

Title Page

Title: Cumulative genetic risk and *C9orf72* repeat status independently associate with ALS status in two case-control studies

Authors: John F. Dou¹, Kelly M. Bakulski¹, Kai Guo^{2,3}, Junguk Hur⁴, Lili Zhou⁵, Sara Saez-Atienzar⁶, Ali R Stark⁶, Ruth Chia⁶, Alberto García-Redondo^{7,8}, Ricardo Rojas-García^{8,9}, Juan Francisco Vázquez-Costa^{8,10,11}, Ruben Fernandez Santiago^{12,13,14}, Sara Bandres-Ciga¹⁵, Pilar Gómez-Garre^{12,16}, Maria Teresa Periñán^{12,16}, Pablo Mir^{12,16,17}, Jordi Pérez-Tur^{12,18,19}, Fernando Cardona^{12,18,19}, Manuel Menendez-Gonzalez^{20,21,22}, Javier Riancho^{23,24}, Daniel Borrego-Hernández^{7,8}, Lucía Galán-Dávila²⁵, Jon Infante Ceberio²⁴, Pau Pastor^{26,27}, Spanish Neurological Consortium^a, Bryan J. Traynor⁶, Eva L. Feldman^{2,3}, and Stephen A Goutman^{2,3}

^aMembers of the Spanish Neurological Consortium: Jesús Esteban-Pérez^{7,8}, Pilar Cordero-Vázquez^{7,8}, Sevilla Teresa^{8,10,28}, Adolfo López de Munain^{12,29,30}, Julio Pardo-Fernández^{23,31}, Ivonne Jericó-Pascual^{32,33}, Oriol Dols-Icardo^{12,34}, Ellen Gelpi Mantius^{35,36}, Janet Hoenicka^{8,37}, Victoria Alvarez Martinez^{38,39}, Francisco Javier Rodríguez de Rivera Garrido⁴⁰, Katrin Beyer^{41, 42}, Jordi Clarimón Echevarría^{12, 34}

Affiliations:

- ¹Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI
- ²Department of Neurology, University of Michigan, Ann Arbor, MI
- ³NeuroNetwork for Emerging Therapies, University of Michigan, Ann Arbor, MI
- ⁴Department of Biomedical Sciences, University of North Dakota, Grand Forks, ND
- ⁵Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI
- ⁶Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA.
- ⁷ALS Unit, Instituto de Investigación Sanitaria 'i + 12' del Hospital Universitario 12 de Octubre de Madrid, SERMAS, Madrid, Spain
- ⁸CIBERER, Center for Networked Biomedical Research into Rare Diseases, Madrid, Spain
- ⁹Neuromuscular Disorders Unit, Neurology Department and Sant Pau Biomedical Research Institute, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain
- ¹⁰Neuromuscular Unit, Hospital Universitario y Politécnico la Fe, IIS La Fe, Valencia, Spain
- ¹¹Department of Medicine, Universitat de València, Valencia, Spain
- ¹²Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Madrid, Spain
- ¹³Lab of Parkinson's disease and Other Neurodegenerative Movement Disorders, IDIBAPS-Institut d'Investigacions Biomèdiques, Barcelona, Catalonia, Spain
- ¹⁴Unitat de Parkinson i Trastorns del Moviment. Servicio de Neurología, Hospital Clínic de Barcelona and Institut de Neurociències de la Universitat de Barcelona (Maria de Maetzu Center), Catalonia, Spain
- ¹⁵Center for Alzheimer's and Related Dementias, National Institute on Aging, Bethesda, Maryland, USA
- ¹⁶Unidad de Trastornos del Movimiento, Servicio de Neurología y Neurofisiología Clínica, Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Sevilla, Spain
- ¹⁷Departamento de Medicina, Universidad de Sevilla, Sevilla, Spain
- ¹⁸Neurology and Molecular Genetics Mixed Investigation Unit. Instituto de Investigación Sanitaria La Fe, Valencia, Spain

- ¹⁹Molecular Genetics Unit. Institut de Biomedicina de València-CSIC, Valencia, Spain
- ²⁰Department of Medicine, Universidad de Oviedo, Oviedo, Spain
- ²¹Department of Neurology, Hospital Universitario Central de Asturias, Oviedo, Spain
- ²²Instituto de Investigación Sanitaria del Principado de Asturias, Oviedo, Spain
- ²³Service of Neurology. Hospital Sierrallana. IDIVAL University of Cantabria, Barrio Ganzo s/n. 39300. Torrelavega. Spain
- ²⁴Instituto de Investigación Marqués de Valdecilla, Santander, Spain
- ²⁵Department of Neurology, ALS Unit, Hospital Clínico Universitario 'San Carlos', Madrid, Spain
- ²⁶Unit of Neurodegenerative diseases, Department of Neurology, University Hospital Germans Trias i Pujol, Badalona, Barcelona, Spain
- ²⁷Neurosciences, The Germans Trias i Pujol Research Institute (IGTP) Badalona, Barcelona, Spain
- ²⁸Universitat de Valencia, Valencia, Spain
- ²⁹Neuroscience Area, Institute Biodonostia, and Department of Neurosciences, University of Basque Country EHU-UPV, San Sebastian, Spain
- ³⁰Neurology Department, Hospital Universitario Donostia, San Sebastian, Spain
- ³¹Neurology Department, Hospital Clinico, Santiago de Compostela, Spain
- ³²Neurology Department. Complejo Hospitalario de Navarra, Pamplona, Spain
- ³³Department of Neurology. ALS Clinic. Hospital Universitario de Navarra. IdisNa (Instituto de Investigación Sanitaria de Navarra), Navarra, Spain
- ³⁴Memory Unit, Neurology Department and Sant Pau Biomedical Research Institute, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain
- ³⁵Neurological Tissue Bank of the Biobank-Hospital Clinic-IDIBAPS, Barcelona, Spain
- ³⁶Institute of Neurology, Medical University of Vienna, Vienna, Austria
- ³⁷Laboratory of Neurogenetics and Molecular Medicine-Pediatric Institute of Rare Diseases, Institut de Recerca Sant Joan de Déu, Barcelona, Spain
- ³⁸Laboratorio de Genética, Hospital Universitario Central de Asturias, Asturias, Spain
- ³⁹Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Asturias, Spain
- ⁴⁰Department of Neurology, ALS Unit, Hospital Universitario La Paz, Madrid, Spain
- ⁴¹Department of Pathology, University Hospital Germans Trias i Pujol, Badalona, Barcelona, Spain
- ⁴²Neurosciences, The Germans Trias i Pujol Research Institute (IGTP) Badalona, Barcelona, Spain

Corresponding Author:

Stephen A. Goutman, MD, MS
 Department of Neurology
 1500 E Medical Center Dr
 Ann Arbor, MI 48109-5223
 Phone: 734-936-8586
 Fax: 734-936-5185
 e-mail: sgoutman@med.umich.edu

Authors

John F. Dou	johndou@umich.edu	0000-0003-4577-8660
Kelly M. Bakulski	bakulski@umich.edu	0000-0002-9605-6337
Kai Guo	kaigu@umich.edu	0000-0002-4651-781X
Junguk Hur	junguk.hur@med.und.edu	0000-0002-0736-2149
Lili Zhou	zhaolili@umich.edu	
Sara Saez-Atienzar	sara.saez@nih.gov	0000-0002-1524-9584
Ali Stark	ali22stark@gmail.com	0000-0002-1271-951X

Ruth Chia	ruth.chia@nih.gov	0000-0002-4709-7423
Alberto García-Redondo	ela@correo.h12o.es	
Ricardo Rojas-García	rrojas@santpau.cat	0000-0003-1411-5573
Juan Francisco Vázquez-Costa	vazquez_juacos@qva.es	
Ruben Fernandez Santiago	ruben.fernandez.santiago@googlemail.com	
Sara Bandres-Ciga	sara.bandresciga@nih.gov	
Pilar Gómez-Garre	mgomez-ibis@us.es	
Maria Teresa Periñán	tperinan-ibis@us.es	
Pablo Mir	pmir@us.es	
Jordi Pérez-Tur	jpereztur@ibv.csic.es	
Fernando Cardona		
Manuel Menendez-Gonzalez	menendezgmanuel@uniovi.es	0000-0002-5218-0774
Javier Riancho	javier.riancho86@gmail.com	0000-0001-7929-1055
Daniel Borrego-Hernández	dborregofernandez.imas12@h12o.es	
Lucia Galán-Dávila	lucgalan@ucm.es	
Jon Infante Ceberio	jon.infante@unican.es	
Pau Pastor	ppastor@unav.es	
Bryan J. Traynor	bryan.traynor@nih.gov	0000-0003-0527-2446
Eva L. Feldman	efeldman@umich.edu	0000-0002-9162-2694
Stephen A. Goutman	sgoutman@med.umich.edu	0000-0001-8780-6637

Spanish Neurological Consortium

Jesús Esteban-Pérez	jesusesteban@h12o.es	
Pilar Cordero-Vázquez	mariadelpilar.cordero@salud.madrid.org	0000-0002-9239-5147
Sevilla Teresa	sevilla_ter@qva.es	
Adolfo Lopez de Munain	adolfo.lopezdemunainarregui@osakidetza.eus	
Julio Pardo-Fernández	julio.pardo@usc.es	0000-0001-8807-1310
Ivonne Jericó-Pascual	ijericop@navarra.es	
Oriol Dols-Icardo	odols@santpau.cat	
Ellen Gelpi Mantius	ellen.gelpimantius@meduniwien.ac.at	
Janet Hoenicka	jhoenicka@fsjd.org	
Victoria Alvarez	victoria.alvarez@sespa.es	0000-0002-1916-2523
Francisco Javier Rodríguez de Rivera Garrido	frrivera.garrido@salud.madrid.org	
Katrin Beyer	katrinbeyer@hotmail.com	
Jordi Clarimón Echevarría	jclarimon@santpau.cat	

Abstract word count: 250

Main text word count: 3411

Abstract

Background: Most amyotrophic lateral sclerosis (ALS) patients lack a monogenic mutation. This study evaluates ALS cumulative genetic risk in an independent Michigan and Spanish replication cohort using polygenic scores.

Methods: ALS (n=219) and healthy control (n=223) participant samples from University of Michigan were genotyped and assayed for the *C9orf72* hexanucleotide expansion. Polygenic scores excluding the C9 region were generated using an independent ALS genome-wide association study (20,806 cases, 59,804 controls). Adjusted logistic regression and receiver operating characteristic curves evaluated the association and classification between polygenic scores and ALS status, respectively. Population attributable fractions and pathway analyses were conducted. An independent Spanish study sample (548 cases, 2,756 controls) was used for replication.

Results: Polygenic scores constructed from 275 single nucleotide polymorphisms had the best model fit in the Michigan cohort. A standard deviation increase in ALS polygenic score associated with 1.28 (95%CI 1.04-1.57) times higher odds of ALS with area under the curve of 0.663 versus a model without the ALS polygenic score (p-value= 1×10^{-6}). The population attributable fraction of the highest 20th percentile of ALS polygenic scores, relative to the lowest 80th percentile, was 4.1% of ALS cases. Genes annotated to this polygenic score enriched for important ALS pathomechanisms. Meta-analysis with the Spanish study, using a harmonized 132 single nucleotide polymorphism polygenic score, yielded similar logistic regression findings (odds ratio: 1.13, 95%CI 1.04-1.23).

Conclusion: ALS polygenic scores can account for cumulative genetic risk in populations and reflect disease-relevant pathways. If further validated, this polygenic score will inform future ALS risk models.

Keywords: amyotrophic lateral sclerosis, polygenic risk, polygenic scores, classification

What is already known on this topic

Amyotrophic lateral sclerosis (ALS) is a complex neurodegenerative disease caused in part by genetic factors, and methods to account for ALS polygenic disease risk are needed.

What this study adds

An ALS polygenic score reflects disease risk in the population and helps ascribe the magnitude of the risk.

How this study might affect research, practice or policy

ALS polygenic scores can assign the overall distribution of genetic risk in a population and can be used to screen individuals at higher risk.

Main Text

Introduction

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterized by rapidly progressive muscle weakness and death within 2 to 4 years from symptom onset^{1,2} with 50% of patients manifesting cognitive or behavioral dysfunction.^{1,2} ALS is traditionally divided into familial and sporadic forms. Familial ALS, comprising 10-15% of persons with ALS, includes individuals with a family history of ALS or related diseases, whereas sporadic ALS represents individuals without a family history. Under a monogenic model, a single risk gene is associated with a greater likelihood of developing ALS³ or contributes to a distinct phenotypic outcome, such as earlier age of disease onset.^{3,4} Since 1994, over 40 genes have been associated with ALS.⁵ The non-coding chromosome 9 open reading frame 72 (*C9orf72*) hexanucleotide expansion is the most common genetic form of ALS and is observed in 40% of familial and 10% of sporadic ALS in mixed European populations.^{6,7} Superoxide dismutase 1 (*SOD1*), TAR DNA binding protein 43 (*TARDBP*), and fused in sarcoma (*FUS*) are the next most common genes with polymorphism frequencies of around 1% or less in sporadic cases.⁸ Importantly, most ALS patients do not carry a single causative ALS risk gene mutation. Further, heritability estimates of ALS range from around 20% for genome-wide complex trait analysis and 60% for twin data⁹ and ALS shares some genetic risk with other diseases and conditions.¹⁰⁻¹² It is increasingly clear that many common polymorphisms may contribute a small amount of disease risk.¹³ Since most ALS patients do not have a monogenic cause, it is crucial to understand the genetic contribution to ALS beyond single highly penetrant mutations.

We hypothesize that cumulative genetic risk for ALS can be summarized using polygenic scores. To our knowledge, the utility of a polygenic score for ALS, independent of *C9orf72* status, has not been tested to understand the associations with ALS and ALS risk prediction. The goals of the current study were to develop a genome-wide ALS polygenic score using an independent ALS cohort of participants not previously included in any genome-wide association study (GWAS) and test the score contribution to ALS risk models independently of *C9orf72* status.

Methods

University of Michigan samples: Study procedures of this Institutional Review Board approved longitudinal case/control study are published (see Supplemental Methods).¹⁴⁻¹⁶ All participants provided informed consent. DNA extraction, genome-wide genotyping, *C9orf72* repeat expansion assay, and data processing followed published protocols and are presented in Supplementary Methods and **Supplementary Figures 1-3**.

Spanish Neurological Consortium: Participants were recruited across several sites in Spain as previously published¹⁷ or as part of the ALS Genetic Spanish Consortium (ALSGESCO) as previously published¹⁸ (see Supplemental Methods). All participants provided informed consent and the study received local ethics board approval. The coordination and use of samples for this publication were approved by the institutional review board of the National Institute on Aging. DNA extraction, genome-wide genotyping, *C9orf72* repeat expansion assay, and processing followed published protocols are presented in Supplementary Methods.

Polygenic scores were developed using published methods¹⁹⁻²¹ and were derived from an independent GWAS of 20,806 ALS cases and 59,804 controls²² (See Supplementary Methods). Developed polygenic scores both included and excluded the *C9orf72* region (**Figure 1**).

Consistent with Polygenic Score Reporting Standards,²³ for each SNP included in the polygenic score, the identifier, chromosome, position, weight, and p-value of association with ALS were provided (**Supplementary Table S1**).

Statistical analysis, including summary of case and control populations, regression models, classification attributable fraction calculation, and sensitivity analyses are presented in Supplementary Methods. Gene pathway analysis examining the overall biological functions of polygenic score genes was performed (Supplementary Methods).

Finally, to assess robustness of the polygenic score in the University of Michigan samples, a second Spanish Neurological Consortium cohort that was not part of the ALS GWAS was examined to test for replication (Supplementary Methods).

Results

Study Participants

The primary analysis included 442 participants (223 controls and 219 ALS cases) (**Table 1**). Family history of ALS was present in 7.8% of ALS cases and 0% of controls. The *C9orf72* repeat was present in 5.9% of ALS cases and 0% of controls. No age differences occurred between ALS and control participants, although the male participant proportion was higher in the ALS (59.0%) versus control (48%, p-value = 0.027) group. There were 24 participants excluded for missing genetic, demographic, or ALS assessment data (**Supplementary Figure S1B**), who had similar characteristics to the included sample (**Supplementary Table S2**).

Genetic Data Characteristics and Polygenic Score Optimization

SNPs were measured at 1,748,250 positions. SNPs missing genomic location data, with missingness frequency of >1% of samples, with minor allele frequency < 5%, or out of Hardy-Weinberg equilibrium (p-value < 10⁻⁶) were removed, leaving 601,350 measured autosomal SNPs (**Supplementary Figure S2**). Imputation resulted in 47,109,465 SNPs. Imputed SNPs with an imputation quality R-squared value less than 0.5 and SNPs with a minor allele frequency < 1% were filtered. The final dataset had 8,179,459 imputed SNPs (**Supplementary Figure S3**).

The *C9orf72* region on chromosome 9 spanned from 27.4 Mb to 27.7 Mb. Following pruning, 5 SNPs were present in this region (**Figure 1**). Of these, one SNP rs3849943, located at position 27,543,382, was associated with *C9orf72 expansion* status (fisher p-value = 0.00001). Because our goal was to estimate the cumulative genetic risk for ALS beyond the *C9orf72 expansion*, the primary polygenic score excluded this entire region out of caution. We constructed polygenic scores, including SNPs based on their association with ALS in an independent GWAS. Polygenic score performance was highest when constructed using a p-value threshold of approximately 10⁻⁴, using 275 SNPs (**Supplementary Figure S4**). At this threshold, the incremental R² for the polygenic score was approximately 1.2%.

For sensitivity analyses, we considered a polygenic score using all available SNPs post pruning (n = 254,307 SNPs), with an incremental R² of approximately 0.4%. As another sensitivity analysis, we included SNPs in the *C9orf72* region and observed polygenic score performance was also highest using a p-value threshold of approximately 10⁻⁴ (n = 280 SNPs) (**Supplementary Figure S5**).

Associations Between Genetic Predictors and ALS Cases Status

In bivariate analyses, ALS cases had higher mean ALS polygenic scores (average standardized score of 0.03) than controls (average standardized score -0.08) (p-value = 0.11) (**Supplemental Figure S6**). We examined the roles of genetic variables and family history in analyses adjusted for age, sex, and five genetic principal components. In the full study sample (n = 442 participants), a one standard deviation increase in ALS polygenic score was associated with 1.28 times higher odds of ALS (95% CI: 1.04, 1.57) (**Table 2**), after also adjusting for *C9orf72* repeat expansion status and family history of ALS. These findings were consistent when we subset the sample to participants lacking a *C9orf72* repeat or family history of ALS (N = 416 participants). A one standard deviation increase in ALS polygenic score was again associated with 1.28 times higher odds of ALS (95% CI: 1.04, 1.57).

ALS Case Classification Performance

Beyond association testing, we were interested in the performance of genetic factors in classifying ALS cases and controls (**Figure 2**). Our base classification model adjusted for sex, age, and five genetic principal components had an area under the curve (AUC) of 0.591. Adding family history of ALS alone to the base model increased AUC to 0.631 and improved classification over the base model (likelihood ratio test p = 0.06). Including *C9orf72* repeat status as a covariate on top of the base model and family history increased the AUC to 0.647 and improved classification (likelihood ratio test p-value < 0.001). Adding the ALS polygenic score following family history and *C9orf72* repeat status further raised AUC to 0.663 and improved classification (likelihood ratio test p-value < 0.001). To assess prediction accuracy, we split datasets into training and testing and performed five-fold cross-validation. These AUC results were 0.539 for the base model, 0.588 adding family history, 0.603 adding *C9orf72* repeat status, and finally 0.620 adding ALS polygenic score (**Supplemental Figure S7**). While the AUCs were attenuated, as a result of the sampling procedure, similar sequential prediction accuracy remained, highlighting the prediction capability.

Attributable Fraction

To assess the fraction of ALS cases attributable to genetic factors, we compared those in the highest 20th percentile of ALS polygenic score to the rest of the sample. We observed that 4.1% (95% CI: -9.1%, 17.3%) of ALS cases would be prevented if the highest 20th percentile of ALS polygenic score were at the level of the rest of the population. For the *C9orf72* expansion, 6.3% (95% CI: -2.7%, 15.3%) of ALS cases would be avoided if they lacked the expansion.

Sensitivity Analyses

Sensitivity analysis (Supplementary Results, **Table 2**), including analysis around the *C9orf72* region, and an analysis restricted to European ancestry participants (**Supplementary Table S3**, **Supplementary Table S3**, **Supplementary Figure S8**), overall showed findings consistent with the main analysis.

Gene Pathway Analysis

In the 275 SNP associated genes, included the polygenic score, *richR* identified 65 highly enriched GO biological process terms, including several related to the neuronal system, such as “neuron differentiation”, “generation of neurons”, “neuron projection morphogenesis”, “neurogenesis” and “neuron development” (**Figure 3**, **Supplementary Table S5**). A total of nine KEGG pathways were significantly enriched at a nominal p-value < 0.05, which included “Glycosphingolipid biosynthesis-ganglio series”, “Fatty acid degradation” and “Pancreatic secretion” (**Figure 4**, **Supplementary Table S6**).

Replication Results

The Spanish cohort had 548 ALS cases and 2,756 controls, after removing 232 participants for missing age or *C9orf72* information. Family history, *C9orf72* expansion, and sex were associated with ALS case status (**Supplementary Table S7**). Due to differences in genotyping arrays and allele frequencies between the Michigan and Spanish cohorts, available SNPs varied between the two cohorts. To harmonize our analyses, we restricted to SNPs available in both cohorts, and the best performance in the Michigan cohort among overlapping SNPs resulted from a polygenic score consisting of 132 SNPs (p-value threshold = 5×10^{-5}). In the Spanish cohort, a one standard deviation increase in the harmonized ALS polygenic score was associated with 1.11 higher odds (95% CI: 1.01, 1.22) of ALS case status (p-value = 0.028), adjusted for sex, age, *C9orf72* expansion, family history, and five genetic principal components. In the Michigan cohort, a one standard deviation increase in the harmonized ALS polygenic score was associated with 1.22 higher odds (95% CI: 1.00, 1.50) of ALS case status (p=0.04) when including all ancestries, mirroring results above with the 275 SNP polygenic score. When limiting to European genetic ancestry in the Michigan cohort, the harmonized 132 SNP polygenic score had a stronger association, where one standard deviation increase in ALS polygenic score was associated with 1.27 higher odds (95% CI: 1.03, 1.57) of ALS case status (p-value = 0.02). Meta-analysis of the Spanish cohort and Michigan cohort (all ancestry) resulted in an estimate of one standard deviation increase in ALS polygenic score being associated with 1.13 higher odds (95% CI: 1.04, 1.23) of ALS case status (p-value = 0.004) (**Supplementary Figure S9**).

Discussion

ALS risk factors are incompletely understood. Models that predict the steps involved in developing ALS²⁴ are necessary to generate ALS risk profiles. Representing this genetic risk²⁵ is critical as most individuals with ALS lack a monogenic ALS risk gene. Since genetic risk may be distributed throughout the genome, identifying polygenic risk facilitates an understanding of the multiple ALS pathological pathways. Here we developed a weighted polygenic score using a large ALS-control GWAS.¹³ This score differed significantly in ALS cases versus controls from an independent Michigan cohort. Further, this polygenic score represents important genes and biological functions in the pathophysiology of ALS.

In the current study, the ALS polygenic score with the best model fit and lowest p-value was represented by 275 SNPs when excluding the region around *C9orf72* and 280 SNPs when including the region. We tested other SNP combinations as determined by default PRSice-2 p-value thresholds and a model including all SNPs. In each case, the model with fewer SNPs outperformed the larger models, suggesting that the genetic contributions to ALS are limited to a smaller subset of genes as opposed to a wide-ranging set of genes across more genomic regions. Next, we showed that a standard deviation increase in the ALS polygenic score raised ALS odds by 1.28 times in both models without and with the *C9orf72* region. Interestingly, all risk increased when the *C9orf72* region was included, even after adjusting for the *C9orf72* expansion, suggesting a possible role for alleles around the *C9orf72* region on disease status, even in the absence of the repeat itself. Unsurprisingly, in these models, ALS risk was disproportionate for individuals with a family history or the *C9orf72* expansion. Removing individuals with an ALS family history or a *C9orf72* expansion did not change the impact of the polygenic score on ALS risk, meaning the polygenic score itself plays an essential role on the overall ALS risk profile. Additionally, findings persisted when restricting to a European genetic ancestry population.

Polygenic scores summarize the combined effects that common and low-frequency alleles have on disease risk, thereby summarizing the genetic architecture of that disease.²⁶ Multiple medical specialties utilize polygenic scores to explain risks such as cardiovascular disease, cancers, neurodegenerative diseases,^{19,26} and other phenotypic traits.¹⁰ While polygenic scores are gaining traction for ALS,²⁷ few studies propose an ALS-specific polygenic score that can stratify populations at risk for ALS.

In contrast to our methods, McCann and colleagues leveraged a list of 853 genetic variants with a changed amino acid sequence from a comprehensive literature search.²⁸ After screening the population, 43 genetic variants from 18 genes were retained in the model, affecting 35.4% of their ALS population. However, the authors did not further develop polygenic scores.²⁸ Wainberg et al. identified individuals in the Arivale Scientific Wellness cohort at elevated genetic risk for ALS using polygenic risk scores developed through literature and sought linkages to proteomics, metabolomics, and other clinical laboratory information. This group found that increased Ω -3 and decreased Ω -6 fatty acid levels and higher IL-13 levels correlated with ALS genetic risk.²⁷ Based on KEGG analysis of the polygenic score developed herein, we found enrichment of the fatty acid degradation pathway.²⁹

Placek and colleagues used sparse canonical correlation analysis to identify a polygenic score of cognitive dysfunction in an ALS population.³⁰ Like our methods, the authors focused on SNPs achieving genome-wide significance in the Nicolas study¹³ and with risk loci in ALS and frontotemporal dementia. Of the 45 SNPs used in their models, 27 were associated with cognitive performance in their ALS population, involving the genes *MOBP*, *NSF*, *ATXN3*, *ERGIC1*, and *UNC13A*. Our polygenic score also included SNPs in *MOBP*, *ATXN3*, and *UNC13A*, thereby supporting its validity. Additional uses of polygenic scores in ALS include examining polygenic traits for other diseases that overlap with ALS.^{10,31} Although this was not our approach, such studies have yielded linkages between ALS and traits of schizophrenia, cognitive performance, and educational attainment.^{10,31}

Polygenic scores have shown utility in other neurodegenerative conditions, such as Alzheimer's disease, to find those at high and low genetic risk.³² For example, a polygenic score derived from the International Genomics of Alzheimer's Disease Project GWAS showed it could predict participants that would transition from mild cognitive impairment to late-onset Alzheimer's disease.³³ A similar approach using a polygenic score created from an Alzheimer's cohort GWAS dataset associated with incident dementia in a large Swedish birth cohort.³⁴

Our disease classification model further supports the utility of our polygenic score. Our polygenic score improved model performance, even one that included the most prevalent ALS risk gene, the *C9orf72* expansion. In Alzheimer's, similar findings are noted, where a polygenic risk score was able to classify Alzheimer's cases versus controls with an AUC of 0.83, even when excluding *APOE4* carriers.³⁵ This indicates that these genetic models are beneficial in case classification, even when considering strong genetic risks, which superimpose on polygenic risk. Another analysis similarly showed that polygenic scores in Alzheimer's disease could classify patients accurately and that the prediction improved when incorporating additional variables such as sex and age.³⁶

Since polygenic scores often overlap in persons with and without a disease of interest, focusing on patients with polygenic scores in distribution tails may offer better predictive power.³⁷ Thus, to add further perspective to this polygenic risk, we showed that 4.1% of ALS cases could be avoided for individuals with the highest 20% of polygenic score if an intervention were possible. While this population attributable risk approach considers the fraction of disease caused by

exposure, this idea can also be applied to genetic data.^{38 39} For example, a study of polygenic scores in cutaneous squamous cell carcinoma showed that removing all risk alleles from a population would decrease the risk of this cancer by 62%.⁴⁰ The authors argue that identifying those at the highest genetic risk could inform programs for skin cancer screening, with the caveat that interactions of SNPs with environmental factors⁴¹ are not included in the model. A parallel approach is also proposed for breast cancer to help identify populations that would benefit from targeted risk reduction strategies.⁴² A similar analysis has shown changes in the prevalence of type 2 diabetes, breast cancer, hypertension, and myocardial infarctions, if a proportion of polygenic risk is removed or enhanced in the population.⁴³ Currently, there is no biomarker or tool that can definitively predict who will develop ALS later in life. Therefore, even if the polygenic score can only explain a small number of individuals at risk, it could be an important screening method for risk reduction.

Replication of these findings is important to determine the generalizability of the results. We used genotype and ALS phenotype data from an independent Spanish cohort as a replication cohort. Although the SNPs included in the polygenic score were adjusted due to the available overlap of SNPs in both datasets, there was consistency in the magnitude of the polygenic score effect, further providing support for our proposed polygenic score. Replication of polygenic scores is critical to ascertain that the methods and population background used to develop the score is generalizable.³⁷ Further, replication cohorts can determine which risk variants are applicable across diverse populations.⁴⁴ Replicating polygenic scores in ALS remains important, although this requires large numbers of samples from participants not included in GWAS used to derive SNP weights.

We next queried how this set of SNPs impacts disease pathobiology. Through gene enrichment and pathway analysis, we showed that this polygenic score selects multiple pathways relevant to ALS biology, including synaptic signaling, regulation of protein metabolic process, neuron projection, and axon guidance. Using KEGG pathways, we also identified important ALS biological functions, including glycosphingolipid biosynthesis and fatty acid degradation.^{16 45} Saez-Atienzar et al. used a cohort of 78,500 individuals to develop a polygenic score for biological pathways and cell types to determine involvement in ALS.⁴⁶ Significant pathways included those involved in neuronal development and differentiation with an emphasis on the cytoskeleton. Of these pathways, the cytoskeleton pathway was significant for individuals both with and without the *C9orf72* repeat expansion, whereas the autophagosome pathway was only significant for *C9orf72* carriers. Overlapping enriched GO pathways in our polygenic score with those of Saez-Atienzar et al. included neuron projection morphogenesis, cell morphogenesis involved in differentiation, neuron development, cellular component morphogenesis, cell development, and cell projection organization. The overall overlap shows that these two different methods for developing a polygenic score selects similar pathways. Other studies of gene expression in ALS have also identified dysregulated metabolic pathways and cytoskeletal pathways.⁴⁷

This study has limitations. Due to cost and a research interest in common genetic variants, we performed genome-wide genotyping instead of whole genome sequencing. While whole genome sequencing would allow us to better account for genetic background, the method we used are validated across many studies. In addition, the study population size is small compared to the number of individuals impacted by ALS. However, the sample size here was limited to participants not included in prior GWASs and is thus a strength. This is important since developing polygenic scores from participants that are already in the reference GWAS may lead to biased results. Also, since we did utilize a lower-cost genotyping strategy imputed to maximize overlap with the ALS GWAS used for weights, these methods could be beneficial

for population screening where the cost of whole genome sequencing is not economically feasible.

Conclusion

In conclusion, we find that a polygenic score for ALS can account for cumulative genetic risk in the population and reflect cellular processes that are relevant to ALS. If further validated, this polygenic score can be a valuable tool for ALS risk models and the design of ALS prevention studies.

Acknowledgments

We thank Crystal Pacut, Caroline Piecuch, and Stacey Sakowski, PhD for study support, Masha Savelieff, PhD and Emily Koubek, PhD, for editorial assistance, and the following members of the Spanish Neurological Consortium for study assistance: Mónica Povedano Panadés, Antonio Guerrero-Sola, Tania García-Sobrino, Marilina Puente Hernández, María Jesús Sobrido-Gómez, Ivonne Jericó-Pascual, Carmen Paradas López, Adriano Jimenez-Escrig, Mario Ezquerra, and Ana Gorostidi Pagola. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the NIH.

Funding

National ALS Registry/CDC/ATSDR (1R01TS000289); National ALS Registry/CDC/ATSDR CDCP-DHHS-US (CDC/ATSDR 200-2013-56856); NIEHS K23ES027221; NIEHS R01ES030049; NINDS R01NS127188, ALS Association (20-IIA-532), the Dr. Randall W. Whitcomb Fund for ALS Genetics, the Peter R. Clark Fund for ALS Research, the Scott L. Pranger ALS Clinic Fund, and the NeuroNetwork for Emerging Therapies at the University of Michigan. This work was supported in part by the Intramural Research Program of the NIH, National Institute on Aging (Z01-AG000949-02).

Project “ALS Genetic study in Madrid Autonomous Community” funded by “ESTRATEGIAS FRENTE A ENFERMEDADES NEURODEGENERATIVAS” from Spanish Ministry of Health

Competing interests

JFV-C receives payment for lectures and presentations from Biogen. PM receives payments for honoraria or lectures from Abbvie, Abbott, and Zambon. LG-D receives consulting fees, payment, or honoraria from Akcea, Alnylan, Genzyme, Sobi, Pfizer and equipment donation from Pfizer. JIC receives payment for lectures and presentations from Abbvie, Bial, and Zambon. BJT holds a patent for “Diagnostic and therapeutic implications for the C9orf72 repeat expansion” and has collaborative research agreements with Ionis Pharmaceuticals, Roche, and Optimeos. ELF receives consulting fees from Novartis and is an inventor on a patent held by University of Michigan titled, “Methods for Treating Amyotrophic Lateral Sclerosis.” SAG is an inventor on a patent held by University of Michigan titled, “Methods for Treating Amyotrophic Lateral Sclerosis.” JFD, KMB, KG, JH, LZ, SS-A, ARS, RC, AG-R, RR-G, RFS, SB-C, PG-C, MTP, JP-T, FC, MM-G, JR, DB-H, and PP declare no competing interests.

Authors’ contributions

Substantial contributions to the conception or design of the work: KMB, SS-A, BJT, ELF, and SAG. Acquisition, analysis, or interpretation of data for the work: JFD, KMB, KG, JH, LZ, SS-A, ARS, RC, AG-R, RR-G, JFV-C, RFS, SB-C, PG-C, MTP, PM, JP-T, FC, MM-G, JR, DB-H, LG-D, JIC, PP, BJT, ELF, and SAG. Drafting the work or revising it critically for important intellectual content: JFD, KMB, KG, JH, LZ, SS-A, ARS, RC, AG-R, RR-G, JFV-C, RFS, SB-C, PG-C, MTP, PM, JP-T, FC, MM-G, JR, DB-H, LG-D, JIC, PP, BJT, ELF, and SAG.

Data and Code Availability

Data may be shared by qualified investigators by reasonable request to the corresponding author. A data request proposal is reviewed and approved by a review panel, and a signed data-sharing agreement will then be approved. Code to perform preprocessing and analyses is available (<https://github.com/bakulskilab>).

References

1. Goutman SA, Hardiman O, Al-Chalabi A, et al. Emerging insights into the complex genetics and pathophysiology of amyotrophic lateral sclerosis. *The Lancet Neurology* 2022 doi: 10.1016/s1474-4422(21)00414-2 [published Online First: 2022/03/26]
2. Goutman SA, Hardiman O, Al-Chalabi A, et al. Recent advances in the diagnosis and prognosis of amyotrophic lateral sclerosis. *The Lancet Neurology* 2022 doi: 10.1016/s1474-4422(21)00465-8 [published Online First: 2022/03/26]
3. McCann EP, Henden L, Fifita JA, et al. Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. *J Med Genet* 2020 doi: 10.1136/jmedgenet-2020-106866 [published Online First: 20200514]
4. Sheppard SR, Parker MD, Cooper-Knock J, et al. Value of systematic genetic screening of patients with amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* 2021;92(5):510-18. doi: 10.1136/jnnp-2020-325014 [published Online First: 2021/02/17]
5. Gregory JM, Fagegaltier D, Phatnani H, et al. Genetics of Amyotrophic Lateral Sclerosis. *Current Genetic Medicine Reports* 2020;8(4):121-31. doi: 10.1007/s40142-020-00194-8
6. Renton AE, Majounie E, Waite A, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 2011;72(2):257-68. doi: 10.1016/j.neuron.2011.09.010
7. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* 2011;72(2):245-56. doi: 10.1016/j.neuron.2011.09.011
8. Mejjini R, Flynn LL, Pitout IL, et al. ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Front Neurosci* 2019;13:1310. doi: 10.3389/fnins.2019.01310 [published Online First: 2019/12/24]
9. McLaughlin R, Vajda A, Hardiman O. Heritability of amyotrophic lateral sclerosis: Insights from disparate numbers. *JAMA neurology* 2015 doi: 10.1001/jamaneurol.2014.4049
10. Bandres-Ciga S, Noyce AJ, Hemani G, et al. Shared polygenic risk and causal inferences in amyotrophic lateral sclerosis. *Annals of neurology* 2019;85(4):470-81. doi: 10.1002/ana.25431 [published Online First: 2019/02/07]
11. Nakamura R, Misawa K, Tohnai G, et al. A multi-ethnic meta-analysis identifies novel genes, including ACSL5, associated with amyotrophic lateral sclerosis. *Commun Biol* 2020;3(1):526. doi: 10.1038/s42003-020-01251-2 [published Online First: 20200923]
12. Li C, Ou R, Wei Q, et al. Shared genetic links between amyotrophic lateral sclerosis and obesity-related traits: a genome-wide association study. *Neurobiol Aging* 2021;102:211 e1-11 e9. doi: 10.1016/j.neurobiolaging.2021.01.023 [published Online First: 2021/03/01]
13. Nicolas A, Kenna KP, Renton AE, et al. Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* 2018;97(6):1268-83.e6. doi: 10.1016/j.neuron.2018.02.027
14. Su FC, Goutman SA, Chernyak S, et al. Association of Environmental Toxins With Amyotrophic Lateral Sclerosis. *JAMA neurology* 2016;73(7):803-11. doi: 10.1001/jamaneurol.2016.0594
15. Goutman SA, Boss J, Patterson A, et al. High plasma concentrations of organic pollutants negatively impact survival in amyotrophic lateral sclerosis. *Journal of neurology, neurosurgery, and psychiatry* 2019;90(8):907-12. doi: 10.1136/jnnp-2018-319785 [published Online First: 2019/02/15]

16. Goutman SA, Boss J, Guo K, et al. Untargeted metabolomics yields insight into ALS disease mechanisms. *Journal of neurology, neurosurgery, and psychiatry* 2020 doi: 10.1136/jnnp-2020-323611 [published Online First: 2020/09/16]
17. Bandres-Ciga S, Ahmed S, Sabir MS, et al. The Genetic Architecture of Parkinson Disease in Spain: Characterizing Population-Specific Risk, Differential Haplotype Structures, and Providing Etiologic Insight. *Mov Disord* 2019;34(12):1851-63. doi: 10.1002/mds.27864 [published Online First: 2019/10/30]
18. García-Redondo A, Dols-Icardo O, Rojas-García R, et al. Analysis of the C9orf72 gene in patients with amyotrophic lateral sclerosis in Spain and different populations worldwide. *Human mutation* 2013;34(1):79-82. doi: 10.1002/humu.22211 [published Online First: 2012/09/01]
19. Bakulski KM, Vadari HS, Faul JD, et al. Cumulative Genetic Risk and APOE ϵ 4 Are Independently Associated With Dementia Status in a Multiethnic, Population-Based Cohort. *Neurol Genet* 2021;7(2):e576. doi: 10.1212/nxg.0000000000000576 [published Online First: 2021/03/11]
20. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* 2019;8(7) doi: 10.1093/gigascience/giz082 [published Online First: 2019/07/16]
21. Ware EB, Faul JD, Mitchell CM, et al. Considering the APOE locus in Alzheimer's disease polygenic scores in the Health and Retirement Study: a longitudinal panel study. *BMC Med Genomics* 2020;13(1):164. doi: 10.1186/s12920-020-00815-9 [published Online First: 2020/11/05]
22. Nicolas A, Kenna KP, Renton AE, et al. Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* 2018;97(6):1268-83.e6. doi: 10.1016/j.neuron.2018.02.027
23. Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 2021;591(7849):211-19. doi: 10.1038/s41586-021-03243-6 [published Online First: 2021/03/12]
24. Al-Chalabi A, Calvo A, Chio A, et al. Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. *The Lancet Neurology* 2014;13(11):1108-13. doi: 10.1016/s1474-4422(14)70219-4
25. Yanes T, McInerney-Leo AM, Law MH, et al. The emerging field of polygenic risk scores and perspective for use in clinical care. *Human molecular genetics* 2020;29(R2):R165-r76. doi: 10.1093/hmg/ddaa136 [published Online First: 2020/07/06]
26. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* 2018;19(9):581-90. doi: 10.1038/s41576-018-0018-x
27. Wainberg M, Magis AT, Earls JC, et al. Multiomic blood correlates of genetic risk identify presymptomatic disease alterations. *Proceedings of the National Academy of Sciences of the United States of America* 2020;117(35):21813-20. doi: 10.1073/pnas.2001429117 [published Online First: 2020/08/21]
28. McCann EP, Henden L, Fifita JA, et al. Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. *Journal of medical genetics* 2021;58(2):87-95. doi: 10.1136/jmedgenet-2020-106866
29. Shi N, Kawano Y, Tateishi T, et al. Increased IL-13-producing T cells in ALS: positive correlations with disease severity and progression rate. *J Neuroimmunol* 2007;182(1-2):232-5. doi: 10.1016/j.jneuroim.2006.10.001 [published Online First: 2006/11/14]
30. Placek K, Benatar M, Wu J, et al. Machine learning suggests polygenic risk for cognitive dysfunction in amyotrophic lateral sclerosis. *EMBO molecular medicine* 2021;13(1):e12595. doi: 10.15252/emmm.202012595 [published Online First: 2020/12/04]
31. Restuadi R, Garton FC, Benyamin B, et al. Polygenic risk score analysis for amyotrophic lateral sclerosis leveraging cognitive performance, educational attainment and

- schizophrenia. *European journal of human genetics* : *EJHG* 2021 doi: 10.1038/s41431-021-00885-y [published Online First: 2021/04/29]
32. Leonenko G, Baker E, Stevenson-Hoare J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nature Communications* 2021;12(1):4506. doi: 10.1038/s41467-021-24082-z
 33. Chaudhury S, Brookes KJ, Patel T, et al. Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. *Translational Psychiatry* 2019;9(1):154. doi: 10.1038/s41398-019-0485-7
 34. Najar J, van der Lee SJ, Joas E, et al. Polygenic risk scores for Alzheimer's disease are related to dementia risk in APOE ϵ 4 negatives. *Alzheimers Dement (Amst)* 2021;13(1):e12142-e42. doi: 10.1002/dad2.12142
 35. Escott-Price V, Myers A, Huentelman M, et al. Polygenic Risk Score Analysis of Alzheimer's Disease in Cases without APOE4 or APOE2 Alleles. *J Prev Alzheimers Dis* 2019;6(1):16-19. doi: 10.14283/jpad.2018.46 [published Online First: 2018/12/21]
 36. Escott-Price V, Shoai M, Pither R, et al. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiology of aging* 2017;49:214.e7-14.e11. doi: <https://doi.org/10.1016/j.neurobiolaging.2016.07.018>
 37. Baker E, Escott-Price V. Polygenic Risk Scores in Alzheimer's Disease: Current Applications and Future Directions. *Frontiers in Digital Health* 2020;2 doi: 10.3389/fdgth.2020.00014
 38. Witte JS, Visscher PM, Wray NR. The contribution of genetic variants to disease depends on the ruler. *Nat Rev Genet* 2014;15(11):765-76. doi: 10.1038/nrg3786 [published Online First: 2014/09/17]
 39. Arnold N, Koenig W. Polygenic Risk Score: Clinically Useful Tool for Prediction of Cardiovascular Disease and Benefit from Lipid-Lowering Therapy? *Cardiovasc Drugs Ther* 2021;35(3):627-35. doi: 10.1007/s10557-020-07105-7 [published Online First: 2020/11/06]
 40. Sordillo JE, Kraft P, Wu AC, et al. Quantifying the Polygenic Contribution to Cutaneous Squamous Cell Carcinoma Risk. *J Invest Dermatol* 2018;138(7):1507-10. doi: 10.1016/j.jid.2018.01.031 [published Online First: 2018/02/13]
 41. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res Treat* 2012;132(2):365-77. doi: 10.1007/s10549-011-1818-2 [published Online First: 2011/11/01]
 42. Maas P, Barrdahl M, Joshi AD, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncol* 2016;2(10):1295-302. doi: 10.1001/jamaoncol.2016.1025
 43. Lello L, Raben TG, Hsu SDH. Sibling validation of polygenic risk scores and complex trait prediction. *Scientific Reports* 2020;10(1):13190. doi: 10.1038/s41598-020-69927-7
 44. Bogumil D, Conti DV, Sheng X, et al. Replication and Genetic Risk Score Analysis for Pancreatic Cancer in a Diverse Multiethnic Population. *Cancer Epidemiol Biomarkers Prev* 2020;29(12):2686-92. doi: 10.1158/1055-9965.Epi-20-0963 [published Online First: 2020/09/23]
 45. Goutman SA, Guo K, Savelieff MG, et al. Metabolomics identifies shared lipid pathways in independent amyotrophic lateral sclerosis cohorts. *Brain : a journal of neurology* 2022 doi: 10.1093/brain/awac025 [published Online First: 2022/01/29]
 46. Saez-Atienzar S, Bandres-Ciga S, Langston RG, et al. Genetic analysis of amyotrophic lateral sclerosis identifies contributing pathways and cell types. *Sci Adv* 2021;7(3) doi: 10.1126/sciadv.abd9036 [published Online First: 2021/02/02]
 47. Maniatis S, Aijo T, Vickovic S, et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science (New York, NY)* 2019;364(6435):89-93. doi: 10.1126/science.aav9776 [published Online First: 2019/04/06]

Tables

Table 1. Included study sample characteristics by ALS case and control status for shared ancestry cohort.

Characteristic	Control N = 223 ¹	ALS N = 219 ¹	p-value ²
ALS polygenic score with C9orf72 region removed	-0.08 (-0.73, 0.64)	0.03 (-0.55, 0.74)	0.11
ALS polygenic score with C9orf72 region included	-0.13 (-0.74, 0.65)	0.08 (-0.56, 0.75)	0.080
C9orf72 Expansion Status			<0.001
Negative	223 (100%)	206 (94%)	
Positive	0 (0%)	13 (5.9%)	
Family History of ALS	0 (0%)	17 (7.8%)	<0.001
ALS Onset Segment			-
Bulbar	0 (0%)	59 (27%)	
Cervical	0 (0%)	80 (37%)	
Lumbar	0 (0%)	73 (33%)	
Respiratory	0 (0%)	1 (0.5%)	
Thoracic	0 (0%)	4 (1.8%)	
Generalized	0 (0%)	2 (0.9%)	
Not applicable	209 (100%)	0 (0%)	
Age (years)	65 (58, 71)	67 (59, 73)	0.4
Sex			0.027
Female	115 (52%)	90 (41%)	
Male	108 (48%)	129 (59%)	
Self-Reported Race/Ethnicity			0.037
Asian	2 (0.9%)	1 (0.5%)	
Black or African American	11 (5.0%)	2 (0.9%)	
Hispanic or Latino	5 (2.3%)	3 (1.4%)	
White or Caucasian	203 (92%)	213 (97%)	
Missing	2	0	
Multi-Ancestry Genetic PC1	0.0080 (0.0077, 0.0082)	0.0080 (0.0078, 0.0082)	0.9
Multi-Ancestry Genetic PC2	-0.020 (-0.020, -0.020)	-0.020 (-0.020, -0.019)	0.6
Multi-Ancestry Genetic PC3	-0.0079 (-0.0086, -0.0071)	-0.0080 (-0.0085, -0.0071)	0.7
Multi-Ancestry Genetic PC4	-0.0095 (-0.0101, -0.0089)	-0.0098 (-0.0102, -0.0090)	0.066
Multi-Ancestry Genetic PC5	-0.004 (-0.007, 0.000)	-0.004 (-0.007, -0.001)	0.8

¹Median (25th percentile, 75th percentile); n (%)

²Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test
ALS, Amyotrophic lateral sclerosis; N, number; PC, principal component

Table 2. Regression results in the full sample used in sensitivity analyses (n=223 controls, n=219 ALS cases).

Regression results provided are odds ratios and 95% confidence intervals within parentheses for association with ALS status. All Firth penalized logistic regression models were also adjusted for participant age, sex, and 5 genetic ancestry principal components. Polygenic scores for ALS are based on weights in an independent genome-wide association study (Nicholas et al. 2018).¹³

Variable	C9orf72 region SNPs excluded from polygenic score			C9orf72 region included
	Polygenic score ($P_{\text{threshold}}=0.0001$, N= 275 SNPs)	Polygenic score ($P_{\text{threshold}}=0.0001$, N= 275 SNPs)	Polygenic score ($P_{\text{threshold}}=1.0$, N= 254,280 SNPs)	Polygenic score ($P_{\text{threshold}}=0.001$, N= 280 SNPs)
	N = 442 participants	N = 416 participants (without family history and/or C9orf72 expansion)	N = 442 participants	N = 442 participants
Polygenic score (one standard deviation increase)	1.28 (1.04, 1.57)	1.28 (1.04, 1.57)	1.13 (0.77, 1.66)	1.28 (1.05, 1.58)
C9orf72 repeat (positive)	22.8 (2.8, 2954)	-	20.7 (2.6, 2674)	21.4 (2.7, 2775)
Family history of ALS (yes)	33.2 (4.3, 4268)	-	32.6 (4.3, 4184)	32.7 (4.2, 4209)
Age (10-year increase)	1.1 (0.91, 1.33)	1.1 (0.91, 1.33)	1.09 (0.9, 1.32)	1.09 (0.9, 1.32)
Sex (male)	1.52 (1.02, 2.27)	1.52 (1.02, 2.28)	1.47 (0.99, 2.19)	1.53 (1.03, 2.29)

SNP: single nucleotide polymorphism

Figures

Figure 1. *C9orf72* region of chromosome 9 visualized as a locus zoom plot.

Single nucleotide polymorphisms (SNPs) are plotted by genomic position. The y-axis corresponds to $-\log_{10}(\text{p-values})$ from the ALS genome-wide association study (Nicholas et al. 2018). We considered the *C9orf72* region to span from 27.4 Mb to 27.7 Mb on chromosome 9 as illustrated with the blue dashed box. In an independent sample, our primary polygenic score excluded the *C9orf72* region, and a sensitivity polygenic score included these SNPs. The SNP highlighted by the green diamond (rs3849943, located chr9:27543382) was associated with *C9orf72* repeat status (fisher p-value = 0.00001). Below the plot, positions of *C9orf72* as well as other genes in the region are shown.

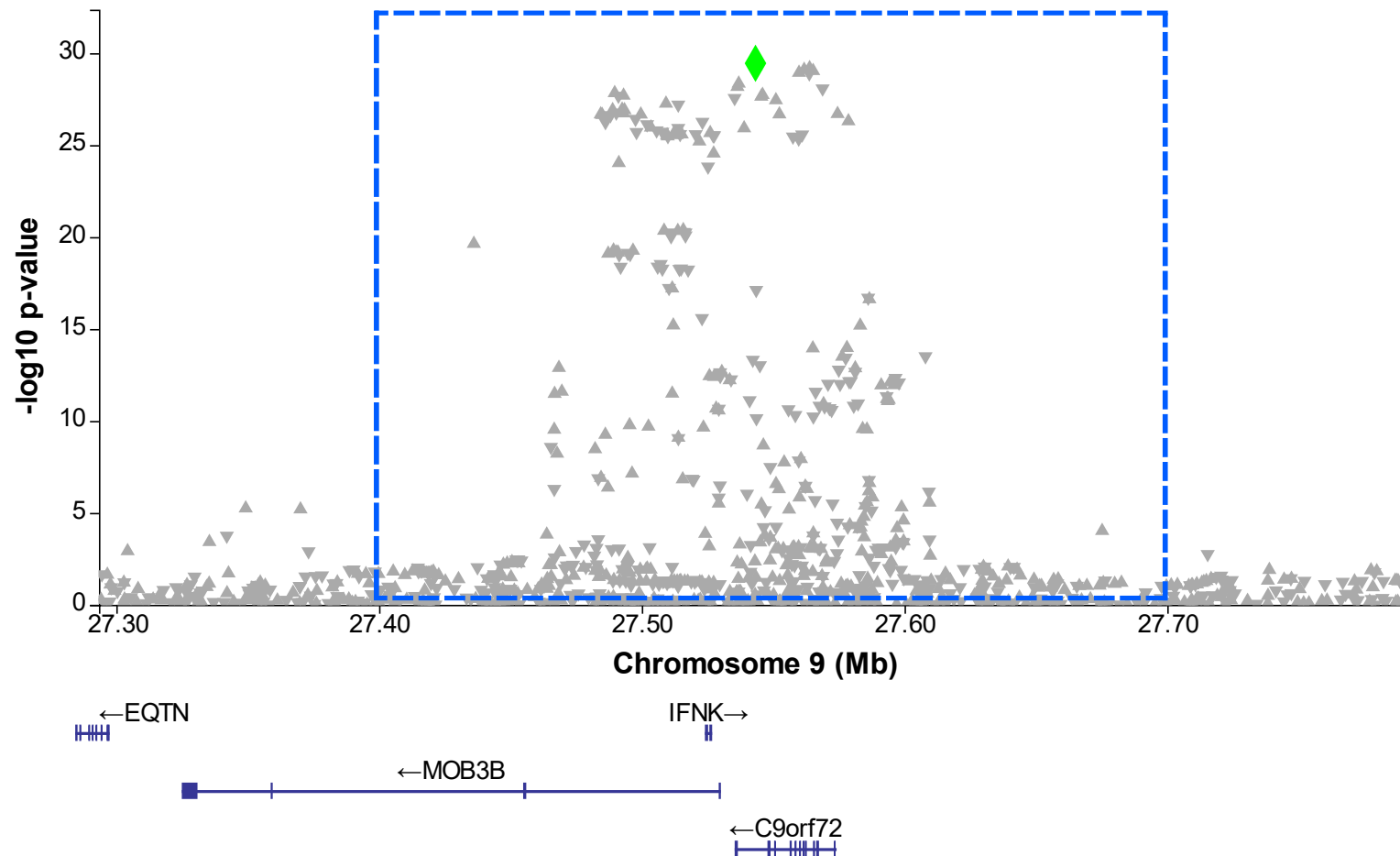


Figure 2. ROC Curve. Base model has sex, age, ancestry principal components. (n = 442)

Receiver operating characteristic curve (ROC) for classification of ALS and control participants. The base model includes sex, age, and 5 genetic principal components and has an area under the curve (AUC) of 0.591. Adding family history to the base model increases the AUC to 0.631 (likelihood ratio test p-value = 0.06). Adding *C9orf72* expansion in addition to family history increases the AUC to 0.647 (likelihood ratio test p-value < 0.001). Adding polygenic score (PGS, region around *C9orf72* removed) in addition to family history and *C9orf72* expansion improves the AUC to 0.663 (likelihood ratio test p-value < 0.001).

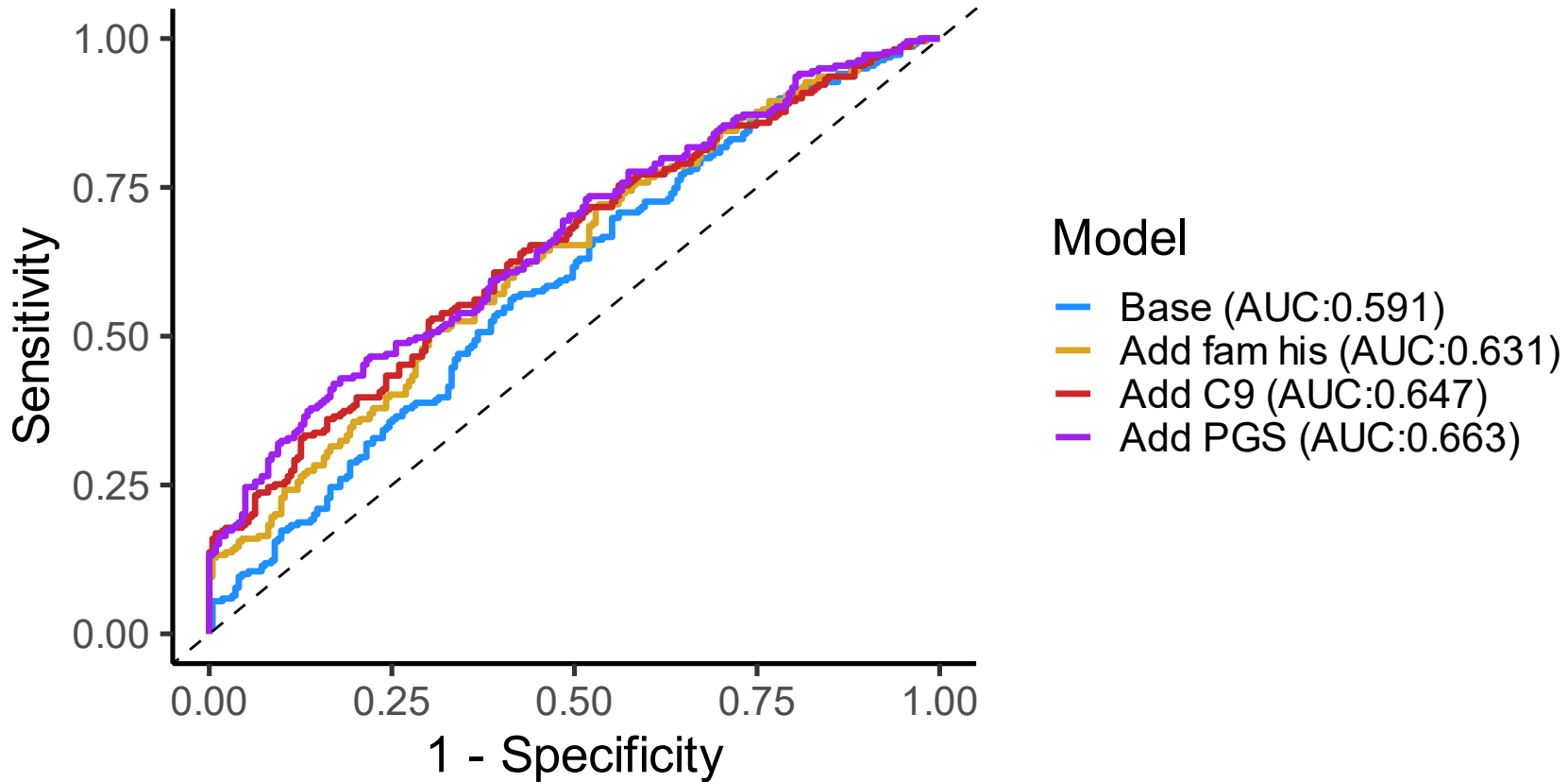


Figure 3. Highly enriched gene ontology (GO) biological processes

The 50 most significantly enriched biological functions using GO are illustrated in dot plots. Rich Factor refers to the proportion of single nucleotide polymorphism (SNP) associated genes belonging to a specific term. The color indicates the level of significance ($-\log_{10}(P_{adj})$). The numbers correspond to the number of SNP associated genes belong to the term.

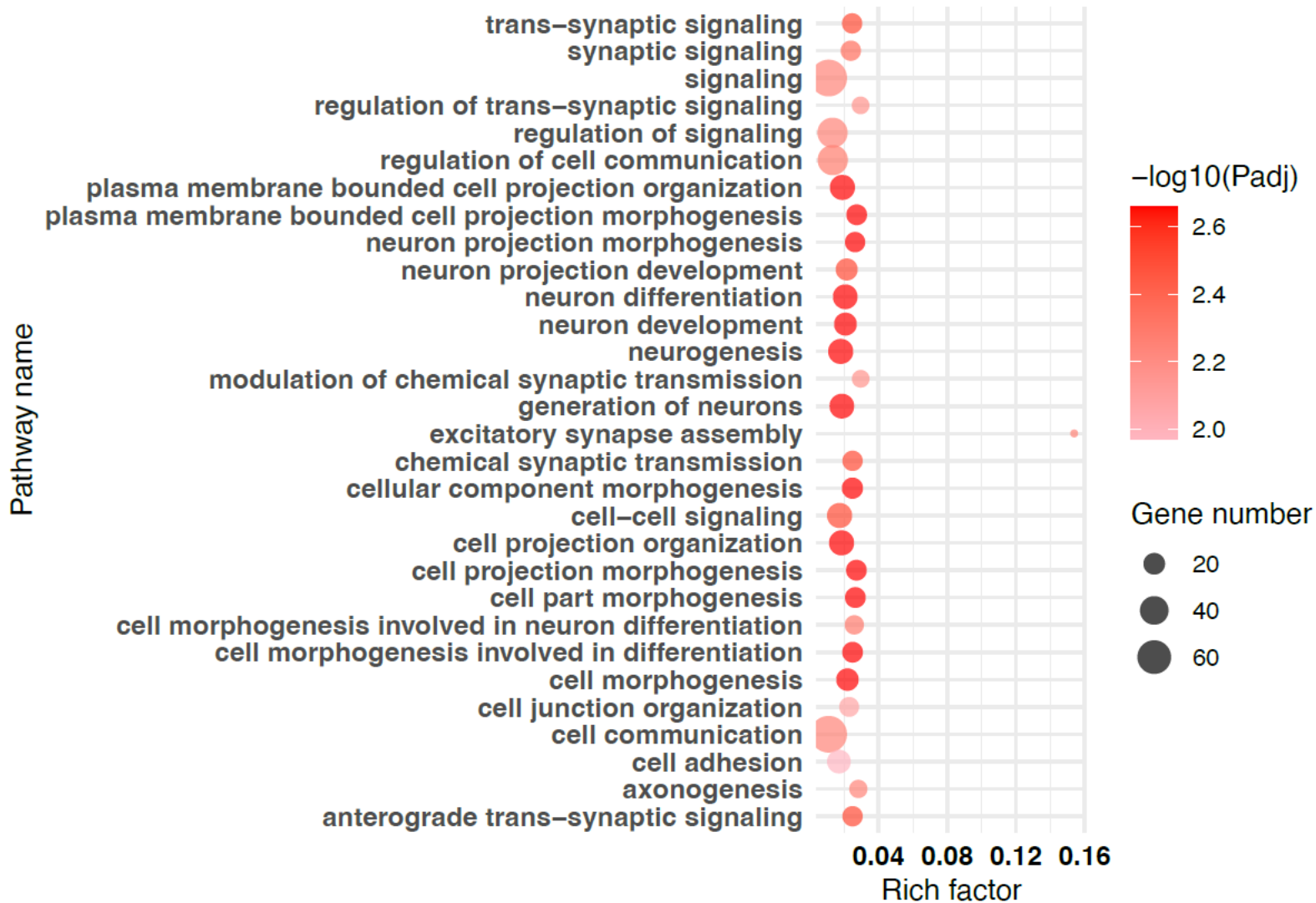


Figure 4. Highly enriched KEGG pathways.

The significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways are illustrated in dot plots. Rich Factor refers to the proportion of single nucleotide polymorphism (SNP) associated genes belonging to a specific term. Node size (Gene number) refers to the number of SNP associated genes within each term and node color indicates the level of significance ($-\log_{10}p$ -value).

