clustering replicates. To select representative genes for each subtype, genes were ranked by their contributions to the identified subtypes. We selected the top 2,000 genes for each subtype and identified unique and common genes of the NMF-derived subtypes. Subsequently, we conducted functional enrichment analysis on the unique and common genes to explore the biological annotations associated with the two nsNMF-derived subtypes.

**Statistical procedures**

The statistical analyses and plotting were conducted with R statistical software (version 4.2.0; R Foundation for Statistical Computing, Vienna, Austria). The normality of variable distribution was tested using the Shapiro-Wilk normality test. Continuous variables with normal distribution were compared using two-sample t-test, while Mann-Whitney test was utilized for comparing variables with non-normal distribution. For comparison of categorical variables, chi-squared test or Fisher exact test was employed. After identifying subtypes, we compared clinical and

genetic features, cell-type proportions, MEs of modules across transcriptomic subtypes. Correlation analyses were conducted between SuStaIn stages and clinical features, cell-type proportions, MEs of modules. All statistical analyses will be considered significant at a threshold of $p < 0.05$.

## Data availability

All raw RNA-seq data can be accessed via the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) database (GSE137810, GSE124439, GSE116622 and GSE153960). Two snRNA-seq datasets used are available in the Human Brain Cell Atlas (https://data.humancellatlas.org/hca-bio-networks/nervous-system/atlases/brain-v1-0), and GEO database (GSE190442). The brain cell type specific gene lists can be accessed at doi: 10.1038/s41598-018-27293-5. All RNA-seq data generated by the NYGC ALS Consortium are made immediately available to all members of the Consortium and with other consortia with which we have a reciprocal sharing arrangement. To request immediate access to new and ongoing data generated by the NYGC ALS Consortium and for samples provided through the Target ALS Postmortem Core, complete a genetic data request form at CGND_help@nygenome.org. Other datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request and approval from the Penn Neurodegenerative Data Sharing Committee. Requests may be submitted using a webform request: https://www.pennbindlab.com/data-sharing.

## Code availability

All packages and software described in the Methods section were used according to standard protocols. Software packages used herein include featureCounts v2.0.2 (https://subread.sourceforge.net/), GTEx gene model pipeline (https://github.com/broadinstitute/gtex-pipeline), Limma package (https://bioconductor.org/packages/release/bioc/html/limma.html), DESeq2 (https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html), WGCNA (https://cran.r-project.org/web/packages/WGCNA/index.html), gprofiler2 (https://cran.r-project.org/web/packages/gprofiler2/vignettes/gprofiler2.html), SuStaIn (https://github.com/ucl-pond/pySuStaIn), Scaden (https://github.com/KevinMenden/scaden), NMF package (https://cran.r-project.org/web/packages/NMF/index.html), Python v3.8.3 (https://www.python.org/), and R v4.3.1 (https://www.r-project.org/).

## Acknowledgements

This work was supported by NIH funding (P01AG066597, R01NS109260, P30AG072979), Penn Institute on Aging and DeCrane Fund for Primary Progressive Aphasia. The NYGC ALS Consortium was supported by ALSA Grant 19-SI-459 and by the Tow Foundation. Dr. Pavel P. Kuksa was supported by NIA funding including U24-AG041689 (NIAGADS) and U54-AG052427 (GCAD).

## Author contributions

T.S., and C.T.M. conceived and designed the study. B.S., P.P.K., and E.B.L. contributed to RNA-seq data analysis and cell-type deconvolution. V.V.D. contributed to genetic and clinical data analysis. H.P. contributed to data collection. T.S. and C.T.M. wrote and reviewed the paper. All authors edited the paper.

## Competing interests

The authors declare no competing interests.

## References


1 Taylor, J. P., Brown, R. H., Jr. & Cleveland, D. W. Decoding ALS: from genes to mechanism. *Nature* **539**, 197-206 (2016). https://doi.org:10.1038/nature20413

2 Wood, A., Gurfinkel, Y., Polain, N., Lamont, W. & Lyn Rea, S. Molecular Mechanisms Underlying TDP-43 Pathology in Cellular and Animal Models of ALS and FTLD. *Int J Mol Sci* **22** (2021). https://doi.org:10.3390/ijms22094705

3 Brown, A. L. *et al.* TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature* **603**, 131-137 (2022). https://doi.org:10.1038/s41586-022-04436-3

4 Liu, E. Y. *et al.* Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons. *Cell Rep* **27**, 1409-1421 e1406 (2019). https://doi.org:10.1016/j.celrep.2019.04.003

5 Zhao, W. *et al.* Characterization of Gene Expression Phenotype in Amyotrophic Lateral Sclerosis Monocytes. *JAMA Neurol* **74**, 677-685 (2017). https://doi.org:10.1001/jamaneurol.2017.0357

6 D'Erchia, A. M. *et al.* Massive transcriptome sequencing of human spinal cord tissues provides new insights into motor neuron degeneration in ALS. *Sci Rep* **7**, 10046 (2017). https://doi.org:10.1038/s41598-017-10488-7

7 Daneshafrooz, N., Bagherzadeh Cham, M., Majidi, M. & Panahi, B. Identification of potentially functional modules and diagnostic genes related to amyotrophic lateral sclerosis based on the WGCNA and LASSO algorithms. *Sci Rep* **12**, 20144 (2022). https://doi.org:10.1038/s41598-022-24306-2

8 Humphrey, J. *et al.* Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. *Nat Neurosci* **26**, 150-162 (2023). https://doi.org:10.1038/s41593-022-01205-3

9 Hasan, R. *et al.* Transcriptomic analysis of frontotemporal lobar degeneration with TDP-43 pathology reveals cellular alterations across multiple brain regions. *Acta Neuropathol* **143**, 383-401 (2022). https://doi.org:10.1007/s00401-021-02399-9

10 Li, J. *et al.* Divergent single cell transcriptome and epigenome alterations in ALS and FTD patients with C9orf72 mutation. *Nat Commun* **14**, 5714 (2023). https://doi.org:10.1038/s41467-023-41033-y

11 Yazar, V. *et al.* Impaired ATF3 signaling involves SNAP25 in SOD1 mutant ALS patients. *Sci Rep* **13**, 12019 (2023). https://doi.org:10.1038/s41598-023-38684-8

12 Tan, H. H. G. *et al.* MRI clustering reveals three ALS subtypes with unique neurodegeneration patterns. *Ann Neurol* (2022). https://doi.org:10.1002/ana.26488

13 Bede, P., Murad, A., Lope, J., Hardiman, O. & Chang, K. M. Clusters of anatomical disease-burden patterns in ALS: a data-driven approach confirms radiological subtypes. *J Neurol* **269**, 4404-4413 (2022). https://doi.org:10.1007/s00415-022-11081-3

14 Young, A. L. *et al.* Data-driven neuropathological staging and subtyping of TDP-43 proteinopathies. *Brain* **146**, 2975-2988 (2023). https://doi.org:10.1093/brain/awad145

15 Eshima, J. *et al.* Molecular subtypes of ALS are associated with differences in patient prognosis. *Nat Commun* **14**, 95 (2023). https://doi.org:10.1038/s41467-022-35494-w

16 Whitwell, J. L. *et al.* Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: a cluster analysis study. *Brain* **132**, 2932-2946 (2009). https://doi.org:10.1093/brain/awp232

17 Noh, Y. *et al.* Anatomical heterogeneity of Alzheimer disease: based on cortical thickness on MRIs. *Neurology* **83**, 1936-1944 (2014). https://doi.org:10.1212/WNL.0000000000001003

18 Fonteijn, H. M. *et al.* An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage* **60**, 1880-1889 (2012).

https://doi.org:10.1016/j.neuroimage.2012.01.062
19 Eshaghi, A. *et al.* Progression of regional grey matter atrophy in multiple sclerosis. *Brain* **141**, 1665-1677 (2018). https://doi.org:10.1093/brain/awy088
20 Oxtoby, N. P. *et al.* Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* **141**, 1529-1544 (2018). https://doi.org:10.1093/brain/awy050
21 Young, A. L. *et al.* Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* **9**, 4273 (2018). https://doi.org:10.1038/s41467-018-05892-0
22 Shen, T. *et al.* Novel data-driven subtypes and stages of brain atrophy in the ALS-FTD spectrum. *Transl Neurodegener* **12**, 57 (2023). https://doi.org:10.1186/s40035-023-00389-3
23 Menden, K. *et al.* Deep learning-based cell composition analysis from tissue expression profiles. *Sci Adv* **6**, eaba2619 (2020). https://doi.org:10.1126/sciadv.aba2619
24 Le Gall, L. *et al.* Molecular and Cellular Mechanisms Affected in ALS. *J Pers Med* **10** (2020). https://doi.org:10.3390/jpm10030101
25 van Rheenen, W. *et al.* Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. *PLoS One* **13**, e0198874 (2018). https://doi.org:10.1371/journal.pone.0198874
26 Grima, N. *et al.* RNA sequencing of peripheral blood in amyotrophic lateral sclerosis reveals distinct molecular subtypes: Considerations for biomarker discovery. *Neuropathol Appl Neurobiol* **49**, e12943 (2023). https://doi.org:10.1111/nan.12943
27 Namboori, S. C. *et al.* Single-cell transcriptomics identifies master regulators of neurodegeneration in SOD1 ALS iPSC-derived motor neurons. *Stem Cell Reports* **16**, 3020-3035 (2021). https://doi.org:10.1016/j.stemcr.2021.10.010
28 Beland, L. C. *et al.* Immunity in amyotrophic lateral sclerosis: blurred lines between excessive inflammation and inefficient immune responses. *Brain Commun* **2**, fcaa124 (2020). https://doi.org:10.1093/braincomms/fcaa124
29 Zhang, W., Xiao, D., Mao, Q. & Xia, H. Role of neuroinflammation in neurodegeneration development. *Signal Transduct Target Ther* **8**, 267 (2023). https://doi.org:10.1038/s41392-023-01486-5
30 Morello, G. *et al.* Integrative multi-omic analysis identifies new drivers and pathways in molecularly distinct subtypes of ALS. *Sci Rep* **9**, 9968 (2019). https://doi.org:10.1038/s41598-019-46355-w
31 Liu, J. & Wang, F. Role of Neuroinflammation in Amyotrophic Lateral Sclerosis: Cellular Mechanisms and Therapeutic Implications. *Front Immunol* **8**, 1005 (2017). https://doi.org:10.3389/fimmu.2017.01005
32 Young, A. L. *et al.* Characterizing the Clinical Features and Atrophy Patterns of MAPT-Related Frontotemporal Dementia With Disease Progression Modeling. *Neurology* **97**, e941-e952 (2021). https://doi.org:10.1212/WNL.0000000000012410
33 Vogel, J. W. *et al.* Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med* **27**, 871-881 (2021). https://doi.org:10.1038/s41591-021-01309-6
34 Tam, O. H. *et al.* Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Rep* **29**, 1164-1177 e1165 (2019). https://doi.org:10.1016/j.celrep.2019.09.066
35 Prudencio, M. *et al.* Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. *J Clin Invest* **130**, 6080-6092 (2020). https://doi.org:10.1172/JCI139741
36 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014). https://doi.org:10.1093/bioinformatics/btt656
37 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data

with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org:10.1186/s13059-014-0550-8

38 Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023). https://doi.org:10.1126/science.add7046

39 Yadav, A. *et al.* A cellular taxonomy of the adult human spinal cord. *Neuron* **111**, 328-344 e327 (2023). https://doi.org:10.1016/j.neuron.2023.01.007

40 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). https://doi.org:10.1186/1471-2105-9-559

41 McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci Rep* **8**, 8868 (2018). https://doi.org:10.1038/s41598-018-27293-5

42 Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is my network module preserved and reproducible? *PLoS Comput Biol* **7**, e1001057 (2011). https://doi.org:10.1371/journal.pcbi.1001057

43 Kolberg, L. *et al.* g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res* **51**, W207-W212 (2023). https://doi.org:10.1093/nar/gkad347

44 Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193-200 (2007). https://doi.org:10.1093/nar/gkm226

45 Mol, M. O., Miedema, S. S. M., van Swieten, J. C., van Rooij, J. G. J. & Dopper, E. G. P. Molecular Pathways Involved in Frontotemporal Lobar Degeneration with TDP-43 Proteinopathy: What Can We Learn from Proteomics? *Int J Mol Sci* **22** (2021). https://doi.org:10.3390/ijms221910298

46 Cary, G. A. *et al.* Genetic and multi-omic risk assessment of Alzheimer's disease implicates core associated biological domains. *Alzheimers Dement (N Y)* **10**, e12461 (2024). https://doi.org:10.1002/trc2.12461

47 Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**, 4164-4169 (2004). https://doi.org:10.1073/pnas.0308531101

48 Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010). https://doi.org:10.1186/1471-2105-11-367

## Table, Figures, Legends

**Table 1.** Comparison of clinical characteristics between SuStaIn subtypes.

| | **S0: Normal-appearing group** | **S1: Immune/Apoptosis/ Proteostasis subtype** | **S2: Synapse/RNA-Metabolism subtype** | **Missing data** | $p_{S0vS1}$ | $p_{S0vsS2}$ | $p_{S1vsS2}$ |
|---|---|---|---|---|---|---|---|
| **Subject, n** | 28 | 75 | 37 | - | - | - | - |
| **Age at Death, years** | 66.6 (8.1) | 63.8 (10.1) | 62.0 (12.3) | 0 (0.0%) | 0.10 | **0.03** | 0.44 |
| **Age at symptom onset, years** | 62.3 (10.1) | 60.8 (10.5) | 58.7 (12.6) | 4 (%) | 0.30 | 0.14 | 0.40 |
| **Sex, male%** | 20 (71.4%) | 42 (56.0%) | 16 (43.2%) | 0 (0.0%) | 0.15 | **0.02** | 0.20 |
| **Disease duration, months** | 50.9 (36.9) | 34.9 (24.4) | 42.2 (30.3) | 5 (%) | **0.04** | 0.33 | 0.28 |
| **Clinical phenotypes** | n=28 | n=75 | n=37 | 0 (0.0%) | 1.00 | 0.45 | 0.38 |
| ALS | 24 (85.7%) | 64 (85.3%) | 34 (91.9%) | | | | |
| ALS-FTD | 4 (14.3%) | 11 (14.7%) | 3 (8.1%) | | | | |
| **Symptom onset site** | n=28 | n=72 | n=34 | 6 (%) | 0.77 | 0.18 | 0.20 |
| Bulbar | 11 (39.3%) | 26 (36.1%) | 8 (23.5%) | | | | |
| Non-bulbar | 17 (60.7%) | 46 (63.9%) | 26 (76.5%) | | | | |
| ***C9orf72*** | n=26 | n=54 | n=32 | 28 (%) | 0.10 | 0.12 | 0.97 |
| Positive | 3 (11.5%) | 15 (27.8%) | 9 (28.1%) | | | | |
| Negative | 23 (88.5%) | 39 (72.2%) | 23 (71.9%) | | | | |
| **SuStaIn stage** | 0 (0.0) | 5.5 (4.8) | 7.4 (5.0) | 0 (0.0%) | - | - | **0.03** |

Note. Continuous variables are presented as mean (standard deviation), and compared with two-sample t-test or Mann-Whitney test. Categorical variables are presented as number (percentage), and compared with Chi-square test or Fisher's exact test. Missing data indicates the percentage of individuals with missing data. P-values < 0.05 are bolded. ALS, amyotrophic lateral sclerosis; ALS-FTD, amyotrophic lateral sclerosis-frontotemporal degeneration; SuStaIn, Subtype and Stage Inference.

## Figures

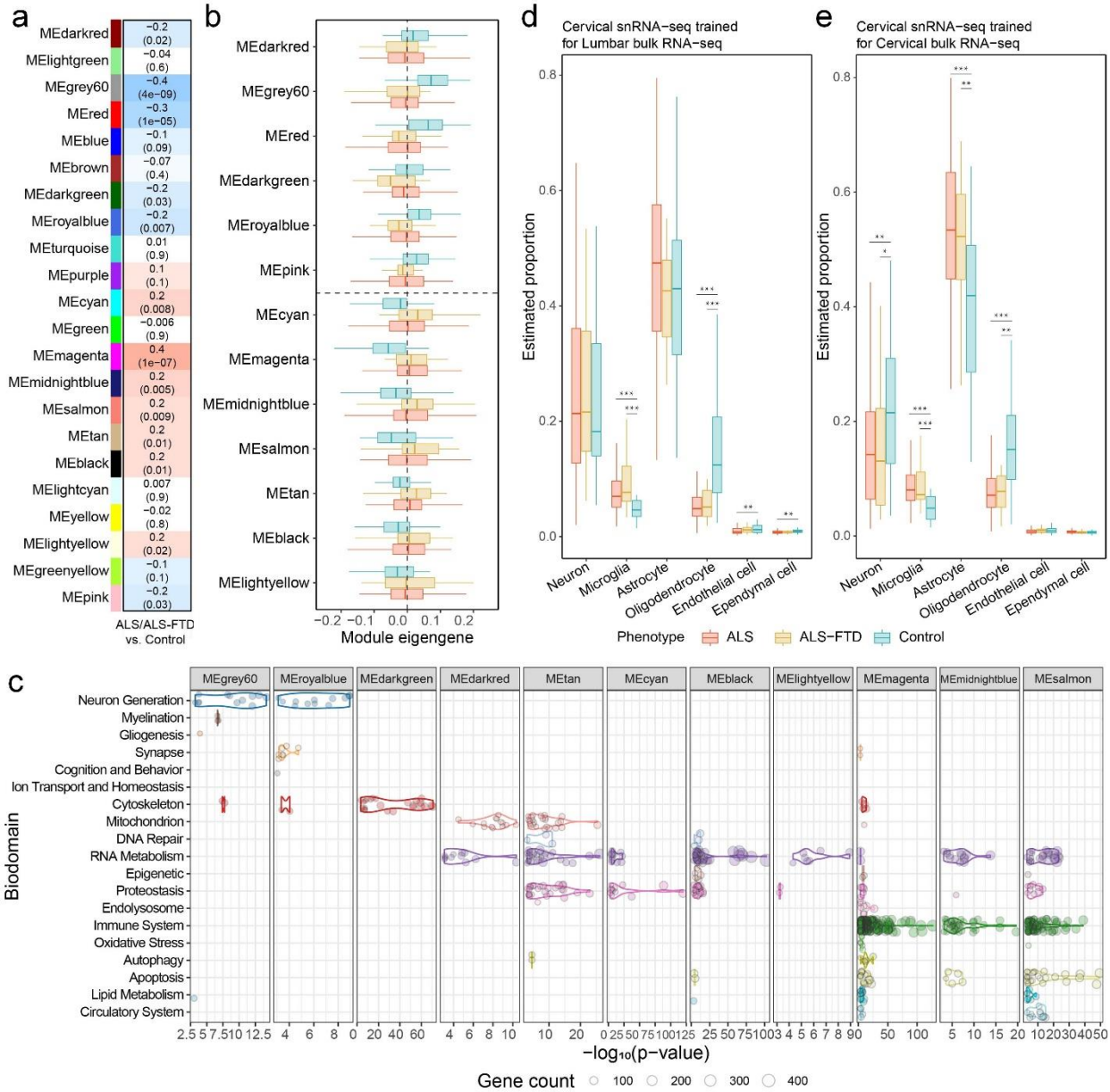


**Fig.1 Results of co-expression module construction and cell-type deconvolution. a,** Expression patterns of the 22 modules in ALS/ALS-FTD patients compared to controls. The numbers at the top denote the t-statistics derived from the comparative analysis, whereas the numbers at the bottom signify the p-values. Blue represents down-regulated and red indicates up-regulated modules in ALS/ALS-FTD relative to controls. **b,** Module eigengenes of 13 significantly differentially expressed modules across different clinical phenotypes. These modules exhibit significant expression differences (p-value < 0.05) between ALS/ALS-FTD and controls. **c,** Biological annotation of each module. Each dot represents a GO term within a given biodomain. Color of the dots represent different biodomains. The x-axis indicates the significance of term enrichment, and the dot size indicates the number of genes in that GO term and in the module. **d-e,** Cell type proportions estimated by using cervical snRNA-seq reference in bulk RNA-seq sample from different regions across different clinical phenotypes. * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001.

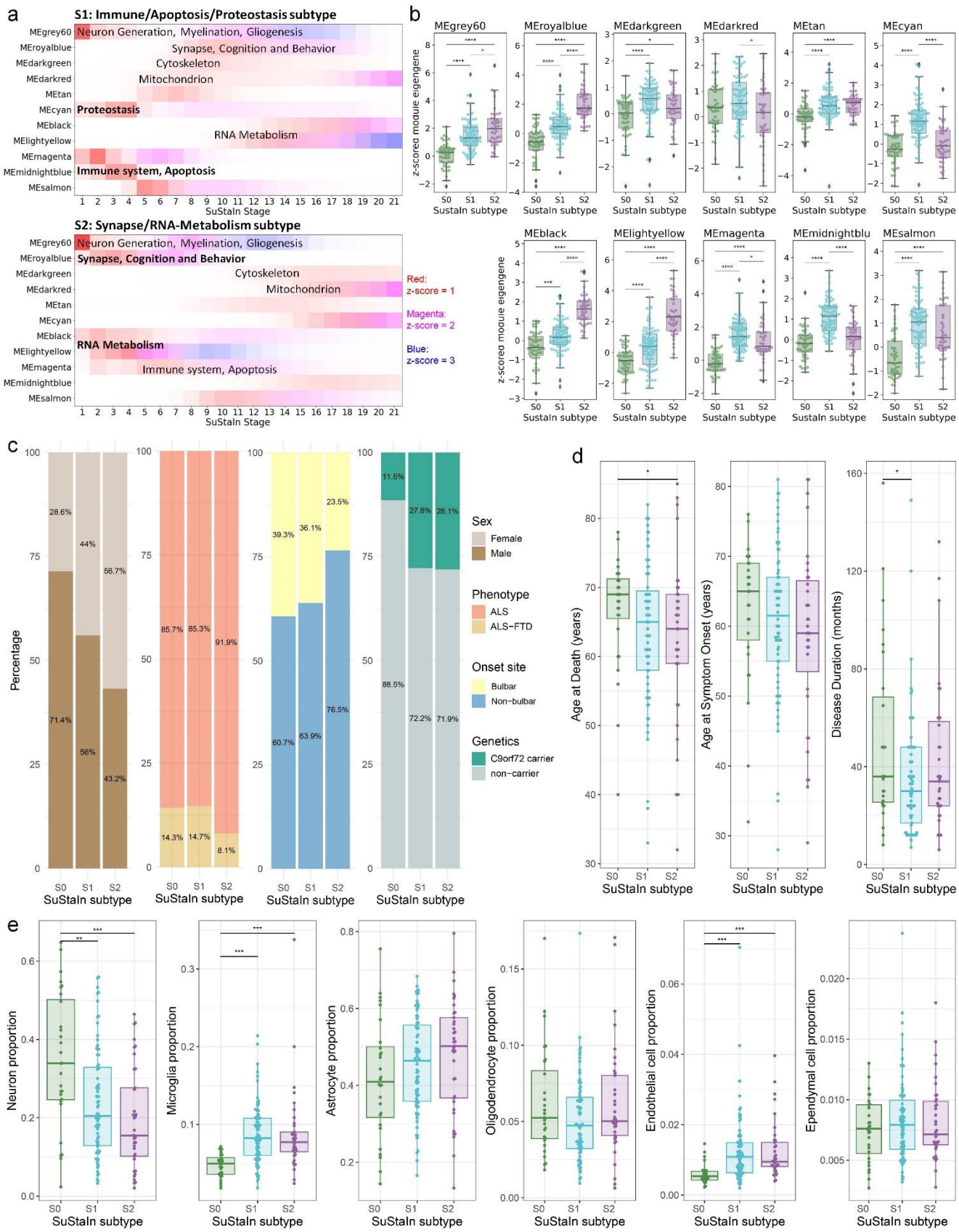


**Fig.2 The two transcriptomic subtypes identified by SuStaIn algorithm. a,** Progression trajectories of modules across subtypes. Z-scores represent the dynamic changes of each module along SuStaIn stages. Color gradient indicates the probability of abnormality severity: red (mild), magenta (moderate),

blue (severe). **b,** Comparisons of z-scored module eigengenes across SuStaIn subtypes for each module. The z-scores of down-regulated modules were multiplied by -1 to ensure that all biomarkers used as inputs for the SuStaIn model are elevated in disease. Each module is labeled with its predominant and specific bidomain characteristics. **c,** Bar chart illustrating the distinct distribution of clinical and genetic features across subtypes. **d,** Transcriptomic subtypes show significantly younger age and shorter disease duration. **e,** Cellular composition differences across subtypes. Cell type proportions shown in this figure were estimated by using cervical snRNA-seq reference. S0, Normal-appearing group; S1, Immune/Apoptosis/Proteostasis subtype; S2, Synapse/RNA-Metabolism subtype. ALS, amyotrophic lateral sclerosis; ALS-FTD, amyotrophic lateral sclerosis-frontotemporal degeneration; SuStaIn, Subtype and Stage Inference. * indicates $p<0.05$, ** indicates $p<0.01$, *** indicates $p<0.001$, **** indicates $p<0.0001$.

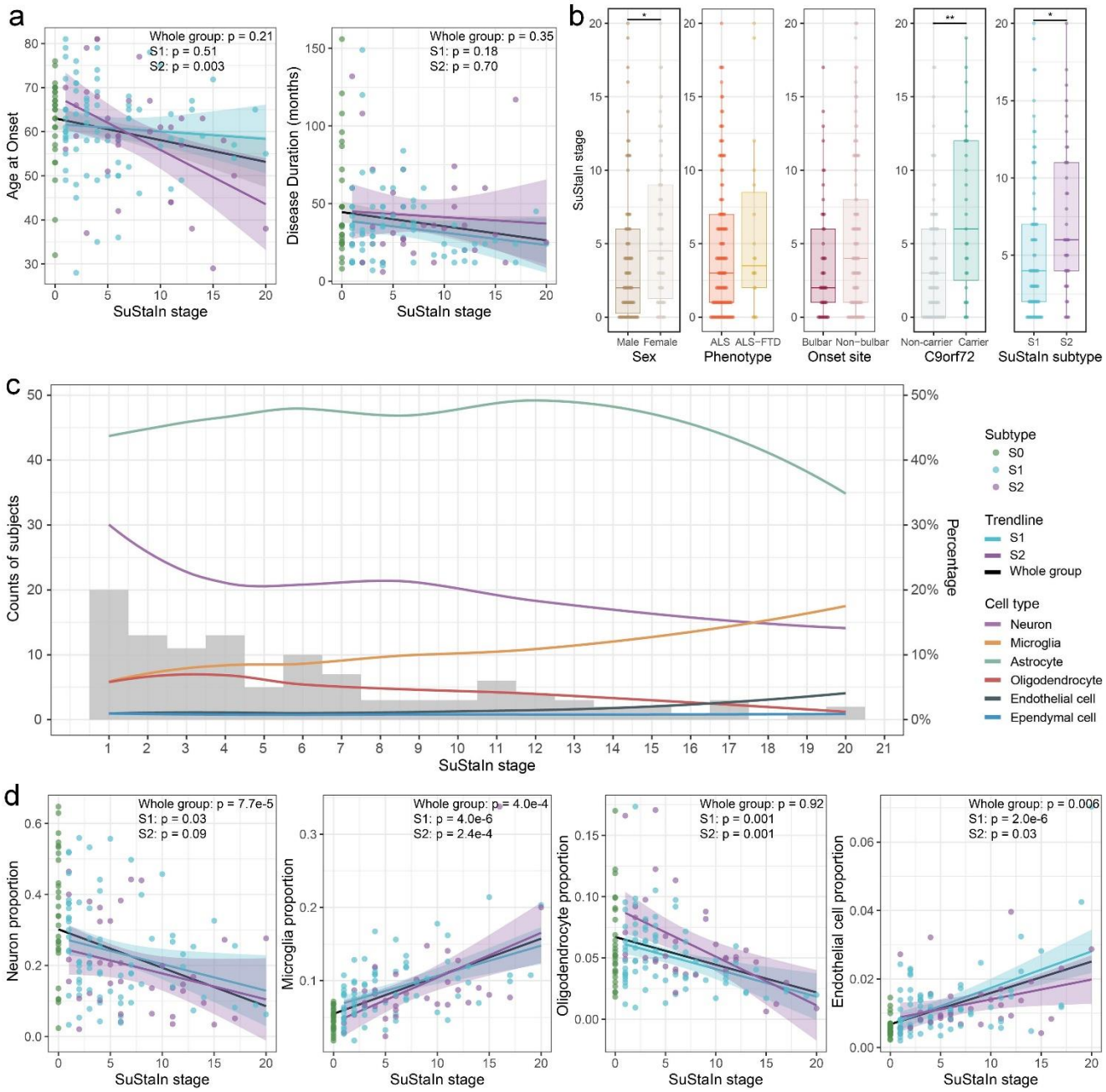


**Fig.3 Description of SuStaIn progression trajectory. a,** SuStaIn stage correlated with age at onset. **b,** Boxplots comparing SuStaIn stage across various demographic and clinical subgroups: sex, phenotype, onset site, *C9orf72* genotype, and SuStaIn subtype. * indicates p<0.05, ** indicates p<0.01. **c,** Distribution of subjects across SuStaIn stages and cellular composition at different stages. The grey bars represent the count of subjects at each stage. The line graphs illustrate the trend lines for the whole group, indicating the percentage of each cell type across the stages. **d,** SuStaIn stage correlated with different cell type proportions. S0, Normal-appearing group; S1, Immune/Apoptosis/Proteostasis subtype; S2, Synapse/RNA-Metabolism subtype. SuStaIn, Subtype and Stage Inference.

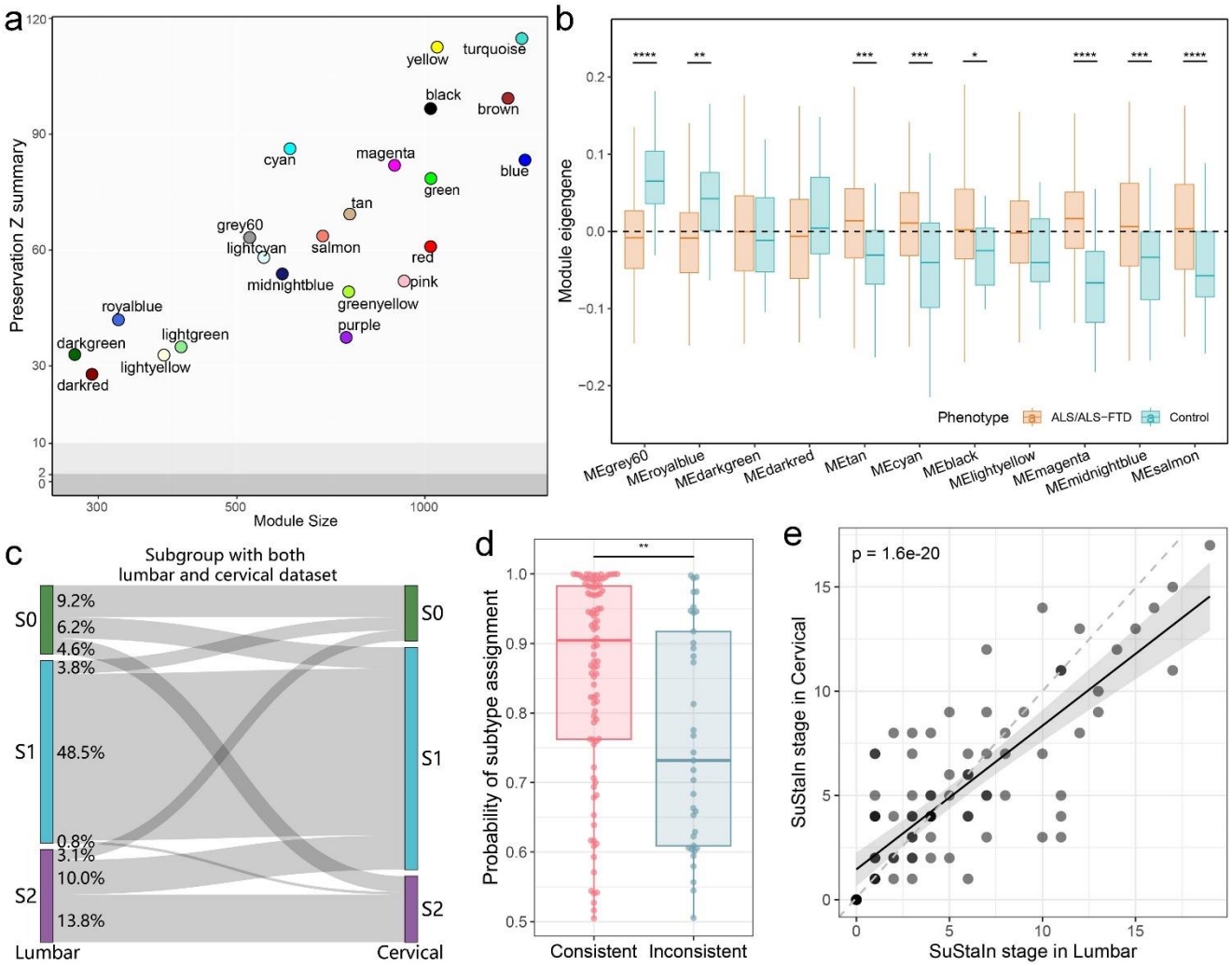


**Fig.4 Evaluation of Lumbar-trained SuStaIn model in testing dataset. a,** Preservation of the Lumbar-constructed modules in the cervical spinal cord. The Zsummary statistics indicate the degree of module preservation, with modules having Zsummary > 10 considered highly preserved. **b,** Comparisons of module eigengenes derived from cervical dataset across phenotypes. **c,** Subtype assignment between lumbar and cervical datasets. **d,** The subtype probability is higher in individuals with consistent subtype assignments across lumbar and cervical datasets. **e,** SuStaIn stages in the lumbar are highly correlated with stages in cervical samples for individuals with consistent subtype assignments. S0, Normal-appearing group; S1, Immune/Apoptosis/Proteostasis subtype; S2, Synapse/RNA-Metabolism subtype. * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001, **** indicates p<0.0001.

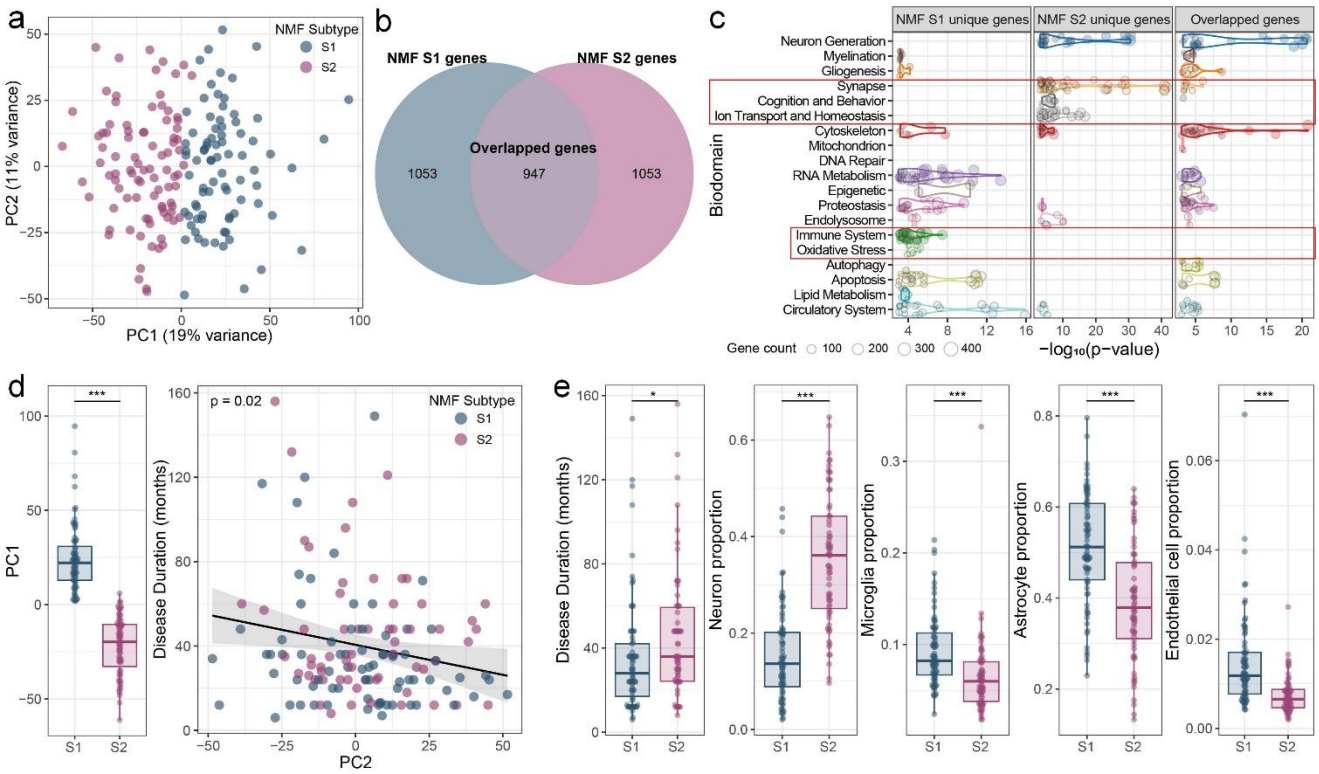


**Fig.5 The two transcriptomic subtypes identified by subtype-only model. a,** PCA shows the clustering of samples based on NMF subtypes. **b,** Venn diagram illustrates the unique and shared top contributing genes of two NMF subtypes. **c,** Biological annotation of subtype specific and shared genes. Each dot represents a GO term within a given biodomain. Color of the dots represent different biodomains. The x-axis indicates the significance of term enrichment, and the dot size indicates the number of genes in that GO term and in the top contributing genes of two subtypes. **d,** Relationships of principal components with NMF subtyping and disease duration. **e,** Distinct clinical features and cell type proportions between NMF subtypes. PCA, principal component analysis; NMF, non-negative matrix factorization; S1, Immune subtype; S2, Synapse subtype. * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001, **** indicates p<0.0001.

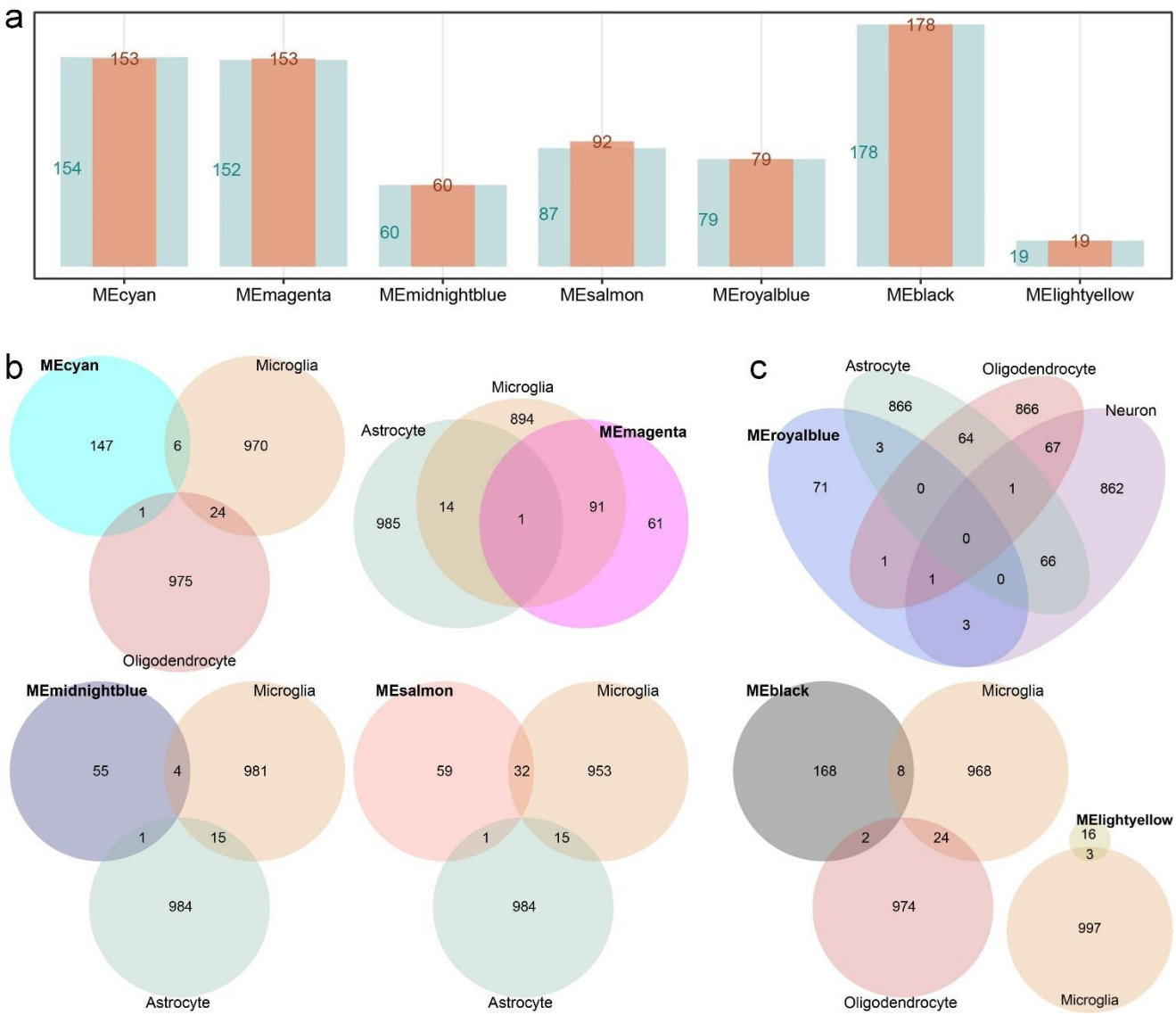


**Fig.6 Hub genes within core modules associated with subtyping and staging. a,** A substantial overlap between hub genes associated with subtyping and staging in each module. The light blue bars represent hub genes associated with subtyping, and the orange bars represent hub genes associated with staging. The overlap indicates a significant intersection between the two sets of genes in each module, with the numbers adjacent to each bar corresponding to the count of hub genes within that module. **b,** Venn diagrams illustrating the overlap between hub genes and cell type-specific gene markers across the core modules.

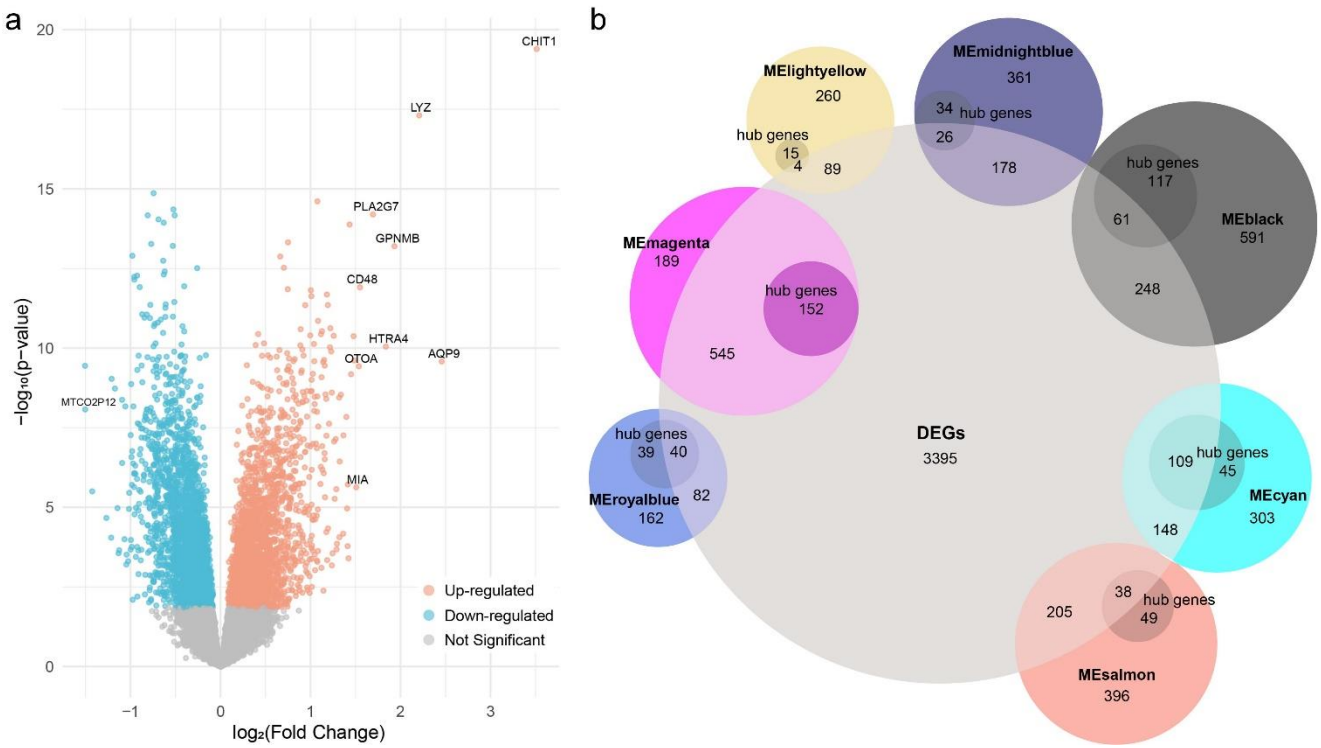


**Fig.7 Differentially expressed genes (DEGs) in ALS. a,** Volcano plot showing DEGs identified in individuals with ALS/ALS-FTD relative to controls. **b,** Overlap of DEGs with core modules. For each module, the outer circle represents the total number of genes in the module, while the inner circle represents the hub genes of that module.