TITLE

Rare variant analyses validate known ALS genes in a multi-ethnic population and identifies *ANTXR2* as a candidate in PLS

AUTHORS: Tess D. Pottinger[1,2], Joshua E. Motelow[1,3], Gundula Povysil[1], Cristiane A. Martins Moreno[4], Zhong Ren[1], Hemali Phatnani[5,6,7], The New York Genome Center ALS Sequencing Consortium, Timothy J. Aitman[8], Javier Santoyo-Lopez[9], Scottish Genomes Partnership, GTAC Investigators, David B. Goldstein[1], Matthew B. Harms[1,5,6]

[1]Institute for Genomic Medicine, Columbia University Irving Medical Center, New York, New York, United States of America.

[2]Department of Internal Medicine, Columbia University Irving Medical Center, New York, New York, United States of America.

[3]Department of Pediatrics, Columbia University Irving Medical Center, New York, New York, United States of America.

[4]Department of Neurology, School of Medicine, Universidade de São Paulo, São Paul, Brazil.

[5]Department of Neurology, Columbia University Irving Medical Center, New York, New York, United States of America.

[6]Center for Motor Neuron Biology and Disease, Columbia University Irving Medical Center, New York, New York, United States of America.

[7]New York Genome Center, New York, New York, United States of America.

[8]Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, Scotland.

[9]Edinburgh Genomics, University of Edinburgh, Edinburgh, Scotland.

ABSTRACT

Rare genetic variants have been previously associated with amyotrophic lateral sclerosis (ALS). In a large multi-ethnic cohort of 6,970 ALS patients, rare variant burden testing validated known ALS genes and demonstrated the utility of intolerant domain-based collapsing methods. A first-ever analysis in 166 patients with primary lateral sclerosis (PLS), a motor neuron disease related to ALS, found a candidate association to loss-of-function mutations in *ANTXR2*.

MAIN TEXT

ALS is a rare neurodegenerative disease characterized by the progressive loss of upper motor neurons in the cortex and lower motor neurons of the brainstem and spinal cord. Even with FDA-approved disease modifying medication and palliation by artificial nutrition and ventilation, the prognosis is poor and death from accumulating paralysis occurs a median of 32 months after symptoms first manifest[1]. Over the last 30 years, genetic study of the 5-10% of ALS patients with family history[2] have securely implicated ~20 monogenic causes and showed possible association to a similar number of genes (https://clinicalgenome.org/affiliation/40096/). Causative mutations in the most prevalent ALS genes (*C9ORF72, SOD1*, *TARDBP*, and *FUS)* explain ~70% of familial ALS[3]. Due in part to incomplete penetrance, 10% of simplex ALS cases also carry mutations in these same genes[4].

A paucity of unsolved ALS pedigrees for family studies has intersected with falling sequencing costs for large-scale sequencing to allow gene discovery studies based on rare variant burden or collapsing methods on cohorts using predominantly simplex patients. Since our group first used this approach to implicate *TBK1* and *NEK1,* others have also identified *DNAJC7, TUBA4A* and several candidates[4-7]. These analyses utilized the entire gene or recognizable functional domains as the regions for burden testing [4,5] and were restricted to

cohorts with European ancestry, or with less than 5% non-European ALS cases. Recognizing that power for discovery could be improved by a) increasing case and control numbers, b) diversifying the ancestries of participants, and c) collapsing on domains known to be intolerant of variation, we conducted both standard gene and intolerant domain-based collapsing analyses on 6,970 multi-ethnic ALS cases (87% European) and 22,534 ancestry-matched controls.

Standard gene collapsing analyses were conducted to identify case enrichment of rare variants (minor allele frequency of 0.001) in an ALS cohort with 12 sub-population groups (Supplemental Figure 1A) that correspond to ancestry-based clusters (Supplemental Figure 1B; Supplemental Tables 1). Analyses were conducted on clusters with at least 3 cases. Controls were drawn from individuals sequenced for phenotypes/diseases with no known association with ALS (Supplemental Table 3). As expected, a negative control analysis for rare synonymous variants found no case-enrichment (Supplemental Figure 3). Collapsing analysis of all rare functional variants (missense and protein truncating variants) (Supplemental Table 5) found genome-wide and study-wide significant ($p < 4.9 \times 10^{-7}$) case-enrichment for known ALS genes *SOD1*, *TARDBP*, *TBK1* (OR=19.18, $p = 3.67 \times 10^{-39}$; OR=4.73, $p = 2 \times 10^{-10}$; OR=2.3, $p = 7.49 \times 10^{-9}$, respectively) and control-enrichment for *ALKBH3* (OR=0.26, $p = 4.88 \times 10^{-7}$) (Figure 1A). Although *SOD1, TBK1* and *TARDBP* are definitive ALS genes, we were intrigued by the identification of control-enriched *ALKBH3.* Control-enrichment was not explained by sequencing methodology, ancestry cluster, or specific phenotype/disease population within the control cohort. Because *ALKBH3* plays a role in DNA repair[8], a mechanism increasingly implicated in ALS pathogenesis[9] , we attempted to replicate this novel association by analyzing summary statistics from the Project MinE cohort, which is similar in size to ours [10]. None of the available models focused on variation as rare as in our analyses, but at a higher minor allele frequency (MAF) for qualifying variants (0.005), a minor degree of control-enrichment was in fact observed (OR= 0.56, $p = 3.96 \times 10^{-4}$). This raises the possibility that rare missense and

protein truncating variants (PTVs) in *ALKBH3* could protect from ALS, a finding that requires validation in large cohorts.

Because gene-based collapsing considers variation across the entire gene, regions that are tolerant of variation could swamp case-enrichment signals originating from regions that are intolerant of variation [5]. To overcome this limitation, we conducted rare variant collapsing on domains that are intolerant to variation as defined as a subRVIS domain score threshold of 25, a cutoff based on threshold testing. This analysis implicated the same three known ALS genes (*SOD1*, *TARDBP*, and *TBK1* at OR=20.63, $p = 1.68 \times 10^{-38}$; OR=10.08, $p = 3.62 \times 10^{-16}$; and OR=3.15, $p = 8.38 \times 10^{-11}$, respectively) (Figure 1B). The intolerant domain analysis did not improve over the gene-based analysis for *SOD1* or *TBK1* (Supplemental Figure 4; Supplemental Figure 5) but doubled the odds ratio and significantly lowered the p-value obtained for *TARDBP.* The improvement of the intolerant domain model (Figure 1C, 1D) stemmed from a significant drop (one-tailed z-score p=0.031) in the number of qualifying variants found in controls dispersed across tolerant regions, while highlighting qualifying variants in ALS cases predominantly in the intolerant C-terminal region.

Primary lateral sclerosis (PLS) is also a neurodegenerative disease of motor neurons with clinical features, neuropathology, and some genetics that overlaps with ALS[11] . PLS is nearly always simplex and 20 times rarer than ALS[12]. Because large-scale sequencing studies of ALS often include PLS patients we were able to conduct a gene-based collapsing analysis in 166 PLS multi-ethnic cases (88% European) and 17,695 ancestry matched controls (Supplemental Figure 2; Supplemental Tables 2). We knew the study would be underpowered for securely implicating causative genes but used this as an opportunity to generate candidates for future study. Although most models showed no significant genes, the dominant PTV model showed significant case enrichment for *ANTXR2* (OR=174.57, $p=8.38 \times 10^{-6}$) (Figure 2; Supplemental Table 6), a gene associated with brain connectivity changes and Alzheimer's

disease[13]. Currently, there are no additional large sequencing studies of PLS in which we could attempt replication.

In summary, we performed the largest rare variant analyses of a multi-ethnic population of patients with ALS to date. Our analysis did not identify new ALS genes but demonstrated that collapsing models informed by regions of intolerance can be useful for identifying genes where disease-associated variation is limited to regions with low background variation. This analysis also confirmed the association of the C-terminal domain of *TARDBP.* We also identified *ALKBH3* as a potentially protective gene that warrants further study in additional and larger cohorts. Finally, we conducted the first rare variant collapsing analysis in PLS, identifying PTVs in *ANTXR2* as a candidate disease gene. This association and potential mechanisms for PTVs in this gene will need to be investigated further in larger PLS cohorts.

It is important to note that this analysis doubled the number of ALS cases and quadrupled the number of controls from our first study[4] but remained underpowered for the identification of new ALS genes. A recently published rare variant burden analysis with a similar number of ALS cases did not identify new genes[10] either, emphasizing the need for increasingly large genomically characterized ALS cohorts, especially in non-European populations.

Figures

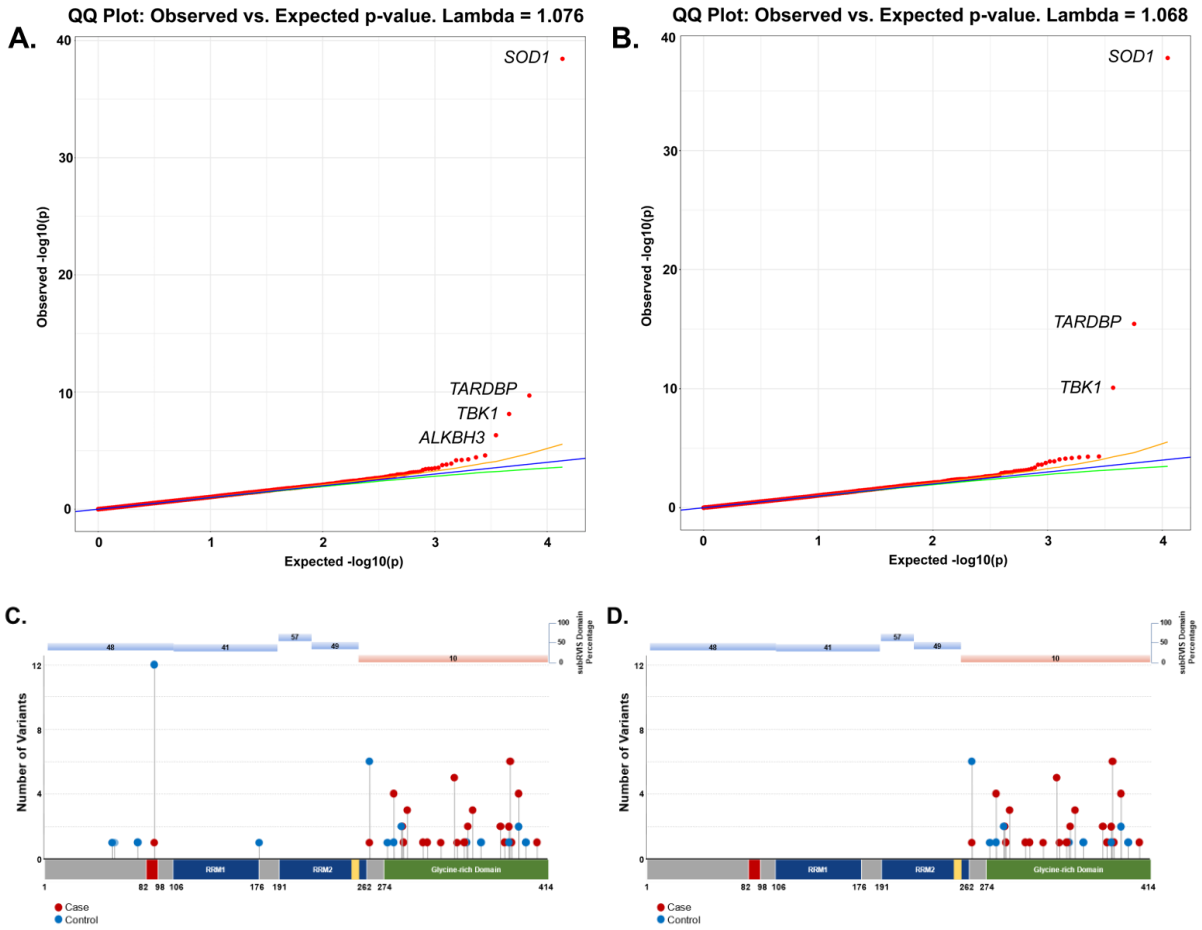


Figure 1. Q-Q plots of gene- and domain-level collapsing of ALL functional coding variants in ALS cohort. (A) The results for a standard gene-level collapsing of 6,970 ALS cases and 22,524 controls. P-values were generated using an exact two-sided Cochran-Mantel-Haenszel (CMH) by gene by cluster. The genes with the top associations that achieved study-wide significance of $p<4.9\times10^{-7}$ (*SOD1* (OR=19.18), *TARDBP* (OR=4.73), *TBK1* (OR=2.3), and *ALKBH3* (OR=0.26)) are labeled. *SOD1, TARDBP, TBK1* have been previously implicated in rare variant association studies of ALS. Yellow and green lines indicate the 2.5th and 97.5th percentile of expected p-values, respectively. (B) The results for the domain-based collapsing restricting qualifying variants to those with subRVIS domain percentage score < 25 of 6,970 cases and 22,524 controls. P-values were generated using an exact two-sided Cochran-Mantel-

Haenszel (CMH) by gene by cluster. The genes with the top associations (*SOD1* (OR=20.63), *TARDBP* (OR=10.08), and *TBK1* (OR=3.15)) are labeled. (C) Standard gene-level collapsing model showed 44 qualifying variants in cases (red circles) and 31 in controls (blue circles) for *TARDBP* (D) subRVIS domain collapsing improved association by removing control variants (cases = 43; controls = 15). Regions with subRVIS domain percentage below 25 are highlighted in orange while those above this threshold are highlighted in blue. A one tailed z-score showed that there were significantly less controls in the intolerant domain as indicated by subRVIS domain percentage score < 25 (p=0.031).

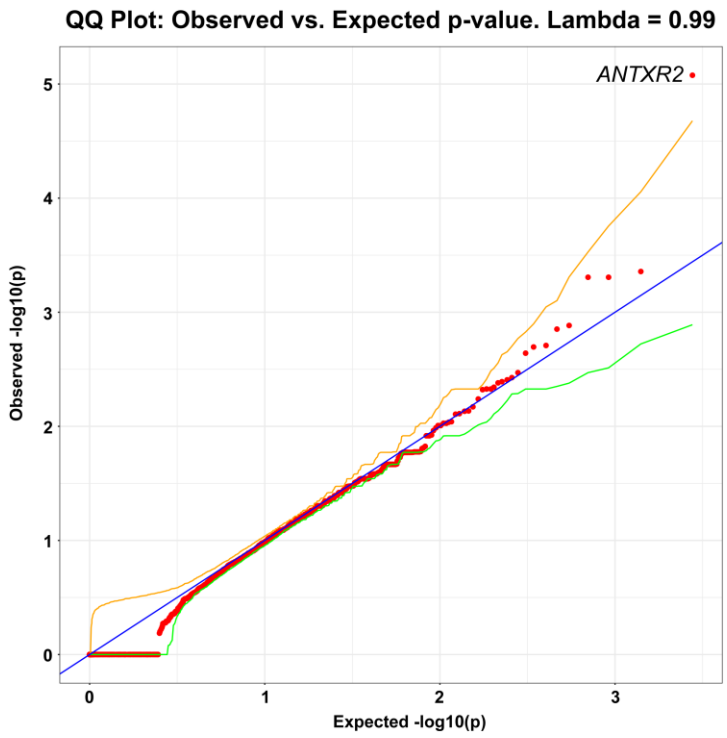


Figure 2. Q-Q plot of gene-level collapsing of protein truncating variants (PTV) in PLS cohort. The results for a standard gene-level collapsing of 166 PLS cases and 17,695 controls. P-values were generated using an exact two-sided Cochran-Mantel-Haenszel (CMH) by gene by cluster. The gene with the top associations that achieved genome-wide significance of $p<8.38\times10^{-6}$ (*ANTXR2* (OR=174.57)) is labeled. *ANTXR2* has not been previously implicated in rare variant association studies of PLS. Yellow and green lines indicate the 2.5th and 97.5th percentile of expected p-values, respectively.

FUNDING SOURCES:

The collection of ALS samples and data was funded in part by: The Scottish Genomes Partnership (Chief Scientist Office of the Scottish Government Health Directorates (SGP/1) and The Medical Research Council Whole Genome Sequencing for Health and Wealth Initiative (MC/PC/15080); The New York Genome Center ALS Consortium (ALS Association 15-LGCA-234, 19-SI-459, and the Tow Foundation; The GTAC study (ALS Association 16-LGCA-310 and Biogen Idec); ALS Exome Sequencing Consortium (Biogen Idec).

The collection of control samples and data was funded in part by: Bryan ADRC NIA P30 AG028377; NIH RO1 HD048805; Gilead Sciences, Inc.; D. Murdock; National Institute of Allergy and Infectious Diseases Center for HIV/AIDS Vaccine Immunology (CHAVI) (U19-AI067854); National Institute of Allergy and Infectious Diseases Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery (UM1-AI100645); Bill and Melinda Gates Foundation; NINDS Award# RC2NS070344; New York-Presbyterian Hospital; The Columbia University College of Physicians and Surgeons; The Columbia University Medical Center; NIH U54 NS078059; NIH P01 HD080642; The J. Willard and Alice S. Marriott Foundation; The Muscular Dystrophy Association; The Nicholas Nunno Foundation; The JDM Fund for Mitochondrial Research; The Arturo Estopinan TK2 Research Fund; UCB; Epilepsy Genetics Initiative, A Signature Program of CURE; Epi4K Gene Discovery in Epilepsy study (NINDS U01-NS077303) and The Epilepsy Genome/Phenome Project (EPGP - NINDS U01-NS053998); The Ellison Medical Foundation New Scholar award AG-NS-0441-08; National Institute Of Mental Health of the National Institutes of Health under Award Number K01MH098126; B57 SAIC-Fredrick Inc. M11-074; OCD Rare 1R01MH097971-01A1. This research was supported in part by funding from Funding from the Duke Chancellor's Discovery Program Research Fund 2014; an American Academy of Child and Adolescent Psychiatry (AACAP) Pilot Research Award; NIMH Grant RC2MH089915; Endocrine Fellows Foundation Grant; The NIH Clinical and Translational Science Award Program (UL1TR000040); NIH U01HG007672; The Washington Heights Inwood Columbia Aging Project; The Stanley Institute for Cognitive Genomics at Cold Spring Harbor Laboratory and the Utah Foundation for Biomedical Research. Data collection and sharing for the WHICAP project (used as controls in this analysis) was supported by The Washington Heights–Inwood Columbia Aging Project (WHICAP, PO1AG07232, R01AG037212, RF1AG054023) funded by the National Institute on Aging (NIA) and by The National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001873. This manuscript has been reviewed by WHICAP investigators for scientific content and consistency of data interpretation with previous WHICAP Study publications. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ACKNOWLEDGEMENTS:

We thank the following groups for contributing ALS samples, sequencing, or clinical data:

New York Genome Center ALS Consortium: J. Kwan, D. Sareen, J.R. Broach, Z. Simmons, X. Arcila-Londono, E.B. Lee, V.M. Van Deerlin, E. Fraenkel, L.W. Ostrow, F. Baas, N. Zaitlen, J.D. Berry, A. Malaspina, P. Fratta, G.A. Cox, L.M. Thompson, S. Finkbeiner, E. Dardiotis, T.M. Miller, S. Chandran, S. Pal, E. Hornstein, D.J. MacGowan, T. Heiman-Patterson, M.G. Hammell, N.A. Patsopoulos, J. Dubnau, and A. Nath.

ALS Exome Sequencing Consortium: S.H. Appel, R.H. Baloh, R.S. Bedlack, R. Brown, W.K. Chung, S. Gibson, J.D. Glass, A. Gitler, D.B. Goldstein, T.M. Miller, R.M. Myers, S.M. Pulst, J.M. Ravits, G. Rouleau, E. Greene, N. Shneider, and W.W. Xin;

Genomic Translation for ALS Care (GTAC) study: S.H. Appel, R.H. Baloh, R.S. Bedlack, S. Chandran, L. Foster, S. Goutman, E. Green, C. Karam, D. Lacomis, G. Manousakis, T.M. Miller, S. Pal, D. Sareen, A. Sherman, Z. Simmons, L. Wang.

We thank The Washington Heights–Inwood Columbia Aging Project (WHICAP) for the contribution of control samples. We also thank the WHICAP study participants and the WHICAP research and support staff for their contributions to this study: K. Welsh-Bomer, C. Hulette, and J. Burke; D. Valle, J. Hoover-Fong, and N. Sobriera; A. Poduri; S. Palmer; R. Buckley; K. Newby; The Murdock Study Community Registry and Biorepository Pro00011196; National Institute of Allergy and Infectious Diseases Center for HIV/AIDS Vaccine Immunology (CHAVI) (U19-AI067854); National Institute of Allergy and Infectious Diseases Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery (UM1-AI100645); CHAVI Funding; R. Ottman; V. Shashi; S. Berkovic, I. Scheffer, and B. Grinton; The Epi4K Consortium and Epilepsy Phenome/Genome Project; C. Depondt, S. Sisodiya, G. Cavalleri, and N. Delanty; S. Hirose; C. Woods, C. Village, K. Schmader, S. McDonald, M. Yanamadala, and H. White; G. Nestadt, J. Samuels, and Y. Wang; D. Levy; E. Pras, D. Lancet, and Z. Farfel; S. Chen; R. Wapner; C. Moylan, A. Mae Diehl, and M. Abdelmalek; DUHS (Duke University Health System) Nonalcoholic Fatty Liver Disease Research Database and Specimen Repository; M. Winn and R. Gbadegesin; M. Hauser; S. Delaney; A. Need and J. McEvoy; M. Walker; M. Sum; Undiagnosed Diseases Network; National Institute on Aging (R01AG037212, P01AG007232).

References

1. Dorst J, Ludolph AC, Huebers A. Disease-modifying and symptomatic treatment of amyotrophic lateral sclerosis. *Ther Adv Neurol Disord.* 2018;11:1756285617734734.
2. Chen S, Sayana P, Zhang X, Le W. Genetics of amyotrophic lateral sclerosis: an update. *Mol Neurodegener.* 2013;8:28.
3. Chia R, Chiò A, Traynor BJ. Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. *Lancet Neurol.* 2018;17(1):94-102.
4. Cirulli ET, Lasseigne BN, Petrovski S, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science.* 2015;347(6229):1436-1441.
5. Gelfman S, Dugger S, de Araujo Martins Moreno C, et al. A new approach for rare variation collapsing on functional protein domains implicates specific genic regions in ALS. *Genome Res.* 2019;29(5):809-818.
6. Jih KY, Tsai PC, Tsai YS, Liao YC, Lee YC. Rapid progressive ALS in a patient with a DNAJC7 loss-of-function mutation. *Neurol Genet.* 2020;6(5):e503.
7. Smith BN, Ticozzi N, Fallini C, et al. Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. *Neuron.* 2014;84(2):324-331.
8. Kuang W, Jin H, Yang F, et al. ALKBH3-dependent m(1)A demethylation of Aurora A mRNA inhibits ciliogenesis. *Cell Discov.* 2022;8(1):25.
9. Konopka A, Atkin JD. DNA Damage, Defective DNA Repair, and Neurodegeneration in Amyotrophic Lateral Sclerosis. *Front Aging Neurosci.* 2022;14:786420.
10. van Rheenen W, van der Spek RAA, Bakker MK, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat Genet.* 2021;53(12):1636-1648.
11. Tartaglia MC, Rowe A, Findlater K, Orange JB, Grace G, Strong MJ. Differentiation between primary lateral sclerosis and amyotrophic lateral sclerosis: examination of symptoms and signs at disease onset and during follow-up. *Arch Neurol.* 2007;64(2):232-236.
12. Turner MR, Barohn RJ, Corcia P, et al. Primary lateral sclerosis: consensus diagnostic criteria. *J Neurol Neurosurg Psychiatry.* 2020;91(4):373-377.
13. Elsheikh SSM, Chimusa ER, Mulder NJ, Crimi A. Genome-Wide Association Study of Brain Connectivity Changes for Alzheimer's Disease. *Sci Rep.* 2020;10(1):1433.

ONLINE METHODS

Methods

Study Population

All samples and data came from participants that provided written, informed consent for genetic studies that had been IRB-approved at each contributing center. The study cohort includes participants from the Genomic Translation for ALS Care (GTAC study), the Columbia University Precision Medicine Initiative for ALS, the New York Genome Consortium, and the ALS Sequencing Consortium (IRB-approved genetic studies from Columbia University Medical Center, including the Coriell NINDS repository), University of Massachusetts at Worchester, Stanford University (including samples from Emory University School of Medicine, the Johns Hopkins University School of Medicine, and the University of California, San Diego), Massachusetts General Hospital Neurogenetics DNA Diagnostic Lab Repository, Duke University, McGill University (including contributions from Saint-Luc and Notre-Dame Hospital of the Centre Hospitalier de l'Université de Montréal [CHUM], [University of Montreal]), Gui de Chauliac Hospital of the CHU de Montpellier (Montpellier University), Pitié Salpêtrière Hospital, Fleurimont Hospital of the Centre Hospitalier Universitaire de Sherbrooke (CHUS) (University of Sherbrooke), Enfant Jésus Hospital of the Centre hospitalier affilié universitaire de Québec (CHA) (Laval University), Montreal General Hospital, Montreal Neurological Institute and Hospital of the McGill University Health Centre, the University of Edinburgh Scotland, and Washington University in St. Louis (including contributions from Houston Methodist Hospital, Virginia Mason Medical Center, University of Utah, and Cedars Sinai Medical Center). Participants were determined to have ALS or PLS by neuromuscular specialists at tertiary motor neuron disease care centers with expertise in distinguishing between the two. ALS diagnoses were based on the El Escorial Criteria in all cases. For PLS, explicit criteria requiring >3 years of symptoms without conversion to ALS were used for 79 of the 172 PLS participants. The criteria used for the remaining PLS diagnoses were not available.

Controls were selected from >100,000 whole-exome or -genome sequenced individuals housed in the IGM Data Repository. Individuals with known neurodegenerative disease were excluded. However, none of the controls were screened for neurodegenerative disease. All participants consented to the use of DNA in genetic research.

Whole exome and genome sequencing

DNA sequencing was performed at Columbia University, the New York Genome Center, Duke University, McGill University, Stanford University, HudsonAlpha, and University of Massachusetts, Worcester. Kits used to conduct whole-exome capture are as follows: Agilent All Exon kits (50MB, 65MB, and CRE), Nimblegen SeqCap EZ Exome Enrichment kits (V2.0, V3.0, VCRome, and MedExome), and IDT Exome Enrichment panel. There were 2,185 participants with ALS who were sequenced using Nimblegen SeqCap EZ Exome Enrichment kits and 51 who were sequenced using the IDT Exome Enrichment panel (Supplemental Table 4). While 1,272 controls were evaluated using the Aligent All Exon kits, 8,498 with the IDT Exome Enrichment panel, and 11,201 with the Nimblegen SeqCap EZ Exome Enrichment kits. Sequencing was performed using Illumina GAIIx, HiSeq 2000, HiSeq 2500, and NovaSeq 6000 sequencers according to standard protocols. Whole genome sequencing was conducted at the New York Genome Center and in-house at the IGM. Sample-level BAM files were transferred from the New York Genome Center to the IGM (n = 3,418). An additional 1,316 genomes were processed by the IGM. There were 1,553 genomes in our control cohort (Supplemental Table 4). Data were aligned to the human reference genome (NCBI Build 37) using DRAGEN (Edico Genome, San Diego, CA, USA). Picard (http://picard.sourceforge.net) was used to remove duplicate reads and to process lane-level BAM files to create a sample-level BAM file. GATK was used to recalibrate base quality scores, realign around indels, and call variants utilizing the Best Practices recommendations v3.6 [1]. Variants were annotated using ClinEff and the Analysis Tool for Annotated Variants (ATAV), an in-house IGM annotation tool [2]. Variants were annotated

with gnomAD v2.1 frequencies, regional-intolerance metrics, and the clinical annotations by the Human Gene Mutation Database (HGMD), ClinVar, and Online Mendelian Inheritance in Man (OMIM). Exonic regions were retained for downstream statistical analyses.

Sample and Variant Quality Control

Samples reporting >2% contamination according to verifyBamID [3] and those with consensus coding sequence (CCDS release 20) <90% were excluded from these analyses. KING [4] was used to test for relatedness. Only unrelated (up to second-degree) individuals were included in these analyses. For related pairs, samples were chosen to prefer cases. Samples where X:Y coverage ratios did not match expected sex were excluded.

Only variants within the CCDS or the 2 bp canonical sites were included in these analyses. These variants were also required to have a quality score of at least 50, a quality by depth score of at least 5, genotype quality score of at least 20, read position rank sum of at least −3, mapping quality score of at least 40, mapping quality rank sum greater than −10, and a minimum coverage of at least 10. SNVs had a maximum Fisher's strand bias of 60, while indels had a maximum of 200. For heterozygous genotypes, the alternative allele ratio was required to be greater than or equal to 30%. Only variants with the GATK Variant Quality Score Recalibration filter "PASS", "VQSRTrancheSNP90.00to99.00", or "VQSRTrancheSNP99.00to99.90" were included. Variants were excluded if they were marked by EVS, ExAC, or gnomAD as being failures (http://evs.gs .washington.edu/EVS/).

Clustering, ancestry, and coverage harmonization

A neural network pre-trained on samples of known ancestry was used to calculate probability estimates for six ancestry groups (African, East Asian, European, Hispanic, Middle Eastern, and South Asian). Methods for characterizing samples into clusters has been previously described [5].

To ensure balanced sequencing coverage of evaluated sites between cases and controls, we imposed a statistical test of independence between the case/control status and coverage as previously described [6]. Sites were removed where the absolute difference in percentages of cases and controls with at least 10x coverage was greater than 7%. Samples were then pruned using this method on a cluster-by-cluster basis. Through this approach, approximately 7- 11% were removed. Clusters with less than 5 participants were not included in these analyses, thereby removing 6 participants with PLS but none with ALS.

Variant-level statistical analysis

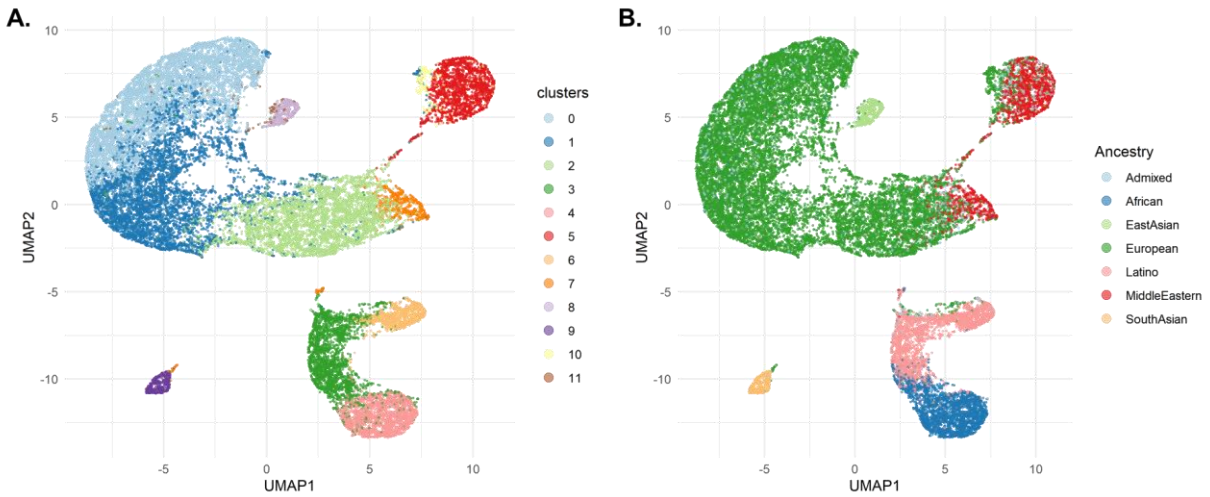
The models that were used to test for associations of nonsynonymous coding or canonical splice variants with outcome included variants with MAF <0.1% for each population represented in gnomad and internal AF of <0.1%. Models tested were a standard gene-unit collapsing analysis, and a domain-unit analysis. The models used for these analyses were previously described [7]. A domain-based approach utilizing subRVIS domain percentage with a threshold of 25 was also used to evaluate case enrichment of rare variants. The full list of 18,653 CCDS genes was analyzed for each model. Genes with at least one qualifying variant were included for analyses. As we are meta analyzing across clusters an exact 2-sided Cochran-Mantel-Haenszel test was used (using the stats package in R v3.6). Study-wide significance was determined by accounting for 6 nonsynonymous models- multiplicity-adjusted significance threshold $\alpha = 4.9 \times 10^{-7}$ (Supplemental Table 5). Model inflation was calculated using empirical (permutation-based) expected probability distributions as described by Povysil and colleagues [5].

References

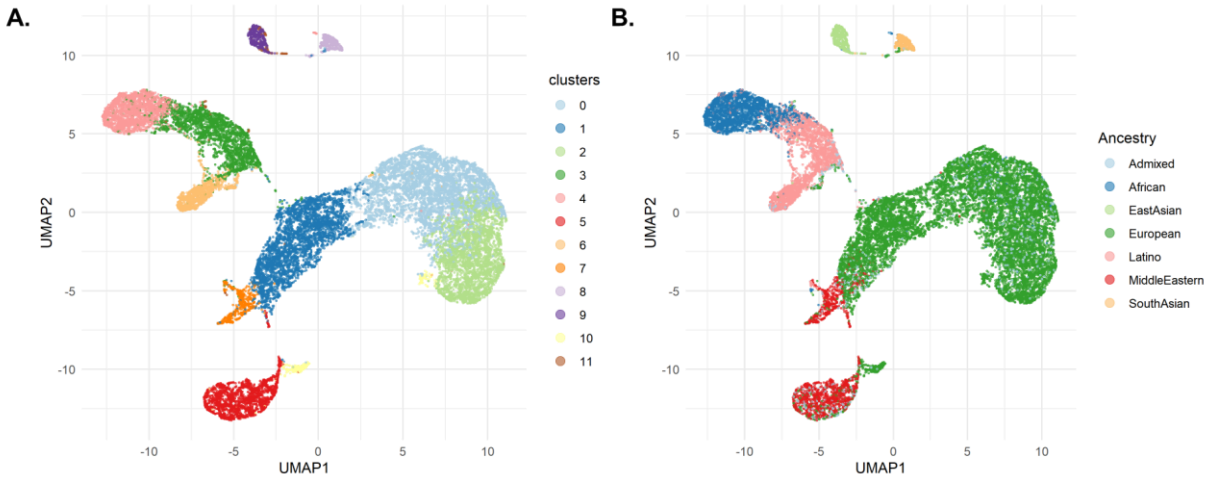
1. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303.
2. Ren Z, Povysil G, Hostyk JA, Cui H, Bhardwaj N, Goldstein DB. ATAV: a comprehensive platform for population-scale genomic analyses. *BMC Bioinformatics.* 2021;22(1):149.
3. Jun G, Flickinger M, Hetrick KN, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet.* 2012;91(5):839-848.
4. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics.* 2010;26(22):2867-2873.
5. Povysil G, Chazara O, Carss KJ, et al. Assessing the Role of Rare Genetic Variation in Patients With Heart Failure. *JAMA Cardiol.* 2021;6(4):379-386.
6. Petrovski S, Todd JL, Durheim MT, et al. An Exome Sequencing Study to Assess the Role of Rare Genetic Variation in Pulmonary Fibrosis. *Am J Respir Crit Care Med.* 2017;196(1):82-93.
7. Gelfman S, Dugger S, de Araujo Martins Moreno C, et al. A new approach for rare variation collapsing on functional protein domains implicates specific genic regions in ALS. *Genome Res.* 2019;29(5):809-818.

Supplemental Figures and Tables

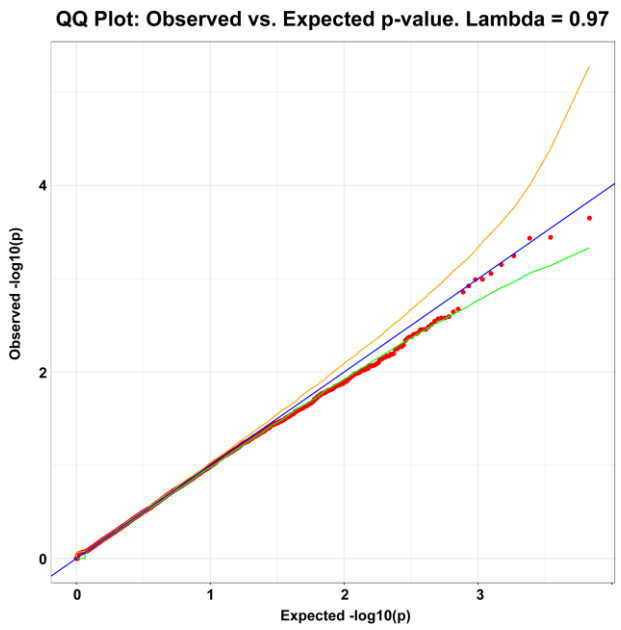
Figures



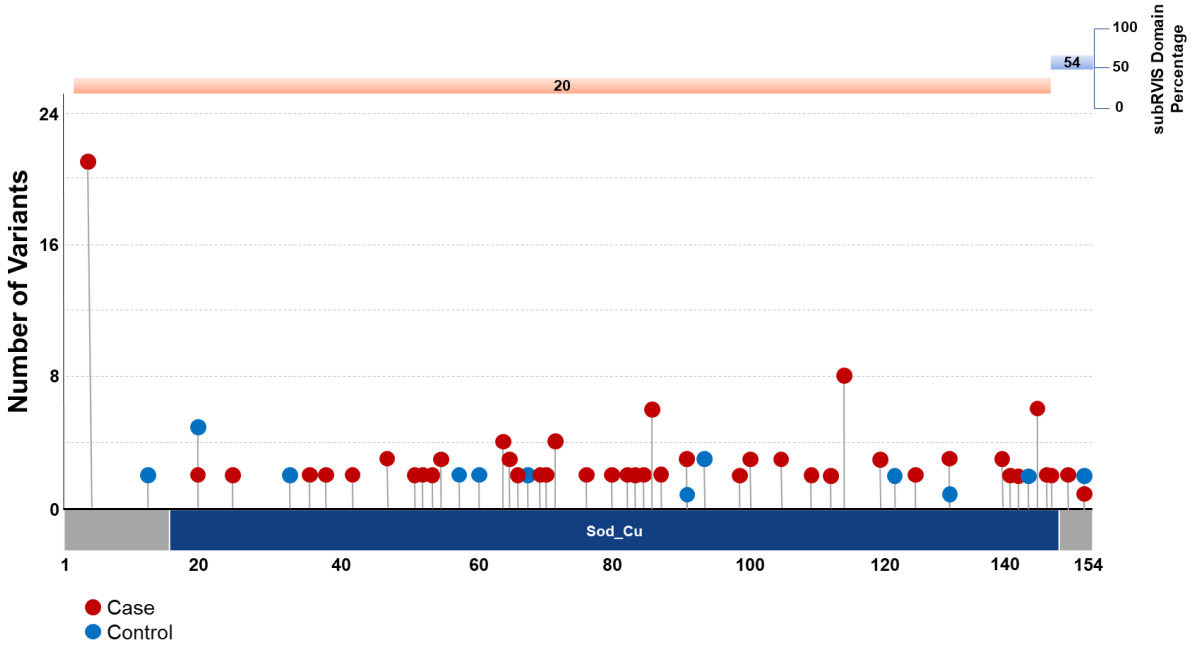
Supplemental Figure 1. **UMAP plot of ALS participants** (A) Cluster assignment (B) Predicted ancestry



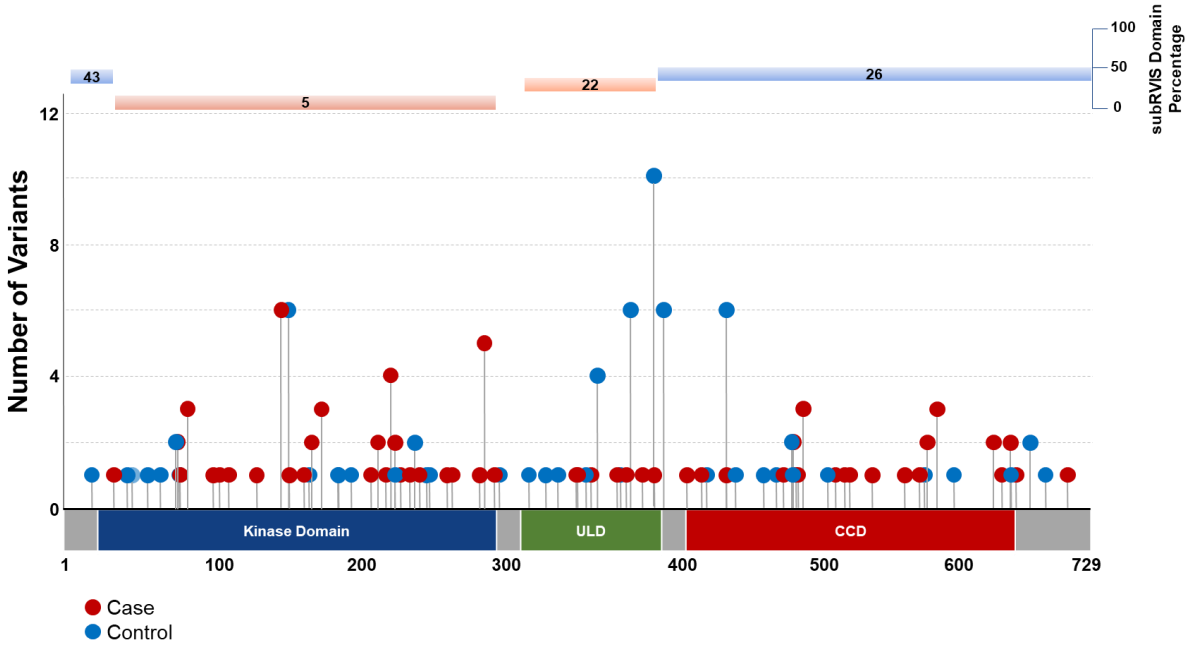
Supplemental Figure 2. **UMAP plot of PLS participants** (A) Cluster assignment (B) Predicted ancestry



Supplemental Figure 3. Q-Q plots of gene level collapsing in a Synonymous model. Yellow and green lines indicate the 2.5th and 97.5th percentile of expected p-values, respectively. The genomic inflation factor, lambda (λ), is 0.97 indicating no inflation. We generated p values from the exact two-sided Cochran-Mantel-Haenszel (CMH) test by gene by cluster to indicate a different carrier status of affected individuals in comparison to control individuals.



Supplemental Figure 4. Plot of gene- and domain-level collapsing of ALL *SOD1* functional coding variants. Standard gene-level collapsing model showed 93 qualifying variants in cases (red circles) and 18 in controls (blue circles) for *SOD1*. subRVIS domain collapsing improved association by removing control variants (cases = 90; controls = 16). Regions with subRVIS domain percentage below 25 are highlighted in orange while those above this threshold are highlighted in blue. However, a one tailed z-score showed that the differences in the number of controls in the intolerant domain was not significantly lower than those in the entire gene as indicated by subRVIS domain percentage score < 25 (p=0.4).



Supplemental Figure 5. Plot of gene- and domain-level collapsing of ALL *TBK1* functional coding variants. Standard gene-level collapsing model showed 73 qualifying variants in cases (red circles) and 143 in controls (blue circles) for *TBK1*. subRVIS domain collapsing improved association by removing control variants (cases = 47; controls = 72). Regions with subRVIS domain percentage below 25 are highlighted in orange while those above this threshold are highlighted in blue. However, a one tailed z-score showed that the differences in the number of controls in the intolerant domain was not significantly lower than those in the entire gene as indicated by subRVIS domain percentage score < 25 (p=0.3).

Tables

Supplemental Table 1. Cluster sizes for ALS cohort

Cluster	Ancestry	Cases	Controls	Ratio
0	European1	2,620	4,792	0.55
1	European2	2,502	3,989	0.63
2	European3	922	3,868	0.24
3	Latino1	104	2,788	0.037
4	African1	204	2,102	0.097
5	MiddleEastern1	294	1,924	0.15
6	Latino2	110	1,068	0.1
7	MiddleEastern2	86	669	0.13
8	EastAsian1	35	496	0.071
9	SouthAsian1	40	457	0.088
10	European4	36	265	0.14
11	Admixed1	17	106	0.15
Total		6,970	22,524	0.31

Supplemental Table 2. Cluster sizes for PLS cohort

Cluster	Ancestry	Cases	Controls	Ratio
0	European1	90	4,792	0.017
1	European2	19	3,989	0.0047
2	European3	41	3,868	0.012
3	Latino1	1	2,788	0.00039
4	African1	3	2,102	0.0013
5	MiddleEastern1	10	1,924	0.0052
6	Latino2	3	1,068	0.0027
7	MiddleEastern2	1	669	0.0016
8	SouthAsian1	2	496	0.004
9	EastAsian1	1	457	0.002
10	European4	1	265	0.0028
11	Admixed1	0	106	0
Total		172	22,530	0.0076

Supplemental Table 3. Phenotypes of control participants

Control Phenotype	Number of Participants
Kidney and urological disease	8,416 (37.36%)
Healthy family member of a different study	6,297 (27.96%)
Control of different study	3,156 (14.01%)
Epilepsy	2,892 (12.84%)
Obsessive compulsive disorder	721 (3.2%)
Pulmonary disease	403 (1.79%)
Hematological disease	256 (1.13%)
Schizophrenia	161 (0.71%)
Infectious disease	107 (0.48%)
Liver disease	82 (0.36%)
Primary immune deficiency	33 (0.15%)
Total	22,524

Supplemental Table 4. Sequencing kits used for cohort

Sample Type	Capture Kit	Cases	Controls
Exome	Agilent All Exon kits	0	1,272
	Nimblegen SeqCap EZ Exome Enrichment kits	2,185	11,201
	IDT Exome Enrichment panel	51	8,498
Genome		4,734	1,553
Total		6,970	22,524

Supplemental Table 5. Description of collapsing models

	Synonymous	PTV	Missense&PTV	Missense&PTV subRVIS
Inheritance	Dominant	Dominant	Dominant	Dominant
Functions	Synonymous	PTV	Missense + PTV	Missense + PTV
LOO MAF	0.001	0.001	0.001	0.001
External MAF	0.001	0.001	0.001	0.001
Additional Filters				subRVIS Domain Percentage <25

Supplemental Table 6. Description of PLS participants with *ANTXR2* PTV Qualifying Variants

ID	Sex	Age Range at Onset	Phenotype	Location Hg19	Effect	Depth	MAF	Alleles in Gnomad 2.1
PLS1	M	66-70	Leg onset	4-80990653-A-G	p.Met1?	39x	0.44	0
PLS2	F	41-45	Leg onset	4-80905972-C-T	c.1086+1G>A	29x	0.48	0
PLS3	M	61-65	Leg onset	4-80905984-CA-C	p.Ala359fs	20x	0.35	3
Ctrl1	F	Unk	OCD	4-80957125-C-T	c.697+1G>A	18x	0.67	1
Ctrl2	M	Unk	Kidney disease	4-80977127-TA-T	p.Ser113fs	20x	0.6	0