The Project MinE databrowser: bringing large-scale whole-genome sequencing in ALS to researchers and the public.

Authors
Project MinE ALS Sequencing Consortium[#]

Correspondence
J.H.Veldink@umcutrecht.nl

A full list of Project MinE GWAS Consortium members appears at the end of the paper.

Keywords: Amyotrophic Lateral Sclerosis, whole-genome sequencing, WGS, open-access, databrowser, genetics

SUMMARY

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive fatal neurodegenerative disease affecting 1 in 350 people. The aim of Project MinE is to elucidate the pathophysiology of ALS through whole-genome sequencing at least 15,000 ALS patients and 7,500 controls at 30X coverage. Here, we present the Project MinE data browser (databrowser.projectmine.com), a unique and intuitive one-stop, open-access server that provides detailed information on genetic variation analyzed in a new and still growing set of 4,366 ALS cases and 1,832 matched controls. Through its visual components and interactive design, the browser specifically aims to be a resource to those without a biostatistics background and allow clinicians and preclinical researchers to integrate Project MinE data into their own research. The browser allows users to query a transcript and immediately access a unique combination of detailed (meta)data, annotations and association statistics that would otherwise require analytic expertise and visits to scattered resources.

INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive fatal neurodegenerative disease affecting 1 in 350 people. While research over the past years has revealed an increasing number of genetic variants contributing to ALS risk, the bulk of heritability in ALS remains to be elucidated. In addition to known rare variants, there is evidence for a central role of low-frequency and rare genetic variation in ALS susceptibility (van Rheenen et al., 2016). Well-powered genetic studies enabled through large-scale collaboration are crucial for identify these variants and improving our understanding of ALS pathophysiology (McLaughlin et al., 2017; Nicolas et al., 2018).

Project MinE, an international collaboration, was initiated precisely with the challenge of sample aggregation in mind. The Project MinE ALS sequencing Consortium has set out to collect whole-genome sequencing (WGS) of 15,000 ALS patients and 7,500 controls (Project MinE ALS Sequencing Consortium, 2018a). Currently, the Project MinE initiative has sequenced 4,366 ALS patients and 1,832 age- and sex-matched controls. Project MinE is a largely crowd-funded initiative. As such, we are committed to sharing data and results with the scientific and healthcare communities, as well as the public more broadly. Data sharing within the genetics community facilitated large-scale genome-wide association studies and ignited initiatives such as the Gene Atlas, LDhub GWAShare Center, and MRbase, places where people can share, explore and analyze data with few restrictions (Canela-Xandri et al., 2017; Zheng et al., 2017). In this same spirit, we aim to share raw sequence data, provide results from our analyses, and facilitate interpretation through integration with existing datasets to serve translational researchers and the public across disciplines.

We therefore created the Project MinE databrowser (databrowser.projectmine.com). We integrated multi-level association statistics, metadata, and public resources including gnomAD, GTEx and ClinVar in an intuitive and flexible framework (GTEx Consortium, 2013; Landrum et al., 2016; Lek et al., 2016), Figure 1. These data are freely available through the browser for any research initiative. We aim for the data to serve several purposes, including providing a backbone for new gene discovery, serving as a costless replication dataset, facilitating translational efforts and aiding clinical interpretation of individual ALS patient genomes or specific genetic variants.

RESULTS

Dataset

The databrowser currently comprises 4,366 ALS cases and 1,832 age- and sex-matched controls whole-genome sequenced and quality control processed as part of the broader Project MinE effort.

Quality control and association analysis

The quality controlled dataset includes 6,198 individuals and describes more than 105 million SNVs and indels. In this sample we have limited power to detect genome-wide significant association in a single variant framework and as a result we did not find any variants reaching genome-wide significance. In our rare-variant burden framework we find that the excess of disruptive and damaging variants at MAF < 1% in the canonical transcript of *NEK1* in ALS patients compared to controls reaches exome-wide significance ($p = 2.31 \times 10^{-7}$, odds ratio = 3.55 [95% confidence

interval = 2.02 – 6.26], Figure 2). We also noticed that some genes might contain a transcript specific burden, most notably in *TARDBP* (Table S3).

Next, we aggregated all variants across the exome. We observed no difference in the exome-wide burden of *synonymous* variants between cases and controls, which provides no indication for systemic confounding of burden analyses using higher-order variant aggregation strategies. Therefore, we proceeded to test a genome-wide excess of rare *non-synonymous* variants among ALS patients. In contrast to similar analyses in schizophrenia (Genovese et al., 2016) and educational attainment (Ganna et al., 2016), we found no evidence for such excess in any variant set combining all allele frequency cut-offs and functional classification. Furthermore, we do not find any protein families, druggable categories significantly enriched for rare variants after collapsing allele-frequency cut-offs and variant classification. All association analysis results are available for download at the browser website.

Databrowser

By entering a gene or transcript in the databrowser you will be shown a visualisation of the rare-variant burden tests, as well as several other components (Figure 3).

Transcript details. Here we describe the elementary transcript details for the gene of interest. This includes the Ensembl transcript ID, Ensembl Gene ID, number of exons and genomic coordinates as described in the GRCh37 build.

Coverage information (Figure 3a). To illustrate whether a particular gene/transcript or exon has been adequately covered to detect variation, we have included a graphical representation of average depth of coverage. This graph also includes the coverage information from the ExAC database to illustrate the difference

in coverage between genome- and exome-sequencing. Optionally, the coverage across introns can be visualized.

Genic burden results (Figure 3b). Burden testing, by definition, aggregates many variants. This approach can increase statistical power to find an association, but can obscure which variant(s) are driving a potential association. Therefore, we have included an interactive graphical representation of the gene indicating where variants are located and whether these variants are case or control-specific. Hovering over a specific variant will reveal the position, alleles, heterozygous and homozygous allele counts in ALS cases and controls, and functional annotation of the variant. We additionally provide the burden test statistics. To further facilitate interpretation, we describe the burden test properties and relevant references in a dropdown menu “Burdentest Info.” We have performed genic burden results for all transcripts.

Geneset burden results (Figure 3c). Here, we show burden test results for genesets such as protein families and druggable targets to which the selected gene belongs. This includes a mini-Manhattan plot generated to indicate which genes might be driving an association signal in the geneset by plotting their individual genic burden results.

Tissue-specific gene expression (Figure 3d). This panel shows gene expression levels across all general tissues included in GTEx.

Variant annotation table (Figure 3e). Each variant has been extensively annotated and aggregated in a customizable table. By default, only allele frequency in cases and controls, comparison to gnomAD genomes and exomes, and amino acid change, impact and functional consequence are shown. All information can be downloaded in tabular form.

Gene-specific literature. To provide background information on the gene's function and disease association from literature, we have included an iframe linking to PubMed, UCSC, GeneCards, Ensembl, WikiGenes, GTEx and the GWAScatalog. This allows a user to rapidly extract information from various resources while staying on the same page.

Group and individual level data sharing

The summary statistics for the latest GWAS, WGS single variant association and all WGS burden analyses can be downloaded directly. Access to individual-level data can be requested by providing a digital form with a brief research proposal (https://www.projectmine.com/research/data-sharing/).

Duplicate and relatedness checks

We have created sumchecks for each individual in our dataset. Sumchecks are hashes which have been created, based on a small subset of SNPs, which allow for the identification of duplicates without sharing the genetic data itself. If researchers wish to check duplicates with our dataset, they can simply request the sumchecks, create hashes for their own data and compare the hashes. The code to generate the hashes and the list of SNPs used is available on the Project MinE Bitbucket. These hashes only identify duplicate samples, and in some instances relatedness information can be valuable, e.g., extending pedigrees or meta-analyses. Therefore, we will perform the relatedness checks when a statement is uploaded that this information will be used for academic purposes only and will not be used to re-identify individuals without consent. These checks do not require a data-access request nor approval.

Technical details

The whole website, including data storage, runs on a dual core server with 4Gb RAM and needs <50Gb of storage. As of July 2018, we have had over 6,200 sessions from over 1,400 users.

DISCUSSION

Both fundamental research, translational research and clinical work increasingly rely on open-access databases to find newly-associated variants and interpret genetic findings when counselling patients (Gleeson et al., 2017; Middleton, 2018). Therefore, sharing de-identified data is instrumental to ensuring scientific and clinical progress, and patient-derived data should not be regarded as intellectual property nor as trade secret (Acmg Board Of Directors, 2017). Also, most genetic browsers are based on healthy individuals, or unselected individuals who might carry specific rare genetic variants which hampers adequate comparison to a sample of patients from another geographical region. With exactly this in mind, we developed a unique, publicly-available, disease-specific databrowser which serves as a transparent framework for sharing data and results in ALS genetics. The Project MinE Databrowser contains an unprecedented amount of WGS data from ALS patients, more than doubling the currently-available exome based databases, and provides (meta)data in far greater detail. The intuitive design facilitates interpretation of robust statistical association analyses by presenting detailed metadata and through

integration with population-based observations, biological/functional context and literature. As a result, we make our data and results accessible to a broad public of diverse backgrounds and for any research initiative. The databrowser provides an easy framework for other consortia who are generating similar genetic data and results in ALS and other diseases.

The data has already provided a backbone for new gene discovery and variant interpretation in ALS. For example, subsets of the current dataset have been incorporated in previous publications which identified *C21orf2*, *NEK1* and *KIF5A (Kenna et al., 2016; Nicolas et al., 2018; van Rheenen et al., 2016)*. The resource will continue to grow as the Project MinE consortium does, and will thus increasingly allow for more reliable identification of true positives (Project MinE ALS Sequencing Consortium, 2018b; Van Der Spek et al., 2017). The growth in both sample size and ancestral diversity will increasingly reflect the ALS mutation spectrum and yield increasingly accurate estimations of effect sizes in the general population. The browser can also offer researchers quick, easy to access to a reliable dataset for significant improvement in statistical power without financial burden.

One of the major goals of the databrowser is to allow cross-disciplinary interrogation and interpretation of the data with minimal effort. We enable this through the intuitive display of individual variant level data, statistical results and through the integration with databases including GTEx and gnomAD. The databrowser ensures transparency and continued reevaluation of established associations, vitally important for clinical laboratories to make appropriate variant classifications (Project MinE ALS Sequencing Consortium, 2018b). Furthermore, we aim to facilitate the design of functional experiments by showing which variants,

might be driving a genic burden signal and if these are located in specific exons and therefore specific protein domains.

Project MinE is largely crowd-funded and the ALS-community is highly engaged in the scientific progress in our field. Consequently, we feel an obligation to give something back to the community and promote data sharing in general. We hope that our databrowser will inspire similar efforts in other fields. The Project MinE databrowser is a light-weight and open-source R script that can easily be adapted to serve other consortia and thus share similarly important data. Further, we aim to improve data sharing by encouraging fellow researchers to gain access to individual-level data by submitting an analysis proposal to the consortium. After access is granted, analyses can be performed on the compute facilities of SURFsara, a supercomputer based in Amsterdam, The Netherlands . Researchers will only need to pay a minimal fee to compensate costs for their core hours and data storage requirements.

Project MinE continues to work forward to its ultimate goal of whole-genome sequencing 15,000 cases and 7,500 matched controls, as well as combining the data with publicly-available control data. Current efforts also focus on single SNV and aggregated SNV analyses of autosomal chromosomes. Future efforts will aim to include sex chromosomes, indels, structural variation (in particular, repeat expansions (Dolzhenko et al., 2017)) and non-coding burden analyses. Additionally, Project MinE is collecting methylation data on all samples using the Infinium Human Methylation 450K and EPIC BeadChip. These data and analyses will also be shared expeditiously through our databrowser prior to publication. As the project proceeds and data generation continues apace, we intend for the browser to pave the way for

more accurate diagnosis and prognosis, aid in the identification of novel disease-associated genes, and elucidate potential novel therapeutic targets.

ACKNOWLEDGEMENTS

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

AUTHOR CONTRIBUTIONS

R.A.A.v.d.S wrote source-code for the databrowser and together with W.v.R. and J.H.V. designed the databrowser, performed and discussed all analyses and wrote the manuscript. S.P performed QC on the WGS dataset and was involved is revising the manuscript. K.P.K, R.L.McL, M.M., A.D., G.T. contributed to data collection and discussions to improve the design of the databrowser. L.H.v.d.B. advised and assisted in study design. Unmentioned authors contributed to data collection and funding in their countries.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

PROJECT MINE ALS SEQUENCING CONSORTIUM

Rick A.A. van der Spek[1#], Wouter Van Rheenen[1#], Sara L. Pulit[2], Kevin P. Kenna[1], Russell L. McLaughlin[3], Matthieu Moisse[4,5,6], Annelot M. Dekker[1], Gijs H.P. Tazelaar[1], Brendan Kenna[1], Kristel R. Van Eijk[1], Joke J.F.A. Van Vugt [1], Perry T.C. Van Doormaal[1], Bas Middelkoop[1], Raymond D. Schellevis[1], William J. Brands[1], Ross Byrne[3], Johnathan Cooper-Knock[7], Ahmad Al Khleifat [8], Yolanda Campos[9], Atay Vural[10], Jonathan D. Glass[11,12], Alfredo Iacoangeli[13], Aleksey Shatunov[8], William Sproviero[8], Ersen Kavak[14], Tuncay Seker[10], Fulya Akçimen [10], Cemile Kocoglu [10], Ceren Tunca[10], Nicola Ticozzi[15,16], Maarten Kooyman[17], Alberto G. Redondo[18], Ian Blair[19], Naomi R. Wray[20], Matthew C. Kiernan[21], Mamede de Carvalho[22], Vivian Drory[23], Marc Gotkine[24], Peter M. Andersen[25,26], Philippe Corcia[27,28], Philippe Couratier[27,28], Vera Fominyh[29], Mayana Zatz[30], Miguel Mitne-Neto[30], Adriano Chio[31,32], Vincenzo Silani[15,16], Boris Rogelj[33,34], Blaž Koritnik[35], Janez Zidar[35], Markus Weber[36], Guy Rouleau[37], Nicolas Dupre[38,39,40], Ian Mackenzie[41], Ekaterina Rogaeva[42,43], Gabriel Miltenberger-Miltenyi[44], Lev Brylev[29], Ervina Bilić[45], Ivana Munitic[46], Victoria López Alonso[47], Karen E. Morrison[48], Stephen Newhouse[49,50], Johnathan Mill[51,52], Pamela J. Shaw[53], Christopher E. Shaw[8], Monica P. Panades[54], Jesus S. Mora[55], Wim Robberecht[4,5,6], Philip Van Damme[4,5,6], A. Nazli Basak[10], Orla Hardiman[56,57], Michael A. Van Es[1], Ammar Al-Chalabi[8], John E. Landers[58], Leonard H. Van den Berg[1§], Jan H. Veldink[1§*]

1. Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
2. Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands
3. Population Genetics Laboratory, Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Republic of Ireland
4. KU Leuven - University of Leuven, Department of Neurosciences, Experimental Neurology and Leuven Research Institute for Neuroscience and Disease (LIND), B-3000 Leuven, Belgium
5. VIB, Vesalius Research Center, Laboratory of Neurobiology, Leuven, Belgium
6. University Hospitals Leuven, Department of Neurology, Leuven, Belgium
7. Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK
8. Maurice Wohl Clinical Neuroscience Institute, King's College London, Department of Basic and Clinical Neuroscience, London, UK
9. Mitochondrial pathology Unit, Instituto de Salud Carlos III, Madrid, Spain
10. Neurodegeneration Research Laboratory, Bogazici University, Istanbul, Turkey
11. Department Neurology, Emory University School of Medicine, Atlanta, GA, USA
12. Emory ALS Center, Emory University School of Medicine, Atlanta, GA, USA
13. Department of Biostatistics, IoPPN, King's College London, London, US
14. Genomize Inc. Bogazici University, Technology Transfer Region, ETAB, Istanbul, Turkey
15. Department of Neurology and Laboratory of Neuroscience, IRCCS Istituto Auxologico Italiano, Milano, Italy
16. Department of Pathophysiology and Transplantation, Dino Ferrari Center, Università degli Studi di Milano, Milano, Italy
17. SURFsara, Amsterdam, the Netherlands
18. Hospital Carlos III, Madrid, Spain
19. Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, New South Wales, Australia
20. Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia
21. Brain and Mind Centre, The University of Sydney, New South Wales 2050, Australia
22. Physiology Institute, Faculty of Medicine, Instituto de Medicina Molecular, University of Lisbon, Lisbon, Portugal
23. Department of Neurology Tel-Aviv Sourasky Medical Centre , Israel
24. Hadassah University Hospital, Jerusalem, Israel
25. Department of Neurology, Ulm University, Ulm, Germany
26. Department of Pharmacology and Clinical Neurosience, Umea University, Umea, Sweden
27. Centre SLA, CHRU de Tours, Tours, France
28. Federation des Centres SLA Tours and Limoges, LITORALS, Tours, France
29. Neurology Department, Bujanov Moscow City Clinical Hospital
30. Human Genome and stem-cell center, Biosciences Institute, Universidade de São Paulo, Brazil
31. Rita Levi Montalcini, Department of Neuroscience, ALS Centre, University of Torino, Turin, Italy
32. Azienda Ospedaliera Citta della Salute e della Scienza, Torino, Italy
33. Department of Biotechnology, Jozef Stefan Institute, Ljubljana, Slovenia
34. Biomedical Research Institute BRIS, Ljubljana, Slovenia
35. Ljubljana ALS Centre, Institute of Clinical Neurophysiology, University Medical Centre Ljubljana, SI-1000 Ljubljana, Slovenia
36. Neuromuscular Diseases Unit/ALS Clinic, Kantonsspital St. Gallen, 9007, St. Gallen, Switzerland
37. Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
38. Department of Human Genetics, McGill University, Montreal, Quebec, Canada
39. Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
40. Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada
41. Department of Pathology and Laboratory Medicine, University of British Columbia, Canada
42. Tanz Centre for Research in Neurodegenerative Diseases, University of Toronto, Toronto, Ontario, Canada
43. Department of Medicine, Division of Neurology, University of Toronto, Toronto, Ontario, Canada
44. Physiology Institute, Faculty of Medicine, Instituto de Medicina Molecular, University of Lisbon, Lisbon, Portugal
45. Department of Neurology Clinical Hospital Center Zagreb, University of Zagreb School of Medicine
46. Department of Biotechnology, University of Rijeka
47. Computational Biology Unit, Instituto de Salud Carlos III, Madrid, Spain
48. Faculty of Medicine, University of Southampton, Southampton, UK
49. Department of Biostatistics, IoPPN, King's College London, London, US
50. Biomedical Research Centre for Mental Health, IoPPN, King's College London, London, UK
51. Maurice Wohl Clinical Neuroscience Institute, King's College London, Department of Basic and Clinical Neuroscience, London, UK
52. University of Exeter Medical School, Exeter University, St Luke's Campus, Magdalen Street, Exeter EX1 2LU, UK
53. Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK
54. Neurology Department, Hospital Universitari de Bellvitge, Barcelona, Spain
55. Hospital San Rafael, Madrid, Spain

56. Academic Unit of Neurology, Trinity College Dublin, Trinity Biomedical Sciences Institute, Dublin, Republic of Ireland
57. Department of Neurology, Beaumont Hospital, Dublin, Republic of Ireland
58. Department of Neurology, University of Massachusetts Medical School, Worcester, MA, USA

shared first, § shared last, *Corresponding and lead author: Jan H. Veldink, Department of Neurology and Neurosurgery, University Medical Centre Utrecht, Department of Neurology G03.228, P.O. Box 85500, 3508 GA Utrecht, The Netherlands , J.H.Veldink@umcutrecht.nl

REFERENCES

Acmg Board Of Directors (2017). Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. Genet. Med. *19*, 721–722.

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2017). An atlas of genetic associations in UK Biobank.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80–92.

Cotto, K.C., Wagner, A.H., Feng, Y.-Y., Kiwala, S., Coffman, A.C., Spies, G., Wollam, A., Spies, N.C., Griffith, O.L., and Griffith, M. (2017). DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. Nucleic Acids Res.

Dolzhenko, E., van Vugt, J.J.F.A., Shaw, R.J., Bekritsky, M.A., van Blitterswijk, M., Narzisi, G., Ajay, S.S., Rajan, V., Lajoie, B.R., Johnson, N.H., et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. *27*, 1895–1903.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika *80*, 27–38.

Ganna, A., Genovese, G., Howrigan, D.P., Byrnes, A., Kurki, M.I., Zekavat, S.M., Whelan, C.W., Kals, M., Nivard, M.G., Bloemendal, A., et al. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. Nat. Neurosci.

Gaudet, P., Michel, P.-A., Zahn-Zabal, M., Britan, A., Cusin, I., Domagalski, M., Duek, P.D., Gateau, A., Gleizes, A., Hinard, V., et al. (2017). The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res. *45*, D177–D182.

Genovese, G., Fromer, M., Stahl, E.A., Ruderfer, D.M., Chambert, K., Landén, M., Moran, J.L., Purcell, S.M., Sklar, P., Sullivan, P.F., et al. (2016). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat. Neurosci. *19*, 1433–1441.

Gleeson, P., Davison, A.P., Silver, R.A., and Ascoli, G.A. (2017). A Commitment to Open Source in Neuroscience. Neuron *96*, 964–965.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Heinze, G., and Ploner, M. (2004). A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems. Medical University of

Vienna, Vienna.

Kenna, K.P., van Doormaal, P.T.C., Dekker, A.M., Ticozzi, N., Kenna, B.J., Diekstra, F.P., van Rheenen, W., van Eijk, K.R., Jones, A.R., Keagle, P., et al. (2016). NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. Nat. Genet. *48*, 1037–1042.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum. Mutat. *37*, 235–241.

McLaughlin, R.L., Schijven, D., van Rheenen, W., van Eijk, K.R., O'Brien, M., Kahn, R.S., Ophoff, R.A., Goris, A., Bradley, D.G., Al-Chalabi, A., et al. (2017). Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. Nat. Commun. *8*, 14774.

Middleton, A. (2018). Society and personal genome data. Hum. Mol. Genet. *27*, R8–R13.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267.

Nicolas, A., Kenna, K.P., Renton, A.E., Ticozzi, N., Faghri, F., Chia, R., Dominov, J.A., Kenna, B.J., Nalls, M.A., Keagle, P., et al. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. Neuron *97*, 1268–1283.e6.

Project MinE ALS Sequencing Consortium (2018a). Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. Eur. J. Hum. Genet.

Project MinE ALS Sequencing Consortium (2018b). CHCHD10 variants in Amyotrophic Lateral Sclerosis: Where is the evidence? Ann. Neurol.

Pulit, S.L., With, S., and Bakker, P. (2017). Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. Genet. Epidemiol.

van Rheenen, W., Shatunov, A., Dekker, A.M., McLaughlin, R.L., Diekstra, F.P., Pulit, S.L., van der Spek, R.A.A., Võsa, U., de Jong, S., Robinson, M.R., et al. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat. Genet. *48*, 1043–1048.

Ruden, D., Cingolani, P., Patel, V., Coon, M., Nguyen, T., Land, S., and Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front. Genet. *3*, 35.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. *43*, W589–W598.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Van Der Spek, R.A., Van Rheenen, W., Pulit, S.L., Kenna, K.P., Ticozzi, N., Kooyman, M., McLaughlin, R.L., Moisse, M., Van Eijk, K.R., Van Vugt, J.J.F.A., et al. (2017). Reconsidering the causality of TIA1 mutations in ALS. Amyotroph. Lateral Scler. Frontotemporal Degener. 1–3.

Wilson, D., Madera, M., Vogel, C., Chothia, C., and Gough, J. (2007). The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res. *35*, D308–D313.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

Zheng, J., Erzurumluoglu, A.M., Elsworth, B.L., Kemp, J.P., Howe, L., Haycock, P.C., Hemani, G., Tansey, K., Laurin, C., Pourcain, B.S., et al. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics *33*, 272–279.

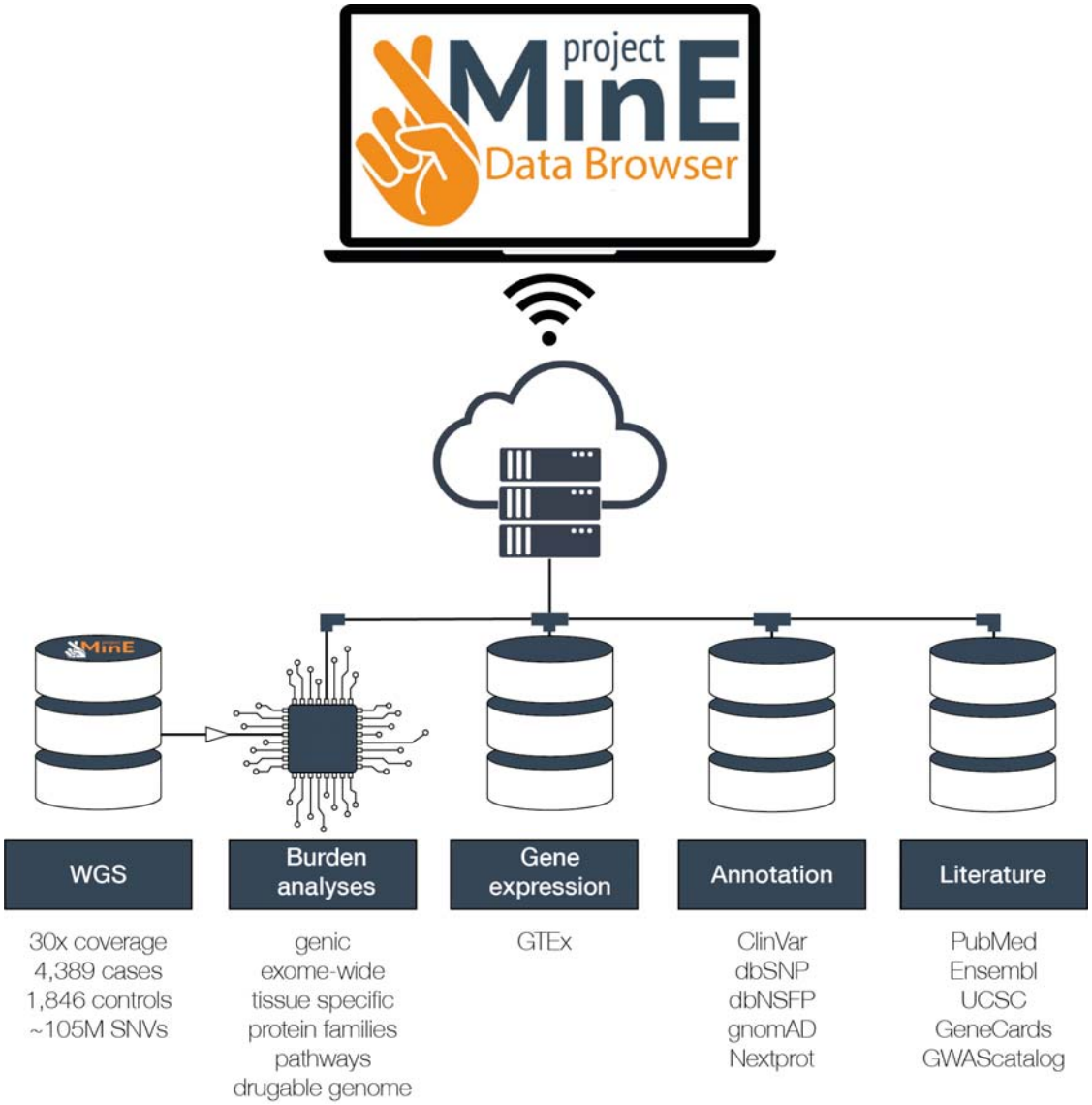

Figure 1. Schematic representation of the databrowser Whole genomes generated by Project MinE are openly available for research and the public. The databrowser does not have a login requirement. It integrates multiple public resources and provides a wide range of robust statistical analyses.

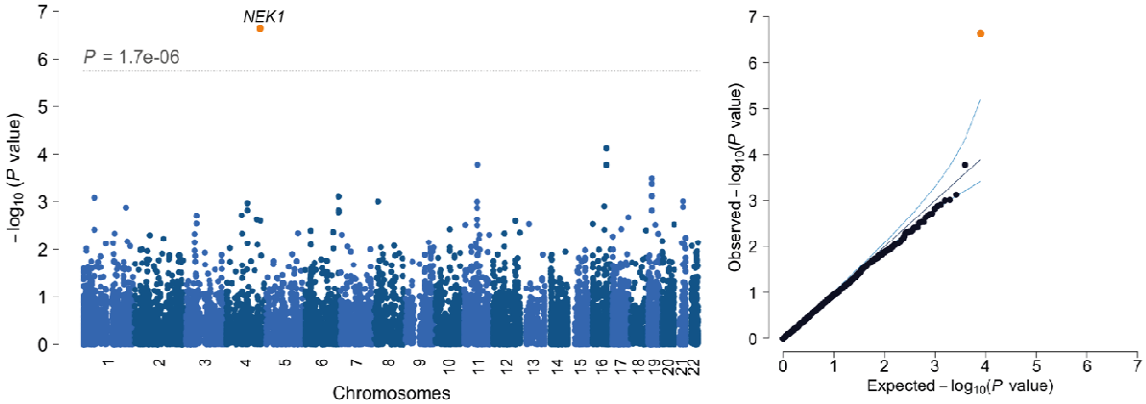

Figure 2. Manhattan and QQ-plot Results are shown for genic (canonical transcripts only) firth logistic regression including variants with a MAF<1% and categorized as disruptive and damaging. $\lambda_{GC} = 0.907$, $\lambda_{1000} = 0.964$.

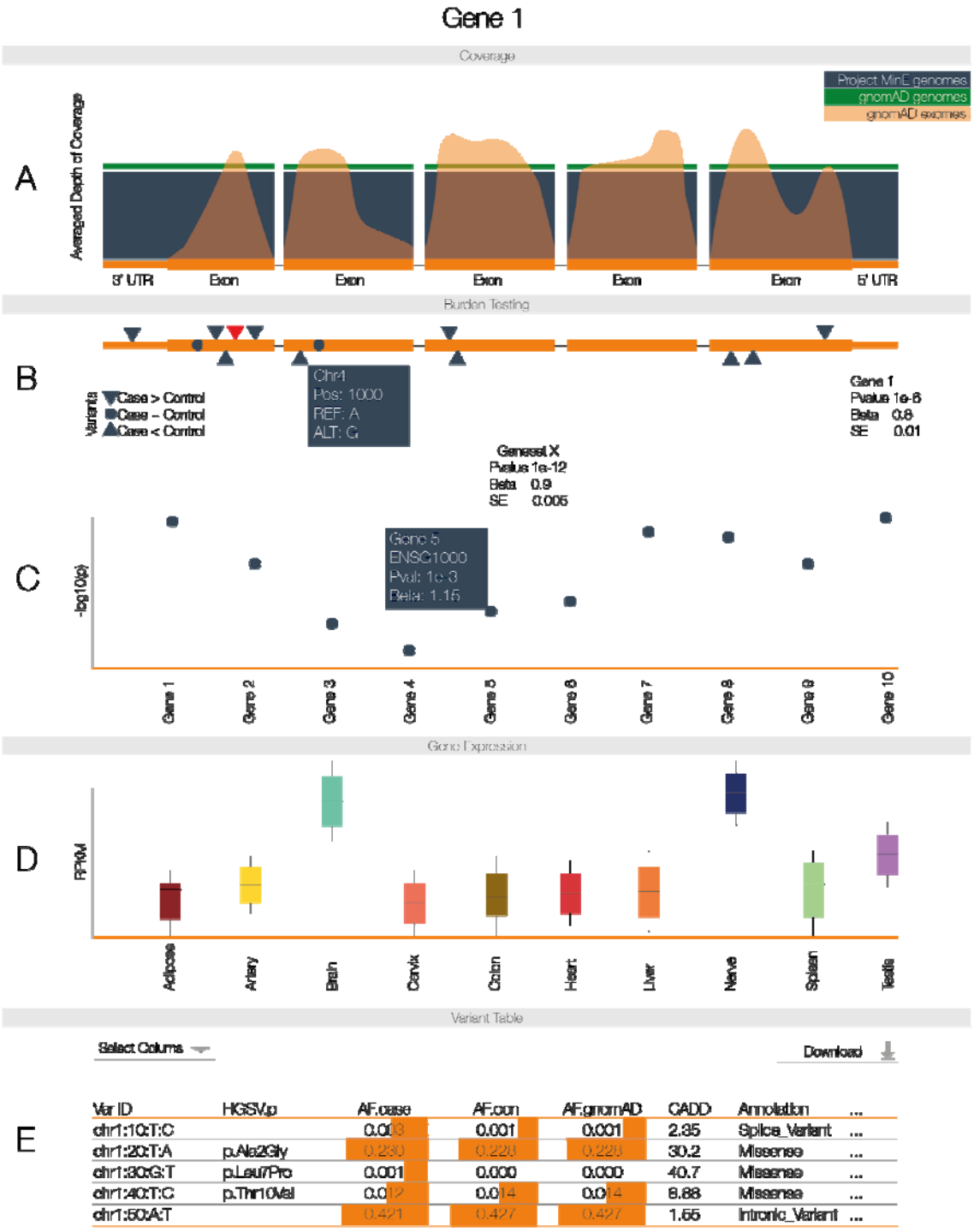

Figure 3. Databrowser After entering the gene name (HGNC, Ensembl gene (ENSG) or transcript (ENST) identifier) in the search box on the homepage, you will be directed to the gene-specific page. **A** Averaged depth of coverage in the Project MinE dataset, compared to public data and indicating quality of coverage in the region. **B** Firth logistic regression-based genic burden tests. Triangles indicate variant locations. Red triangles reach nominal significance in the single variants association test. Hovering over the triangles to obtain more information about that variant. **C** Firth logistic regression-based geneset burden test. Tests are based on pathways, gene families or druggable gene categories. To elucidate the gene or genes genes driving a signal in the geneset, a Manhattan plot indicates the genic burden results for each of the genes included in the geneset. Hovering over individual genes will reveal more information about that gene. **D** Gene expression profiles extracted

from GTEx. **E** Variant table. By default, a subset of variant information is shown; columns of interest can be selected from the dropdown menu. Minor allele frequency is based on all unrelated and QC passing samples in the Project MinE dataset (6,198 genomes). Frequency information is also stratified by phenotypic status and compared to public exome and whole genome data. For comparison, we have indicated the allele frequency on a log scale with orange bars; the longer the bar, the higher the allele frequency. Variant filtering can be customized using the search boxes below the header of each column. All data, including case/control frequencies, are available for download in a tab-delimited file. For a more detailed view of the databrowser, see Figure S1.

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jan Veldink (j.h.veldink@umcutrecht.nl).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample selection and WGS

The first batch of samples (1,935 cases and controls collected in the Netherlands) were sequenced on the Illumina HiSeq 2000 platform (Project MinE ALS Sequencing Consortium, 2018a). All remaining samples (4,644 cases and controls) were sequenced on the Illumina HiSeq X platform. All samples were sequenced to ~35X coverage with 100bp reads for the HiSeq 2000 and ~25X coverage with 150bp reads for the HiSeq X. Both sequencing sets used PCR-free library preparation. Samples were also genotyped on the Illumina 2.5M array. Sequencing data was then aligned to GRCh37 using the iSAAC Aligner, and variants called using the iSAAC variant caller; both the aligner and caller are standard to Illumina's aligning and calling pipeline.

Informed consent

All participants gave written informed consent and the relevant institutional review boards approved this study. The informed consent clearly indicates that there is no duty to hunt for clinically actionable results and that participants will not be re-contacted for genotyping results.

METHOD DETAILS

Data merging and initial site filtering

Per individual, WGS data was stored in both BAM and Illumina gVCF format. These gVCFs contain single-nucleotide variants (SNVs), short insertions and deletions (indels), and large structural variants (SVs). To begin quality control, we merged all sample into a single file using the Illumina 'agg' tool v0.3.4 (https://github.com/Illumina/agg). 'Agg' first generates metadata across all samples, typically batched into groups (e.g., n = 50) to minimize CPU time, and then proceeds to extract genotypes for all samples across all sites containing at least one non-reference allele in the full case-control dataset. This process results in a VCF of all samples and all variants with minor allele count ≥ 1.

The resulting merged VCF contained all possible sites, regardless of whether or not they passed the Isaac pipeline set of variant filters. We therefore applied basic site filtering to the initial merged VCF. Specifically, we set sites with a genotype quality (GQ) < 10 to missing using bcftools and single-nucleotide variants (SNVs) and indels with quality (QUAL) scores < 20 and < 30, respectively, were removed. We then removed variants with missingness > 10% (typically induced by setting genotypes with GQ < 10 to missing). To ensure unique marker identifiers at all sites, particularly for those multiallelic sites that were devolved by processing the data through 'agg,' we labeled all variants using the following nomenclature: chromosome:position:reference_allele:alternate_allele.

Sample-level quality control

We fixed genotype ploidy on the sex by first inferring biological sex from the available SNP array data, and then using the 'fix-ploidy' option in bcftools.

We then performed sample-level quality control (QC). We calculated the transition-transversion ratio in each sample using SnpSift 4.3p. In WGS data, the expected transition-transversion ratio is ~2.0; a number much lower than this (i.e. approaching 0.5, in accordance with the expected number of transitions and transversions if genotypes were called randomly) indicates an enrichment for false-positive genotype calls. We removed two samples with a Ti/Tv ratio > 6 standard deviations (sd) from the full distribution of samples.

Per sample, we also calculated (a) the total number of SNVs, (b) total number of indels, and (c) total number of singletons . We removed samples with a total number of SNPs > 6 sd from the mean. The shift in sequencing platforms from HiSeq 2000 to HiSeq X (which occurred in parallel with a change in the calling pipeline, to improve indel detection) caused an obvious shift in observed indels per sample. Samples were thus filtered by platform (HiSeq 2000 or HiSeq X) and removed samples with number of indels > 6 sd from the mean of their respective group. Finally, we identified samples with an excess number of singletons (calculated by cohort, to avoid overly-stringent filtering due to population stratification); samples with a total number of singletons > 6 sd from the sample distribution were removed.

Next, we calculated sample-level missingness and removed samples with > 5% missingness. We calculated average sample depth and again observed noticeable differences between those samples sequenced on the HiSeq 2000 and the HiSeq X, where average depth of coverage was somewhat higher (35X, on average) for samples sequenced on HiSeq 2000 compared to the samples sequenced on the HiSeq X (25X, on average). We removed no samples at this step. We then subsetted the sequence data down to those markers that overlapped with the 2.5M array genotyping data. Across the intersect of markers, we calculated the

sample concordance between the sequence and array data, and removed all samples with concordance < 96%.

Using X chromosome variants, we tested to see if biological sex (inferred from the X chromosome data) was concordant with the sex as annotated in the available phenotype information. We excluded 62 (of 6,579) samples with mismatching information.

We performed the remaining sample QC on a high-quality set of ~100,000 autosomal variants with: minor allele frequency (MAF) > 10%; genotype missingness < 0.1%; residing outside four complex regions (the major histocompatibility complex (MHC) on chromosome 6; the lactase locus (*LCT*), on chromosome 2; and inversions on chromosomes 8 and 17); excluding A/T and C/G variants. We used this set of markers to calculate inbreeding in two ways: first, by calculating inbreeding coefficients using Plink 1.9 (--het); and secondly, by calculating the ratio of heterozygous to homozygous non-reference genotypes per sample. In the first instance, we removed samples > 6 sd from the full sample distribution. In the second instance, we filtered samples for inbreeding coefficients on a cohort-by-cohort basis and excluded individuals > 6 sd from the cohort distribution.

We estimated kinship coefficients (i.e., relatedness) using the KING method, as implemented in the SNPRelate package in R. As samples were ascertained from a number of countries, we used the KING method, as it calculates kinship in the presence of potential population stratification (a potential confounder in other identity-by-descent approaches, such as that implemented in Plink). In some instances, research groups had intentionally ascertained related samples. We identified all pairs of related individuals (kinship > 0.0625). Of these, several pairs included one sample appearing to be related to several other samples in the data

(likely due to sample contamination). Samples related to > 100 other samples in the data were dropped; true related pairs were left in the data. For burden testing, we excluded these related samples (kinship coefficient > 0.0625). For single variant association analysis, we used a linear mixed model in GCTA, including a genetic relationship matrix and the first 20 principal components (PCs), alleviating the need to exclude related samples.

Lastly, we used principal component analysis (PCA) implemented in EIGENSTRAT to visualize potential structure in the data, induced by population stratification or other variables. Projection onto the HapMap 3 populations indicated that the samples were primarily of European ancestry, though some were of African or East Asian ancestry, while other samples appeared to be admixed. PCA across the dataset alone revealed structure induced not only by population but also by sequencing platform/calling algorithm (e.g., principal component 2). However, because of a balanced case-control ratio in both batches, we observed a very small effect of platform/calling algorithm when including it as a covariate in association testing. A summary of all sample QC, including thresholds and removed samples, is provided in Supplementary Table 2.

Variant-level quality control

To clean variants, we first inferred a set of QC thresholds from the set of SNVs falling on chromosomes 1-22 and then extrapolated these thresholds to filter all variants, including indels and variants on the sex and mitochondrial chromosomes.

We calculated Hardy-Weinberg equilibrium (HWE) in controls only, on a cohort-specific basis (to avoid potential population confounding) and removed all variants

with HWE $p < 1 \times 10^{-6}$. We calculated differential missingness between cases and controls and removed any variants with $p < 1 \times 10^{-6}$.

Next, we binned the variants by a number of metrics: depth of coverage, minor allele frequency, missingness, quality (QUAL) score, and passing rate. The last metric, passing rate, indicates the proportion of samples for which the variant was annotated as 'PASS'ing variant filters in the original, per-sample gVCF data. For example, a passing rate of 70% indicates that a variant is annotated as 'passing' the Isaac thresholds in 70% of all samples.

Once we had stratified the variants by these metrics, we calculated (for each bin) the transition/transversion (Ti/Tv) ratio and the ratio of heterozygous to homozygous non-reference genotypes (het/hom-non-ref) and then plotted the bins according to these metrics. From these visualizations of the data, we could infer the following QC thresholds and remove: variants with total depth < 10,000 reads (i.e., ~1.53X per sample) or > 226,000 reads (i.e., ~34.8X per sample), variants with missingness > 5%, and variants with a passing rate < 70%. We did not filter variants on minor allele frequency or QUAL score.

Scripts used for performing data merging and sample- and variant-level quality control are available through the Project MinE BitBucket (https://bitbucket.org/ProjectMinE/databrowser). Scripts include calls to PLINK, bcftools, SnpSift, EIGENSTRAT, and SNPRelate, as well as the relevant command-line options used for the QC steps described here.

Data integration and annotation

After quality control, we performed functional annotation of all variants using snpEff V4.3T and SnpSift using the GRCh37.75 database (including Nextprot and Motif),

dbSNFP v2.9, dbSNP b150 GRCh37p13 and ClinVar GRCh37 v2.0(Cingolani et al., 2012; Gaudet et al., 2017; Landrum et al., 2016; Liu et al., 2016; Ruden et al., 2012; Sherry et al., 2001). We obtained population frequency estimates from gnomAD (Lek et al., 2016). To visualize the variant-level coverage from Project MinE and external sources, we included coverage information from Project MinE samples, gnomAD database (123,136 exome sequences plus 15,496 genome sequences). We further integrated tissue-specific gene expression profiles for 53 tissues from the GTEx resource (https://gtexportal.org/home/datasets) (GTEx Consortium, 2015). Finally, the available literature on each gene is presented through an iframe linking to either PubMed, UCSC, GeneCards, Ensembl, WikiGenes, GTEx or the GWAScatalog.

QUANTIFICATION AND STATISTICAL ANALYSIS

Association analyses

The main association analysis consists of several rare-variant burden analyses for an association with ALS risk. For quality control we have performed single variant association analysis using a mixed linear model, including a genetic relationship matrix and the first 20 PCs, as implemented in GCTA (Yang et al., 2011). We set genome-wide significance in single variant association analyses at $p < 5 \times 10^{-9}$ (Pulit et al., 2017), to account for the increased number of independent SNVs tested in sequence data.

We performed rare-variant burden tests using firth logistic regression in R, adjusting for the first 10 PCs, sex and platform (Firth, 1993; Heinze and Ploner, 2004). Variants for the rare-variant burden tests have been aggregated on multiple levels; gene, protein superfamilies, pathways, druggable categories and exome-wide. Genic regions were defined as all transcripts in the GRCh37.p13 version of

Ensembl Biomart (Smedley et al., 2015). Higher level aggregation for burden analysis was performed by creating genesets. These genesets are based on: (a) protein superfamilies (Wilson et al., 2007); (b) drugable categories as defined by the drug-gene interaction database (Cotto et al., 2017); and (c) pathways downloaded from GSEA, using curated genesets v6.1 from KEGG, BioCarta or Reactome (Mootha et al., 2003; Subramanian et al., 2005).

We tested genes or genesets when we could identify ≥ 5 individuals with ≥ 1 variant. We used three definitions for 'rare': minor allele frequency cutoffs 1% and 0.5%. and variants not observed in ExAC (Lek et al., 2016). We classified variants based on their functional annotation (disruptive, damaging, missense-non-damaging, and synonymous , and described previously (Genovese et al., 2016)). Briefly, frame-shift, splice site, exon loss, stop gained, stoploss, startloss and transcription ablation variants were regarded as disruptive variants. We defined damaging variants as missense variants (resulting in an amino-acid change) predicted as damaging by *all* of seven methods: SIFT, Polyphen-2, LRT, Mutation Taster, Mutations Assessor, and PROVEAN (Genovese et al., 2016). Missense-non-damaging variants are missense variants that are not classified as damaging. Synonymous variants do not result in an amino-acid change. From these annotations, we created three variant sets for burden testing: (1) disruptive variants, (2) disruptive + damaging variants, (3) disruptive + damaging + missense-non-damaging variants. The synonymous category functions as a null category to check for biases when testing for association. We set the threshold for exome-wide significance in genic rare-variant burden analyses at $p < 1.7 \times 10^{-6}$. We acknowledge that this threshold does not fully

account for the multiple testing burden introduced by the different variant sets, allele-frequency cut-offs, and various burden testing approaches.

DATA AND SOFTWARE AVAILABILITY

Existing ALS datasets

The browser also includes freely-available summary-level data for the 2016 ALS GWAS (van Rheenen et al., 2016) for download . Additionally, downloadable SKAT and SKAT-O burden testing results from 610 ALS cases and 460 controls with Chinese ancestry (Gratten et al, 2017) are available.

Language

The data browser can be accessed at http://databrowser.projectmine.com/. The interface is based on the statistical programming language R (v3.4.1, https://www.r-project.org/) together with the interactive web application framework Shiny (v1.0.5, https://shiny.rstudio.com/). Interactive visualisations have been created using base R and the Plotly library (v4.7.1,https://plot.ly/r/). The code is open-source and can be downloaded from https://bitbucket.org/ProjectMinE/databrowser.

SUPPLEMENTAL INFORMATION

Country	N
Belgium	568
Ireland	408
Netherlands	2,995
Spain	359
Turkey	224
United Kingdom	1,468
United States	557
Total	6,579

Table S1. Total number of samples before quality control, stratified by country.

	Belgium	Ireland	Netherlands	Spain	Turkey	UK	USA	Total
Total SNPs	0	1	17	2	0	6	11	37
Total indels	1	1	12	2	0	6	10	32
Total singletons	1	1	2	1	0	1	1	35
Ti/Tv ratio	1	0	1	0	0	0	0	2
Missingness	2	0	0	0	0	0	0	2
Geno-Seq concordance	0	0	4	0	0	0	0	4
Depth of cov	0	0	0	0	0	0	0	0
Relatedness	0	1	23	1	1	15	2	43
Het/hom-non-ref ratio	0	1	15	1	1	7	1	26
Inbreeding	3	1	17	1	1	7	1	31
Sex check	13	1	17	15	8	7	1	62
PCA	0	0	0	0	0	0	0	0
Total failures including sex check	18	3	49	16	9	25	13	133
Total failures excluding sex check	6	2	41	2	1	20	12	84

Table S2. Quality control. Summary of the number of samples that have failed quality control steps. Please note that some samples fail multiple steps; therefore, the sum of a column will not be equal to the number of samples that actually failed.

Ensembl Transcript ID	variants	Controls*	Cases*	P	Beta	SE
ENST00000439080	19	0	28	3.4×10^{-5}	3.53	1.64
ENST00000240185[C]	21	6	37	0.017	0.93	0.44
ENST00000315091	5	6	12	0.71	-0.19	0.50
ENST00000473118	5	6	12	0.71	-0.19	0.50
ENST00000473869	5	6	12	0.71	-0.19	0.50
ENST00000476201	5	6	12	0.71	-0.19	0.50
ENST00000472476	3	0	3	-	-	-
ENST00000477447	0	-	-	-	-	-
ENST00000496840	0	-	-	-	-	-

Table S3. **Transcript-specific burden for *TARDBP*.** All non-synonymous variants with a MAF < 0.5% were included. Firth logistic regression was performed with the first 10PCs, sex and platform were included as covariates. Transcript ENST00000439080 clearly shows an increased burden as compared to the other transcripts. **Figure S2** shows the burdenplot of the first 3 transcripts. [C] Canonical transcript. * Number of controls or cases carrying at least 1 variant.

Figure S1. Screenshot of the databrowser

Figure S2. Transcript specific burden in *TARDBP*. All non-synonymous variants (indicated by triangles and circles) with a MAF < 0.5% were included. This overview clearly shows an increased burden in ENSE00001766796 / ENSE00001145815 while the effect might be diluted by exon ENSE00003543832.