

Supplementary Information for

Characterising the motif composition and allele length distribution of *ZFHX3* GGC repeat expansions in amyotrophic lateral sclerosis

Zoe N. Zussa, Andrew N. Smith, Joke van Vugt, Daniel S. O'Shaughnessy, Natalie Grima, Sandrine Chan Moi Fat, Ian P. Blair, Dominic B Rowe, Roger Pamphlett, Garth A. Nicholson, Matthew C Kiernan, Wouter van Rheenen, Jan Veldink, Project MinE ALS Sequencing Consortium, Kelly L. Williams⁺ and Lyndal Henden⁺*

*Corresponding Author. Email: lyndal.henden@mq.edu.au

⁺Contributed equally to this work

This PDF file includes:

Supplementary Methods

Supplementary Figures 1–3

Supplementary Tables 1–6

Members of Project Mine Consortium

Supplementary References

Supplementary Methods

Whole Genome Sequencing Data Processing

Australian ALS cohort

Genomic DNA was extracted using whole blood according to standard protocols. Samples underwent library preparation using the TruSeq PCR free library prep kit (Illumina V2.5). Prepped libraries underwent multiplex 150-bp paired-end whole genome sequencing (WGS) on the Illumina HiSeq X Ten instrument (Kinghorn Centre for Clinical Genomics, Sydney, Australia and Macrogen Oceania). Sequencing data was aligned to the GRCh38 reference genome using BWA-MEM¹. Variant calling and joint genotyping were generated using GATKs Haplotypecaller and genotypeGVCF pipeline².

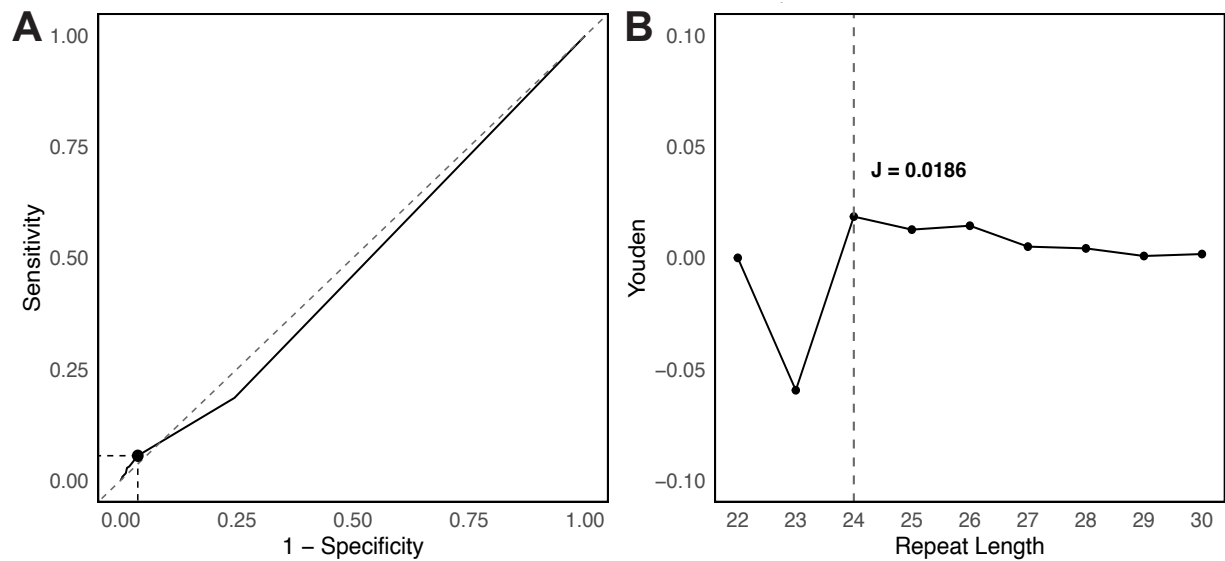
Project MinE Data Freeze 2

Sequencing and data processing have been described previously³⁻⁵. Briefly, case and control samples were sequenced using PCR-free library preparation on the Illumina HiSeq 2000 and HiSeq X platforms to ~35-fold coverage with 100 bp reads and ~25-fold coverage with 150 bp reads, respectively. Sequencing data was aligned to the GRCh38 reference genome and variant calling were performed using the Illumina Isaac pipeline⁶.

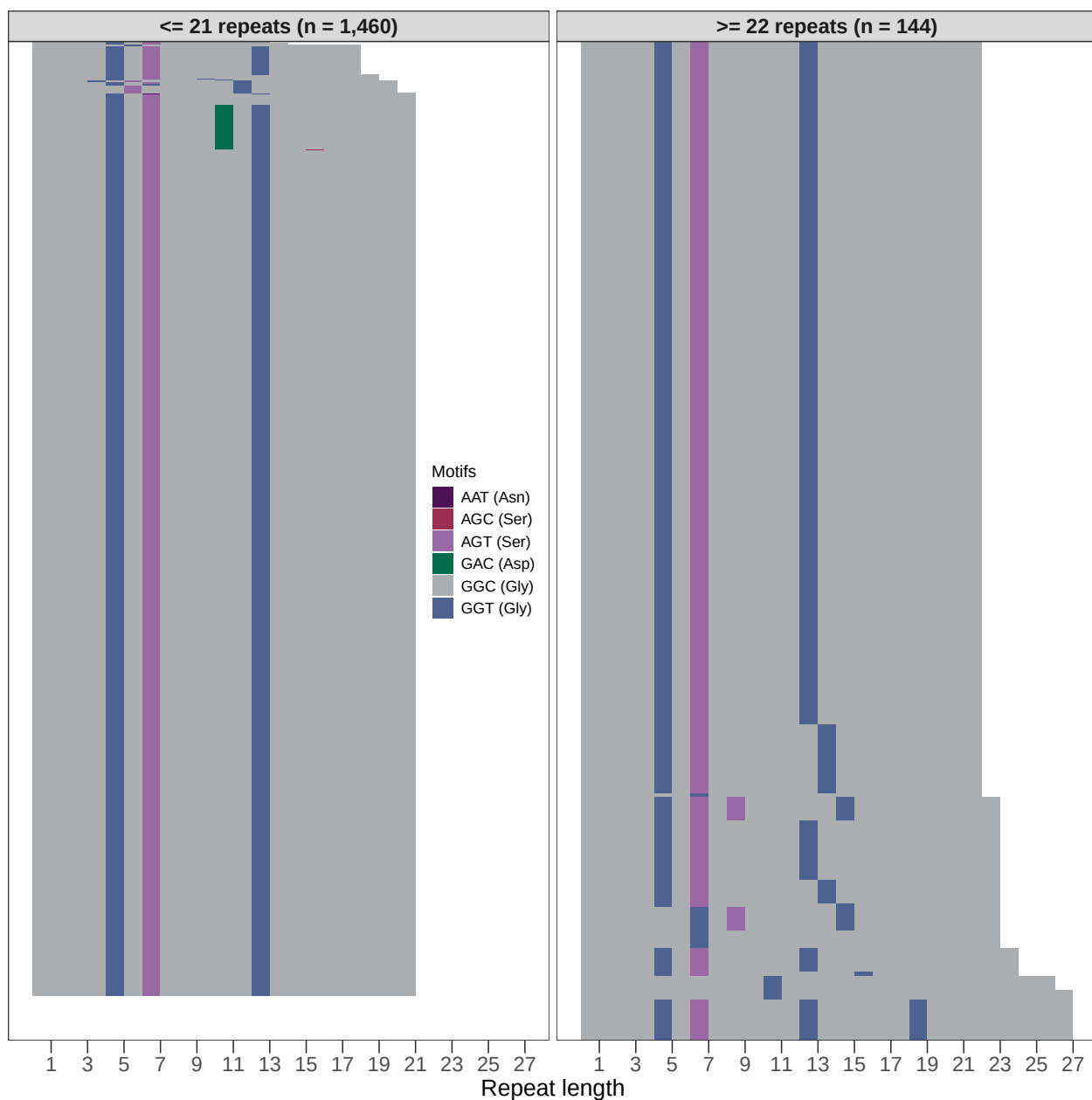
Variant Catalogue for ZFH3 GGC Expansion Detection using Expansion-Hunter v5

```
{  "LocusId": "ZFHX3",
  "LocusStructure": "(GCC)*",
  "VariantType": "Repeat",
  "ReferenceRegion": "chr16:72787694-72787757",
  "RepeatUnit": "GCC",
  "Gene": "ZFHX3",
  "GeneRegion": "coding:polyglycine",
  "GeneId": "ENST00000268489",
  "DiscoveryMethod": "L,WGS",
  "DiscoveryYear": 2023,
  "Diseases": [
    {
      "Symbol": "SCA4",
      "Name": "Spinocerebellar ataxia 4",
      "Inheritance": "AD",
      "OMIM": "600223",
      "NormalMax": 26,
      "PathogenicMin": 41
    }
  ],
  "MainReferenceRegion": "chr16:72787694-72787757",
  "Inheritance": "AD"
}
```

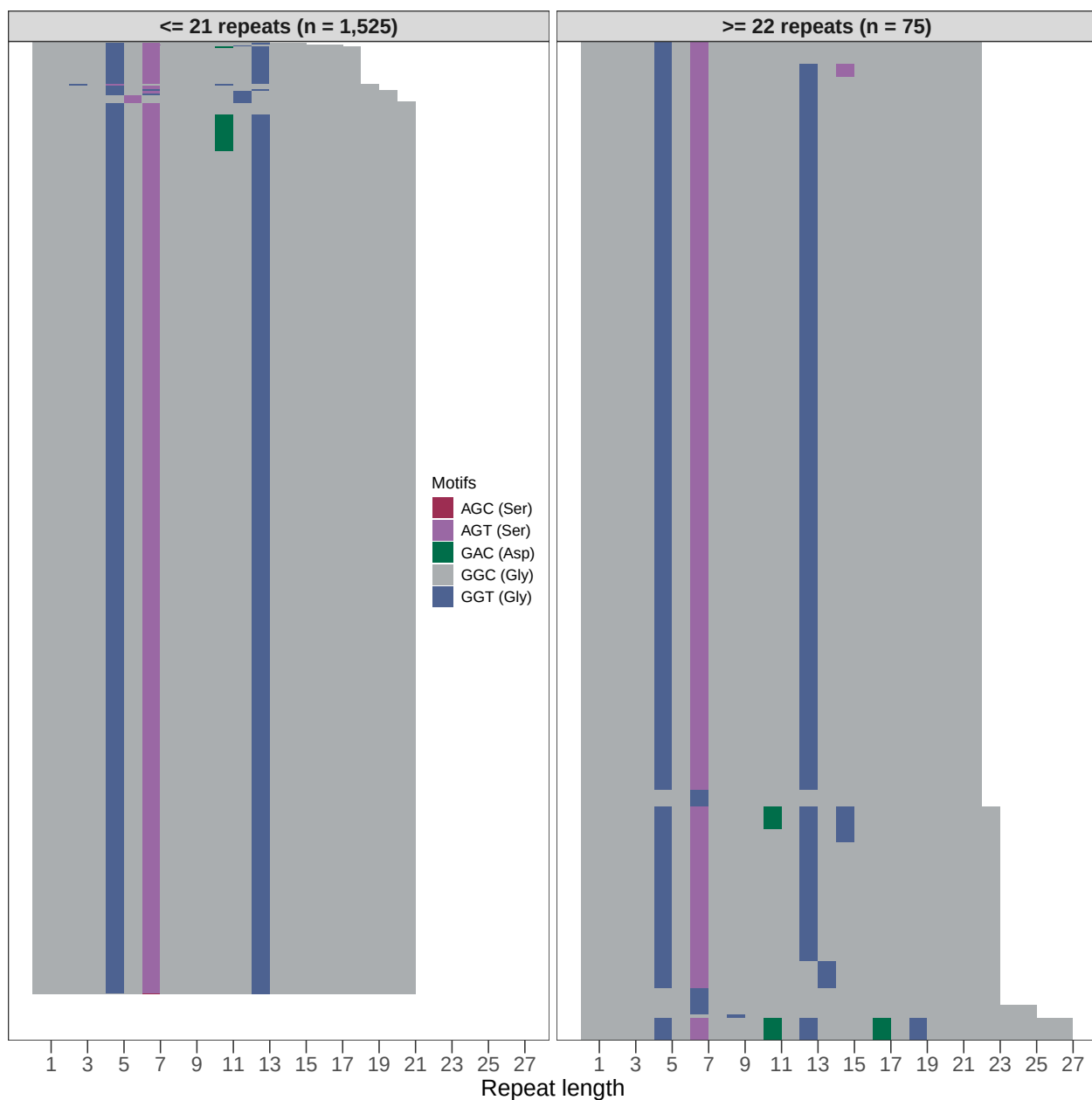
Supplementary Figures 1–3



Supplementary Figure 1: ROC analysis and Youden's J statistic of *ZFH3* repeat expansions (22-30). A) ROC analysis of *ZFH3* GGC repeat expansions (22-30). Sensitivity is the true positive rate and specificity is the true negative rate. The optimal sensitivity and specificity is indicated by a black dot (n=24 repeats). B) Youden's J statistic calculated for each repeat length (22-27). The J statistic assess the sensitivity and specificity from the ROC analysis to find the optimal threshold that can discriminate cases from controls. The maximum Youden's J statistic was observed at a *ZFH3* repeat length of 24 ($J = 0.0186$), indicating the optimal threshold.



Supplementary Figure 2: Motif composition of *ZFH3* repeat alleles in the Australian ALS cohort ($n = 802$). Motif composition of the long and short alleles from 802 Australian ALS patients. Alleles are ordered by increasing repeat length. The left panel shows alleles with 21 repeats or fewer. The right panel shows alleles with 22 repeats or more. Colours denote distinct repeat motifs, with grey indicating the canonical GGC repeat motif. The y-axis has been scaled according to the number of alleles with repeats either ≤ 21 ($n = 1,460$) or ≥ 21 ($n = 144$).



Supplementary Figure 3: Motif composition of *ZFX3* repeat alleles in the project MinE control cohort ($n = 800$). Motif composition of the long and short alleles from 800 healthy controls sourced from project MinE. Alleles are ordered by increasing repeat length. The left panel shows alleles with 21 repeats or fewer. The right panel shows alleles with 22 repeats or more. Colours denote distinct repeat motifs, with grey indicating the canonical GGC repeat motif. The y-axis has been scaled according to the number of alleles with repeats either ≤ 21 ($n = 1,525$) or ≥ 21 ($n = 75$).

Supplementary Tables 1–6

Supplementary Table 1 Summary of participants from Project MinE DF2 and contributing countries.

Country	ALS	Controls
Belgium	552	183
France	246	39
Ireland	462	232
Israel	100	-
Italy	55	-
Netherlands	586	393
Portugal	45	16
Spain	370	165
Sweden	241	112
Switzerland	44	-
Turkey	608	141
United Kingdom	1,409	443
United States of America	450	77

Supplementary Table 2 ALS-causal or -associated genetic variants present in ALS cases in this cohort⁷

Gene	Accession number	cDNA change	Protein change	sALS		fALS	
				Male	Female	Male	Female
<i>C9orf72</i>	NM_001256054	GGGGCC ^{RE}		197	155	-	1
<i>TARDBP</i>	NM_007375	c.146A>G	p.Gln49Arg	-	1	-	-
<i>TARDBP</i>	NM_007375	c.269C>T	p.Ala90Val	2	1	-	-
<i>TARDBP</i>	NM_007375	c.859G>A	p.Gly287Ser	2	1	-	-
<i>TARDBP</i>	NM_007375	c.928G>A	p.Gly310Arg	-	1	-	-
<i>TARDBP</i>	NM_007375	c.1028A>G	p.Gln343Arg	1	-	-	-
<i>TARDBP</i>	NM_007375	c.1055A>G	p.Asn352Ser	3	1	-	-
<i>TARDBP</i>	NM_007375	c.1144G>A	p.Ala382Thr	2	-	-	-
<i>TARDBP</i>	NM_007375	c.1169A>G	p.Asn390Ser	-	1	-	-
<i>FUS</i>	NM_004960	c.14A>G	p.Asp5Gly	-	1	-	-
<i>FUS</i>	NM_004960	c.188A>G	p.Asn63Ser	2	3	-	-
<i>FUS</i>	NM_004960	c.730C>T	p.Arg244Cys	-	1	-	-
<i>FUS</i>	NM_004960	c.731G>A	p.Arg244His	-	1	-	-
<i>FUS</i>	NM_004960	c.820C>T	p.Arg274Cys	1	-	-	-
<i>FUS</i>	NM_004960	c.1292C>T	p.Pro431Leu	2	1	-	-
<i>FUS</i>	NM_004960	c.1375C>G	p.Pro459Ala	1	-	-	-
<i>FUS</i>	NM_004960	c.1390A>G	p.Met464Val	1	-	-	-
<i>FUS</i>	NM_004960	c.1520G>A	p.Gly507Asp	1	-	-	-
<i>FUS</i>	NM_004960	c.1561C>T	p.Arg521Cys	-	2	-	1
<i>FUS</i>	NM_004960	c.1574C>T	p.Pro525Leu	1	2	-	-
<i>SOD1</i>	NM_000454	c.14C>T	p.Ala5Val	-	2	-	-
<i>SOD1</i>	NM_000454	c.25C>G	p.Leu9Val	1	-	-	-
<i>SOD1</i>	NM_000454	c.44T>G	p.Val15Gly	1	-	-	-
<i>SOD1</i>	NM_000454	c.59A>G	p.Asn20Ser	2	-	-	-
<i>SOD1</i>	NM_000454	c.65A>G	p.Glu22Gly	-	2	-	-
<i>SOD1</i>	NM_000454	c.68A>T	p.Gln23Leu	1	2	-	-
<i>SOD1</i>	NM_000454	c.112G>A	p.Gly38Arg	2	-	-	-
<i>SOD1</i>	NM_000454	c.123A>C	p.Glu41Asp	-	1	-	-
<i>SOD1</i>	NM_000454	c.197A>G	p.Asn66Ser	-	1	-	-
<i>SOD1</i>	NM_000454	c.260A>G	p.Asn87Ser	-	1	-	-
<i>SOD1</i>	NM_000454	c.280G>T	p.Gly94Cys	2	2	-	-
<i>SOD1</i>	NM_000454	c.281G>C	p.Gly94Ala	1	-	-	-
<i>SOD1</i>	NM_000454	c.302A>G	p.Glu101Gly	-	-	1	1
<i>SOD1</i>	NM_000454	c.317C>T	p.Ser106Leu	1	-	-	-
<i>SOD1</i>	NM_000454	c.341T>C	p.Ile114Thr	4	6	9	8
<i>SOD1</i>	NM_000454	c.358G>C	p.Val120Leu	1	1	-	-
<i>SOD1</i>	NM_000454	c.403A>G	p.Ser135Gly	1	-	-	-
<i>SOD1</i>	NM_000454	c.442G>A	p.Gly148Ser	1	-	-	-
<i>SOD1</i>	NM_000454	c.446T>G	p.Val149Gly	-	-	2	-
<i>SOD1</i>	NM_000454	c.457G>A	p.Ala153Thr	1	-	-	-

Supplementary Table 3 Reported thresholds for *ZFHX3* GGC repeat expansions in literature.

Resource	Normal	Intermediate	Pathogenic
gnomAD Tandem Repeats: <i>ZFHX3</i> ⁸	≤ 26	-	≥ 41
STRchive: SCA4 <i>ZFHX3</i> ⁹	14–26	27–45	46–74
STRipy: SCA4 <i>ZFHX3</i> ¹⁰	14–31	-	≥ 42
Wallenius et al., 2024 ¹¹	14–26	-	42–74
Figuerola et al., 2024 ¹²	≤21	-	42–74
Paucar et al., 2024 ¹³	≤26	-	46–64

Supplementary Table 4 Distribution of *ZFHX3* GGC allele lengths in ALS cases and controls, including overall allele counts and allele counts stratified by sex.

<i>ZFHX3</i> repeat length	Total		Males		Females	
	ALS	Control	ALS	Control	ALS	Control
4	1	0	1	0	0	0
5	0	2	0	1	0	1
6	1	2	0	1	1	1
7	1	0	0	0	1	0
8	1	1	1	0	0	1
9	1	1	1	1	0	0
13	13	15	5	9	8	6
14	17	15	9	12	8	3
15	16	16	13	9	3	7
16	7	22	3	15	4	7
17	7	10	5	6	2	4
18	323	458	204	260	117	198
19	50	91	24	52	26	39
20	195	263	124	151	71	112
21	10352	14295	6255	8322	4085	5973
22	476	584	290	343	186	241
23	76	161	49	89	27	72
24	15	14	12	6	3	8
25	2	4	1	0	1	4
26	7	2	3	2	4	0
27	5	6	2	6	3	0
28	2	0	2	0	0	0
29	1	2	1	1	0	1
30	1	0	1	0	0	0
Total	11570	15964	7006	9286	4550	6678

Supplementary Table 5 Distribution of longest ZFH3 allele lengths in ALS cases and controls, including overall counts and counts stratified by sex.

ZFHX3 long allele repeat length	Total		Males		Females	
	ALS	Control	ALS	Control	ALS	Control
15	1	0	0	0	1	0
18	4	15	2	8	2	7
19	4	2	3	1	1	1
20	12	17	8	6	4	11
21	5202	7195	3139	4194	2056	3001
22	457	565	282	331	175	234
23	72	160	47	88	25	72
24	15	14	12	6	3	8
25	2	4	1	0	1	4
26	7	2	3	2	4	0
27	5	6	2	6	3	0
28	2	0	2	0	0	0
29	1	2	1	1	0	1
30	1	0	1	0	0	0
Total	5785	7982	3503	4643	2275	3339

Supplementary Table 6 Association of ZFH3 GGC repeat expansion status with ALS and clinical phenotypes after excluding carriers of pathogenic ALS variants.

Outcome	Statistical model	Effect measure for ZFH3 ≥ 24 repeats	Effect estimate	95% CI	P-value
ALS case-control status	Logistic regression	OR	1.62	0.97–2.71	0.064
Age at disease onset	Linear regression	β , years	4.71	0.24–9.18	0.039
Site of disease onset	Logistic regression	OR	1.1	0.48–2.34	0.814
Disease duration	Cox proportional hazards regression	HR	0.85	0.56–1.30	0.46

Note: Abbreviations: β , regression coefficient; CI, confidence interval; HR, hazard ratio; OR, odds ratio. Analyses excluded ALS cases carrying a pathogenic C9orf72 repeat expansion or pathogenic variants in SOD1, TARDBP or FUS. Individuals were classified as ZFH3 repeat carriers when their longest allele contained ≥ 24 GGC repeat units. All models were adjusted for sex. For the site-of-onset model, the OR represents the odds of bulbar rather than spinal onset, assuming spinal onset was the reference category.

Members of Project Mine ALS Sequencing Consortium

Philip van Damme¹, Philippe Corcia², Philippe Couratier³, Patrick Vourc'h⁴, Orla Hardiman⁵, Russell McLaughlin⁶, Marc Gotkine⁷, Yossef Lerner⁷, Yehuda Shovman⁷, Vivian Drory⁸, Nicola Ticozzi⁹, Vincenzo Silani⁹, Jan H. Veldink¹⁰, Leonard H. van den Berg¹⁰, Mamede de Carvalho¹¹, Teresa Salas¹², Jesus S. Mora Pardina¹³, Monica Povedano¹⁴, Peter Andersen¹⁵, Markus Weber¹⁶, Nazli A. Başak¹⁷, Ammar Al-Chalabi¹⁸, Chris Shaw¹⁸, Pamela J. Shaw¹⁹, Karen E. Morrison²⁰, John E. Landers²¹, Jonathan D. Glass²², Clifton L. Dalgard²³

¹ KU Leuven - University of Leuven, Department of Neurosciences

² Centre SLA, CHRU de Tours, Tours, France; UMR 1253, iBrain, Université de Tours, Inserm, Tours, France.

³ Centre SLA CHU Dupuytren Limoges France.

⁴ Service de Biochimie et Biologie moléculaire, CHU de Tours, Tours, France

⁵ Academic Unit of Neurology, Trinity College Dublin, Trinity Biomedical Sciences Institute, Dublin, Republic of Ireland.

⁶ Complex Trait Genomics Laboratory, Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Republic of Ireland.

⁷ Department of Neurology, Hadassah Medical Organization and Faculty of Medicine, Hebrew University of Jerusalem, Israel

⁸ Department of Neurology Tel-Aviv Sourasky Medical Centre, Israel.

⁹ Department of Neurology and Laboratory of Neuroscience, IRCCS Istituto Auxologico Italiano, Milano, Italy.

¹⁰ Department of Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.

¹¹ Instituto de Fisiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

¹² ALS Unit, Hospital Universitario La Paz-Carlos III, Madrid, Spain

¹³ ALS Unit, Hospital San Rafael, Madrid, Spain.

¹⁴ la Unitat Funcional de Motoneurona, Cap de Secció de Neurofisiologia, Servei de Neurologia, Hospital Universitario de Bellvitge-IDIBELL

¹⁵ Department of Clinical Science, Neurosciences, Umeå University, Sweden.

¹⁶ Neuromuscular Diseases Unit/ALS Clinic, HOCH, Health Ostschweiz, Kantonsspital St. Gallen, 9007, St. Gallen, Switzerland.

¹⁷ Koç University, School of Medicine, Molecular Biology and Genetics- KUTTAM, Suna and Inan Kiraç Foundation, Istanbul Turkey.

¹⁸ Maurice Wohl Clinical Neuroscience Institute, King's College London, Department of Basic and Clinical Neuroscience, London, UK.

¹⁹ Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK.

²⁰ School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, UK.

²¹ Department of Neurology, University of Massachusetts Chan Medical School, Worcester, MA, USA.

²² Department Neurology, Emory University School of Medicine, Atlanta, GA, USA.

²³ The American Genome Center, Uniformed Services University - "America's Medical School", Bethesda, MD, USA.

Supplementary References

- [1] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 5. Available from: <http://arxiv.org/abs/1303.3997>.
- [2] der Auwera GAV, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current protocols*. 2013;43:11.10.1-11.10.33.
- [3] Rheenen WV, Pulit SL, Dekker AM, Khleifat AA, Brands WJ, Iacoangeli A, et al. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics*. 2018 10;26:1537-46.
- [4] van Rheenen W, van der Spek RAA, Bakker MK, van Vugt JJFA, Hop PJ, Zwamborn RAJ, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nature Genetics*. 2021 12;53:1636-48.
- [5] Moisse M, Zwamborn RAJ, van Vugt J, van der Spek R, van Rheenen W, Kenna B, et al. The Effect of SMN Gene Dosage on ALS Risk and Disease Severity. *Annals of Neurology*. 2021 4;89:686-97.
- [6] Mehta PR, Iacoangeli A, Opie-Martin S, Vugt JJFAV, Khleifat AA, Bredin A, et al. The impact of age on genetic testing decisions in amyotrophic lateral sclerosis. *Brain*. 2022 12;145:4440-7.
- [7] McCann EP, Grima N, Fifita JA, Fat SCM, Lehnert K, Henden L, et al. Characterising the Genetic Landscape of Amyotrophic Lateral Sclerosis: A Catalogue and Assessment of Over 1,000 Published Genetic Variants. *Journal of Neuromuscular Diseases*. 2023;10(6):1127-41.
- [8] Genome Aggregation Database (gnomAD) Short Tandem Repeats: ZFHX3; 2026. Accessed: 12 February 2026. https://gnomad.broadinstitute.org/short-tandem-repeat/ZFHX3?dataset=gnomad_r4.
- [9] Hiatt L, Weisburd B, Dolzhenko E, Rubinetti V, Avvaru AK, VanNoy GE, et al. STRchive: a dynamic resource detailing population-level and locus-specific insights at tandem repeat disease loci. *Genome Medicine*. 2025;17:29.
- [10] Halman A, Dolzhenko E, Oshlack A. STRipy: a graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. *Human Mutation*. 2022;43(7):859-68.
- [11] Wallenius J, Kafantari E, Jhaveri E, Gorcenco S, Ameer A, Karremo C, et al. Exonic trinucleotide repeat expansions in ZFHX3 cause spinocerebellar ataxia type 4: A poly-glycine disease. *American Journal of Human Genetics*. 2024 1;111:82-95.
- [12] Figueroa KP, Gross C, Buena-Atienza E, Paul S, Gandelman M, Kakar N, et al. A GGC-repeat expansion in ZFHX3 encoding polyglycine causes spinocerebellar ataxia type 4 and impairs autophagy. *Nature Research*; 2024.
- [13] Paucar M, Nilsson D, Engvall M, Laffita-Mesa J, Söderhäll C, Skorpil M, et al. Spinocerebellar ataxia type 4 is caused by a GGC expansion in the ZFHX3 gene and is associated with prominent dysautonomia and motor neuron signs. *Journal of Internal Medicine*. 2024 9;296:234-48.