Running Title: Polygenic Contributions to ALS

Machine learning suggests polygenic contribution to cognitive dysfunction in amyotrophic lateral sclerosis (ALS)

Katerina Placek[1], Michael Benatar[2], Joanne Wuu[2], Evadnie Rampersaud[3], Laura Hennessy[1], Vivianna M. Van Deerlin[4], Murray Grossman[1], David J. Irwin[1], Lauren Elman[1], Leo McCluskey[1], Colin Quinn[1], Volkan Granit[2], Jeffrey M. Statland[5], Ted. M Burns[6], John Ravits[7], Andrea Swenson[8], Jon Katz[9], Erik Pioro[10], Carlayne Jackson[11], James Caress[12], Yuen So[13], Samuel Maiser[14], David Walk[14], Edward B. Lee[4], John Q. Trojanowski[4], Philip Cook[15], James Gee[15], Jin Sha[16,17],Adam C. Naj[4,16,17], Rosa Rademakers[18], The CReATe Consortium[19], Wenan Chen[3], Gang Wu[3], J. Paul Taylor[3,20], & Corey T McMillan[1*]

[1] Department of Neurology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[2] Department of Neurology, University of Miami, Leonard M. Miller School of Medicine, Miami, FL
[3] Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN
[4] Department of Pathology & Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[5] Department of Neurology, University of Kansas Medical Center, Kansas City, KS
[6] Department of Neurology, University of Virginia Health System, Charlottesville, VA
[7] Department of Neurosciences, University of California San Diego, San Diego, CA
[8] Department of Neurology, University of Iowa
[9] Department of Neurology, Washington University in St. Louis, Missouri.
[10] Department of Neurology, Cleveland Clinic, Cleveland, OH
[11] Department of Neurology, University of Texas Health Science Center, San Antonio, San Antonio, TX
[12] Department of Neurology, Wake Forest University School of Medicine, Winston-Salem, NC
[13] Department of Neurology, Stanford University Medical Center, San Jose, CA
[14] Department of Neurology, University of Minnesota Medical Center, Minneapolis, MN
[15] Penn Image Computing Science Laboratory (PICSL), Department of Radiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[16] Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[17] Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[18] Department of Neuroscience, Mayo Clinic, Jacksonville, FL
[19] Rare Diseases Clinical Research Network, National Institutes of Health, Bethesda, MD
[20] The Howard Hughes Medical Institute, Chevy Chase, MS

[±] Refer to Appendix for full CReATe Consortium author list.
[*] **Correspondence:**
Penn Frontotemporal Degeneration Center, Department of Neurology
3700 Hamilton Walk, Suite 606B
Philadelphia, PA,19104
mcmillac@pennmedicine.upenn.edu; (215) 614 0987

Word/Figure Counts: Abstract=163 words; Body=6,841 words; References=65; Figures=4; Tables=2; Supplementary Figures=8; Supplementary Tables=3

Abstract

Amyotrophic lateral sclerosis (ALS) is a multi-system disease characterized primarily by progressive muscle weakness. Cognitive dysfunction is commonly observed in patients, however factors influencing risk for cognitive dysfunction remain elusive. Using sparse canonical correlation analysis (sCCA), an unsupervised machine-learning technique, we observed that single nucleotide polymorphisms collectively associate with baseline cognitive performance in a large ALS patient cohort from the multicenter Clinical Research in ALS and Related Disorders for Therapeutic Development (CReATe) Consortium (N=327). We demonstrate that a polygenic risk score derived using sCCA relates to longitudinal cognitive decline in the same cohort, and also to *in vivo* cortical thinning in the orbital frontal cortex, anterior cingulate cortex, lateral temporal cortex, premotor cortex, and hippocampus (N=114) as well as *post mortem* motor cortical neuronal loss (N=88) in independent ALS cohorts from the University of Pennsylvania Integrated Neurodegenerative Disease Biobank. Our findings suggest that common genetic polymorphisms may exert a polygenic contribution to the risk of cortical disease vulnerability and cognitive dysfunction in ALS.

Introduction

As many as half of patients with amyotrophic lateral sclerosis (ALS) manifest progressive decline in cognition consistent with extra-motor frontal and temporal lobe neurodegeneration, including 14% also diagnosed with frontotemporal dementia (FTD) [1,2]. Comorbid cognitive dysfunction is a marker of poorer prognosis in this fatal disease and confers risk for more rapid functional decline, shorter survival, and greater caregiver burden [3-6]. While linkage analysis and genome-wide association studies (GWAS) have identified rare causal mutations [7-10] and common risk loci [11-15] suggested shared genetic architecture between ALS and FTD, whether and how identified variants relate to phenotypic heterogeneity, including in cognition, remain largely unexplored.

The genetic landscape of ALS is largely characterized by ‘apparently sporadic’ disease occurring in 90% of patients with neither a known family disease history nor an identifiable pathogenic mutation [16]. Population-based studies estimate that only 5-10% of non-familial and 40-50% of familial ALS cases can be attributed to known pathogenic mutations [17] (e.g. *C9ORF72* [7,8], *NEK1* [18], *SOD1* [19]), but GWAS have revealed many loci of common genetic variation that confer risk for ALS and FTD. Indeed, recent evidence, supports a polygenic contribution to disease risk from common genetic variants [20,21]. These include the largest ALS GWAS to-date which newly identified risk variants in the *KIF5A* gene [12] and genome-wide conjunction and conditional false discovery rate (FDR) analyses demonstrating shared genetic

contributions between ALS and FTD from common single nucleotide polymorphisms (SNPs) at known and novel loci [15].

An accumulating body of research suggests that SNPs associated with risk of ALS and FTD demonstrate quantitative trait modification of patient phenotype. For example, a SNP identified as a risk locus for ALS and FTD was found to contribute to cognitive decline, *in vivo* cortical degeneration in the prefrontal and temporal cortices, and *post mortem* pathologic burden of hyperphosphorylated TAR-DNA binding protein [43 kDa] (TDP-43) in the middle frontal, temporal, and motor cortices [22]. Another study found that a SNP identified as a risk locus for FTD with underlying TDP-43 pathology was additionally associated with cognition in patients with ALS [23]. Others have recently demonstrated shared polygenic risk between ALS and other traits (e.g. smoking, education) and diseases (e.g. schizophrenia) [20,21,24], suggesting that a single variant is unlikely to fully account for observed disease phenotype modification. However, there are presently no published studies evaluating polygenic contribution to cognitive dysfunction in ALS.

Here we employed an unsupervised machine-learning approach, sparse canonical correlation analysis (sCCA), to identify and evaluate a potential polygenic contribution to cognitive dysfunction in ALS. Traditional approaches for constructing polygenic scores identify variants associated with disease risk through GWAS in a univariate manner, and then compute the sum of alleles at each identified variant weighted by their effect sizes. In this study, we used sCCA to identify polygenic associations with a continuous

phenotype of cognitive performance in ALS. This data-driven method employs sparsity to select maximally-contributing variants and assigns corresponding weights based on model contribution with minimal *a priori* assumptions. We used sCCA to derive a polygenic risk score for cognitive dysfunction in a large longitudinal cohort of cognitively-characterized patients with ALS or a related disorder participating in the Phenotype-Genotype-Biomarker (PGB) study of the Clinical Research in ALS and Related Disorders for Therapeutic Development (CReATe) Consortium. We then evaluated independent neuroimaging and autopsy ALS patient cohorts from the University of Pennsylvania Integrated Neurodegenerative Disease Biobank (UPenn Biobank) [25] to evaluate whether polygenic risk for cognitive dysfunction also relates to *in vivo* cortical neurodegeneration and *ex vivo* cortical neuronal loss and TDP-43 pathology. We focused our investigation on SNPs achieving genome-wide significance in the largest published ALS GWAS [12] and SNPs identified as shared risk loci for both ALS and FTD [15]. We hypothesized that a sparse multivariate approach would reveal a subset of genetic loci associated with cognitive dysfunction profiles in ALS in a polygenic manner, and that follow-up analyses in independent neuroimaging and autopsy cohorts would converge to characterize quantitative traits associated with polygenic risk from identified loci.

Results

Heterogeneity of baseline cognitive and motor phenotype in ALS patients.

Smaller-scale studies have shown that ALS patients have impairments in executive, verbal fluency, and language domains, but with relative sparing of memory and

visuospatial function [4]. The Edinburgh Cognitive and Behavioral ALS Screen (ECAS) was developed to measure cognitive function minimally confounded by motor disability and includes an “ALS-Specific” score that captures impairments in language, executive function, and verbal fluency domains that are frequently observed in ALS patients, and an “ALS-Non-Specific” score that captures less frequently observed impairments in memory and visuospatial function, in addition to overall performance (ECAS Total score) [26]. To quantify heterogeneity in cognitive dysfunction, we evaluated 327 patients with ALS or a related disorder (e.g., ALS-FTD, primary lateral sclerosis (PLS), progressive muscular atrophy (PMA)) participating in the PGB study of the CReATe Consortium (NCT02327845) (*Table 1*). We used linear mixed-effects (LME) to model variability between individuals in baseline performance and rate of decline on the ECAS (Total, ALS-Specific, and ALS-Non-Specific scores, and scores for each individual cognitive domain), on the ALS Functional Rating Scale – Revised (ALSFRS-R), and on clinician ratings of upper motor neuron (UMN) and lower motor neuron (LMN) signs (UMN and LMN burden scores); each model included covariate adjustment for potential confounders including age, education, bulbar onset, and disease duration. We confirmed that cognitive and motor performance at baseline are heterogeneous across individuals (*Figure 1A*), and correlation analyses suggested that this is independent of disability in physical function or clinical burden of UMN/LMN signs (all $R<0.2$; *Figure 1B*). Together this establishes the heterogeneity of baseline and longitudinal cognitive and motor phenotypes within the PGB cohort.

Multivariate analyses indicate polygenic contributions to baseline cognitive performance.

To identify potential polygenic contributions to cognitive impairment in ALS we employed sCCA [27], an unsupervised machine-learning approach enabling identification of multivariate relationships between a dataset of one modality (e.g. genetic variables including allele dosage of SNPs) and another modality (e.g. clinical measures of cognitive and motor function). Traditional CCA identifies a linear combination of all variables that maximize the correlation between datasets, resulting in an association of variables from one dataset (e.g., SNPs) and variables from another dataset (e.g., clinical scores) [27]. The “sparse” component of sCCA additionally incorporates an L1 penalty that shrinks the absolute value of the magnitude of coefficients to yield sparse models (i.e. models with fewer variables) such that some coefficients are zero, and the variables associated with them are effectively eliminated from the model. As a result, variables that contribute little variance to the model are dropped and instead of a linear combination of all model variables, we are able to identify a data-driven subset of variables from one dataset that relate to a subset of variables from another dataset. Unstandardized regression coefficients resulting from sCCA serve as canonical weights indicating the direction and strength of the relationships between selected variables.

We evaluated an allele-dosage dataset comprised of 33 SNPs identified as shared risk loci for both ALS and FTD [15], and 12 SNPs identified as risk loci for ALS from the largest published case-control GWAS [12], with the latter chosen to include loci

associated with ALS but not specifically with FTD (*Figure 1C*). We included the first two principle components from a PCA conducted in the PGB cohort and binary variables for sex, *C9ORF72* repeat expansion status, and other mutation status (e.g. *SOD1*) in this dataset to account for inter-individual genetic differences in population structure, sex, and mutation status. We then used sCCA to examine the association between this genetic dataset and a dataset comprised of adjusted baseline performance on clinical measures of cognitive and motor performance extracted from the LME models.

After optimizing model sparsity parameters (*Supplementary Figure 1*), we ran sCCA 10,000 times and employed random bootstrapped subsamples of 75% of participants in each iteration (*Supplementary Figure 2*). We then calculated the median canonical correlation between the clinical and genetic datasets, the median canonical weight for each variable in the genetic dataset, and the proportion of times (as a percentage) each variable from the clinical dataset was chosen out of 10,000 iterations. We report percentages rather than median canonical weight for clinical features because the optimized L1 parameter for the clinical dataset was the most stringent (i.e. 0.1), thus resulting in only one variable from the clinical dataset being chosen in each of the 10,000 iterations.

To assess model performance under the null hypothesis (no association between genetic factors and clinical phenotypes), we similarly ran 10,000 bootstrapped sCCAs using the same L1 and subsampling parameters; however, we randomly permuted each dataset 100 times in each model iteration. We examined the proportion of times each

variable in the clinical and genetic datasets was selected by this null model (i.e. achieving a non-zero canonical weight). We used the null model to define a *p* value for the true, unpermuted model by calculating the probability under the null hypothesis of observing a canonical correlation greater than or equal to the median canonical correlation under sCCA modeling of the true data.

We observed that a subset of 29 genetic variables were correlated with a single clinical variable, achieving a median canonical correlation between the two datasets of R=0.35 (95% Confidence Interval: 0.23, 0.42); *p*=0.019) (*Figure 2*, *Supplementary Figure 3*). Over the 10,000 iterations, the most frequently selected clinical variable was the ECAS ALS-Specific score (percentage of times selected: 37%), followed by the ECAS Total (29%), Executive Function (17%), Language (9.5%), Verbal Fluency (2.3%), ALS-Specific (2.2%), Memory (2%), and Visuospatial (0.34%) scores. The ALSFRS-R and UMN and LMN burden scores were each selected in less than 0.05% of the model iterations. By contrast, performance of sCCA modeling under the null hypothesis demonstrated that each clinical variable was selected in a largely equal proportion of iterations (all variables ranging 5.9% to 9.4%), demonstrating that the true sCCA modeling selected cognitive and not motor features beyond what would be expected by chance (*Supplementary Figure 4A*).

Of the 29 selected genetic variables, the 12 most highly weighted were rs1768208 and rs9820623 (*MOBP*), rs7224296 (*NSF*), rs538622 (*ERGIC1*), rs10143310 (*ATXN3*), rs6603044 (*BTBD1*), rs4239633 (*UNC13A*), rs2068667 (*NFASC*), rs10488631

(*TNPO3*), rs11185393 (*AMY1A*), rs3828599 (*GPX3*), and sex. Twenty-seven of the 29 genetic variables selected were SNPs, and 85% of model-selected SNPs (23/27) were shared risk loci for ALS and FTD [15]. Modeling under the null hypothesis revealed that each genetic variable achieved a largely equal median weight, and thus there were no stronger model contributions from any subset of genetic variables (*Supplementary Figure 4B*). The association of genetic variables most frequently with the ECAS ALS-Specific score suggests polygenic contribution to impairment in domains of cognition frequently impaired in patients with ALS (e.g. language, verbal fluency, and executive function), that are also the most impaired domains of cognition observed in FTD.

Polygenic score captures baseline cognition as well as longitudinal rate of cognitive decline, but not motor decline.

Next we investigated potential polygenic contributions to rate of decline in cognitive and motor performance in the PGB cohort. Investigation of baseline performance may only capture differences at a single (somewhat arbitrary) point in time, but not differences in the trajectory of performance over time.

To evaluate association with longitudinal performance, we first calculated a weighted polygenic score (wPGS) by computing a sum of allele dosage for each individual genetic variable multiplied by their median canonical weights from sCCA modeling. Spearman rank-order correlations between the wPGS and adjusted baseline estimates of the four clinical features selected in 10% or more of the 10,000 iterations (e.g. ALS-Specific, Total, Executive Function, and Language scores from the ECAS) resulted in

correlation values similar to the median canonical correlation observed from sCCA modeling (e.g. for ECAS ALS-Specific: *rs(329)*=-0.34, p=2.4×10^{-10}) (*Figure 3A*), suggesting construct validity.

We then conducted Spearman's rank order correlations between the wPGS and adjusted rate of decline on each clinical measure of cognitive and motor performance using a Bonferroni family-wise error correction. To obtain adjusted rates of decline, we extracted individual slope estimates from prior LME (see above) for the 277 individuals (85%) from the PGB cohort with 2 or more observations on the ECAS, ALSFRS-R, and UMN and LMN burden scores. We observed significant negative relationships between the wPGS and adjusted rate of decline on ECAS ALS-Specific (*rs*(277)=-0.21, p=5.3×10^{-3}), ALS-Non Specific (*rs*(277)=-0.19, p=0.016), and Total scores (*rs*(277)=-0.26, p=8.1×10^{-5}) (*Figure 3B*), but not on the ALSFRS-R or UMN and LMN burden scores (all p >0.9) (*Supplementary Figure 5*). These findings suggest polygenic contribution to rate of cognitive – but not motor – decline from the SNPs associated with risk of ALS or joint risk of ALS and FTD that were included in this analysis.

Polygenic score associates with cortical thinning in the UPenn Biobank.

Cognitive dysfunction in ALS, including performance on the ECAS, has previously been attributed to sequential disease progression rostrally and caudally from the motor cortex [28-30] and to advancing disease stage [4]. To evaluate the neuroanatomic basis for polygenic contribution to cognitive performance in patients with ALS, we applied the wPGS score derived in the CReATe PGB Cohort to an independent cohort of patients

with ALS from the UPenn Biobank. We used voxel-wise *in vivo* measures of reduced cortical thickness (in mm^3) to quantify cortical neurodegeneration. Cross-sectional measurements of cortical thickness were derived from T1-weighted magnetic resonance imaging (MRI) in 114 patients with ALS and 114 age, sex, and education-matched healthy controls who were recruited for research from UPenn (*Table 2*). Nonparametric modeling using 10,000 random permutations revealed extensive reduction of cortical thickness bilaterally in the frontal and temporal cortices of patients relative to controls (*Table 2, Supplementary Figure 6*).

After identifying regions of reduced cortical thickness in patients with ALS, we investigated whether the wPGS derived from sCCA modeling in the CReATe PGB cohort contributed to magnitude of reduced cortical thickness in the independent UPenn Biobank neuroimaging cohort. Nonparametric modeling using 10,000 random permutations with adjustments for potential confounds in age, disease duration, and scanning acquisition revealed that a higher wPGS (i.e. greater risk) associated with greater reduction of cortical thickness in the orbital prefrontal cortex, anterior cingulate cortex, premotor cortex, lateral temporal cortex, and hippocampus (*Figure 4A*; *Supplementary Table 3*). The frontal and temporal lobe cortical regions identified in this analysis are known to support the domains of cognitive dysfunction characterized by the ECAS [28]. These findings provide a potential neuroanatomical basis for the observed polygenic relationships between the wPGS and baseline cognitive performance and rate of decline, and are consistent with prior associations of cortical neurodegeneration with cognitive dysfunction in patients with ALS [29].

Polygenic score associates with neocortical neuronal loss in the UPenn Biobank. To complement these *in vivo* neuroanatomical data, we also explored whether polygenic risk for cognitive dysfunction associated with *post-mortem* anatomical distribution of neuronal loss and TDP-43 pathology. We assessed the magnitude of neuronal loss and TDP-43 pathological inclusions on an ordinal scale in tissue sampled from the middle frontal, cingulate, motor, and superior / middle temporal cortices and from the cornu ammonus 1 (CA1) / subiculum of the hippocampus in 88 autopsy cases from the UPenn Biobank with confirmed ALS due to underlying TDP-43 pathology (*Table 2*). We conducted ordinal logistic regression with covariate adjustment for age at death and disease duration and found that ALS cases with higher wPGS were 2.05 times more likely (95% CI: 1.05, 4.10; p=0.0043) to have greater neuronal loss in the motor cortex relative to ALS cases with a lower wPGS (*Figure 4B*); older age at death and longer disease duration were not found to influence likelihood of greater neuronal loss (p>0.05). We observed no statistically significant associations between the wPGS and neuronal loss in any other region, or between the wPGS and TDP-43 pathology in any other region (all p values>0.1; *Supplementary Figures 7 and 8*). These findings suggest that polygenic risk for cognitive dysfunction is associated with the neuroanatomic distribution of neuronal loss in ALS cases with end-stage disease.

Discussion

In this study, we evaluated polygenic contributions to cognitive dysfunction in patients with ALS by employing machine learning. We identified polygenic risk for cognitive

dysfunction from genetic variables associated with risk of ALS and FTD, which we further investigated through quantitative-trait evaluations of two independent ALS cohorts with *in vivo* neuroimaging and *post-mortem* neuropathology data. Our results indicate a polygenic contribution to the presence and rate of decline of cognitive dysfunction in domains specifically impaired in ALS. Converging evidence from independent cohorts further demonstrates the generalizability of polygenic contribution to biologically-plausible associations including reduced *in vivo* cortical thickness and *post-mortem* cortical neurodegeneration including in the prefrontal, motor, and temporal cortices. These findings contribute novel evidence in support of polygenic contribution to cognitive dysfunction and cortical disease burden in ALS and provide further detailed phenotypic evidence for genetic overlap between ALS and FTD. Below, we highlight clinical, biological, and methodological implications for our observations.

Our findings add to an increasing body of evidence for genetic contribution to phenotypic variability in ALS and support the idea that polygenic variation accounts for a portion of variability in cognitive dysfunction and cortical disease burden in ALS. While cognitive dysfunction has been more frequently linked to genetic mutations causally associated with ALS, such as *C9ORF72* repeat expansions [31], studies examining individual SNPs have demonstrated quantitative-trait modification of cognitive performance and cortical disease burden [22,23]. However, mounting evidence suggests that there are polygenic, rather than single allele, modifiers of disease risk and phenotype in ALS and related neurodegenerative diseases [20,21,24]. Our observation of polygenic association between of 27 SNPs and the ECAS ALS-Specific score, a

combined measure of executive, language, and verbal fluency domains most commonly affected in ALS, is consistent with the idea of polygenic contribution to phenotypic variability in ALS. Notably, our observed polygenic association in the CReATe PGB cohort appears specific to cognitive variability: we demonstrate relative independence of cognitive performance and motor disease severity (i.e. UMN or LMN burden scores, functional performance on the ALSFRS-R) and observe no evidence for polygenic association with motor disease severity. This suggests that, in this study, polygenic risk for cognitive dysfunction does not appear to be confounded by motor disease severity.

The majority (85%) of the 27 SNPs selected by our machine learning modeling for association with cognitive dysfunction are shared risk loci for ALS and FTD [15]. The selection frequency of these ALS and FTD risk variants outweighed the selection of ALS-only risk variants, emphasizing the contribution of genetic overlap between ALS and FTD to polygenic risk associated with cognitive dysfunction in ALS. SNPs in or near the *MOBP, NSF, ATXN3, ERGIC1,* and *UNC13A* genes were among those with the strongest model contributions (i.e. with the highest canonical weights). Our group has previously shown that SNPs mapped to *MOBP*, including rs1768208, relate to regional neurodegeneration in sporadic FTD and to shorter survival in FTD with underlying tau or TDP-43 pathology [32,33]. Our group has also demonstrated that rs12608932 in *UNC13A* relates to *in vivo* prefrontal cortical thinning, *post mortem* frontal cortical burden of TDP-43 pathology, and executive dysfunction [22]. rs538622 near *ERGIC1*, originally identified as a shared risk locus for ALS and FTD, has also been previously demonstrated to contribute to quantitative trait modification in ALS by relating to

reduced expression of the protein BNIP1 in ALS patient motor neurons [15]. Other top-weighted variants near *NSF* and *ATXN3* indicate potential biological plausibility. rs10143310 is found near *ATXN3* which encodes a de-ubiquitinating enzyme, and polyglutamine expansions in *ATXN3* cause spinocerebellar ataxia – type 3 [34]. rs7224296 near *NSF* tags the *MAPT* H1 haplotype [35] and is associated with increased risk for FTD syndromes including progressive supranuclear palsy and corticobasal degeneration [36], as well as Alzheimer's and Parkinson's diseases [37].

While the mechanism of polygenic contribution to cognitive dysfunction in ALS requires further investigation, we speculate based on our findings that identified SNPs may contribute to neuroanatomic disease burden. A weighted polygenic risk score derived from the observed multivariate genotype-phenotype correlation in the CReATe PGB cohort showed robust relationships in independent cohorts from the UPenn Biobank to both *in vivo* cortical thinning and *post-mortem* cortical neuronal loss. Anatomically, these findings were largely consistent with prior *in vivo* structural imaging studies of neurodegeneration associated with cognitive dysfunction and with *post mortem* investigations of cortical thinning in ALS [28,29,38]. Thus, in addition to indicating polygenic contribution to cognitive dysfunction in ALS, our findings suggest a possible mechanism of observed findings via disease pathophysiology.

Beyond the potential biological mechanism of identifying polygenic contributions to ALS disease heterogeneity, we additionally suggest that sCCA may provide a tool for defining polygenic factors of disease risk. While sCCA has been widely applied to

imaging-genetic studies [39], we are unaware of prior applications using sCCA to define a polygenic score based on rich clinical phenotypic and biomarker data. Traditional approaches to the generation of polygenic scores include using data from established, typically case-control GWAS, but practical considerations involve the selection of how many variants to include in a model and how to define the weights of an appropriate statistical model [40]. Critically, rather than an arbitrary selection of variants and their weights, the sparsity parameter of sCCA facilitates an unsupervised, data-driven method to select the number of variants to include and also provides data-driven canonical weights to define the statistical model. The positive or negative direction of model-derived weights is potentially biologically informative, and could reflect 'risk' (i.e. positive weight) or 'protective' (i.e. a negative weight) effects. Further investigation is necessary to clarify the relationships between model-selected SNPs and model-derived canonical weights from both biological (e.g., some SNPs and/or genes may contribute more strongly to risk factors) and mathematical (e.g. weights may be constrained by minor allele frequency) perspectives. Nonetheless, sCCA may provide a promising method for future studies of polygenic variation and may direct research efforts towards model-selected variants.

Several limitations should be considered in the present study. Here, we focus our analysis on a relatively small set of SNPs selected *a priori* from previous large-scale GWAS based on genome-wide association with ALS [12] or shared risk between ALS and FTD [15]. Other genetic variants not included in the present study may also contribute to cognitive dysfunction in ALS and related disorders, and future genome-

wide analyses or broad genotype selection strategies (e.g., targeted pathways) are necessary to elucidate discovery of novel genetic contributions to cognition that have not been identified through prior case-control studies. However, such larger scale studies will require validation in independent cohorts, many of which are lacking the rich phenotype data needed to identify cognitive dysfunction. We derived a weighted polygenic score from sCCA modeling to further investigate polygenic associations with longitudinal cognitive and motor performance, and with *in vivo* and *post-mortem* cortical disease burden in independent ALS cohorts from the UPenn Biobank. While we define our polygenic score from sCCA using adjusted estimates of baseline cognitive and motor performance, future work using longitudinal data as the starting point to define polygenic associations may further elucidate genetic risk for cognitive dysfunction in ALS. However, our finding that polygenic risk associated with baseline cognitive dysfunction also relates to longitudinal cognitive decline in the CReATe PGB cohort and relevant cortical disease anatomy in independent cohorts from the UPenn Biobank suggests its relevance to longitudinal cognitive phenotypes in ALS. Previous critique of polygenic scores suggests that 1) calculation based on GWAS-defined odds ratios for univariate risk loci, and 2) undue influence by population variance, limit their use in clinical and prognostic settings [41]. To avoid these potential confounds, our computation of a weighted polygenic risk score is based on model-selected parameters derived from an analysis including all genetic variants and, in addition, covariates for genetic mutation status and sex in an effort to account for multivariate genetic relationships. We also included the first two principal components in our model from a

PCA conducted in the PGB CReATe cohort in an effort to account for differences in population substructure [42].

Our analyses focused on the investigation of genetic contribution to cognitive dysfunction in ALS, yet it is well established that behavioral impairment is also part of the ALS spectrum disease [43]. Further research is necessary to investigate polygenic risk for behavioral dysfunction in ALS, and whether loci included in our calculated polygenic score confer risk for both cognitive and behavioral dysfunction. While this study demonstrates converging, multimodal evidence for polygenic risk in independent neuroimaging and autopsy cohorts, replication in additional, large cohorts that allow for robust cross-validation is warranted. However, alternative datasets for ALS that contain detailed genotyping and cognitive phenotyping are currently lacking and the CReATe PGB cohort represents the largest of its kind. Future research investigating additional large-scale patient cohorts is necessary.

With these limitations in mind, our research demonstrates converging clinical, neuroimaging, and pathologic evidence for polygenic contribution to cognitive dysfunction and cortical neurodegeneration in ALS. These findings should stimulate further investigation into polygenic risk for cognitive disease vulnerability in ALS and suggest their importance in prognostic consideration and treatment trials. More broadly, this work provides insight into genetic contribution to heterogeneous phenotypes in neurodegenerative disease and supports evidence for polygenic architecture in these conditions.

Materials and Methods

Participants: CReATe Consortium

Participants consisted of 339 individuals clinically diagnosed by a board-certified neurologist with a sporadic or familial form of amyotrophic lateral sclerosis (ALS), amyotrophic lateral sclerosis with frontotemporal dementia (ALS-FTD), progressive muscular atrophy (PMA), or primary lateral sclerosis (PLS) who were enrolled and evaluated through the CReATe Consortium's Phenotype-Genotype-Biomarker (PGB) study. All participants provided written informed consent. The PGB study is registered on clinicaltrials.gov (NCT02327845) and the University of Miami Institutional Review Board (IRB) (the central IRB for the CReATe Consortium) approved the study. This study entails participant blood DNA samples available for genetic screening and longitudinal evaluation at regularly-scheduled visits (ALS, ALS-FTD, and PMA: 0 (baseline), 3, 6, 12, and 18 months; PLS: 0 (baseline), 6, 12, 18, and 24 months). Participants were evaluated at each visit using the ALSFRS-R [44] and alternate versions of the Edinburgh Cognitive and Behavioural ALS Screen (ECAS) [26] designed for longitudinal use. UMN and LMN burden scores were calculated from a detailed elemental neuromuscular examination by summing within and across each spinal region resulting in a score ranging from 0 (none) to 10 (worst). Site (e.g. limb, bulbar) and date of motor symptom onset were recorded for each participant. We excluded nine individuals with missing or incomplete data that precluded subsequent analysis and, in an effort to avoid confounds associated with clear outliers, three individuals with extreme values at baseline on the ECAS Visuospatial Score (i.e. >5 standard deviations from group mean), resulting in a total of 327 participants. Of the nine excluded

individuals with missing or incomplete data, one had no genotyping data available, one had no information for UMN burden score, and seven had no information for date of motor symptom onset.

Genotyping: CReATe Consortium

Peripheral blood mononuclear cell DNA was extracted using the QIAamp DNA Blood Mini Kit Qiagen #51106 and quantified using the Quant-iT dsDNA Assay Kit (Life Technologies cat#Q33130). The DNA integrity was verified by agarose gel electrophoresis (E-Gel, Life Technologies, cat#G8008-01). Unique samples were barcoded and whole genome sequencing (WGS) was performed at the HudsonAlpha Institute for Biotechnology Genomic Services Laboratory (Huntsville, Alabama) (HA) using Illumina HiSeq X10 sequencers to generate approximately 360 million paired-end reads, each 150 base pairs (bp) in length. Peripheral DNA was extracted from participant blood samples and screened for known pathogenic mutations associated with ALS and related diseases.

Screening included repeat-primed polymerase chain reaction (PCR) for *C9ORF72* repeat expansions and WGS curated and validated via Sanger sequencing for pathogenic mutations associated with ALS and/or FTD in *ANG, CHCHD10, CHMP2B, FUS, GRN, hnRNPA1, hnRNPA2B1, MAPT, MATR3, OPTN, PFN1, SETX, SOD1, SPG11, SQSTM1, TARDBP, TBK1, TUBA4A, UBQLN2, VCP* (see *Table 1* for participant mutation status). The PGB study also includes patients with hereditary spastic paraplegia (HSP) that were excluded in the current analysis, but we additionally

screened individuals for pathogenic mutations in 67 additional genes associated with HSP and 7 genes associated with distal hereditary motor neuropathy, and all cases were negative for pathogenic mutations in these genes.

Whole genome sequencing (WGS) data were generated using paired-end 150 bp reads aligned to the GRCh38 human reference using the Burrows-Wheeler Aligner (BWA-ALN v0.7.12) [45] and processed using the Genome Analysis Toolkit (GATK) best-practices workflow implemented in GATK v3.4.0 [46]. Variants for individual samples were called with HaplotypeCaller, producing individual variant call format files (gVCFs) that we combined using a joint genotyping step to produce a multi-sample VCF (pVCF). Variant filtration was performed using Variant Quality Score Recalibration (VQSR), which assigns a score to each variant and a pass/fail label and evaluated this in the context of hard filtering thresholds (Minimum Genotype Quality (GQ)≥ 20, minimum mean depth value (DP)≥ 10). Variant annotation was performed using Variant Effect Predictor (VEP) [47] and in-house pipelines including non-coding variant allele frequencies from Genome Aggregation Database (gnomAD) [48]. In-house scripts were used to identify false positives resulting from paralogous mapping or/and gaps in the current human genome assembly. VCFs were further decomposed prior to analyses using the Decompose function of Vt [49].

To control for population substructure, we additionally derived the first two principal components scores for each in the CReATe PGB cohort using principal components analysis (PCA) implemented using Eigenstrat [42].

From the WGS data we extracted 45 hypothesized variants from WGS that previously achieved genome-wide significance for association with ALS [12] or joint association with ALS and FTD [15]. Proxy loci were genotyped (linkage disequilibrium (LD) R^2 > 0.80) when genetic data were not available for previously-published loci (see *Supplementary Table 1* for a complete list). One locus, rs12973192, was common to both references, and another locus (rs2425220 [15]) was excluded from analysis due to high level of missingness across samples; no LD proxy was identified. We then used PLINK software [50] to recode participant genotypes according to additive genetic models (e.g. 0 = no minor allele copies, 1 = one minor allele copy, 2 = two minor allele copies), since the dominant or recessive nature of the loci included in this study remains unknown.

Linear Mixed-Effects Modeling of the ECAS and clinical measures

We conducted linear mixed-effects modeling of performance on the ECAS, ALSFRS-R, and UMN and LMN burden scores using the *nlme* package in R. Each model was fit using maximum likelihood. In addition to the ECAS Total Score, we analyzed Executive Function, Language, Verbal Fluency, Memory, and Visuospatial sub-scores and ALS-Specific and ALS-Non-Specific summary scores each as dependent variables to analyze patient performance in separate cognitive domains and in clinically-grouped cognitive domains. Fixed effects included age at baseline visit (in years), lag between age of symptom onset and age at baseline visit (in years), college education (yes / no), bulbar onset (yes / no) and visit time-point (in months), and we included individual-by-

visit time-point as a random effect. This allowed us to obtain adjusted estimates of baseline performance (i.e. intercept) and rate of decline (i.e. slope) per individual, having regressed out potential confounding variables as fixed effects.

We conducted Spearman's rank-order correlations between baseline performance and rate of decline using a Bonferroni family-wise error correction for multiple comparisons (see *Figure 1B*).

Sparse Canonical Correlation Analysis

We conducted sparse canonical correlation analysis (sCCA) to select a parsimonious linear combination of variables that maximize the correlation between two multivariate datasets using the *PMA* package in R [27]. The first dataset comprised scaled intercepts from each clinical variable per participant (i.e. adjusted baseline performance on the ALSFRS-R, UMN and LMN assessments, and ECAS). The second comprised minor allele counts per individual for each of the 45 SNPs (e.g. 0 = no minor allele copies, 1 = one minor allele copy, 2 = two minor allele copies), binary variables for sex (0 = Female, 1 = Male), *C9ORF72* repeat expansion status (0 = noncarrier, 1 = carrier), and other mutation status (0 = noncarrier, 1 = carrier) and, in an effort to account for potential population differences in population substructure, we also included the raw estimates for the first two principle components per participant derived from a PCA conducted in the CReATe PGB cohort; this method has previously been demonstrated to account for the majority of population structure [42].

We assumed standard (e.g. unordered) organization of each dataset, and selected regularization parameters for the sCCA analysis using a grid search of 100 combinations of L1 values between 0 (most sparse) and 1 (least sparse) in increments of 0.1. We selected the combination of L1 values yielding the highest canonical correlation of the first variate for subsequent analysis, as similarly reported [51].

Using these L1 parameters, we ran 10,000 bootstrap sCCAs and in each iteration employed randomly-generated subsamples comprising 75% of the PGB cohort. We calculated the median canonical correlation for sCCA and the median canonical weights for each variable across all iterations. We utilized the median in these estimates rather than the maximum or mean value in an effort to avoid bias from outliers and to increase the reliability and reproducibility of model estimates.

We next investigated model performance under a null hypothesis (i.e. no association between clinical and genetic datasets) by using randomly-permuted data. Using the same L1 parameters, we again ran 10,000 bootstrap sCCAs and in each iteration employed randomly-generated subsamples of 75% of participants; however, in each iteration we randomly permuted each dataset 100 times using the *randomizeMatrix* function from the *picante* package in R. We calculated a *p* value by reporting the probability under the null of observing a canonical correlation greater than or equal to the median canonical correlation under sCCA modeling of the true data. We also examined the proportion of iterations each variable was selected by the model (i.e. achieving a non-zero canonical weight).

Polygenic Score

We used the output of sCCA modeling to calculate a weighted polygenic score (wPGS) for each individual. A wPGS for each individual in the PGB cohort, and in the neuroimaging and autopsy UPenn Biobank cohorts, was constructed by multiplying allele dosage or binary coding at each genetic variable by its median canonical weight from sCCA modeling, and summing across all values.

To investigate construct validity, we first conducted Spearman's rank-order correlations between the wPGS and adjusted estimates of baseline performance (i.e. LME-derived intercepts) on the most frequently selected clinical measure(s) selected from sCCA.

Then, to investigate longitudinal performance associated with the wPGS, we conducted Spearman's rank-order correlations between the wPGS and adjusted rates of decline (i.e. LME-derived slopes) on all clinical measures using a Bonferroni family-wise error correction. We restricted this analysis to participants in the CReATe PGB cohort with data at 2 or more timepoints (N=277 out of 327 participants), or 84.7% of the cohort.

Participants: UPenn Biobank neuroimaging cohort

We retrospectively evaluated 114 patients with ALS and 114 healthy controls matched for age, sex, and education from the UPenn Biobank who were recruited for research between 2006 and 2019 from the Penn Comprehensive ALS Clinic and Penn Frontotemporal Degeneration Center (*Table 2*) [25]. Inclusion criteria for ALS patients

consisted of complete genotyping at the 45 analyzed SNPs, screening for genetic mutations (e.g. *C9ORF72, SOD1*), white non-Latino racial and ethnic background (population diversity is known to influence allele frequencies across individuals), disease duration from symptom onset < 2.5 standard deviations from respective group means (to avoid confounds associated with clear outliers), and T1-weighted MRI. All patients were diagnosed with ALS by a board-certified neurologist (L.E., L.M., M.G., D.I) using revised El Escorial criteria [52] and assessed for ALS frontotemporal spectrum disorder using established criteria [53]; those patients enrolled in research prior to 2017 were retrospectively evaluated through chart review. All ALS patients and controls participated in an informed consent procedure approved by an IRB convened at UPenn.

Participants: UPenn Biobank autopsy cohort

We evaluated brain tissue samples from 88 ALS autopsy cases identified from the UPenn Biobank [25] who were diagnosed by a board-certified neuropathologist (J.Q.T., E.B.L.) with ALS due to TDP-43 pathology using immunohistochemistry [54] and published criteria [55]; this cohort included 21 patients from the ALS neuroimaging cohort. Inclusion criteria consisted of complete genotyping at the 45 analyzed SNPs, screening for genetic mutations (e.g. *C9ORF72, SOD1*), white non-Latino racial and ethnic background (population diversity is known to influence allele frequencies across individuals), disease duration from symptom onset < 2.5 standard deviations from respective group means (to avoid confounds associated with clear outliers), and brain tissue samples from the middle frontal, motor, cingulate, and superior / temporal cortices, and the cornu ammonis 1 (CA1) / subiculum of the hippocampus for analysis of

neuronal loss and TDP-43 pathology. Nine individuals were missing neuronal loss or TDP-43 pathology data for at least one sampled region (*Supplementary Table 2*).

Genetic Screening and SNP Genotyping: UPenn Biobank

DNA was extracted from peripheral blood or frozen brain tissue following the manufacturer's protocols (Flexigene (Qiagen) or QuickGene DNA whole blood kit (Autogen) for blood, and QIAsymphony DNA Mini Kit (Qiagen) for brain tissue). All patients were screened for *C9ORF72* hexanucleotide repeat expansions using a modified repeat-primed PCR as previously described [56], and we excluded any patient with > 30 hexanucleotide repeats. Of the remaining individuals, we evaluated family history using a three-generation pedigree history, as previously reported [57]. For cases with a family history of the same disease, we sequenced 45 genes previously associated with neurodegenerative disease, including genes known to be associated with ALS (e.g. *SOD1* [19], TBK1 [10]). Sequencing was performed using a custom-targeted next-generation sequencing panel (MiND-Seq) [25] and analyzed using Mutation Surveyor software (Soft Genetics, State College, PA).

DNA extracted from peripheral blood or cerebellar tissue samples was genotyped for each case using the Illumina Infinium Global Screening Array through the Children's Hospital of Philadelphia (CHOP) Center for Applied Genomics Core according to manufacturer's specifications. PLINK [50] was then used to remove variants with <95% call rate, Hardy-Weinberg equilibrium (HWE) p-value$<10^{-6}$ and individuals with >5% missing genotypes. Using the remaining genotypes from samples passing quality

control, we performed genome-wide imputation of allele dosages with the Haplotype Reference Consortium reference panel r1.1 [58] on the Michigan Imputation Server [59] to predict genotypes at ungenotyped genomic positions, applying strict pre-phasing, pre-imputation filtering, and variant position and strand alignment control.

Neuroimaging Processing and Analyses

High-resolution T1-weighted MPRAGE structural scans were acquired for neuroimaging participants using a 3T Siemens Tim Trio scanner with an 8-channel head coil, with T=1620ms, T=3.09ms, flip angle=15°, 192x256 matrix, and 1mm^3 voxels. T1-weighted MRI images were then preprocessed using Advanced Normalization Tools (ANTs) software [60]. Each individual dataset was deformed into a standard local template space in a canonical stereotactic coordinate system. ANTs provide a highly accurate registration routine using symmetric and topology-preserving diffeomorphic deformations to minimize bias toward the reference space and to capture the deformation necessary to aggregate images in a common space. Then, we used N4 bias correction to minimize heterogeneity [61] and the ANTs Atropos tool to segment images into six tissue classes (cortex, white matter, cerebrospinal fluid, subcortical grey structures, brainstem, and cerebellum) using template-based priors and to generate probability maps of each tissue. Voxel-wise cortical thickness was measured in millimeters (mm^3) from the pial surface and then transformed into Montreal Neurological Institute (MNI) space, smoothed using a three sigma full-width half-maximum Gaussian kernel, and downsampled to 2mm isotropic voxels.

We used *randomise* software from FSL to perform nonparametric, permutation-based statistical analyses of cortical thickness images from the UPenn Biobank neuroimaging cohort. Permutation-based statistical testing is robust to concerns regarding multiple comparisons since, rather than a traditional assessment of two sample distributions, this method assesses a true assignment of factors (e.g. wPGS) to cortical thickness compared to many (e.g., 10,000) random assignments [62].

First, we used *randomise* set to 10,000 permutations to identify reduced cortical thickness in ALS patients relative to healthy controls. We constrained this analysis using an explicit mask restricted to high probability cortex (>0.4) and reported clusters that survive $p<0.05$ threshold-free cluster enhancement (TFCE) [63] corrected for family-wise error.

Next, we again used *randomise* set to 10,000 permutations to identify regions of reduced cortical thickness associated with wPGS in ALS patients, constraining analysis to an explicit mask defined by regions of reduced cortical thickness in ALS patients relative to controls (see above). The statistical model for this analysis included covariate adjustment for age, disease duration, and scanner acquisition. We report clusters that survive uncorrected $p<0.01$ with a cluster extent threshold of 10 voxels; we employ an uncorrected threshold to minimize the chance of Type II error (not observing a true result).

Neuropathology Processing and Analyses

The extent of neuronal loss and of phosphorylated TDP-43 intraneuronal inclusions (dots, wisps, skeins) in sampled regions from the middle frontal, cingulate, motor, and superior / middle temporal cortices, and the CA1 / subiculum of the hippocampus were assessed on an ordinal scale: 0=none/rare, 1=mild, 2=moderate, 3=severe/numerous. All neuropathological ratings were performed by an expert neuropathologist (J.Q.T., E.B.L.) blinded to patient genotype. We conducted ordinal logistic regression using the *MASS* package in *R* to investigate whether extent of neuronal loss rated using Hematoxylin and eosin (H&E) and burden of TDP-43 pathology rated using mAbs p409/410 or 171 [64,65] immunohistochemistry differed according to wPGS, with covariate adjustment for age and disease duration at death.

Acknowledgements
The CReATe Consortium (U54NS092091) is part of the Rare Diseases Clinical Research Network (RDCRN), an initiative of the Office of Rare Diseases Research (ORDR), NCATS. CReATe is funded through a collaboration between NCATS, and the NINDS. Additional research support was provided by the National Institutes of Health (NS106754, AG017586, NS092091, AG054060). The genomics sequencing was funded by St. Jude Children's Research Hospital American Lebanese Syrian Associated Charities (ALSAC), with additional support from the ALS Association for biorepository and sequencing costs (grants 17-LGCA-331 and 16-TACL-242).

Author Contributions

Study concept/design: K.P., M.B., C.T.M.

Acquisition, analysis, or interpretation of data: K.P., M.B., J.W., E.R., L.H., V.V.D., D.J.I., L.E., L.M., C.Q., V.G., J.S., T.B., J.R., A.S., J.K., E.P., C.J., J.C., Y.S., S.M., D.W., E.B.L., J.Q.T., P.C., J.G., J.S., A.C.N., R.R., G.W., J.P.T., and C.T.M.

Drafting/revising manuscript: K.P., M.B., C.T.M., M.G., E.B.L., D.W., V.G., A.N., E.R., G.W., W.C., J.S., T.B.

Competing Interests statement

The following authors declare the following competing interests:

C.T.M. receives financial support from Biogen and has provided consulting for Axon Advisors. M.B. reports grants from National Institutes of Health, the ALS Association, the Muscular Dystrophy Association, the Centers for Disease Control and Prevention, the Department of Defense, and Target ALS during the conduct of the study; personal

fees from Mitsubishi Tanabe Pharma, AveXis, and Genentech, outside the submitted work. In addition, M.B. has a provisional patent entitled ‘Determining Onset of Amyotrophic Lateral Sclerosis,’ and serves as a site investigator on clinical trials funded by Biogen and Orphazyme. All other authors declare no competing interests.

References

1. Montuschi A, Iazzolino B, Calvo A, Moglia C, Lopiano L, Restagno G, Brunetti M, Ossola I, Presti Lo A, Cammarosano S, et al. (2015) Cognitive correlates in amyotrophic lateral sclerosis: a population-based study in Italy. *Journal of Neurology, Neurosurgery & Psychiatry* **86**: 168–173.
2. Beeldman E, Raaphorst J, Twennaar MK, de Visser M, Ben A Schmand, de Haan RJ (2016) The cognitive profile of ALS: a systematic review and meta-analysis update. *Journal of Neurology, Neurosurgery & Psychiatry* **87**: 611–619.
3. Elamin M, Bede P, Byrne S, Jordan N, Gallagher L, Wynne B, O'Brien C, Phukan J, Lynch C, Pender N, et al. (2013) Cognitive changes predict functional decline in ALS: A population-based longitudinal study. *Neurology* **80**: 1590–1597.
4. Crockford C, Newton J, Lonergan K, Chiwera T, Booth T, Chandran S, Colville S, Heverin M, Mays I, Pal S, et al. (2018) ALS-specific cognitive and behavior changes associated with advancing disease stage in ALS. *Neurology* **91**: e1370–e1380.
5. Hu WT, Shelnutt M, Wilson A, Yarab N, Kelly C, Grossman M, Libon DJ, Khan J, Lah JJ, Levey AI, et al. (2013) Behavior Matters—Cognitive Predictors of Survival in Amyotrophic Lateral Sclerosis. *PLoS ONE* **8**: e57584.
6. Caga J, Hsieh S, Lillo P, Dudley K, Mioshi E (2019) The Impact of Cognitive and Behavioral Symptoms on ALS Patients and Their Caregivers. *Front Neurol* **10**: 942.
7. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, Nicholson AM, Finch NA, Flynn H, Adamson J, et al. (2011) Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron* **72**: 245–256.
8. Renton AE, Majounie E, Waite A, Simon-Sanchez J, Rollinson S, Gibbs JR, Schymick JC, Laaksovirta H, Van Swieten JC, Myllykangas L, et al. (2011) A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**: 257–268.
9. Van Deerlin VM, Leverenz JB, Bekris LM, Bird TD, Yuan W, Elman LB, Clay D, Wood EM, Chen-Plotkin AS, Martinez-Lage M, et al. (2008) TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *The Lancet Neurology* **7**: 409–416.
10. Freischmidt A, Wieland T, Richter B, Ruf W, Schaeffer V, Müller K, Marroquin N, Nordin F, Hübers A, Weydt P, et al. (2015) Haploinsufficiency of *TBK1* causes familial ALS and fronto-temporal dementia. *Nat Neurosci* **18**: 631–636.
11. van Rheenen W, Shatunov A, Dekker AM, McLaughlin RL, Diekstra FP, Pulit SL, van der Spek RAA, Võsa U, de Jong S, Robinson MR, et al. (2016) Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet* **48**: 1043–1048.
12. Nicolas A, Kenna KP, Renton AE, Faghri F, Chia R, Dominov JA, Kenna BJ, Nalls MA, Keagle P, Rivera AM, et al. (2018) Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* **97**: 1268–1282.e6.
13. van Es MA, Veldink JH, Saris CGJ, Blauw HM, van Vught PWJ, Birve A, Lemmens R, Schelhaas HJ, Groen EJN, Huisman MHB, et al. (2009) Genome-

wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat Genet* **41**: 1083–1087.
14. Diekstra FP, Van Deerlin VM, Van Swieten JC, Al-Chalabi A, Ludolph AC, Weishaupt JH, Hardiman O, Landers JE, Brown RH, van Es MA, et al. (2014) C9orf72 and UNC13A are shared risk loci for amyotrophic lateral sclerosis and frontotemporal dementia: A genome‐wide meta‐analysis. *Ann Neurol* **76**: 120–133.
15. Karch CM, Wen N, Fan CC, Yokoyama JS, Kouri N, Ross OA, Höglinger G, Müller U, Ferrari R, Hardy J, et al. (2018) Selective Genetic Overlap Between Amyotrophic Lateral Sclerosis and Diseases of the Frontotemporal Dementia Spectrum. *JAMA Neurol* **75**: 860–16.
16. Turner MR, Al-Chalabi A, Chiò A, Hardiman O, Kiernan MC, Rohrer JD, Rowe J, Seeley W, Talbot K (2017) Genetic screening in sporadic ALS and FTD. *Journal of Neurology, Neurosurgery & Psychiatry* **88**: 1042–1044.
17. Umoh ME, Fournier C, Li Y, Polak M, Shaw L, Landers JE, Hu W, Gearing M, Glass JD (2016) Comparative analysis of C9orf72 and sporadic disease in an ALS clinic population. *Neurology* **87**: 1024–1030.
18. Kenna KP, van Doormaal PTC, Dekker AM, Ticozzi N, Kenna BJ, Diekstra FP, van Rheenen W, van Eijk KR, Jones AR, Keagle P, et al. (2016) NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat Genet* **48**: 1037–1042.
19. Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, Donaldson D, Goto J, O'Regan JP, Deng H-X, et al. (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**: 59–62.
20. McLaughlin RL, Schijven D, van Rheenen W, van Eijk KR, O'Brien M, Kahn RS, Ophoff RA, Goris A, Bradley DG, Al-Chalabi A, et al. (2017) Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. *Nat Commun* **8**: 14774.
21. Ciga SB, Noyce AJ, Hemani G, Nicolas A, Calvo A, Mora G, Tienari PJ, Stone DJ, Nalls MA, Singleton AB, et al. (2019) Shared polygenic risk and causal inferences in amyotrophic lateral sclerosis. *Ann Neurol* **85**: 470–481.
22. Placek K, Baer GM, Elman L, McCluskey L, Hennessy L, Ferraro PM, Lee EB, Lee VMY, Trojanowski JQ, Van Deerlin VM, et al. (2019) UNC13A polymorphism contributes to frontotemporal disease in sporadic amyotrophic lateral sclerosis. *Neurobiology of Aging* **73**: 190–199.
23. Vass R, Ashbridge E, Geser F, Hu WT, Grossman M, Clay-Falcone D, Elman L, McCluskey L, Lee VMY, Van Deerlin VM, et al. (2011) Risk genotypes at TMEM106B are associated with cognitive impairment in amyotrophic lateral sclerosis. *Acta Neuropathologica* **121**: 373–380.
24. Hagenaars SP, Radakovic R, Crockford C, Fawns-Ritchie C, IFGC IF-GC, Harris SE, Gale CR, Deary IJ (2018) Genetic risk for neurodegenerative disorders, and its overlap with cognitive ability and physical function. *PLoS ONE* **13**: e0198187.
25. Toledo JB, Van Deerlin VM, Lee EB, Suh E, Baek Y, Robinson JL, Xie SX, McBride J, Wood EM, Schuck T, et al. (2014) A platform for discovery: The University of Pennsylvania Integrated Neurodegenerative Disease Biobank. *Alzheimer's & Dementia* **10**: 477–484.e1.
26. Abrahams S, Newton J, Niven E, Foley J, Bak TH (2014) Screening for cognition

and behaviour changes in ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* **15**: 9–14.
27. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**: 515–534.
28. Lulé D, Böhm S, Müller H-P, Aho-Özhan H, Keller J, Gorges M, Loose M, Weishaupt JH, Uttner I, Pinkhardt E, et al. (2018) Cognitive phenotypes of sequential staging in amyotrophic lateral sclerosis. *CORTEX* **101**: 163–171.
29. Agosta F, Ferraro PM, Riva N, Spinelli EG, Chiò A, Canu E, Valsasina P, Lunetta C, Iannaccone S, Copetti M, et al. (2016) Structural brain correlates of cognitive and behavioral impairment in MND. *Hum Brain Mapp* **37**: 1614–1626.
30. Müller H-P, Kassubek J (2018) MRI-Based Mapping of Cerebral Propagation in Amyotrophic Lateral Sclerosis. *Front Neurosci* **12**: 655.
31. Byrne S, Elamin M, Bede P, Shatunov A, Walsh C, Corr B, Heverin M, Jordan N, Kenna K, Lynch C, et al. (2012) Cognitive and clinical characteristics of patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: a population-based cohort study. *The Lancet Neurology* **11**: 232–240.
32. Irwin DJ, McMillan CT, Suh E, Powers J, Rascovsky K, Wood EM, Toledo JB, Arnold SE, Lee VMY, Van Deerlin VM, et al. (2014) Myelin oligodendrocyte basic protein and prognosis in behavioral-variant frontotemporal dementia. *Neurology* **83**: 502–509.
33. McMillan CT, Toledo JB, Avants BB, Cook PA, Wood EM, Suh E, Irwin DJ, Powers J, Olm C, Elman L, et al. (2014) Genetic and neuroanatomic associations in sporadic frontotemporal lobar degeneration. *Neurobiology of Aging* **35**: 1473–1482.
34. Burnett B, Li F, Pittman RN (2003) The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity. *Human Molecular Genetics* **12**: 3195–3205.
35. Yokoyama JS, Karch CM, Fan CC, Bonham LW, Kouri N, Ross OA, Rademakers R, Kim J, Wang Y, Höglinger GU, et al. (2017) Shared genetic risk between corticobasal degeneration, progressive supranuclear palsy, and frontotemporal dementia. *Acta Neuropathol* **133**: 825–837.
36. Ferrari R, Wang Y, Vandrovcova J, Guelfi S, Witeolar A, Karch CM, Schork AJ, Fan CC, Brewer JB, International FTD-Genomics Consortium (IFGC),, et al. (2017) Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer‘s and Parkinson’s diseases. *Journal of Neurology, Neurosurgery & Psychiatry* **88**: 152–164.
37. Desikan RS, Schork AJ, Wang Y, Witoelar A, Sharma M, McEvoy LK, Holland D, Brewer JB, Chen C-H, Thompson WK, et al. (2015) Genetic overlap between Alzheimer‘s disease and Parkinson’s disease at the MAPT locus. *Mol Psychiatry* **20**: 1588–1595.
38. Prudlo J, König J, Schuster C, Kasper E, Büttner A, Teipel S, Neumann M (2016) TDP-43 pathology and cognition in ALS: A prospective clinicopathologic correlation study. *Neurology* **87**: 1019–1023.
39. Parkhomenko E, Tritchler D, Beyene J (2009) Sparse canonical correlation analysis with application to genomic data integration. *Stat Appl Genet Mol Biol* **8**:

Article1–Article34.
40. Sugrue LP, Desikan RS (2019) What Are Polygenic Scores and Why Are They Important? *JAMA* **321**: 1820–1821.
41. Wald NJ, Old R (2019) The illusion of polygenic disease risk prediction. *Genetics in Medicine 2019* **319**: 1.
42. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.
43. Lillo P, Mioshi E, Zoing MC, Kiernan MC, Hodges JR (2010) How common are behavioural changes in amyotrophic lateral sclerosis? *Amyotrophic Lateral Sclerosis* **12**: 45–51.
44. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, Nakanishi A (1999) The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences* **169**: 13–21.
45. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–595.
46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.
47. Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, et al. (2018) Ensembl variation resources. *Database (Oxford)* **2018**: 1193.
48. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* **49**: 531210.
49. Tan A, Abecasis GR, Kang HM (2015) Unified representation of genetic variants. *Bioinformatics* **31**: 2202–2204.
50. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**: 559–575.
51. Xia CH, Ma Z, Ciric R, Gu S, Betzel RF, Kaczkurkin AN, Calkins ME, Cook PA, la Garza de AG, Vandekar SN, et al. (2018) Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat Commun* 1–14.
52. Brooks BR, Miller RG, Swash M, Munsat TL, World Federation of Neurology Research Group on Motor Neuron Diseases (2000) El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. In pp 293–299.
53. Strong MJ, Abrahams S, Goldstein LH, Woolley S, Mclaughlin P, Snowden J, Mioshi E, Roberts-South A, Benatar M, HortobáGyi T, et al. (2017) Amyotrophic lateral sclerosis - frontotemporal spectrum disorder (ALS-FTSD): Revised diagnostic criteria. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* **18**: 153–174.
54. Neumann M, Sampathu DM, Kwong LK, Truax AC, Micsenyi MC, Chou TT, Bruce

J, Schuck T, Grossman M, Clark CM, et al. (2006) Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. *Science* **314**: 130–133.

55. Mackenzie IRA, Neumann M, Baborie A, Sampathu DM, Plessis Du D, Jaros E, Perry RH, Trojanowski JQ, Mann DMA, Lee VMY (2011) A harmonized classification system for FTLD-TDP pathology. *Acta Neuropathol* **122**: 111–113.
56. Suh E, Lee EB, Neal D, Wood EM, Toledo JB, Rennert L, Irwin DJ, McMillan CT, Krock B, Elman LB, et al. (2015) Semi-automated quantification of C9orf72 expansion size reveals inverse correlation between hexanucleotide repeat number and disease duration in frontotemporal degeneration. *Acta Neuropathologica* **130**: 363–372.
57. Wood EM, Falcone D, Suh E, Irwin DJ, Chen-Plotkin AS, Lee EB, Xie SX, Van Deerlin VM, Grossman M (2013) Development and Validation of Pedigree Classification Criteria for Frontotemporal Lobar Degeneration. *JAMA Neurol* **70**: 1411–1417.
58. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**: 1279–1283.
59. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. (2016) Next-generation genotype imputation service and methods. *Nat Genet* **48**: 1284–1287.
60. Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, Kandel BM, van Strien N, Stone JR, Gee JC, et al. (2014) Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *NeuroImage* **99**: 166–179.
61. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: Improved N3 Bias Correction. *IEEE Trans Med Imaging* **29**: 1310–1320.
62. Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE (2014) Permutation inference for the general linear model. *NeuroImage* **92**: 381–397.
63. Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* **44**: 83–98.
64. Lippa CF, Rosso AL, Stutzbach LD, Neumann M, Lee VMY, Trojanowski JQ (2009) Transactive Response DNA-Binding Protein 43 Burden in Familial Alzheimer Disease and Down Syndrome. *Arch Neurol* **66**: 1483–1488.
65. Neumann M, Kwong LK, Lee EB, Kremmer E, Flatley A, Xu Y, Forman MS, Troost D, Kretzschmar HA, Trojanowski JQ, et al. (2009) Phosphorylation of S409/410 of TDP-43 is a consistent feature in all sporadic and familial forms of TDP-43 proteinopathies. *Acta Neuropathologica* **117**: 137–149.

Figure Legends.

1. **Clinical and genetic heterogeneity in the CReATe PGB cohort.** A) Differences in baseline performance and rate of decline on each clinical measure for each participant; the heatmap indicates each participant's standard deviation (SD) from the group mean. B) Spearman's correlations between baseline performance and rate of decline for all clinical measures. C) Allele dosage or binary status for each genetic variable for each participant.

2. **Sparse, polygenic relationship between clinical and genetic variation in ALS.** Variable selection and median canonical weight strength from bootstrap sparse canonical correlation analysis (sCCA) modeling in the CReATe PGB cohort.

3. **Polygenic risk score correlates with cognitive performance on the ECAS in the CReATe PGB cohort.** Weighted polygenic risk score (wPGS) correlates with A) adjusted baseline performance on the Edinburgh Cognitive and Behavioral ALS Screen (ECAS) ALS-Specific, Total, Executive Function, and Language scores, and B) rate of decline on the ALS-Specific, ALS-Non-Specific, and Total scores.

4. **Reduced cortical thickness and greater cortical neuronal loss relates to higher polygenic risk score in independent validation cohorts.** A) ALS patients (N=114) from the UPenn Biobank neuroimaging cohort with higher weighted polygenic risk score (wPGS) exhibited greater reduction of cortical thickness in the orbital prefrontal cortex, anterior cingulate cortex, premotor cortex, lateral temporal cortex, and hippocampus. The heatmap indicates the associated T-statistic for each voxel, with light blue representing the highest value. B) Magnitude of motor cortex neuronal loss in ALS cases (N=88) from the UPenn Biobank is associated with higher wPGS.

Supplementary Figure Legends.

1. **Gridsearch for sparse canonical correlation analysis (sCCA) L1 parameters.** Each column indicates 1 of 100 unique combinations of L1 parameters (ranging 0.1 to 1) applied to clinical and genetic datasets, and each row lists a variable entered into the sCCA. The heatmap denotes the canonical weight strength for each variable; warmer colors indicate positive weights and cooler colors indicate negative weights.

2. **Bootstrapped sparse canonical correlation analysis (sCCA) modeling.** Each column indicates 1 of 10,000 iterations of sparse canonical correlation analysis (sCCA); in each iteration a randomly-bootstrapped subsample of 75% of participants in the CReATe PGB cohort was employed. Each row lists a variable entered into the sCCA. The heatmap denotes the canonical weight strength for each variable; warmer colors indicate positive weights and cooler colors indicate negative weights.

3. ***p* value calculation for sCCA modeling.** Histogram showing the frequency of canonical correlations achieved from sparse canonical correlation analysis (sCCA) modeling under the null hypothesis. The vertical turquoise line denotes the median canonical correlation achieved under true sCCA modeling, and the *p* value demonstrates the proportion of times the median canonical correlation under true modeling was achieved by sCCA modeling under the null hypothesis.

4. **Variables selected in sCCA modeling.** A) Bar graphs demonstrating the proportion of times out of 10,000 iterations that each of the 11 clinical variables were selected by sparse canonical correlation analysis (sCCA) under true modeling (turquoise) and modeling under the null hypothesis (coral). B) Bar graphs demonstrating the number of times out of 10,000 randomly-bootstrapped sCCAs that each of the 45 SNPs were selected by sCCA under true modeling (turquoise) and modeling under the null hypothesis (coral). SNPs are organized according to prior genome-wide association with ALS or joint association with ALS and FTD.

5. **Polygenic risk score correlates with rate of decline in cognitive performance on the ECAS.** Scatterplots demonstrating relationships between weighted polygenic risk score (wPGS) and rate of decline on each Edinburgh Cognitive and Behavioral ALS Screen (ECAS) score, the ALS Functional Rating Scale – Revised (ALSFRS-R), and upper motor neuron (UMN), lower motor neuron (LMN) burden scores in the CReATe PGB cohort.

6. **Reduced cortical thickness in ALS patients relative to healthy controls.** ALS patients (N=114) from the UPenn Biobank neuroimaging cohort displayed widespread cortical thinning relative to age, sex, and education-matched healthy controls in the frontal and temporal lobes. The heatmap indicates the associated T-statistic for each voxel, with light yellow representing the highest value.

7. **Magnitude of neuronal loss in ALS patients relative to wPGS.** Beeswarm boxplots of ordinal measures of neuronal loss in ALS cases (N=88) from the UPenn Biobank autopsy cohort relative to wPGS in the cingulate cortex, motor cortex, middle frontal cortex, superior / middle temporal cortex, and hippocampus.

8. **Magnitude of TDP-43 pathology in ALS patients relative to wPGS.** Beeswarm boxplots of ordinal measures of TDP-43 pathology in ALS cases (N=88) the UPenn Biobank autopsy cohort relative to wPGS in the cingulate cortex, motor cortex, middle frontal cortex, superior / middle temporal cortex, and hippocampus.

Tables

Table 1: Demographic Characteristics of the CReATe PGB cohort.

		ALS	ALS-FTD	PLS	PMA
N		279	13	22	13
Sex	**Male (%)**	163 (58.4)	11 (84.6)	11 (50.0)	8 (61.5)
Number of Visits	**M (SD)**	3.09 (1.37)	3.00 (1.15)	2.86 (1.28)	3.38 (1.45)
Age at Symptom Onset	**M (SD)**	56.32 (12.56)	64.00 (9.11)	49.68 (7.39)	48.08 (15.31)
Symptom Onset to Baseline (years)	**M (SD)**	3.59 (4.98)	3.62 (2.63)	8.45 (6.12)	7.77 (7.17)
Site of Symptom Onset	**N (%)**				
Bulbar		45 (17.1)	4 (33.3)	5 (22.7)	-
Bulbar & Limb		7 (2.7)	-	3 (13.6)	-
Bulbar & Other		7 (2.7)	1 (8.3)	-	-
Limb		175 (66.5)	3 (25)	13 (59.1)	11 (84.6)
Limb & Other		22 (8.4)	-	1 (4.5)	1 (7.7)
Other		7 (2.7)	4 (33.3)	-	1 (7.7)
College Education or greater	**N (%)**	196 (71.3)	9 (69.2)	20 (90.9)	10 (76.9)
Mutation Carrier	**N (%)**	34 (12.2)	3 (20.0)	0 (0.0)	0 (0.0)
C9ORF72		22 (7.9)	3 (20.0)	-	-
C9ORF72 and UBQLN2		1 (0.4)	-	-	-
SOD1		8 (2.9)	-	-	-
SQSTM1		1 (0.4)	-	-	-
TARDBP		1 (0.4)	-	-	-
TBK1		1 (0.4)	-	-	-
Baseline ALSFRS-R (0-48)	**M (SD)**	35.00 (7.09)	35.00 (5.99)	36.50 (5.95)	33.62 (7.83)
UMN Score (0-10)	**M (SD)**	2.70 (1.68)	2.45 (2.00)	4.54 (1.33)	0.87 (0.73)
LMN Score (0-10)	**M (SD)**	2.54 (1.48)	2.81 (1.76)	0.59 (0.96)	4.84 (1.93)
ECAS	**M (SD)**				
ALS-Specific (0-100)		80.94 (10.85)	52.62 (12.07)	87.95 (7.47)	81.62 (11.61)
Language (0-28)		25.85 (2.66)	21.38 (3.93)	26.82 (1.97)	26.62 (1.26)
Verbal Fluency (0- 24)		16.62 (5.11)	7.83 (5.36)	26.82 (1.97)	16.77 (4.36)
Executive (0-48)		38.47 (5.94)	24.00(10.51)	26.82 (1.97)	38.23 (7.50)
ALS-Non-Specific (0-36)		28.04 (3.78)	19.69 (8.30)	29.73 (2.76)	27.62 (6.31)
Memory (0-24)		16.45 (3.54)	9.46 (7.15)	17.95 (2.84)	15.69 (6.20)
Visuospatial (0- 12)		11.59 (0.79)	11.08 (1.24)	11.77 (0.43)	11.92 (0.28)
Total (0-136)		108.97 (13.02)	72.31 (18.53)	117.68 (9.12)	109.23 (16.47)

PGB = Phenotype-Genotype-Biomarker; CReATe = Clinical Research in ALS and Related Disorders for Therapeutic Development; ALS = Amyotrophic lateral sclerosis, ALS-FTD = ALS-Frontotemporal dementia; PLS = Primary lateral sclerosis, PMA = Progressive muscular atrophy; ALSFRS-R = Revised ALS Functional Rating Scale; UMN = upper motor neuron; LMN = lower motor neuron; ECAS = Edinburgh Cognitive and Behavioral ALS Screen; M= mean, SD = standar deviation.

Table 2. Demographics for independent neuroimaging (A) and autopsy (B) amyotrophic lateral sclerosis (ALS) and healthy control cohorts from UPenn Biobank.

A. Neuroimaging Cohort

	ALS	**Healthy Control**
N (Male)	114 (64)	114 (64)
Age at MRI in Years, M (SD)	59.34 (10.92)	61.87 (12.18)
Education in Years, M (SD)	15.09 (2.98)	15.87 (2.47)
Disease Duration in Years, M (SD)	3.02 (2.52)	-
Mutation Carrier, N (%)		
C9ORF72	14 (12.28)	-
SOD1	1 (0.87)	-
VCP	1(0.87)	-
Site of Symptom Onset, N (%)		
Bulbar	26 (22.81)	-
Limb	79 (69.3)	-
Cognitive	9 (7.89)	-
ALSFRS-R, M (SD)	33.23 (7.32)	-

B. Autopsy Cohort

N (Male)	88 (49)
Age at Death Years, M (SD)	63.72 (10.24)
Disease Duration at Death in Years, M (SD)	4.24 (3.41)
Mutation Carrier, N (%)	
C9ORF72	15 (17.04)
Site of Symptom Onset, N (%)	
Bulbar	23
Limb	60
Cognitive	3
Respiratory	1
Unknown	1

Abbreviations: ALSFRS-R = ALS Functional Rating Scale – Revised, M = Mean, SD = standard deviation

Supplementary Tables

Supplementary Table 1: List of genetic variants analyzed in the CReATe PGB Study.

Marker Name	Nearest Gene	Chr	1000 Genome GMAF	GRCh38 Position	Proxy Marker	Proxy HG19.Position
rs2068667	*NFASC*	1	0.208	chr1:204948552	rs11240317	chr1:204920322
rs11185393	*AMY1A*	1	0.368	chr1:104209379	rs67205957	chr1:104752258
rs515342	*ASB1*	2	0.214	chr2:238458655	rs508986	chr2:239337691
rs9820623	*MOBP*	3	0.406	chr3:39452367	rs6765697	chr3:39493239
rs13079368	*MOBP*	3	0.275	chr3:39471060	rs1464047	chr3:39526874
rs1768208	*MOBP*	3	0.323	chr3:39481512	rs616147	chr3:39534481
rs10463311	*TNIP1*	5	0.431	chr5:151031274	-	-
rs3828599	*GPX3*	5	0.417	chr5:151022235	rs4958872	chr5:150402334
rs538622	*ERGIC1*	5	0.32	chr5:172920676	rs2446192	chr5:172352369
rs17111695	*NAF1*	5	0.183	chr5:151052885	rs12518386	chr5:150438085
rs757651	*REEP2*	5	0.016	chr5:138455779	rs149312547	chr5:137792021
rs10488631	*TNPO3*	7	0.059	chr7:128954129	rs12539741	chr7:128596805
rs17070492	*LOC101927815*	8	0.208	chr8:2563763	-	-
rs7813314	*BC045738*	8	0.2	chr8:2558274	rs6996532	chr8:2417678
rs10869188	*C9ORF72*	9	0.49	chr9:72614090	rs7032232	chr9:75229116
rs870901	*AK097706*	9	0.133	chr9:107086201	rs60743641	chr9:109854824
rs10511816	*MOBKL2B*	9	0.206	chr9:27468463	rs12551344	chr9:27466817
rs3849943	*C9ORF72*	9	0.183	chr9:27543384	-	-
rs3849942	*C9ORF72*	9	0.183	chr9:27543283	-	-
rs13302855	*C9ORF72*	9	0.086	chr9:27595997	rs34460171	chr9:27594491
rs3849943	*C9ORF72*	9	0.183	chr9:27543384	-	-
rs732389	*AK294518*	10	0.205	chr10:78584745	rs7071538	chr10:80338173
rs7118388	*CAT*	11	0.454	chr11:34432600	rs1962369	chr11:34456941
rs12803540	*CAT*	11	0.138	chr11:34471200	rs17881488	chr11:34492443
rs117027576	*KIF5A*	12	0.00913	chr12:56922819	-	-
rs113247976	*KIF5A*	12	0.007	chr12:57581917	-	-
rs142321490	*KIF5A*	12	0.006	chr12:58282349	-	-
rs74654358	*TBK1*	12	0.012	chr12:64488187	-	-
rs118082508	*KIF5A*	12	0.005	chr12:5692503	-	-
rs116900480	*KIF5A*	12	0.006	chr12:58262322	-	-
rs1578303	*HTR2A*	13	0.204	chr13:47389011	rs144877054	chr13:47962781
rs10492593	*PCDH9*	13	0.121	chr13:66919985	rs73208976	chr13:67486924
rs17446243	*TTL/TEL*	13	0.116	chr13:40174794	rs78375967	chr13:40751567

rs10139154	*SCFD1*	14	0.428	chr14:30678292	-	-
rs10143310	*ATXN3*	14	0.339	chr14:92074037	-	-
rs12886280	*NUBPL*	14	0.412	chr14:31829453	rs35875023	chr14:32298974
rs6603044	*BTBD1*	15	0.332	chr15:83015059	rs12904695	chr15:83700365
rs9901522	*PMP22*	17	0.18	chr17:14770617	-	-
rs739439	*KIAA0524*	17	0.105	chr17:28396803	rs35714695	chr17:26719788
rs2240601	*MSI2*	17	0.192	chr17:57673751	rs16942143	chr17:55748611
rs2285642	*GGNBP2*	17	0.407	chr17:36556904	rs10707226	chr17:34916453
rs7224296	*NSF*	17	0.472	chr17:46722680	rs9912530	chr17:44836302
rs12973192	*UNC13A*	19	0.278	chr19:17642430	-	-
rs12608932	*UNC13A*	19	0.43	chr19:17641880	rs12973192	chr19:17753239
rs4239633	*UNC13A*	19	0.28	chr19:17631660	rs71162163	chr19:17744075
rs75087725	*C21orf72*	21	0.003	chr21:44333234	-	-

Abbreviations: GMAF = global minor allele frequency; Chr = chromosome; GRCh38 = Genome Reference Consortium Human Build 38; HG19 = Human Genome Project 19

Supplementary Table 2: Number of UPenn Biobank ALS autopsy cases for each neuropathological measurement in each sampled neuroanatomical region.

Region	Neuropathological Measurement	N
Middle frontal cortex	Neuronal loss	87
Middle frontal cortex	TDP-43	87
Cingulate cortex	Neuronal loss	88
Cingulate cortex	TDP-43	87
Motor cortex	Neuronal loss	84
Motor cortex	TDP-43	86
Superior / middle temporal cortex	Neuronal loss	87
Superior / middle temporal cortex	TDP-43	84
CA1 / subiculum (hippocampus)	Neuronal loss	88
CA1 / subiculum (hippocampus)	TDP-43	85

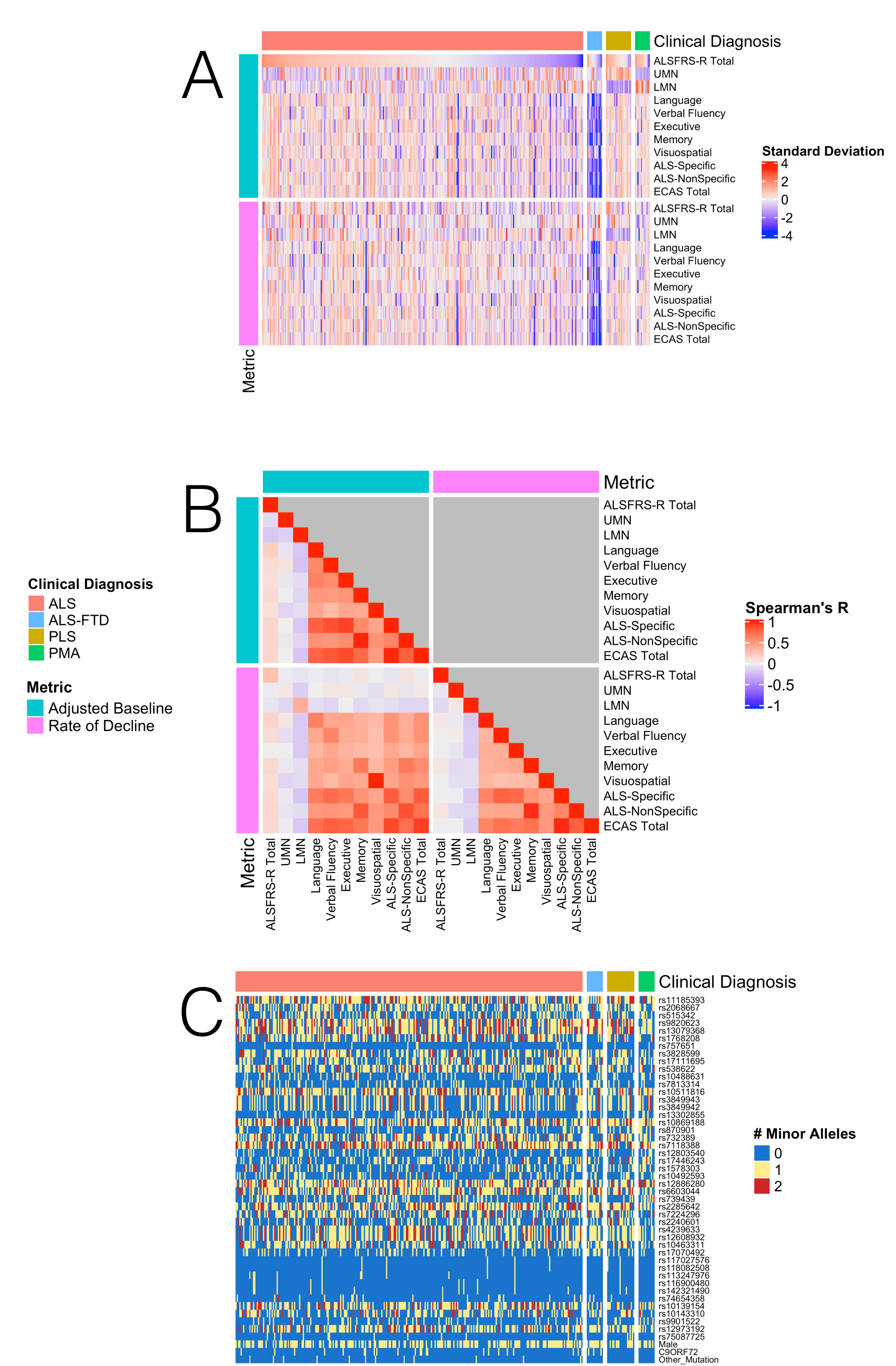
Abbreviations: CA1 = cornu ammonis 1; TDP-43 = TAR DNA-binding protein [43 kDa]

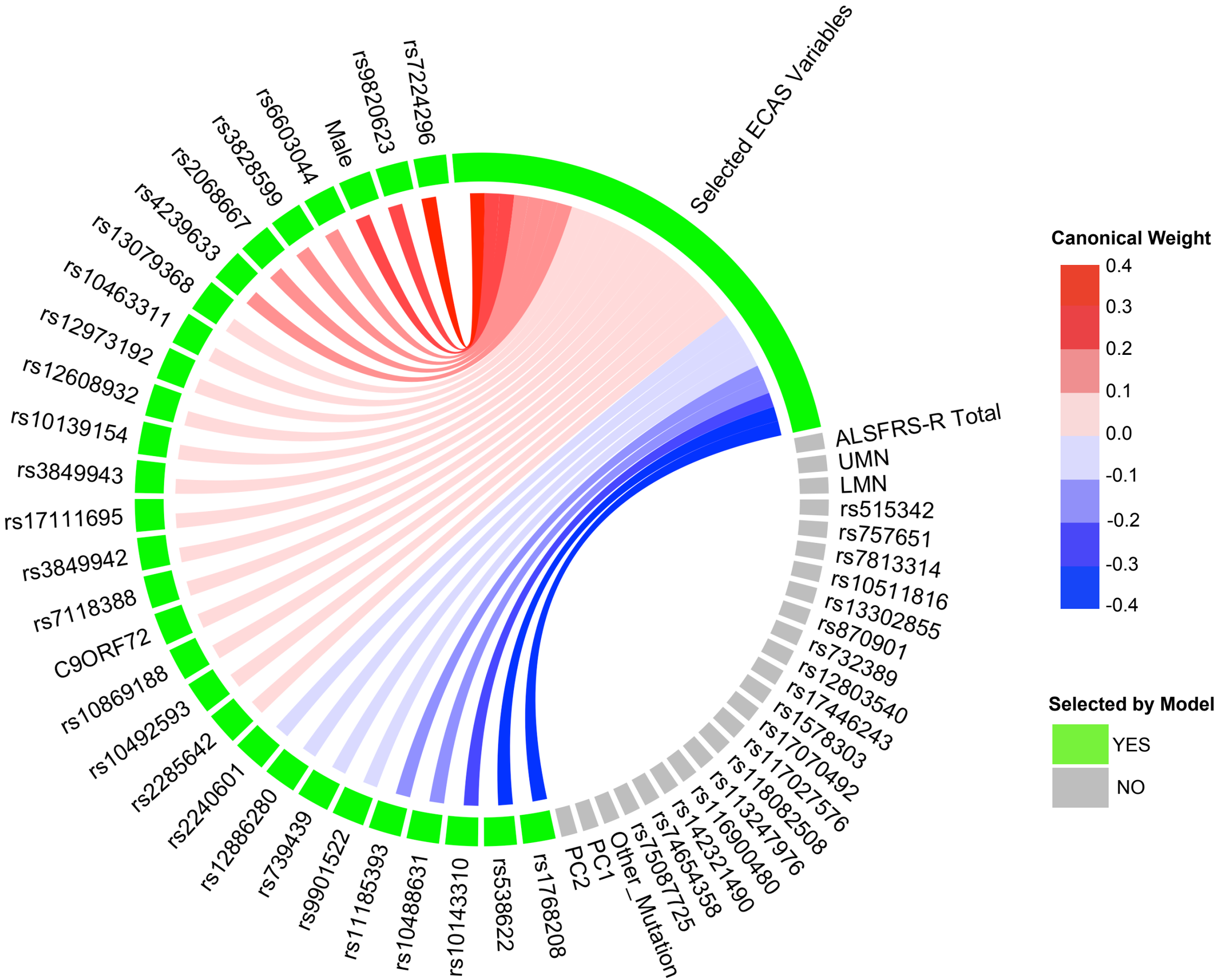
Supplementary Table 3: Peak voxel coordinates for regions of reduced cortical thickness in ALS patients relative to healthy controls, and peak voxel coordinates for regions of reduced cortical thickness associated with higher weighted polygenic score (wPGS) in patients with ALS from the UPenn Biobank neuroimaging cohort.

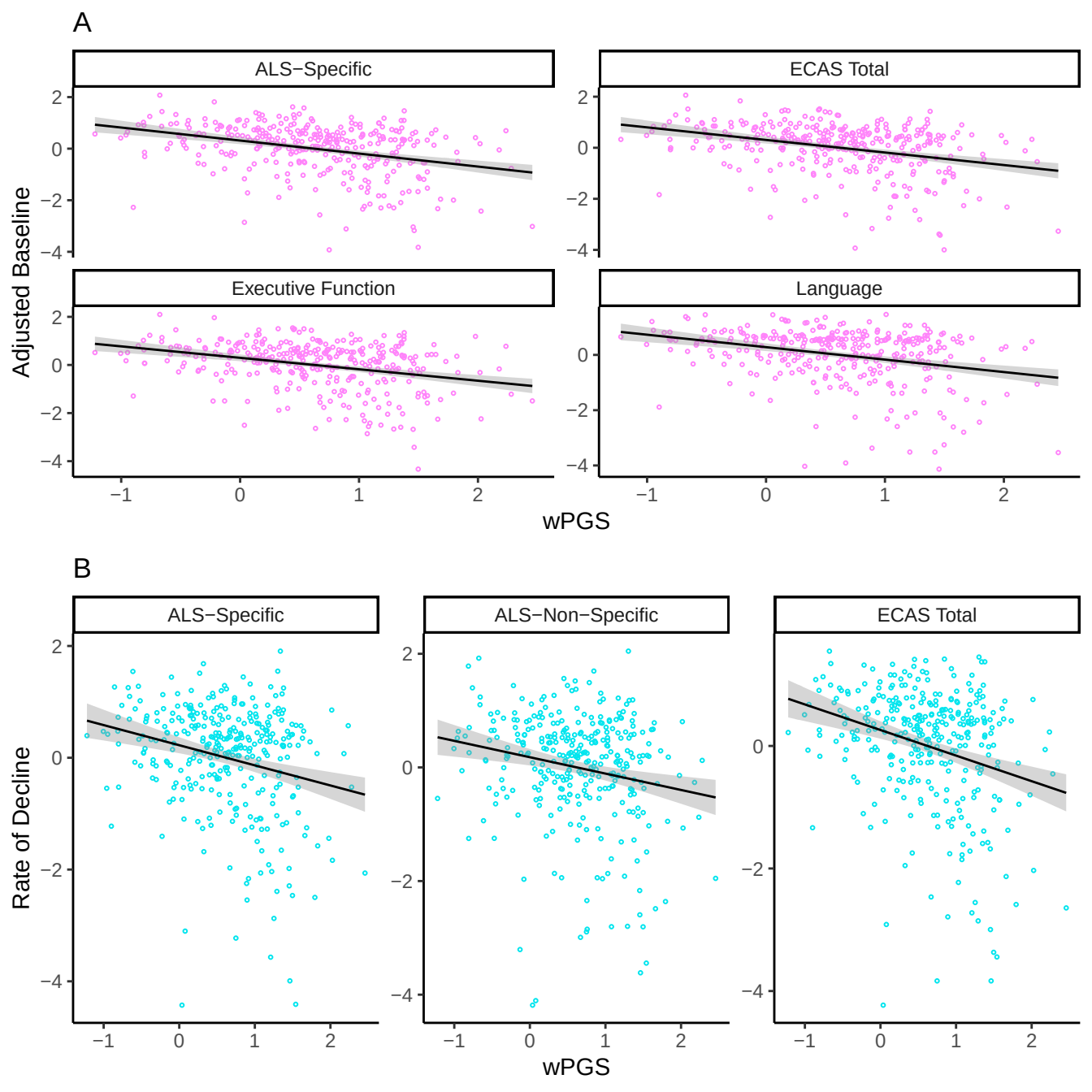
Neuroanatomic region (BA)	**L / R**	**MNI Coordinates x**	**y**	**z**	***T* statistic**	***p* value**	**Voxels**
Reduced cortical thickness in ALS relative to healthy controls[1]:							
						<.001	42994
Anterior cingulate cortex (32)	L	-2	48	10	7.2		
Dorsolateral prefrontal cortex (9)	L	-2	48	18	7.08		
Anterior premotor cortex (8)	L	-2	30	36	6.76		
Orbitofrontal cortex (11)	R	8	26	-26	6.71		
Insula (13)	R	40	16	-12	6.46		
Insula (13)	R	36	22	4	6.37		
Anterior prefrontal cortex (10)	R	26	58	0	6.32		
Insula (13)	R	42	2	0	6.26		
Dorsolateral prefrontal cortex (9)	R	2	48	18	6.22		
Anterior cingulate cortex (32)	R	2	30	22	6		
Reduced cortical thickness associated with wPGS in ALS:							
Lateral temporal cortex (21)	L	-66	-46	-8	3.01	0.003	34
Premotor cortex (6)	R	36	-14	70	3.05	0.001	23
Premotor cortex (6)	L	-14	-8	76	3	0.002	21
Orbital prefrontal cortex (47)	R	34	42	-8	2.67	0.005	18
Lateral temporal cortex (21)	L	-66	-44	6	2.54	0.002	13
Anterior cingulate cortex (32)	R	14	40	0	2.59	0.004	13
Hippocampus (54)	L	-24	-30	-8	2.74	0.004	10

Abbreviations: BA = Brodmann area, L/R = Left/Right, MNI = Montreal Neurological Institute.

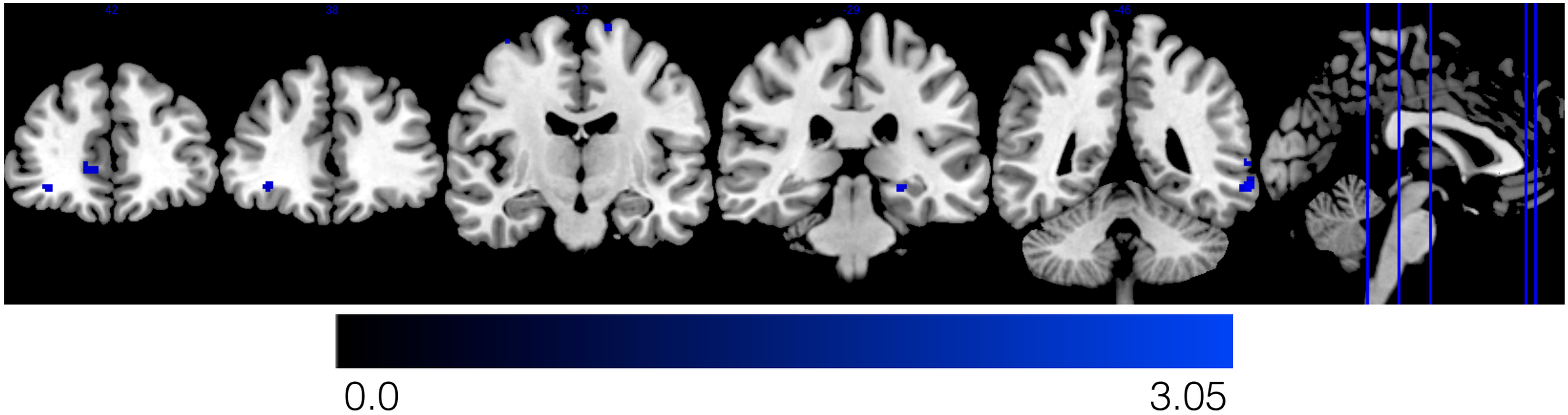
Note.[1] Cortical regions identified from peak voxel coordinates in an effort to describe sub-peaks within a larger, contiguous cluster.


A
Clinical Diagnosis
ALSFRS-R Total
UMN
LMN
Language
Verbal Fluency
Executive
Memory
Visuospatial
ALS-Specific
ALS-NonSpecific
ECAS Total
Metric
Standard Deviation
4
2
0
-2
-4
B
Metric
Spearman's R
1
0.5
0
-0.5
-1
Clinical Diagnosis
ALS
ALS-FTD
PLS
PMA
Metric
Adjusted Baseline
Rate of Decline
C
Clinical Diagnosis
Minor Alleles
0
1
2
Male
C9ORF72
Other_Mutation

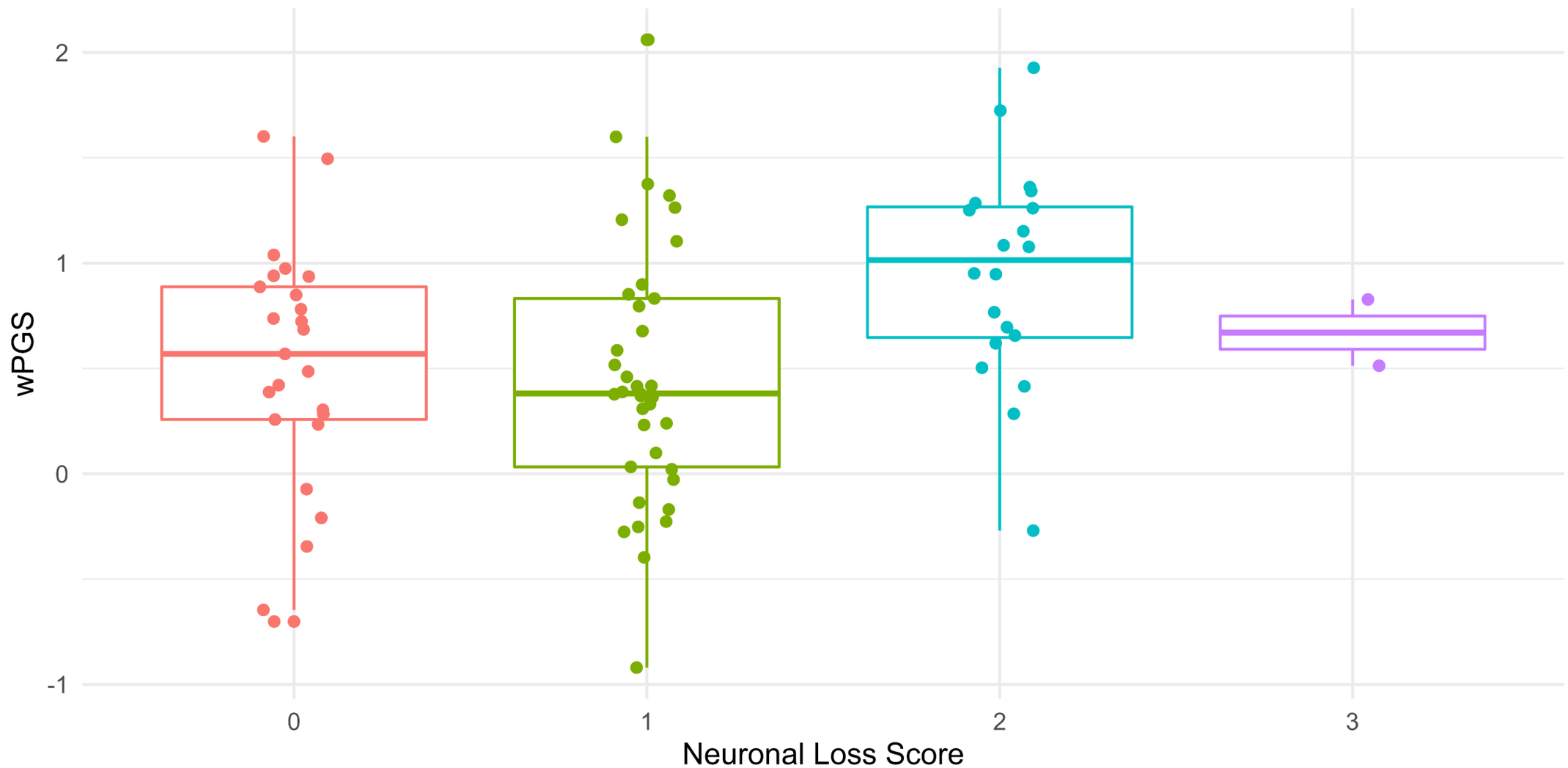

Selected ECAS Variables
rs7224296
rs9820623
Male
rs6603044
rs3828599
rs2068667
rs4239633
rs13079368
rs10463311
rs12973192
rs12608932
rs10139154
rs3849943
rs17111695
rs3849942
rs7118388
C9ORF72
rs10869188
rs10492593
rs2285642
rs2240601
rs12886280
rs739439
rs9901522
rs11185393
rs10488631
rs10143310
rs538622
rs1768208
PC2
PC1
Other_Mutation
rs75087725
rs74654358
rs14232490
rs116900480
rs113247976
rs118082508
rs117027576
rs117070492
rs1578303
rs17446243
rs12803540
rs732389
rs870901
rs13302855
rs10511816
rs7813314
rs757651
rs515342
LMN
UMN
ALSFRS-R Total
Canonical Weight
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
Selected by Model
YES
NO

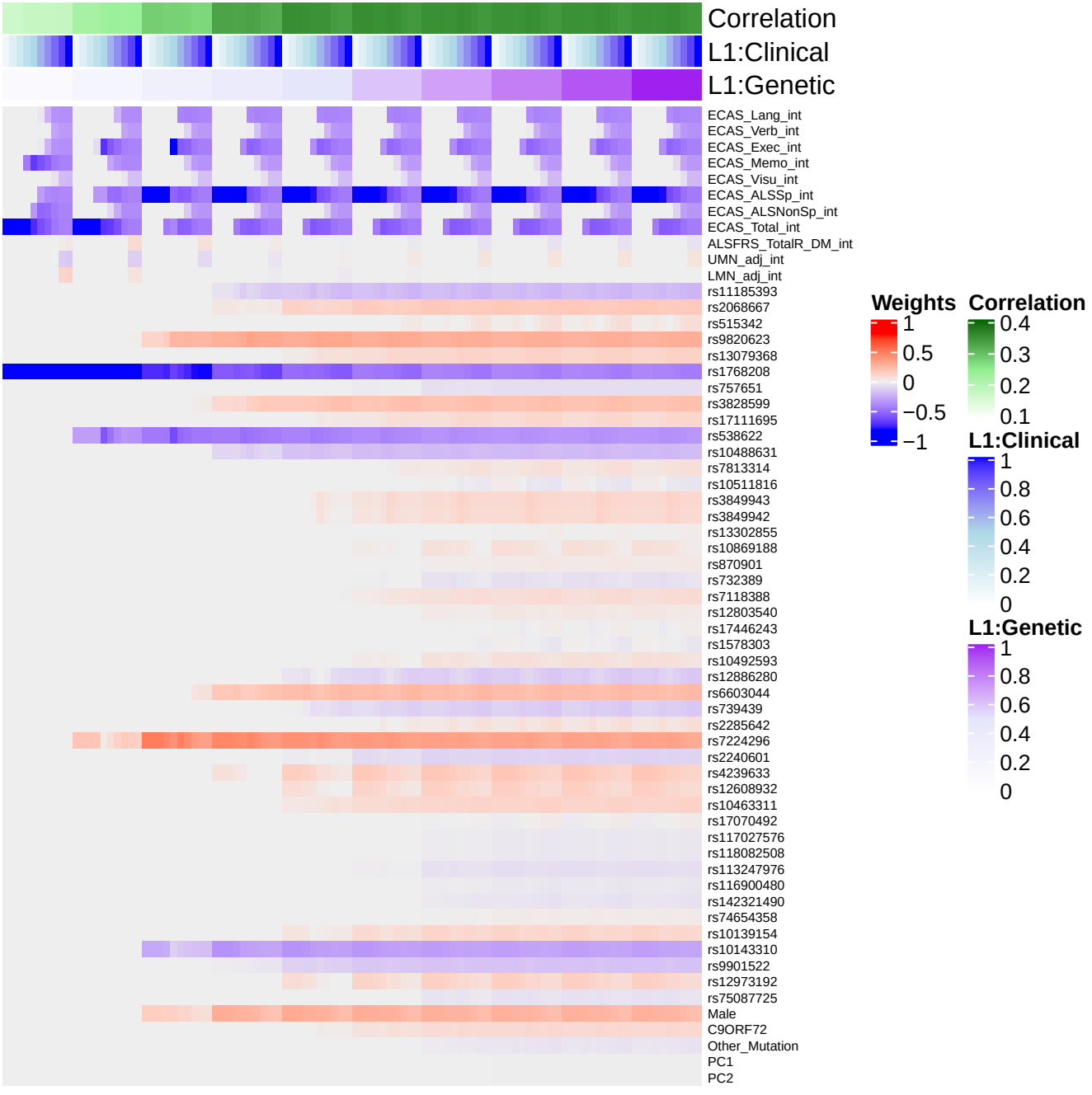

A
ALS−Specific
ECAS Total
Executive Function
Language
Adjusted Baseline
wPGS
B
ALS−Specific
ALS−Non−Specific
ECAS Total
Rate of Decline
wPGS

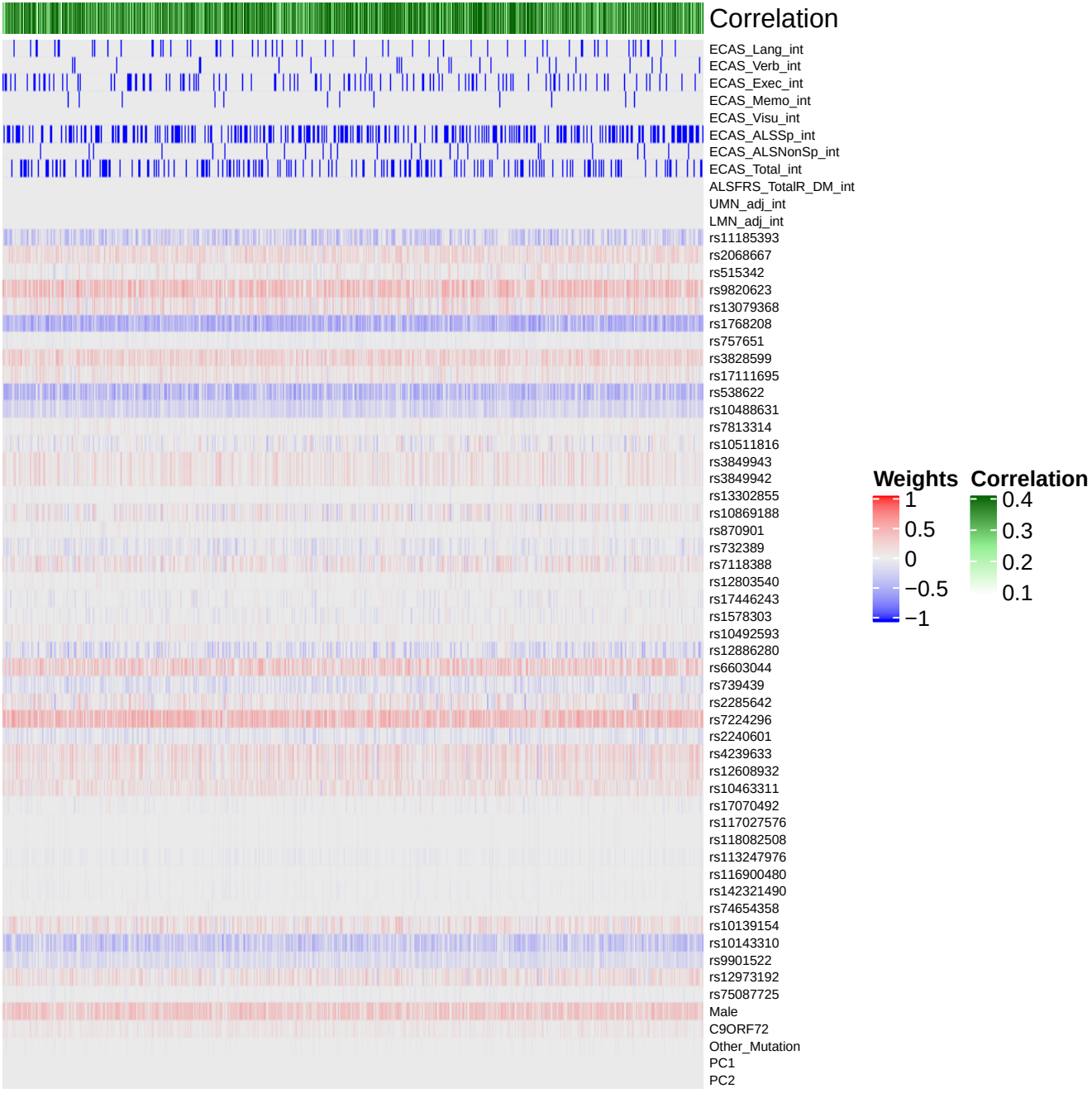
A

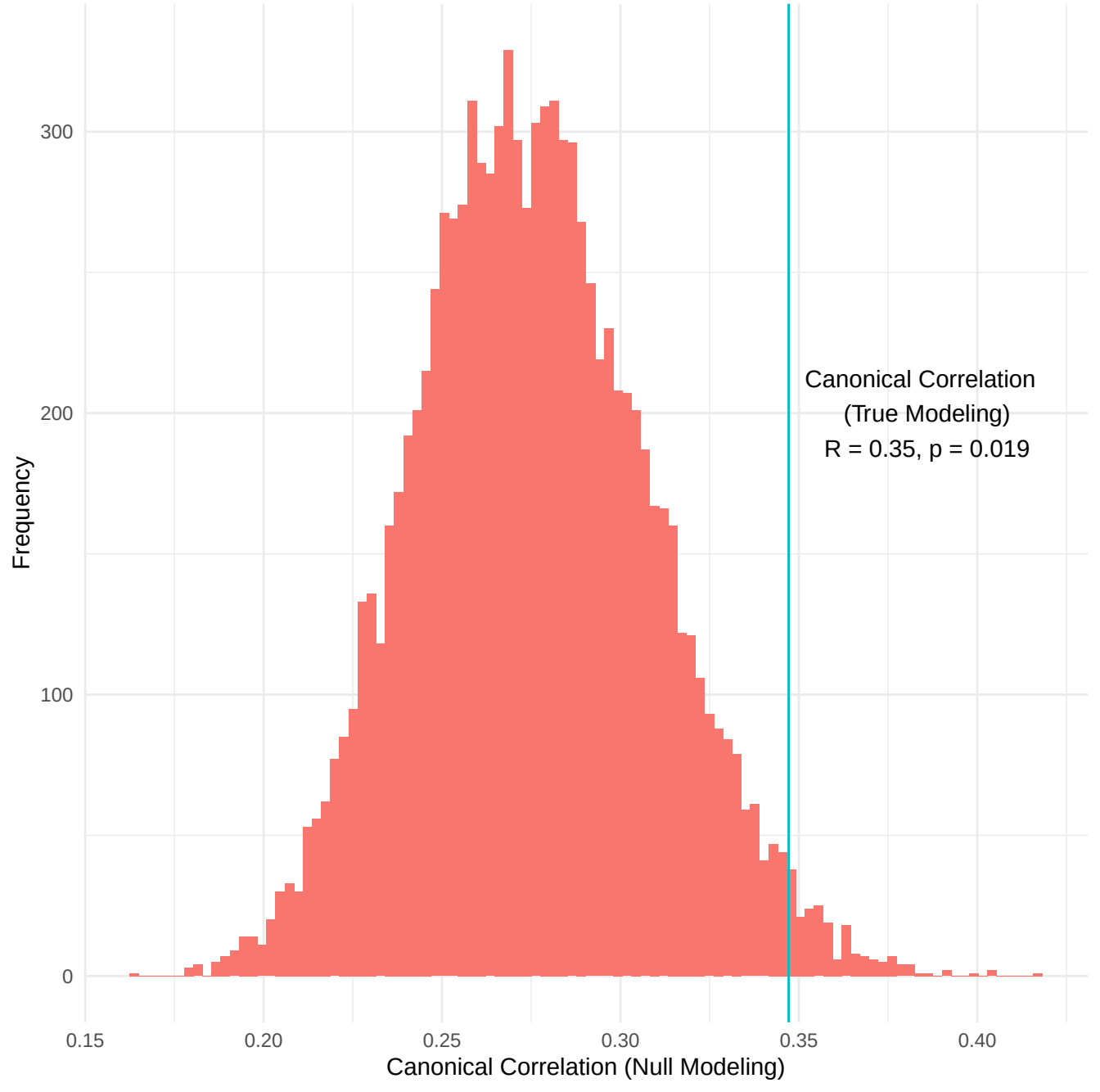


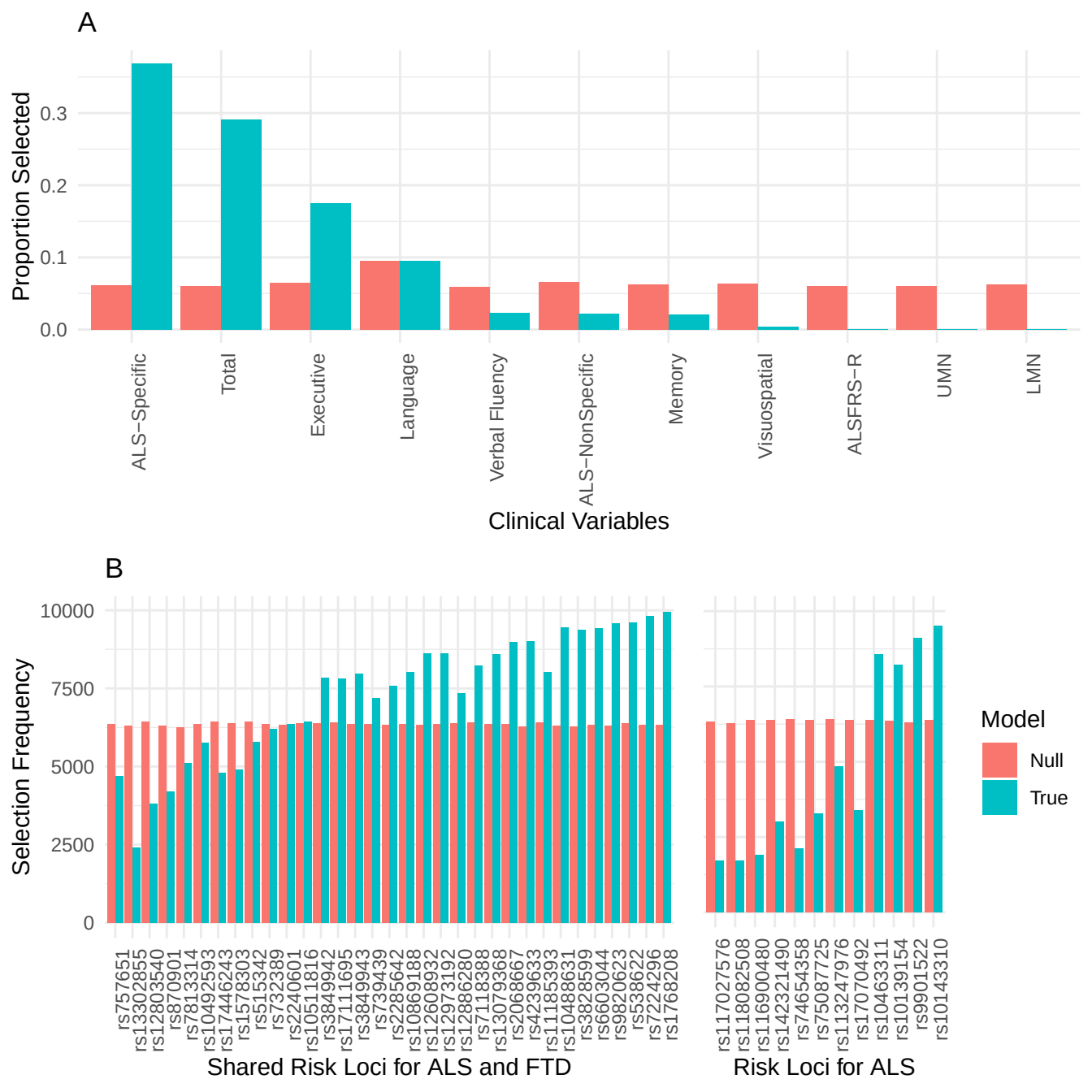
B

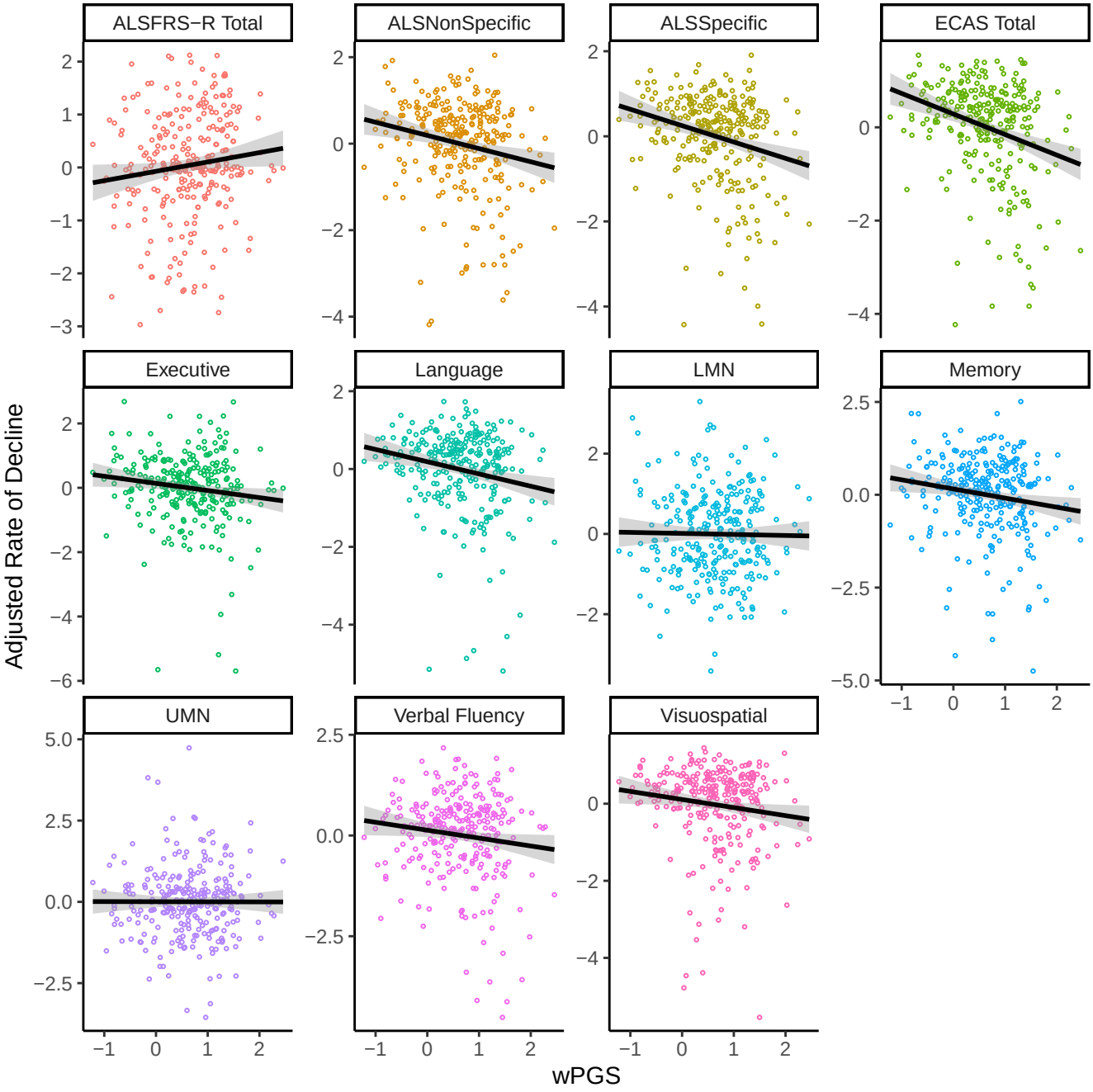


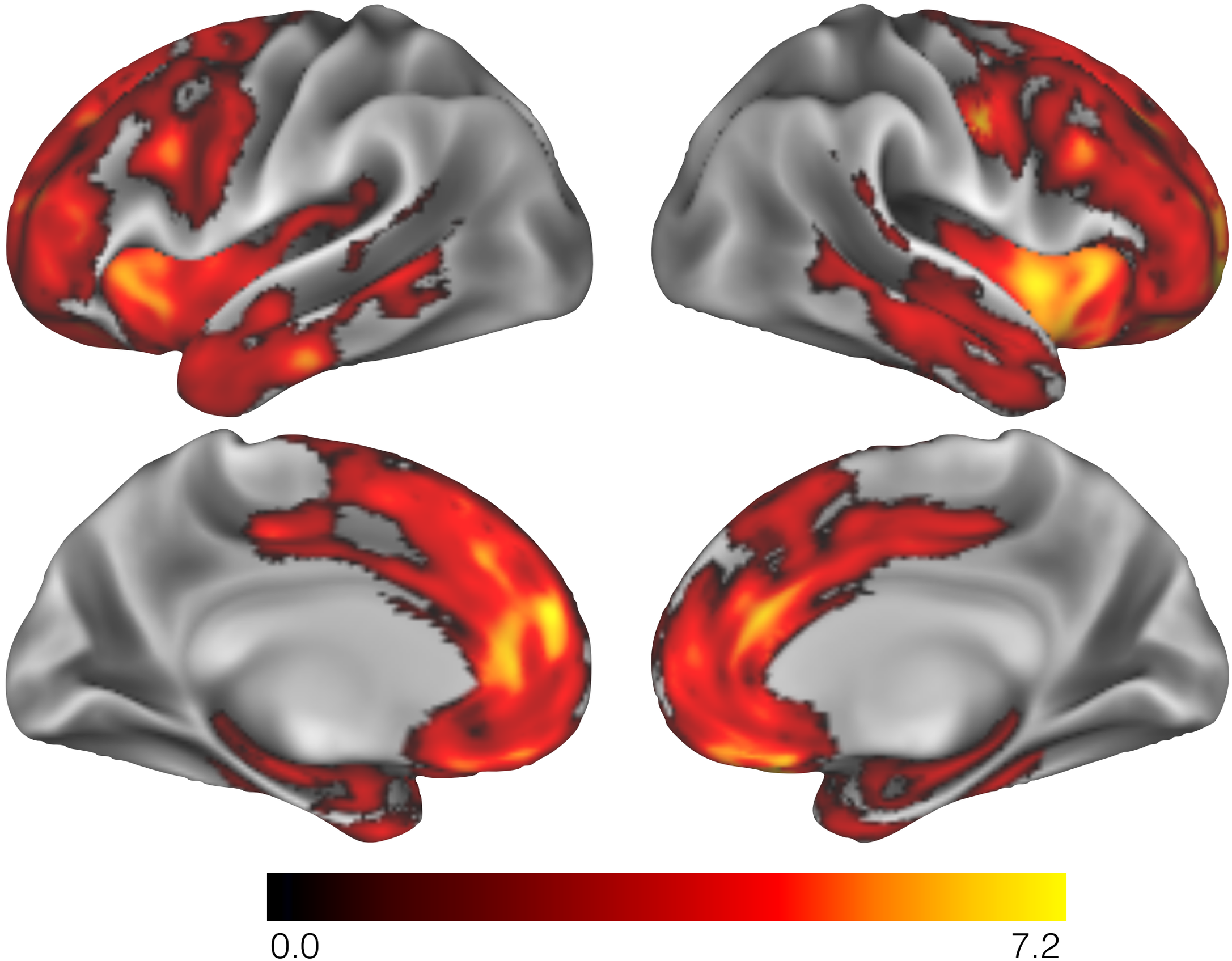

Correlation
L1:Clinical
L1:Genetic
ECAS_Lang_int
ECAS_Verb_int
ECAS_Exec_int
ECAS_Memo_int
ECAS_Visu_int
ECAS_ALSSp_int
ECAS_ALSNonSp_int
ECAS_Total_int
ALSFRS_TotalR_DM_int
UMN_adj_int
LMN_adj_int
rs11185393
rs2068667
rs515342
rs9820623
rs13079368
rs1768208
rs757651
rs3828599
rs17111695
rs538622
rs10488631
rs7813314
rs10511816
rs3849943
rs3849942
rs13302855
rs10869188
rs870901
rs732389
rs7118388
rs12803540
rs17446243
rs1578303
rs10492593
rs12886280
rs6603044
rs739439
rs2285642
rs7224296
rs2240601
rs4239633
rs12608932
rs10463311
rs17070492
rs117027576
rs118082508
rs113247976
rs116900480
rs142321490
rs74654358
rs10139154
rs10143310
rs9901522
rs12973192
rs75087725
Male
C9ORF72
Other_Mutation
PC1
PC2
Weights
1
0.5
0
−0.5
−1
Correlation
0.4
0.3
0.2
0.1
L1:Clinical
1
0.8
0.6
0.4
0.2
0
L1:Genetic
1
0.8
0.6
0.4
0.2
0

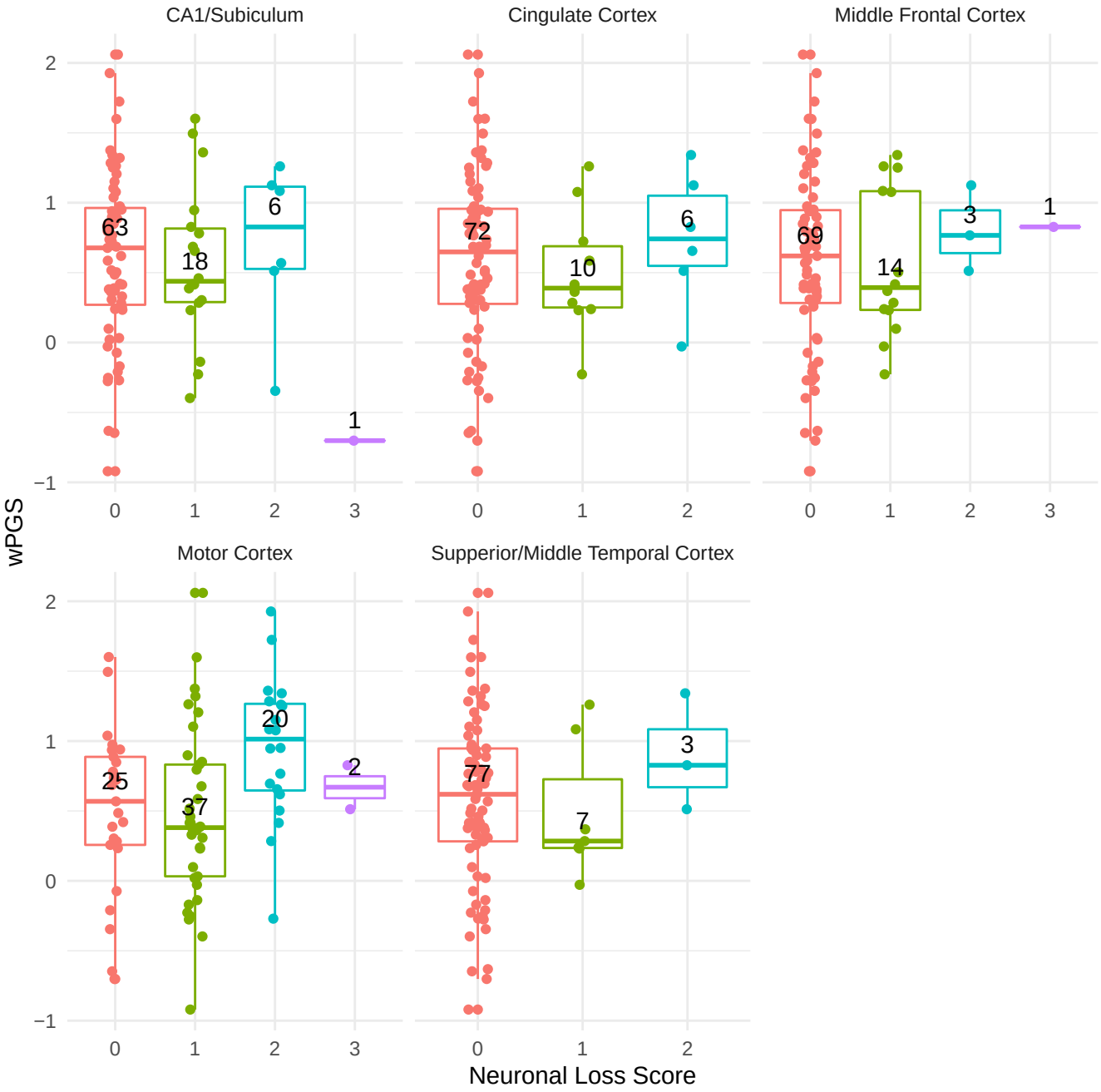

Correlation
ECAS_Lang_int
ECAS_Verb_int
ECAS_Exec_int
ECAS_Memo_int
ECAS_Visu_int
ECAS_ALSSp_int
ECAS_ALSNonSp_int
ECAS_Total_int
ALSFRS_TotalR_DM_int
UMN_adj_int
LMN_adj_int
rs11185393
rs2068667
rs515342
rs9820623
rs13079368
rs1768208
rs757651
rs3828599
rs17111695
rs538622
rs10488631
rs7813314
rs10511816
rs3849943
rs3849942
rs13302855
rs10869188
rs870901
rs732389
rs7118388
rs12803540
rs17446243
rs1578303
rs10492593
rs12886280
rs6603044
rs739439
rs2285642
rs7224296
rs2240601
rs4239633
rs12608932
rs10463311
rs17070492
rs117027576
rs118082508
rs113247976
rs116900480
rs142321490
rs74654358
rs10139154
rs10143310
rs9901522
rs12973192
rs75087725
Male
C9ORF72
Other_Mutation
PC1
PC2
Weights
1
0.5
0
−0.5
−1
Correlation
0.4
0.3
0.2
0.1

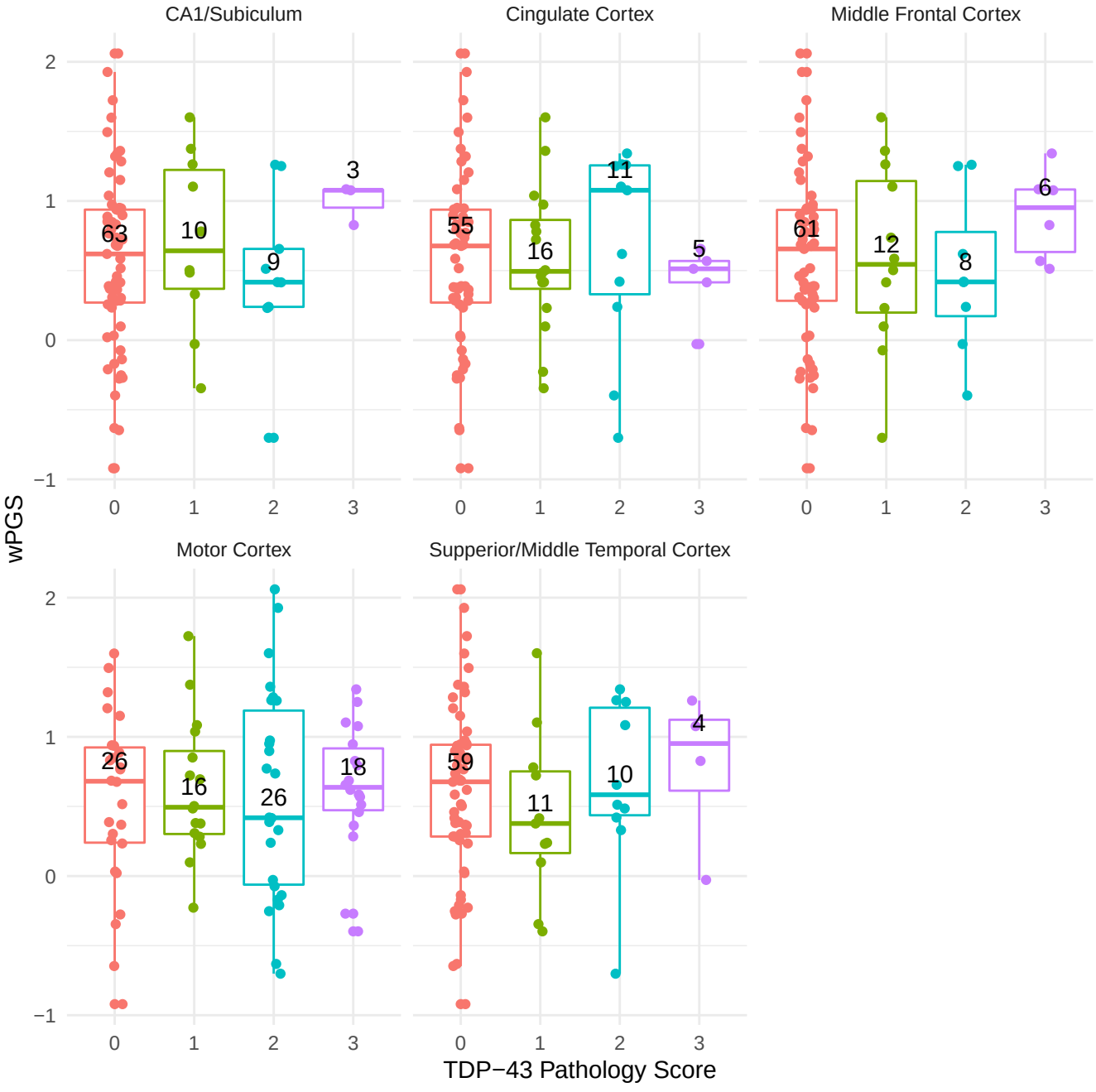

Canonical Correlation
(True Modeling)
R = 0.35, p = 0.019
Frequency
300
200
100
0
0.15
0.20
0.25
0.30
0.35
0.40
Canonical Correlation (Null Modeling)


A
Proportion Selected
0.0
0.1
0.2
0.3
ALS−Specific
Total
Executive
Language
Verbal Fluency
ALS−NonSpecific
Memory
Visuospatial
ALSFRS−R
UMN
LMN
Clinical Variables
B
Selection Frequency
0
2500
5000
7500
10000
rs757651
rs13302855
rs12803540
rs870901
rs7813314
rs10492593
rs17446243
rs1578303
rs515342
rs732389
rs2240601
rs10511816
rs3849942
rs17111695
rs3849943
rs739439
rs2285642
rs10869188
rs12608932
rs12973192
rs12886280
rs7118388
rs13079368
rs2068667
rs4239633
rs11185393
rs10488631
rs3828599
rs6603044
rs9820623
rs538622
rs7224296
rs1768208
Shared Risk Loci for ALS and FTD
rs117027576
rs118082508
rs116900480
rs142321490
rs74654358
rs75087725
rs113247976
rs17070492
rs10463311
rs10139154
rs9901522
rs10143310
Risk Loci for ALS
Model
Null
True


ALSFRS−R Total
ALSNonSpecific
ALSSpecific
ECAS Total
Executive
Language
LMN
Memory
UMN
Verbal Fluency
Visuospatial
Adjusted Rate of Decline
wPGS


0.0
7.2


CA1/Subiculum
Cingulate Cortex
Middle Frontal Cortex
Motor Cortex
Supperior/Middle Temporal Cortex
wPGS
Neuronal Loss Score
63
18
6
1
72
10
6
69
14
3
1
25
37
20
2
77
7
3
2
1
0
−1
0
1
2
3


CA1/Subiculum
Cingulate Cortex
Middle Frontal Cortex
Motor Cortex
Supperior/Middle Temporal Cortex
wPGS
TDP−43 Pathology Score
2
1
0
−1
0
1
2
3
63
10
9
3
55
16
11
5
61
12
8
6
26
16
26
18
59
11
10
4