A network motif-enrichment approach to mapping genetic dysregulation in Amyotrophic Lateral Sclerosis

Samuel Sledzieski[1] and Malek Kabani[2]

[1] Computer Science and Artificial Intelligence Lab., MIT, Cambridge, MA 02139

[2] Department of Biological Engineering, MIT, Cambridge, MA 02139

Abstract. We analyze 21 tissue samples from the NeuroLINCS project to investigate the regulatory mechanisms underlying sporadic amyotrophic lateral sclerosis (ALS). Using a combination of regulatory network reconstruction, differential gene expression analysis, and network motif enrichment analysis, we identify and highlight 14 genes and three pathways which are perturbed in ALS, spanning motor neuron development, myosin binding, and oxidative stress. Our multi-pronged approach considers multiple sources of variation in disease and combines orthogonal sources of information to untangle complicated regulatory dynamics and identify sources of genetic dysregulation.

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) is the most common adult motor neuron disease and is characterized by the degeneration of the upper and lower motor neurons [1]. This selective loss in motor neurons results in muscle atrophy and weakness and makes the onset of ALS one of the most fatal neurodegenerative disorders with a 2 to 4-year survival after diagnosis [2]. Despite decades of intense research, there is currently no cure for ALS, and most treatments focus on symptom management and pain alleviation. To understand this apparent slow progress in developing therapeutics, we need to look at the etiology of ALS. Approximately 5-10% of all ALS cases are familial ALS (fALS), where genetic variants are inherited autosomal dominant [3]. The critical genetic players in fALS have been undercovered and are mainly implicated in RNA regulation, autophagy, and vesicle formation. Of the 50 causatives and disease regulating genes identified, TARDBP, SQSTM1, VCP, FUS, TBK1, and C9orf72 are the most prevalent in fALS cases.

The other 90-95% of the cases are called sporadic ALS (sALS) and occur in individuals with no familial history of ALS. However, due to its apparent clinical heterogeneity, the etiology of sALS remains poorly understood, and the identification of genetic risk factors is still a work in progress [4]. Thus, it is crucial for us to understand the cellular mechanisms of ALS and its causes and functions in sporadic cases. To bridge the gap between sporadic and familial ALS, we need to look at the changes in molecular functions of these gene variations. ALS-associated genes play a crucial role in ALS pathology and alter several cellular processes such as RNA metabolism, DNA repair, oxidative stress, mitochondrial dysfunction, and nucleocytoplasmic transport defects. By constructing gene regulatory networks and identifying differentially expressed genes in ALS patients, we leverage computational biology tools to identify altered pathways in ALS. This procedure is a robust framework well-suited for hypothesis generation regarding ALS etiology and progression.

We propose to embrace the complexity of ALS to analyze the heterogeneity between patients and reveal potential new biomarkers and therapeutic targets. In this study, we analyze publicly available transcriptomic data from the NeuroLINCS initiative. To gain more insight on ALS etiology, we adopt a multi-pronged approach that integrates the study of differentially expressed genes (DEGs) in ALS and control iPSC and induced motor neurons (iMN) samples with a study of regulatory networks, network motifs, and network rewiring. We first reconstruct regulatory networks using RNA-seq from both ALS and control tissue samples. We then search for genes which are differentially expressed in ALS, and for regulatory motifs which are enriched in ALS. Finally, we combine our insights to identify a narrow subset of the regulatory network which significantly varies in disease. We explore three cases studies where network rewiring and dysregulation occur and are linked to known ALS markers in the literature. This integrated approach elucidates ALS-significant pathways, and will allow us to develop a comprehensive list of diagnostic biomarkers and subsequent potential novel therapeutic targets.

2 Methods

2.1 NeuroLINCS ALS Data

We sourced our data from the NeuroLINCS project, a study of neurodegenerative diseases which focuses on ALS and spinal muscular atrophy (SMA). The NeuroLINCS project contains multi-omic data from several cell lines sourced from ALS and SMA affected individuals, as well as control individuals. They measure transcriptomic data with RNA-seq, proteomics using SWATH-MS, and epigenomics with ATAC-seq. From the donor samples, they generated induced pluripotent stem cells (iPSCs), which they then artificially differentiated into induced motor neurons (iMNs) using two different (short and long) protocols. For this study, we will focus on the control and ALS samples from the iPSCs and the iMNs generated using the long differentiation protocol. Table 1 shows the number of different samples obtained from each cell line for each of the three data modalities. We will primarily focus on transcriptomic measurements [5,6] to achieve our aims, leaving proteomic and epigenomic for future multi-omic analysis.

Number of samples		**RNA-seq**	**SWATH-MS**	**ATAC-seq**
Induced Pluripotent Stem Cells (iPSC)	Control	9	9	3
	ALS	12	11	3
Induced Motor Neuron (iMN)	Control	6	8	3
	ALS	8	8	4

Table 1. Summary of available data from the NeuroLINCS initiative. This data set contains transcriptomic (RNA-seq), proteomic (SWATH-MS), and metagenomic (ATAC-seq) samples from induced pluripotent stem cells (iPSCs) and fully differentiated induced motor neurons (iMNs) derived from ALS and control patients. This table shows the number of samples of each cell line and data modality.

2.2 Transcriptomic Pre-processing

Transcriptomic data was retrieved from the NeuroLINCS project as raw counts. We followed standard pre-processing and quality control pipelines to clean our data, which allows for more accurate network reconstruction and differential expression analysis. Specifically, we used Scanpy [7] which, while designed for single-cell analysis, contains standard workflows that are applicable to bulk data and provides useful data structures for analysis. We filter out genes which appear in fewer than 3 samples, perform normalization and log transformation, and annotate mitochondrial and highly-variable genes. Additionally, we perform a mapping from the given Ensembl gene IDs to HGNC identifiers using the Biomart package, which allows for easier downstream interpretation of results.

2.3 Network reconstruction

Understanding regulatory networks and their variance is essential to deciphering the cellular impact of the disease, and allow us to conduct our downstream analyses. We reconstruct four regulatory networks — using transcriptomic data from ALS and control samples of both iPSC and iMN tissues. Networks were reconstructed using GENIE3 [8], which uses tree-based methods to infer relationships between pairs of genes. GENIE3 allows for the specification of a list of transcription factors, which are used to limit which nodes can be sources of regulatory activity. We use the list of transcription factors identified in Lambert et al. [9] for this step, and following network inference we filter the network by these transcription factors for clarity of analysis, resulting in a 2,649 gene network. GENIE3 returns a weighted adjacency matrix M, where element $M_{i,j}$ corresponds to the strength of the inferred regulatory relationship between transcription factor i and gene j. We binarize this matrix at the 99.9^{th} percentile, meaning we keep only the top 0.1% of all pairwise values as true edges within the network. This threshold was chosen to generate reasonably sized networks on which we could feasibly construct our downstream analyses, and to focus on only the strongest regulatory relationships.

2.4 Differential expression analysis

We perform two separate analyses of differential expression, at the individual gene level and at the pathway level. First we use identify the differentially expressed genes in ALS vs. control data using the DESeq2 R package [10]. We use the default settings for DESeq2, and we use the un-normalized counts rather than the pre-processed transcriptomic data as recommended in the DESeq2 manual. To analyze expression at the pathway level, we perform a gene set enrichment analysis using GSEA [11,12], again with the default settings. The individual differentially addressed genes will allow us to narrow our search for interesting sub-networks and motifs, and the gene set enrichment analysis will allow us to link our analysis to the gene ontology and to biological pathways.

2.5 Network motif analysis

Network motifs have been well-characterised as critical and recurring structures in regulatory networks [13]. By conducting a large-scale analysis of the frequency and rewiring of the regulatory motifs that occur in ALS, we aim to understand the key cellular changes that increase disease risk and progression. For each of the four networks, we identify motifs using mFinder [14,15]. mFinder analyzes the input network for motifs and their frequencies, and compares with a probability distribution generated from motif frequencies in random networks. Thus, we can identify motifs which appear in the input network more often than would be expected by random. We used the default settings for mFinder, which generates 100 random networks, and looked for motifs of size 3, 4, and 5. These motif sizes should capture most meaningful variation in network structure while remaining computationally feasible to search — the theoretical number of possible motifs of sizes 3,4, and 5 are 13, 199, and 9,394 respectively, while there are over 1.5 million possible 6-node motifs (OEIS Sequence A003085) [16]. Additionally, we can look for evidence of network rewiring by comparing motifs across out four networks. Once interesting motifs have been identified, we can focus our search further by filtering to differentially expressed genes which are present in interesting motifs. As a result, we bridge both transcriptomic changes and network structure changes that occur during ALS.

3 Results

3.1 Regulatory networks in ALS

We reconstructed four networks with GENIE3 using the RNA-seq expression profiles of iPSC and iMN tissue samples from ALS and control patients. A Cytoscape project with all four networks can be found on the GitHub repository for this study (see Key Resource Table, Table 4). The gene set used to define nodes

Network	Nodes	Edges	Diameter	Components	Clustering coefficient	Density
iPSC, Control	2171	3651	10	53	**0.013**	0.001
iPSC, ALS	2258	4400	12	45	**0.024**	0.001
iMN, Control	2262	3833	10	42	**0.006**	0.001
iMN, ALS	1957	3463	11	81	**0.012**	0.001

Table 2. Statistics of inferred regulatory networks. We reconstructed four regulatory networks using GENIE3. By design, these networks contain similar numbers of nodes and edges because they were constructed over the same node set, and we used a percentile threshold to select positive edges. More interestingly, we find that the iPSC and ALS networks tend to be more highly clustered (bolded) than the iMN and control networks respectively, even at similar densities.

was filtered using the set of human transcription factors, and the highest-scoring regulatory relationships were selected as edges (see Section 2.3). The networks were similar characteristically, with similar numbers of nodes, edges, a similar density (0.001), and similar diameters (10-12). The networks also had mostly similar numbers of connected components, between 42 and 53, with the exception of the ALS iMN network which had 81 connected components. While the number of connected components can vary due to noise in the reconstruction and the set of nodes which are selected, this higher number in the iMN ALS network is potentially worth further investigation, as it may indicate small clusters of regulatory genes which have become disconnected from the primary regulatory mechanism of the cells.

The most notable variance between networks is in their clustering coefficients. In both the iPSC and iMN networks, the ALS network has a clustering coefficient roughly double that of the control network, indicating more tightly clustered regulatory relationships. We hypothesize that this tigheter clustering may be due to network rewiring which creates new disfunctional regulatory relationships. Such rewiring would be revealed by more instances of tightly-connected network motif relationships in the ALS network, which we explore further in Section 3.3. While a comparison of iPSCs vs. iMNs is outside of the scope of this report, we note that the iPSC networks also have clustering coefficients roughly double their iMN counterparts.

3.2 Differentially expressed genes in ALS

The DESeq2 analysis resulted in 5,301 genes which were diferentially expressed with adjusted p-value threshold $p \leq 0.05$, and 470 which were differentially expressed at the more stringent p-value threshold of $p \leq 1e{-}5$. In Figure 1, we show the 30 most significant differentially expressed genes by adjusted p-value, sorted by log-fold change. Of these, four stand out – *NPIPB15*, *CHP2*, and *PVALB*, which have adjusted p-values several orders of magnitude smaller than the next genes, and *GSTM1*, which has a absolute log-fold change more than 4 times the next largest. The downregulated genes, *CHP2* and *PVALB* have previously been identified as ALS-linked in the literature. *PVALB*, which regulates intracellular calcium release, was identified as a candidate marker for ALS-resistant motor neurons in Lederer et al. [18] and *CHP2*, a calcium-binding

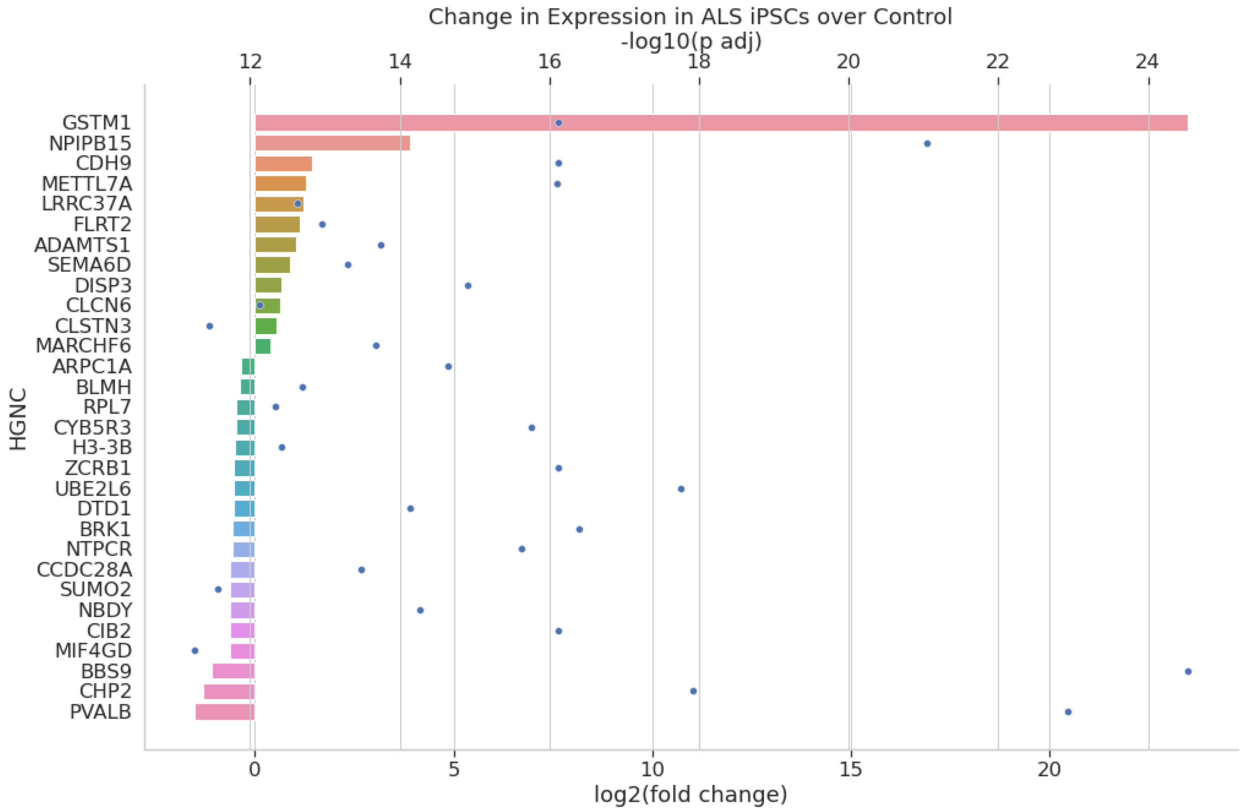

Fig. 1. Top differentially expressed genes in ALS. We show the top 30 differentially expressed ALS genes by adjusted p-value, sorted by their log fold change. Log fold change (LFC) estimation was computed using the `apeglm` R package [17]. We specifically note *GSTM1*, which has an especially high log fold change and *PIPB15*, *CHP2*, and *PVALB*, which have especially significant p-values.

protein which regulates cell pH, is a member of the KEGG Amyotrophic Lateral Sceloriss (ALS) pathway [19]. There is more limited evidence in the literature for an association between ALS and the two upregulated genes. de Faria Santos et al. report that there is no association between *GSTM1* deletion polymorphism and ALS [20], but it is possible that *GSTM1* is linked solely to sporadic ALS through expression and regulatory activity, and not to familial ALS through genetic polymorphism. We cannot find any link between *NPIPB15*, the gene encoding for a protein that interacts with the nuclear pore, and ALS in the literature, so the clear differential expression may be novel, or may be due to noise and uncertainty in our processing and differential expression pipeline.

We also performed a gene set enrichment analysis, and found some enrichment for neuromuscular junction development in ALS (see Appendix A), but there were no gene sets found enriched at 5% FDR threshold, so we did not move forward with the GSEA analysis.

Motif Size	Control	ALS	ALS Unique	ALS (Instances)	ALS Unique (Instances)
3	1	2	1	89	24
4	16	16	0	7,721	0
5	165	183	29	90,375	331

Table 3. Motif enrichment in iPSC regulatory networks. Motif enrichment was computed using mFinder. We show the number of network motifs found enriched in the control and ALS regulatory networks, as well as the number of motifs uniquely enriched in ALS. Additionally, we show the number of instances of each set of enriched motifs that appear overall in the ALS network. We focus our study on the 355 instances of enriched motifs that appear only in the ALS network.

3.3 Motif discovery in ALS

We identified enriched motifs in each of the four constructed networks – we focus here only on the iPSC networks (We analyze the iMN networks in Appendix C). In this analysis, we differentiate between general network motifs and specific motif instances. A general motif is a pattern of connectivity in the network that appears in a re-occurring fashion, regardless of the identities of the genes. One such example is the feed-forward loop. It is well-documented in the literature that the feed-forward loop is a common regulatory motif [14]. In the feed-forward loop, some gene (**A**) has two regulatory children ($\mathbf{A} \rightarrow \mathbf{B}$, $\mathbf{A} \rightarrow \mathbf{C}$), and one of those children also regulates the other ($\mathbf{B} \rightarrow \mathbf{C}$). As a sanity check for our motif enrichment method, we confirmed that the feed-forward loop was reported as a highly enriched motif in both the control and ALS network ($p = 0.0$).

The feedback loop (where **A,B,C** regulate each other in a cycle) was also found as enriched, but only in the ALS network. The feedback loop is typically used in biological systems to regulate homeostasis [21], and so we would expect it to be present in both networks, but especially so in the normally-functioning control network. The lack of enrichment for this motif in the control network is worth noting and is a target for future investigation. For motifs of size 4, the same 16 general motifs were found to be enriched in both control and ALS networks. Because we aim to focus on network rewiring in ALS, we did not further analyze any of these motifs, but a further investigation of enriched four-node motifs in biological networks could be interesting as well. We identified 165 5-node motifs which were significantly enriched in our control network and 183 5-node motifs enriched in our reconstructed ALS network. Of these, 11 are enriched only in the control network and 29 are enriched only in the ALS network.

3.4 Specific motif instances

Here, we turn our attention to the study of specific motif instances — sets of genes in the networks which are connected in the manner of one of the general motifs. Along with a summary of general motif numbers, Table 3 shows the number of specific motif instances enriched in ALS, and of motifs enriched *only* in ALS

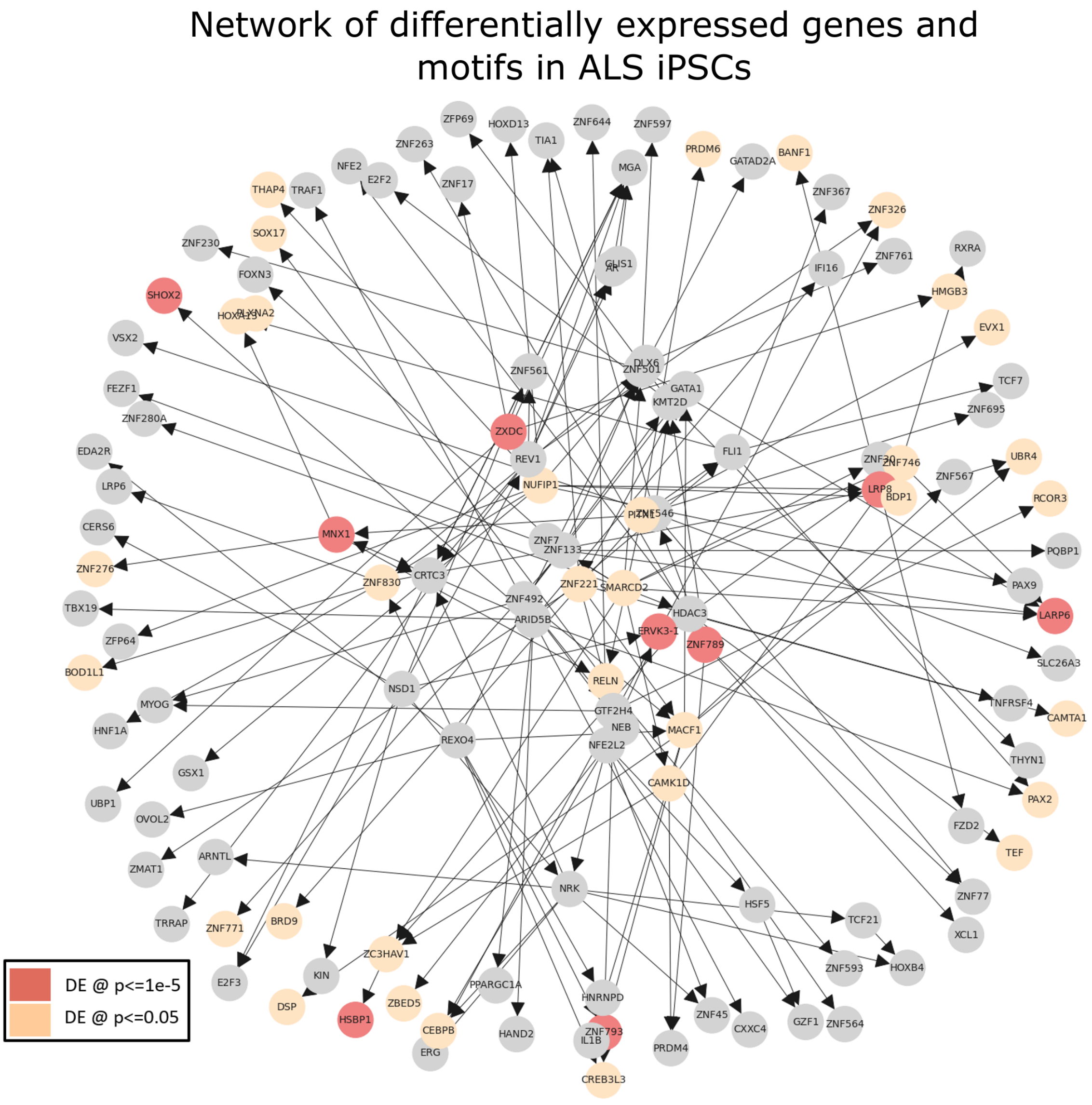

Fig. 2. Disease associated sub-network from gene expression and network motif analysis. This network was generated by the convergence of regulatory network reconstruction, differential gene expression analysis, and motif enrichment analysis. Ten motif instances were identified which were enriched in only the ALS network, and which contained at least one gene differentially expressed at significance level $p \leq 1e-5$. All nodes in these motif instances, as well as any one-hop neighbors in the reconstructed ALS regulatory network, were selected for the disease-specific subnetwork. This more focused network allows us to conduct a more targeted analysis of network rewiring in ALS and identify candidate therapeutic targets.

and not in the control network. We focus on the latter set, of which there are 24 3-gene motif instances and 331 5-gene motif instances. To further refine our search for dysregulated disease pathways, we search for specific instances among these 355 that contain at least one gene which is differentially expressed at the *highly significant* level ($p \leq 1e-5$), which resulted in 10 disease-relevant motifs. The full list of these motifs appears in Appendix B. From this set, we can construct the disease specific ALS sub-network shown in Figure 2. This network contains all genes in one of the 10 disease motifs and any of their one-hop neighbors in the original network. In the following sections (Sections 3.5 - 3.7), we explore three cases studies of ALS-specific gene regulation — two from this iPSC network, and one from the same pipeline run on the induced motor neuron (iMN) samples (the iMN disease sub-network can be found in Appendix C).

3.5 Dysregulation of neuronal development pathway

Our analysis shows evidence for dyregulation of the motor neuron development pathway in ALS patients. There are two ALS-specific motif instances which contain the gene *MNX1* — these motifs are shown in Figure 3. *MNX1*, which was found to be differentially expressed in ALS ($p \leq 1e-5$) is a known marker of motor neurons [22], and has been shown to be a key regulator of several neuronal genes [23]. In the ALS regulatory network, its activity is regulated by the transcription factor *PITX1*. Dixit et al. showed that *PITX1* is specifically up-regulated in the neuromuscular disease facioscapulohumeral muscular dystrophy (FSHD) [24], but do not report increased expression in ALS compared to other neuromuscular diseases. However, our analysis shows that *PITX1* is also up-regulated in ALS samples over control ($p \leq 0.05$). The protein coding gene *RELN*, which encodes the protein reelin, is the third member of a feedback loop wherein its activity is regulated by *MNX1* and it regulates expression of *PITX1*. *RELN* is a regulator of neuron development and migration, and high levels of reelin was shown in Ibi et al. to prevent symptoms of neurodevelopmental disorders in mouse models [25]. Our network also shows that *MNX1* regulates the activity of *HOXA13*, another developmental regulator. Both *RELN* and *HOXA13* were found to be differentially expressed by our analysis ($p \leq 0.05$). These motifs suggest that dysregulation in the motor neuron development pathway, specifically around the motor neuron marker *MNX1*, may contribute to the sclerosis that occurs in ALS.

3.6 Dysregulation of myosin binding pathway

We also identified several genes known to be associated with neuromuscular disease or active in myosin binding pathways, which suggests active dysregulation of mysoin binding in ALS. Figure 4 shows two of three ALS-enriched motif instances containing *LARP6*, which was differentially expressed in ALS ($p \leq 1e-5$). *LARP6*, a protein coding gene which enables RNA binding and myosin binding, is downregulated in ALS,

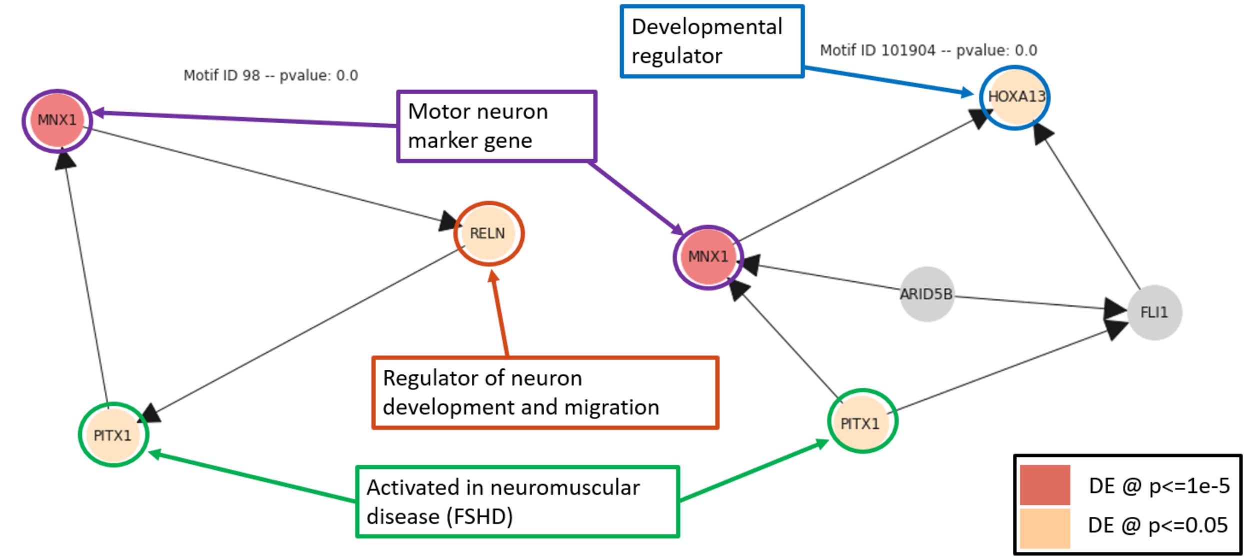
Motif ID 98 -- pvalue: 0.0
MNX1
Motor neuron marker gene
RELN
PITX1
Regulator of neuron development and migration
Activated in neuromuscular disease (FSHD)
Developmental regulator
Motif ID 101904 -- pvalue: 0.0
HOXA13
MNX1
ARID5B
FLI1
PITX1
DE @ p<=1e-5
DE @ p<=0.05

Fig. 3. Differentially expressed MNX1 motifs indicate neuronal development dysregulation in ALS. Two motifs enriched in ALS which contain motor neuron marker *MNX1*, which is differentially expressed in ALS ($p \leq 1e-5$). Several other genes including *RELN*, *PITX1*, and *HOXA13* were differentially expressed ($p \leq 0.05$) and are associated with motor neuron development pathways or activated in neuromuscular disease.

and is the downstream target of a feed-forward loop containing *SMARCD2* and *ZNF746*, where *SMARCD2* regulates both *LARP6* and *ZNF746*. *SMARCD2*, a member of the SWI/SNF family, is a known target of microRNA-206, a candidate biomarker for ALS [26]. *ZNF746*, a member of the zinc finger protein family, is a known biomarker of Parkinson's, another neuromuscular disease [27]. Both of these genes are downregulated in ALS as well ($p \leq 0.05$). This analysis suggests a pathway occuring specifically in ALS where mRNA-206 downregulates the activity of *SMARCD2*, which combines with *ZNF746* to suppress the myosin binding activity of *LARP6*. In addition to mRNA-206, suppression of these three genes may be a candidate biomarker of ALS, and therapeutic stimulation of *SMARCD2* could help recover the myosin binding pathway.

3.7 Innate immunity and oxidative stress in iMNs

While our analysis focused primarily on transcriptional regulation in iPSCs, we performed the same analysis pipeline on tissue samples from induced motor neurons (iMNs). We identified 11 general 5-node motifs enriched only in the ALS network, corresponding to 129 specific instances. Of these, we identified 33 motif instances containing highly significant differentially expressed genes. The disease subnetwork constructed from these 33 motif instances can be found in Appendix C. This disease subnetwork revolves heavily around three differentially expressed genes ($p \leq 1e-5$): *PML*, *KLF15*, and *AGER*. These three genes are linked to altered pathways of RNA metabolism (*FUS*) and oxidative stress (*SOD1*). *PML* is a protein coding gene which encodes for a promyelocytic leukemia (PML) nuclear body scaffold. In the application for NIH

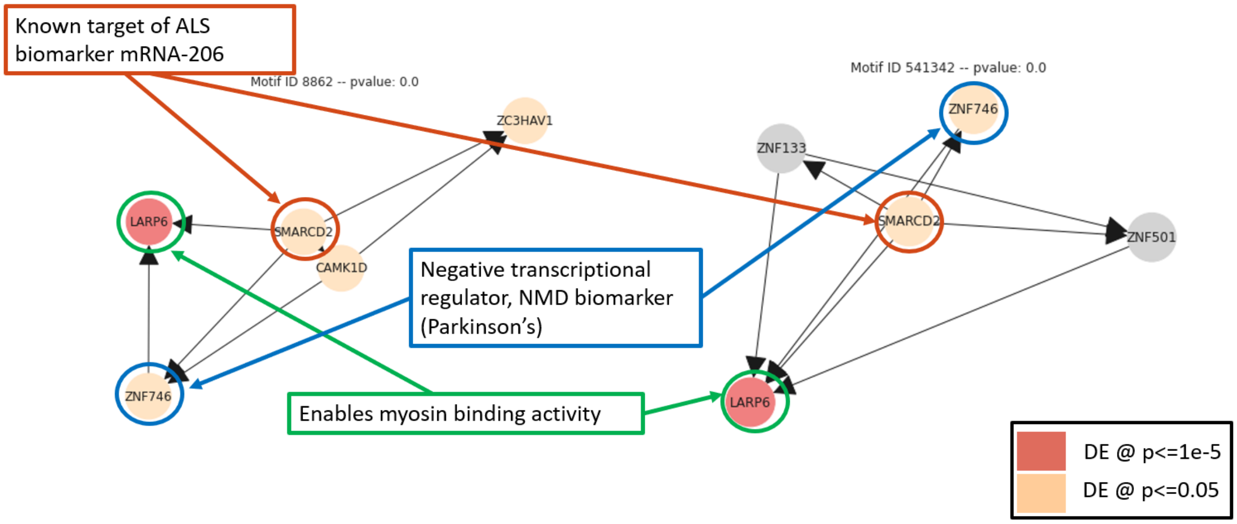
Known target of ALS biomarker mRNA-206
Motif ID 8862 -- pvalue: 0.0
ZC3HAV1
LARP6
SMARCD2
CAMK1D
ZNF746
Negative transcriptional regulator, NMD biomarker (Parkinson's)
Enables myosin binding activity
Motif ID 541342 -- pvalue: 0.0
ZNF746
ZNF133
SMARCD2
ZNF501
LARP6
🟥 DE @ p<=1e-5
🟧 DE @ p<=0.05

Fig. 4. Differentially expressed LARP6 motifs indicate myosin binding dysregulation in ALS. Two motifs enriched in ALS which *LARP6*, a facilitator of myosin binding activity, which is differentially expressed in ALS ($p \leq 1e-5$). *LARP6* is regulated in a feed-forward loop by *SMARCD2* and *ZNF746*, known biomarkers for ALS and the related neuromuscular Parkinson's disease respectively. Both of these regulators are also differentially expressed in ALS ($p \leq 0.05$). These motifs suggest that mRNA-206 suppression of *SMARCD2* leads to deficiency in myosin binding in ALS.

Grant R21-1R21NS095273-01A1, Hayward et al. identify the presence of PML nuclear bodies as a nuclear phenotype caused by ALS-linked *FUS* mutants, and propose further exploration of the role of *PML* in ALS. Our analysis corroborates this proposal, suggesting that *PML* plays a central role in network dysregulation in ALS. *AGER*, which encodes the protein RAGE, has also been linked to ALS progression in *SOD1* G93A transgenic mice by Nowicka et al. [28]. Interestingly, Massopust et al. show that *KLF15* overexpression does *not* extend the lifespan of SOD1G93A mice, nor does it prevent ALS phenotype [29]. As *KLF15* is central in our disease network, we believe that activity of *KLF15* in ALS progression warrants further study.

4 Discussion

In this study, we analyze transcriptomic data from ALS and control samples to identify the mechanism of sporadic ALS and identify candidate ALS biomarkers and therapeutic targets. Through a combination of differential expression analysis and network motif enrichment analysis, we identify several candidate genes which are actively dysregulated in motor neuron development pathways, myosin binding pathways, RNA metabolism, and oxidative stress. While our analysis allowed us to converge quickly on interesting mechanistic pathways, with support in the literature, there are several limitations and possible improvements to the pipeline which could more accurately elucidate the mechanism of ALS. The foremost is our limited statistical power stemming from our data availability. We use bulk RNA transcriptomes from 9 control samples and

12 ALS samples — this small sample size means that any variance observed between the two sets is subject to noise and will be underpowered. Noise in general is a more pervasive limitation – regulatory network reconstruction is inherently noisy, and with limited bulk RNA samples the confidence of any individual edge is low (see Appendix D). Future study in this area would benefit from obtaining more ALS and case samples, potentially at single-cell resolution. Additionally, this study would benefit from explicit modeling of the uncertainty that comes with regulatory network reconstruction, and from batch correction of the tissue samples collected in control and ALS cells.

Beyond collecting more data and modeling or combating noise, there are a few areas of natural extension for this research. The NeuroLINCS project contains not only transcriptomic measurements of control and ALS, but proteomic and epigenomic assays as well. An expansion of this study into multi-omic analysis would provide further evidence for disease genes and help to paint the complete picture of network rewiring and mechanism in sporadic ALS. We conducted our analyses primarily in iPSCs, but also some in iMNs, and we would like to compare the networks and motifs between cell types to determine how the cell type impacts analysis and how ALS regulation varies at different stages of cell development. Finally, motif enrichment analysis with the tools available to us works on simple directed graphs. This approach is powerful, but requires us to treat activating and repressing edges equally. A more rigorous motif enrichment analysis would treat activating and repressing edges, and the different motifs containing them, differently, which would allow for a more fine-grained analysis of motifs which are specifically present in the ALS network, and would help to clarify the types of regulatory relationships which moderate the ALS etiology.

5 Key Resources

Data	
NeuroLINCS Project	`http://neurolincs.org/data/`
iPSC Transcriptomic Data	`http://lincsportal.ccs.miami.edu/datasets/view/LDS-1356`
iMN Transcriptomic Data	`http://lincsportal.ccs.miami.edu/datasets-beta/#/view/LDS-1398`
Software	
Data Processing and Analysis	`https://github.com/samsledje/als_regulation_20440`
GENIE3	`https://bioconductor.org/packages/release/bioc/html/GENIE3.html`
mfinder	`https://www.weizmann.ac.il/mcb/UriAlon/download/network-motif-software`
DESeq2	`https://bioconductor.org/packages/release/bioc/html/DESeq2.html`
GSEA	`https://www.gsea-msigdb.org/gsea/index.jsp`

Table 4. Key Resources Table

References

1. R. L. Redler and N. V. Dokholyan, "The complex molecular biology of amyotrophic lateral sclerosis (ALS)," *Progress in Molecular Biology and Translational Science*, vol. 107, pp. 215–262, 2012.
2. A. Chio, G. Logroscino, O. Hardiman, R. Swingler, D. Mitchell, E. Beghi, B. G. Traynor, E. Consortium, *et al.*, "Prognostic factors in ALS: a critical review," *Amyotrophic Lateral Sclerosis*, vol. 10, no. 5-6, pp. 310–323, 2009.
3. S. Chen, P. Sayana, X. Zhang, and W. Le, "Genetics of amyotrophic lateral sclerosis: an update," *Molecular Neurodegeneration*, vol. 8, no. 1, pp. 1–15, 2013.
4. R. Mejzini, L. L. Flynn, I. L. Pitout, S. Fletcher, S. D. Wilton, and P. A. Akkari, "ALS genetics, mechanisms, and therapeutics: where are we now?," *Frontiers in Neuroscience*, p. 1310, 2019.
5. L. Thompson, "iPSC (Exp 1) - ALS, SMA and Control (unaffected) subject-derived iPSC lines - RNA-seq," 2016.
6. L. Thompson, "iMN (Exp 2) - ALS, SMA and Control (unaffected) iMN cell lines differentiated from iPS cell lines using a long differentiation protocol - RNA-seq," 2017.
7. F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, no. 1, pp. 1–5, 2018.
8. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PloS one*, vol. 5, no. 9, p. e12776, 2010.
9. S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018.
10. M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology*, vol. 15, no. 12, pp. 1–21, 2014.
11. V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, *et al.*, "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature genetics*, vol. 34, no. 3, pp. 267–273, 2003.
12. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.
13. S. Patra and A. Mohapatra, "Review of tools and algorithms for network motif discovery in biological networks," *IET Systems Biology*, vol. 14, no. 4, pp. 171–189, 2020.
14. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.
15. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, 2004.
16. F. Harary and E. Palmer, "Graphical enumeration, acad," *Press, NY*, 1973.
17. A. Zhu, J. G. Ibrahim, and M. I. Love, "Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences," *Bioinformatics*, vol. 35, no. 12, pp. 2084–2092, 2019.

18. C. W. Lederer, A. Torrisi, M. Pantelidou, N. Santama, and S. Cavallaro, "Pathways and genes differentially expressed in the motor cortex of patients with sporadic amyotrophic lateral sclerosis," *BMC genomics*, vol. 8, no. 1, pp. 1–26, 2007.

19. M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

20. K. de Faria Santos, R. M. Azevedo, D. d. C. P. Bento, R. da Silva Santos, and A. A. da Silva Reis, "No association between gstm1 and gstt1 deletion polymorphisms and amyotrophic lateral sclerosis: a genetic study in brazilian patients," *Meta Gene*, vol. 30, p. 100979, 2021.

21. L. Bich, M. Mossio, and A. M. Soto, "Glycemia regulation: from feedback loops to organizational closure," *Frontiers in Physiology*, p. 69, 2020.

22. M. Madill, K. McDonagh, J. Ma, A. Vajda, P. McLoughlin, T. O'Brien, O. Hardiman, and S. Shen, "Amyotrophic lateral sclerosis patient ipsc-derived astrocytes impair autophagy via non-cell autonomous mechanisms," *Molecular brain*, vol. 10, no. 1, pp. 1–12, 2017.

23. M.-a. Sun, S. Ralls, W. Wu, J. Demmerle, J. Jiang, C. Miller, G. Wolf, and T. S. Macfarlan, "Homeobox transcription factor mnx1 is crucial for restraining the expression of pan-neuronal genes in motor neurons," *bioRxiv*, 2021.

24. M. Dixit, E. Ansseau, A. Tassin, S. Winokur, R. Shi, H. Qian, S. Sauvage, C. Mattéotti, A. M. van Acker, O. Leo, *et al.*, "Dux4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of pitx1," *Proceedings of the National Academy of Sciences*, vol. 104, no. 46, pp. 18157–18162, 2007.

25. D. Ibi, G. Nakasai, N. Koide, M. Sawahata, T. Kohno, R. Takaba, T. Nagai, M. Hattori, T. Nabeshima, K. Yamada, *et al.*, "Reelin supplementation into the hippocampus rescues abnormal behavior in a mouse model of neurodevelopmental disorders," *Frontiers in cellular neuroscience*, p. 285, 2020.

26. J. M. Toivonen, R. Manzano, S. Oliván, P. Zaragoza, A. García-Redondo, and R. Osta, "Microrna-206: a potential circulating biomarker candidate for amyotrophic lateral sclerosis," *PLoS One*, vol. 9, no. 2, p. e89065, 2014.

27. A. K. Alieva, E. V. Filatova, A. V. Karabanov, S. N. Illarioshkin, P. A. Slominsky, and M. I. Shadrina, "Potential biomarkers of the earliest clinical stages of parkinson's disease," *Parkinson's Disease*, vol. 2015, 2015.

28. N. Nowicka, K. Szymanska, J. Juranek, K. Zglejc-Waszak, A. Korytko, M. Zalkecki, M. Chmielewska-Krzesinska, K. Wasowicz, and J. Wojtkiewicz, "The involvement of rage and its ligands during progression of als in sod1 g93a transgenic mice," *International journal of molecular sciences*, vol. 23, no. 4, p. 2184, 2022.

29. R. Massopust, D. Juros, D. Shapiro, M. Lopes, S. M. Haldar, T. Taetzsch, and G. Valdez, "Klf15 overexpression in myocytes fails to ameliorate als-related pathology or extend the lifespan of sod1g93a mice," *Neurobiology of disease*, vol. 162, p. 105583, 2022.

30. A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "Upset: visualization of intersecting sets," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1983–1992, 2014.

A Gene set enrichment analysis reveals deficiency in developmental pathways

The GO term GO:0007528, “Neuromuscular Junction Development”, was found enriched in ALS (Figure 5), but was not significant at the 5% FDR correction level.

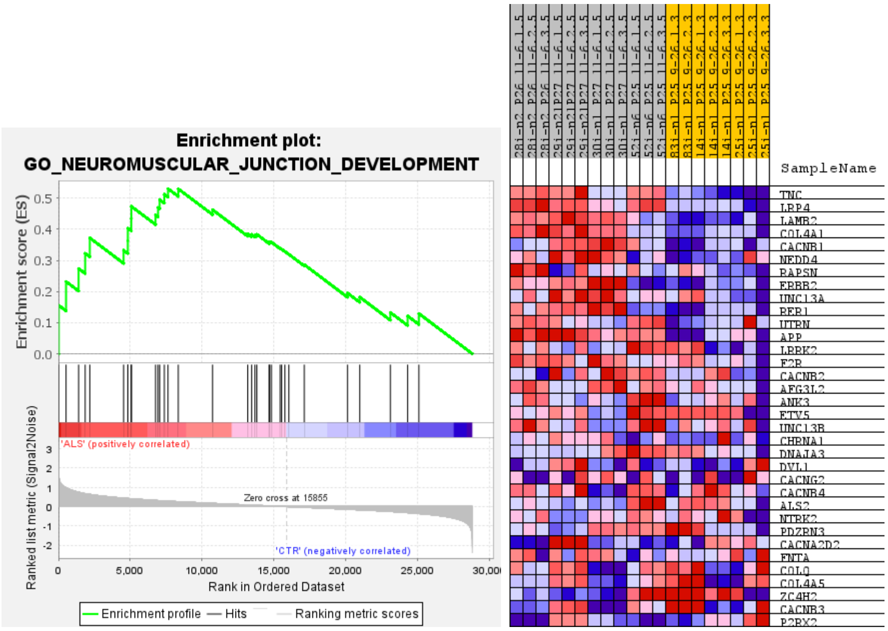

Fig. 5. Gene set enrichment analysis of ALS vs. Control populations. ALS iPSC tissue samples were found to be enriched for GO:0007528, “Neuromuscular Junction Development.”

B Full list of enriched and highly-expressed motifs in ALS iPSCs

Gene Set	Motif Connectivity Matrix
KMT2D, **ZNF789**, ZNF546	[0 1 0 0 0 1 1 0 0]
PITX1, **MNX1**, RELN	[0 1 0 0 0 1 1 0 0]
SMARCD2, CAMK1D, ZNF746, **LARP6**, ZC3HAV1	[0 1 1 1 1 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
SMARCD2, HDAC3, ZNF746, **LARP6**, ZC3HAV1	[0 1 1 1 1 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
ZNF830, NUFIP1, REXO4, DLX6, **LRP8**	[0 0 0 0 1 1 0 0 1 1 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0]
MACF1, GTF2H4, ZNF221, ZNF492, **ZNF793**	[0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0]
NRK, **ERVK3-1**, NFE2L2, NSD1, ERG	[0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0]
MNX1, FLI1, PITX1, ARID5B, HOXA13	[0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0]
SMARCD2, ZNF133, ZNF501, ZNF746, **LARP6**	[0 1 1 1 1 0 0 1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0]
REV1, ZNF561, CRTC3, **ZXDC**, MGA	[0 1 1 1 1 0 0 1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0]

Table 5. List of 10 disease relevant motifs in ALS iPSCs. Disease relevant motifs are those which are enriched only in the ALS regulatory network, and which contain at least one gene differentially expressed at the significance level $p \leq 1e-5$ (bolded). The motif connectivity matrix encodes the shape of the motif — the matrix should be wrapped to be a square the size of the node set (e.g. a length 9 matrix wrapped to 3x3) and corresponds to the adjacency matrix for the given motif.

C Convergence sub-network in iMNs focuses on innate immunity and oxidative stress

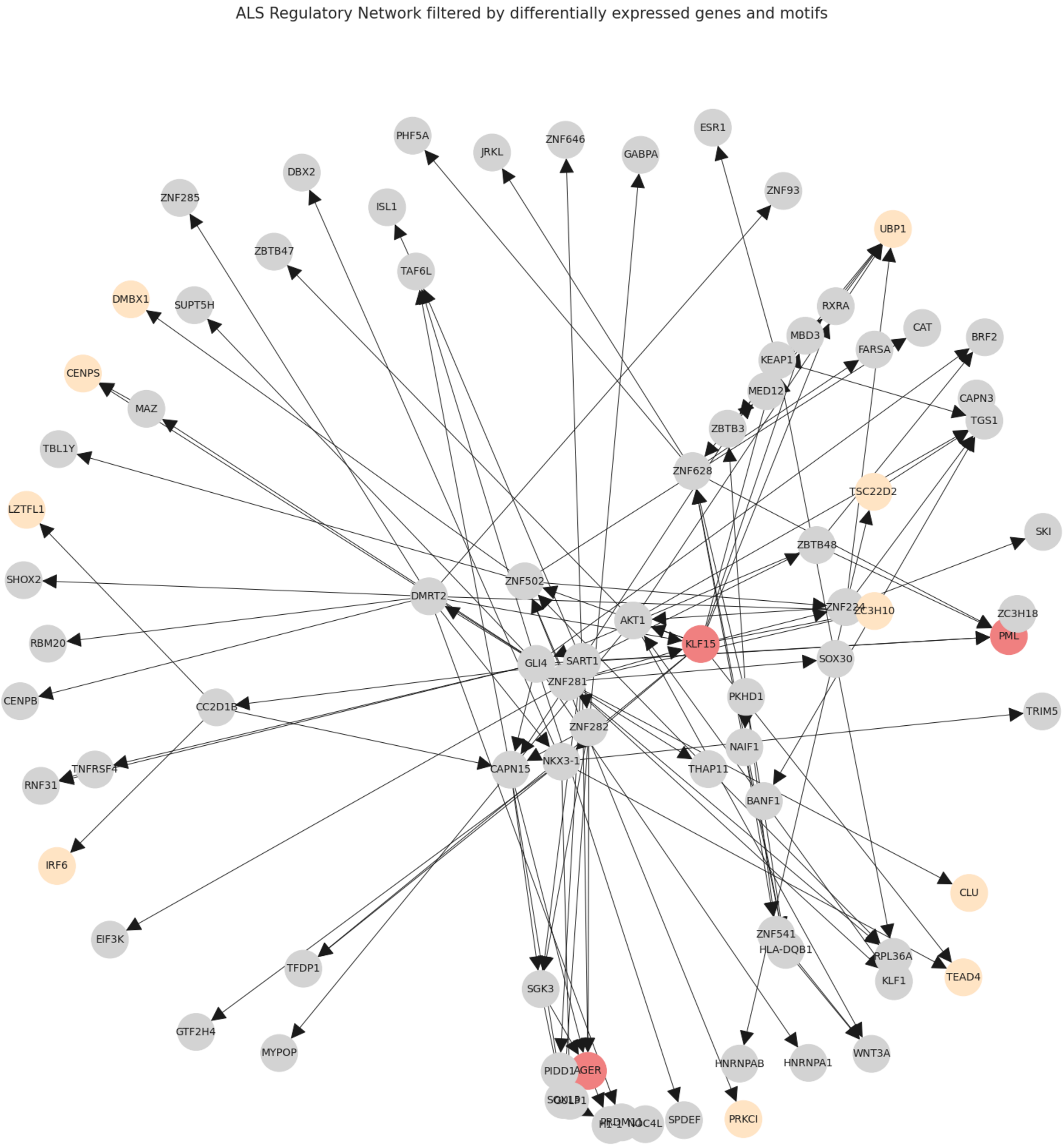

Fig. 6. Innate immunity and oxidative stress pathways enriched in ALS iMNs. This network was generated by the convergence of regulatory network reconstruction, differential gene expression analysis, and motif enrichment analysis in induced motor neurons. 33 motif instances were identified which were enriched in only the ALS network, and which contained at least one gene differentially expressed at significance level $p \leq 1e-5$. All nodes in these motif instances, as well as any one-hop neighbors in the reconstructed iMN ALS regulatory network, were selected for the disease-specific subnetwork. This network highlights the central role of *PML*, *KLF15*, and *AGER* in innate immune pathway regulation in ALS.

D Comparison between reconstructed regulatory networks

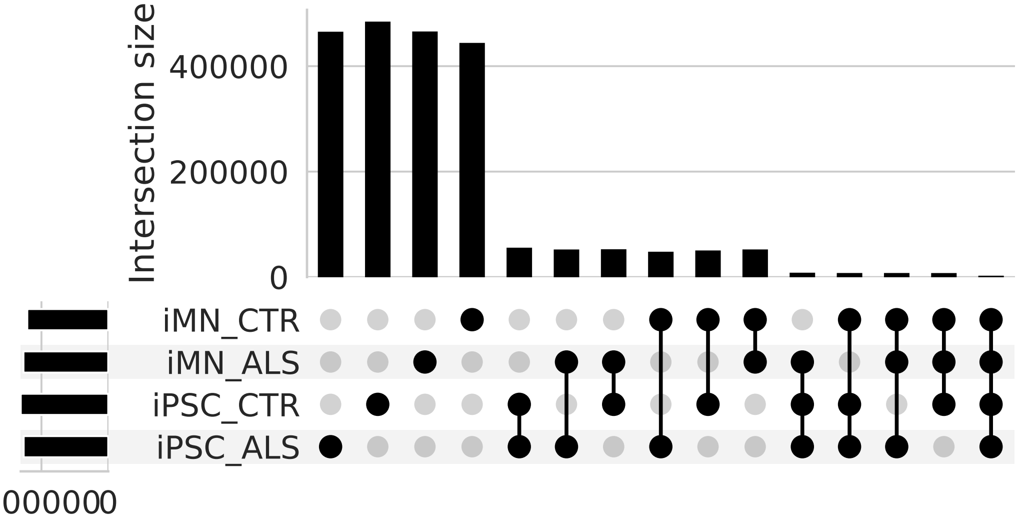

Fig. 7. UpSet [30] plot of edge overlap between 4 networks at 90% threshold. When we compare the edges inferred in four different regulatory networks, we find relatively little overlap in the edges inferred at the very top (highest inferred strength of regulatory relationships). Most edges inferred are unique to their network, which suggests substantial noise in the network inference process.