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

Zoe N. Zussa[1], Andrew N. Smith[1], Joke J.F.A van Vugt[2], Daniel S. O'Shaughnessy[1], Natalie Grima[1], Sandrine Chan Moi Fat[1], Ian P. Blair[1], Dominic B Rowe[1, 3], Roger Pamphlett [4], Garth A. Nicholson[5, 6], Matthew C.K Kiernan[7], Wouter van Rheenen[2], Jan Veldink[2], Project MinE ALS sequencing consortium, Kelly L. Williams[1*] and Lyndal Henden[1*+]

[1] Macquarie University Motor Neuron Disease Research Centre, Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, NSW, Australia

[2] Department of Neurology, UMC Utrecht Brain Center, Utrecht University, 3584 CX Utrecht, The Netherlands

[3] Macquarie University Health Neurology, Macquarie University, Sydney, NSW, Australia

[4] Department of Neuropathology, The University of Sydney, Sydney, NSW, Australia

[5] Molecular Medicine Laboratory, Concord Repatriation General Hospital, Sydney, NSW 2139, Australia

[6] Northcott Neuroscience Laboratory, ANZAC Research Institute, Sydney, NSW 2139, Australia

[7] Neuroscience Research Australia, and the University of NSW, Sydney, NSW, Australia

[*] Contributed equally to this work

[+] Corresponding author: lyndal.henden@mq.edu.au

Abstract

Background and objectives: A pathogenic GGC repeat expansion in the zinc finger homeobox 3 (*ZFHX3*) gene, encoding a pure polyglycine tract, is the cause of spinocerebellar ataxia type 4 (SCA4). Intermediate expansions of other SCA loci contribute to the risk of amyotrophic lateral sclerosis (ALS), a fatal neurodegenerative disease involving the progressive loss of motor neurons. There is increasing awareness of the role of short tandem repeat (STR) motif composition and configuration in disease pathogenicity. Given the genetic pleiotropy between ALS and SCA, this study aimed to evaluate whether *ZFHX3* GGC expansions were associated with ALS and to characterise repeat motif composition.

Methods: ExpansionHunter v5 was used to genotype *ZFHX3* GGC repeat sizes in short-read whole genome sequencing data from people with ALS and healthy controls of European ancestry. Repeat sizes were visually inspected using REViewer v2. Repeat motif configurations of Australian ALS cases were manually derived from REViewer images. Receiver operating characteristic (ROC) curve analysis and Youden's J statistic were performed to find a candidate repeat size threshold for association testing using Fisher's exact test.

Results: Analysis of 5,785 people with ALS and 7,982 healthy controls found no association between *ZFHX3* GGC repeat expansions and disease risk. However, more than 30 unique repeat motif compositions were identified across 802 people with ALS. Of these, seven distinct configurations coded a pure polyglycine tract which, when expanded, is canonical to SCA4.

Discussion: Although no association was observed between *ZFHX3* GGC repeat expansions and ALS, this study established the dynamic nature of *ZFHX3* repeat motif composition and configuration. Unique motif compositions were identified both within and between repeat sizes, including the presence of pure polyglycine repeats in ALS. Consideration of repeat motif composition and configuration, in addition to repeat allele length, may be important for assessing neurodegenerative disease risk.

Introduction

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterised by progressive degeneration of upper and lower motor neurons, resulting in muscle weakness, paralysis and death, typically due to respiratory failure within 3-5 years of symptom onset[1]. Gene mutations are the only proven cause of ALS, accounting for two-thirds of familial (inherited) ALS cases and ~20% of sporadic ALS cases[2]. ALS appears genetically heterogenous, with more than 30 genes and over 1,000 genetic variants implicated in disease risk or causality to date[3]. The most common genetic cause of ALS is a pathogenic short tandem repeat (STR) expansion in the intron of the *C9orf72* gene which accounts for disease in 40% of familial and 10% of sporadic cases[1]. The genetic landscape of ALS is further complicated by pleiotropy where the same mutation can give rise to different phenotypes and different disease manifestations. Notably the *C9orf72* repeat expansion is also the most common genetic cause of frontotemporal dementia (FTD); a neurodegenerative disease characterised by behavioural and cognitive decline[4]. Pleiotropy in ALS extends beyond *C9orf72*, with increasing genetic evidence linking ALS and spinocerebellar ataxia (SCA); a group of neurodegenerative disorders marked by progressive loss of balance, coordination and slurred speech[5]. Pathogenic CAG repeat expansions in *ATXN1* and *ATXN2* cause SCA1 and SCA2, respectively, while intermediate-length CAG expansions in these genes have been associated with increased ALS risk[6,7,8].

Recently, a coding pathogenic GGC repeat expansion in exon 10 of the zinc finger homeobox 3 gene (*ZFHX3*) was identified as the cause of SCA4 in families with Northern European ancestry, placing SCA4 among the growing group of polyglycine (polyG) expansion disorders[9,10,11]. *ZFHX3* encodes a transcription factor for neuronal and myogenic cell differentiation and cell proliferation[12]. Pure polyG repeat expansions in *ZFHX3* exceeding 41 GGC repeats lead to a pathogenic gain of gene function resulting in neuron degeneration, which is a hallmark pathological feature of both SCA and ALS[9]. Interestingly, individuals with SCA4 also show loss of motor neurons in the anterior horns and degradation of posterior tracts of the spinal cord, consistent with ALS pathology[10].

Investigations of STRs in polyG disorders and SCAs have emphasised the importance of motif composition (i.e. the exact sequence of nucleotides or amino acids in the expanded repeat) given that the presence and configuration of synonymous or missense

interrupting motifs can modify disease phenotype and clinical presentation[13,14,15]. Synonymous interruptions are nucleotide changes that alter repeat motif composition without changing the encoded amino acid and have been demonstrated to occur in polyG and polyglutamine (polyQ) disorders. For example, glycine GGC expansions in the *NOTCH2NLC* gene cause neuronal intranuclear inclusion disease (NIID). NIID individuals with the muscle weakness phenotype carried multiple glycine GGA synonymous interruptions, whereas NIID individuals with the parkinsonism-dominant phenotype carried fewer GGA motifs and more AGC (serine) missense interruptions[16]. Additionally, intermediate-sized glutamine CAG expansions in *ATXN2* that specifically confer risk for Parkinson's disease, Lewy body dementia and ALS, contain glutamine GAA synonymous interruptions that alter repeat motif composition without changing the encoded amino acid[17,18]. Interruptions in repeat sequences are primarily recognised for their protective role in maintaining genomic stability by preventing slippage during DNA replication and repair[19]. Beyond this canonical function, interruptions may also modulate variability in disease phenotype and clinical presentation, although the exact mechanism driving these effects remain unclear. This highlights the importance of assessing motif composition and configuration when examining association of STR expansions with disease. Given the phenotypic overlap and genetic pleiotropy between SCA and ALS and increasing awareness of the importance of motif compositions in STRs, we aimed to evaluate whether *ZFHX3* GGC repeat expansions are associated with ALS and characterise the motif composition and configuration in a large ALS cohort of European ancestry.

Methods

Standard protocol approvals, registrations and patient consents

This study was approved by the human research ethics committees of Macquarie University (ID: 10134) and by the institutional review boards of all participating centres within Project MinE. Written informed consent was provided from all participants for investigation and publication at their respective centres and additional approval was obtained from the Medical Ethical Testing Committee NedMec and the Biobanks Testing committee of UMC Utrecht. Patients were included irrespective of their carrier status for variants in known ALS-associated genes.

International ALS cohort and control participants

A total of 6,114 people with ALS and 8,030 healthy controls were initially included in this study. ALS was diagnosed by a neurologist according to the El Escorial criteria or based on progressive weakness in the upper and lower motor neurons, confirmed by electrophysiology of widespread muscle denervation and no other diagnosis being found on neuroimaging[20]. The ALS cohort comprised 943 Australian participants recruited from the Macquarie University Neurodegenerative Disease Biobank and the Australian MND DNA Bank. The remaining 5,171 individuals with ALS were obtained from Project MinE data freeze 2; an international ALS genomics consortium spanning 14 countries[21] summarised in Supplementary Table 1. All individuals with ALS were genotyped for the hexanucleotide repeat expansion in *C9orf72* and known pathogenic ALS variants in *SOD1, TARDBP,* and *FUS.* The control cohort comprised of 1,801 healthy individuals included in Project MinE data freeze 2 and 6,229 controls from the gnomAD European (non-Finnish) short tandem repeat data download v3.1.2.

To retain individuals of European ancestry for this study, ALS genotype call sets for all ALS cases and Project MinE controls were merged with the combined HapMap phase II and III datasets consisting of 11 populations[22]. Principal component analysis was performed using PLINK V1.9[23] to identify samples clustering with the HapMap CEU and TSI populations, representing European ancestry. Duplicate samples were detected and removed using PLINK v1.9. A single proband was selected from ALS families where multiple individuals were available for analysis.

ZFHX3 GGC repeat expansion detection, validation and motif composition

Whole genome sequencing (WGS) data were available for all ALS cases and Project MinE controls of European ancestry. Details on WGS sequencing and data processing are described in the eMethods. The *ZFHX3* GGC repeat expansion was genotyped from GRCh38 aligned BAMS using ExpansionHunter[24] v5 with the gnomAD variant catalogue excluding off target regions (json provided in eMethods). Individual-level ExpansionHunter outputs were merged into a cohort callset and summarised using TRTools[25] mergeSTR and statSTR. For each allele a quality score (Q score) was calculated and then averaged for an individual level Q score, using the ExpansionHunter repeat length and confidence intervals, to align with the Q score metric reported in the

gnomAD callset[26,27]. The *ZFHX3* GGC ExpansionHunter callset derived from WGS data was then merged with the gnomAD control callset. Individuals with a Q score below 0.2 were removed, based on a threshold to optimise sensitivity and specificity[26].

73 alleles (36 cases; 37 controls) with a repeat expansion length ≥24 GGC units were visually validated using REViewer[28] v2. 12 alleles (4 cases; 9 controls) were incorrectly estimated by three or more repeats upon visualisation in REViewer, as these genotypes could be reliably corrected, they were removed from the analysis. The repeat motif composition of all 802 Australian ALS participants was catalogued by manual visualisation of individual REViewer images. RStudio v5.1 was used to generate repeat length distributions and motif composition figures.

Statistical analysis of patient clinical data and association testing

Patient phenotypic features were extracted from clinical records and variability was examined across sex, site of disease onset (bulbar or spinal), age at disease onset and disease duration (from onset until death or last known date alive). All statistical analysis was performed in RStudio v5.1. Associations between sex and site of onset were assessed using a chi-squared test of independence. Welch's two sample t tests were used to evaluate differences in age at disease onset by both sex and site of onset. Kaplan-Meier survival analysis compared disease duration across both sex and site of onset. A linear regression model was fitted to assess the relationship between age at onset and duration for deceased patients only. Alleles with repeats lengths greater than 21 units (the median repeat length) were retained to perform a receiver operating characteristic (ROC) curve analysis and to calculate the Youden's J statistic using the R package, cutpointr[29]. Fisher's exact test was used to evaluate the association between repeats expanded ≥24 GGC units and disease status.

Data availability

Individual-level *ZFHX3* repeat allele size data for 5,785 people with ALS and 7,982 healthy controls plus individual-level motif composition data for the subset of 802 Australian cases with ALS is available at Zenodo (DOI: 10.5281/zenodo.18931240).

Code written in R is available in R Markdown workbooks in a GitLab repository: https://gitlab.com/mq-mnd/grp_williams/zfhx3_analysis_publication_2026.

Results

Cohort summary

After filtering the cohort for ancestry and repeat quality, 5,785 participants with ALS (802 Australian ALS; 4,983 Project MinE ALS) and 7,982 healthy controls (6,215 gnomAD controls; 1,767 Project MinE controls) were retained for downstream analysis. Cohort demographic and clinical features of 13,767 individuals are summarised in Table 1. The ALS cohort comprised 61% males while 74% of controls were male. Of 5,785 people with ALS, 430 carried a known ALS disease-causal variant (Supplementary Table 2).

Differences in site of onset (*P* = 2.3e-27) and age of onset (*P* = 3.8e-13) between male and females were observed. 71% of males presented with spinal onset versus 59% of females (Figure 1A). The average disease onset was 61.2 years in females and 58.7 in males (Fig 1B). Males presented with a longer disease duration than females (Fig 1C; *P* = 1.9e-06). Individuals who presented with ALS at a younger age had a longer disease duration (Fig 1D; *P* = 1.06e-11). 66% of people with ALS had spinal onset. People with spinal onset had a younger age of onset (*P* = 2.1e-47; Fig. 1E) and slower disease progression (*P* = 3.4e-31) compared to bulbar onset cases (Fig 1E and F). This is concordant with existing published ALS cohorts[6, 30].

Frequency of *ZFHX3* GGC repeat lengths

The distribution of *ZFHX3* GGC repeat lengths across 11,270 ALS alleles and 15,964 healthy control alleles was visualised (Figure. 2). Repeat sizes ranged from 4-29 units in people with ALS and 5-29 in healthy controls. No individuals exceeded the pathogenic threshold of 42 GGC repeat units reported in SCA4 families (Supplementary Table 3). ROC analysis and Youden's J statistic indicated that repeat length 24 was the optimal threshold to test for an association between cases and controls (Supplementary Figure 1). Yet a Fisher's exact test revealed no association between expanded GGC repeats ≥24 in people with ALS compared to healthy controls (*P* = 0.088).

Variability of repeat motif compositions and configurations in *ZFHX3*

The frequency of 34 distinct repeat motif compositions from 802 Australian ALS individuals across both alleles was visualised (Figure 3). All repeat motif compositions were interrupted by at least one GGT (glycine) motif that produces a synonymous interruption in the polyglycine tract. Seven motif compositions lacked a missense interruption and coded pure glycine tracts, ranging in size from 13-26 repeat motifs. One individual with ALS carried a pure polyglycine tract on both alleles, with repeat lengths 23 and 26. The most common motif composition (in 84.7% of alleles) was present in alleles with 21 repeats, and comprised 18 GGC (glycine) motifs, interrupted by two GGT (glycine) motifs and one AGT (serine) motif (Figure 3 and Supplementary Figure 2). Each repeat length had multiple configurations in their motif composition, with the exception of repeat lengths 26 and 27, which each had only one configuration. Alleles with a repeat length of 23 had the greatest number of motif compositions, with six distinct configurations identified. Whereas alleles with a repeat length of 21 showed the highest variability in interrupting motifs, with five distinct noncanonical GGC motifs observed across five different motif compositions. One allele with a AGC motif co-occurred with an AGT motif, resulting in a double serine interruption. A detailed breakdown of the motif composition seen in 1,604 alleles is provided in Supplementary Figure 2 and Supplementary Table 4.

Discussion

A pathogenic GGC repeat expansion in *ZFHX3* was recently reported as the cause of SCA4[9,10,11]. Given that intermediate expansions at other SCA loci, including *ATXN1* and *ATXN2*, are associated with increased risk of ALS[6,7,8], and the pathological overlap of SCA4 and ALS[10], we investigated whether *ZFHX3* GGC expansions similarly contribute to ALS risk. Using ExpansionHunter-derived genotypes from whole-genome sequencing data, we found no evidence of association between expanded *ZFHX3* GGC repeats and ALS in a large European cohort, indicating that repeat length does not contribute to the risk of ALS.

Beyond repeat size, motif composition and sequence architecture are increasingly recognised as determinants of pathogenicity in repeat expansion disorders. In polyglutamine (polyQ) and polyglycine (polyG) diseases, interruption pattern, motif order, and tract purity can influence repeat instability, penetrance, and age at disease onset[14,16,31,32]. In our cohort, we identified more than 30 distinct *ZFHX3* repeat motif configurations among 802 individuals with ALS. All contained at least one synonymous or missense interruption, alongside seven compositions that encode uninterrupted polyG tracts. In contrast, pathogenic SCA4 expansions consist of long, pure, uninterrupted polyG tracts[9,10]. Pure repeat tracts are thought to be more prone to replication slippage and somatic instability, making them susceptible to further expansion[19], and conversely, interrupted tracts may confer locus stability. In addition, the interruption pattern and motif configuration of repeats may influence transcription and RNA processing independent of repeat length[33]. Understanding how synonymous or missense interruptions influence and stabilise repeat expansions may inform therapeutic strategies that aim to introduce interruptions to stabilise pathogenic repeat tracts, a strategy currently being explored in other repeat expansion disorders[34]. Together, our findings reinforce the importance of resolving repeat motif configuration, not solely allele size, when assessing the pathogenicity of expanded repeats.

This study had several limitations. Although cohort size was substantial, the analysis was restricted to individuals of European ancestry. Since ALS is genetically heterogeneous with mutation frequencies that vary across ancestral backgrounds[35,36], future studies that incorporate diverse ancestral populations will be important to more definitively evaluate the contribution of *ZFHX3* GGC expansions and motif configurations to ALS risk. In addition, while short-read sequencing datasets are widely available for ALS genomic studies, short-read data has inherent limitations for STR analysis. Repeat lengths that exceed the read length cannot be directly spanned, potentially leading to underestimation of larger repeat sizes. GC-rich motifs can introduce coverage bias and interrupted or structurally complex repeats may not be reliably resolved with short-read data. Indeed, a combination of short- and long-read sequencing was required to initially characterise the pathogenic *ZFHX3* GGC expansion in SCA4 families[9,10,37]. However, once the locus was defined, short-read sequencing was sufficient to reidentify individuals with SCA4 and detect previously unresolved ataxia and sensory neuropathy cases harbouring the *ZFHX3* GGC expansion[9]. Moreover, the repeat length distribution of non-disease alleles

was consistent with that observed here[9,37]. Although repeat length was evaluated in both cases and controls, detailed motif composition was characterised only in a subset of the ALS cohort. The population frequency and broader distribution of these configurations will require systematic assessment in ancestry-matched control populations.

In conclusion, we found no evidence that expanded *ZFHX3* GGC repeats contribute to ALS risk in a large European cohort. However, the marked diversity of repeat motif configurations at this locus highlights the importance of interrogating sequence composition, not solely allele size, when evaluating STRs in neurodegenerative disease. Future studies that integrate long-read sequencing, diverse populations, and detailed phenotypic data will be essential to fully define the genetic and biological relevance of *ZFHX3* repeat variation in ALS and related disorders.

Acknowledgements

Biospecimens and related clinical data from Australian participants were obtained from the Macquarie University Neurodegenerative Disease Biobank and the Australian MND DNA Bank. The authors would like to acknowledge Susan D'Silva, Lorel Adams, Srestha Mazumder and Rosie Fell for compiling and collating clinical and demographic data. The authors would also like to gratefully acknowledge the participants who provided biological samples for this study. The collaboration project is co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. This study was supported by the ALS foundation Netherlands. Project MinE work was sponsored by NOW-Domain science for the use of supercomputer facilities.

Author contributions

Z. N. Zussa: Draft/ revision of manuscript for content, major role in acquisition of data, analysis or interpretation of data. A.N. Smith: major role in acquisition of data, analysis or interpretation of data. J. J.F.A Van Vugt: major role in acquisition of data, analysis or interpretation of data. D. S. O'Shaughnessy: interpretation of data. N. Grima: major role in acquisition of data. S. Chan Moi Fat: interpretation of data. D. B. Rowe: major role in acquisition of data. R. Pamphlett: major role in acquisition of data, G. A. Nicholson: major

role in acquisition of data. M. C. Kiernan: major role in acquisition of data, I. P. Blair: major role in acquisition of data. W. V. Rheenen: major role in acquisition of data J. Veldink: major role in acquisition of data. K. L. Williams: Draft/ revision of manuscript for content, major role in acquisition of data, analysis or interpretation of data, study concept or design, supervision. L. Henden: Draft/ revision of manuscript for content, major role in acquisition of data, analysis or interpretation of data, study concept or design, supervision.

Funding

MotorOn scholarship to Z. N. Zussa. FightMND and the National Health and Medical Research Council of Australia (GNT2033019) fellowships to K. L. Williams. Motor Neuron Disease Australia Innovator Grant to L. Henden. FightMND Angie Cunningham PhD Scholarship and Project Grant-In-Aid Award to N. Grima and I. P. Blair. National Health and Medical Research Council funding to M. C. Kiernan.

Disclosure

DBR reports consultancies or advisory boards for Biogen, Celosia Therapeutics, and Genetic Signatures. JHV reports to have sponsored research agreements with Biogen, Eli Lilly, Trace and Astra Zeneca.

AI statement

During the preparation of this manuscript (February 2025) the authors used ChatGPT 5.2 to improve the readability of content already written by the authors. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility of the content of the publication.

Tables

Table 1. Demographic and clinical information for study cohort.

	ALS	**Controls**
	(n = 5,785)	*(n = 7,982)*
Sex (Males)	3,504 (61%)	4,635 (74%)
Age at disease onset (years)	59.7 ± 12.7	-
Disease duration (months)	9.80 ± 21.1	-
Site of onset		
Bulbar	1,490 (27%)	-
Spinal	3,836 (68%)	-

Age at disease onset and disease duration are presented as mean ± standard deviation. Age at onset was available for 5,563/5,785 individuals, disease duration was available for 5,445/5,785 individuals and site of onset was available for 5,612/5,785 individuals with ALS.

Figure legends

Figure 1. Statistical analysis of clinical variables in the ALS cohort (n = 5,785)

Significant associations are denoted by red *P* values. (**A)** There was a significant difference in site of onset between males and females with ALS, as more males presented with spinal onset. **(B)** Age at disease onset was significantly older in females with ALS. **(C)** Males with ALS have a significantly longer disease duration than females. **(D)** Individuals diagnosed with ALS at a younger age lived significantly longer than those diagnosed later in life. **(E)** People with bulbar onset ALS presented with disease at an older age. **(F)** People with bulbar onset experienced a significantly shorter disease duration than people with spinal onset.

Figure 2. Frequency of *ZFHX3* GGC repeat lengths in ALS cases and controls

ROC analysis and Youden's J statistic identified repeat length 24 as the optimal threshold for association test, indicated by the dashed line. However, the frequency of expanded *ZFHX3* GGC repeats ≥24 in ALS cases was not significantly increased compared to controls. The y-axis is a log scale of percent frequency to allow better depiction of frequency variations.

Figure 3. Frequency of 34 distinct repeat motif compositions in Australian ALS cohort

Frequency of distinct repeat motif compositions across 1,604 alleles from 802 Australian ALS cases, ordered by repeat length. For each motif composition, the first column of pie charts shows the proportional contribution of all repeat motifs, while the second column shows motif proportions after excluding the canonical GGC repeat. An asterisk (*) denotes a repeat motif composition lacking a missense interruption and represents a pure glycine tract. Counts are displayed on a log scale to allow for easier depiction of frequency variability.

References:

1. Shatunov, A., & Al-Chalabi, A. (2021). The genetic architecture of ALS. In Neurobiology of Disease (Vol. 147). Academic Press Inc. https://doi.org/10.1016/j.nbd.2020.105156

2. Grassano, M., Calvo, A., Moglia, C., Sbaiz, L., Brunetti, M., Barberis, M., Casale, F., Manera, U., Vasta, R., Canosa, A., D'Alfonso, S., Corrado, L., Mazzini, L., Dalgard, C., Karra, R., Chia, R., Traynor, B., & Chiò, A. (2022). Systematic evaluation of genetic mutations in ALS: a population-based study. Journal of Neurology, Neurosurgery and Psychiatry, 93(11), 1190–1193. https://doi.org/10.1136/jnnp-2022-328931

3. McCann, E. P., Grima, N., Fifita, J. A., Chan Moi Fat, S., Lehnert, K., Henden, L., Blair, I. P., & Williams, K. L. (2023). Characterising the Genetic Landscape of Amyotrophic Lateral Sclerosis: A Catalogue and Assessment of Over 1,000 Published Genetic Variants. Journal of Neuromuscular Diseases, 10(6), 1127–1141. https://doi.org/10.3233/JND-230148

4. Abramzon, Y. A., Fratta, P., Traynor, B. J., & Chia, R. (2020). The Overlapping Genetics of Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. In Frontiers in Neuroscience (Vol. 14). Frontiers Media S.A. https://doi.org/10.3389/fnins.2020.00042

5. Klockgether, T., Mariotti, C., & Paulson, H. L. (2019). Spinocerebellar ataxia. Nature Reviews Disease Primers, 5(1). https://doi.org/10.1038/s41572-019-0074-3

6. Henden, L., Fearnley, L. G., Grima, N., Mccann, E. P., Dobson-Stone, C., Fitzpatrick, L., Friend, K., Hobson, L., Chan, S., Fat, M., Rowe, D. B., D'silva, S., Kwok, J. B., Halliday, G. M., Kiernan, M. C., Mazumder, S., Timmins, H. C., Zoing, M., Pamphlett, R., … Williams, K. L. (2023). Short tandem repeat expansions in sporadic amyotrophic lateral sclerosis and frontotemporal dementia. https://www.science.org

7. Lattante, S., Pomponi, M. G., Conte, A., Marangi, G., Bisogni, G., Patanella, A. K., Meleo, E., Lunetta, C., Riva, N., Mosca, L., Carrera, P., Bee, M., Zollino, M., & Sabatelli, M. (2018). ATXN1 intermediate-length polyglutamine expansions are associated with amyotrophic lateral sclerosis. Neurobiology of Aging, 64, 157.e1-157.e5. https://doi.org/10.1016/j.neurobiolaging.2017.11.011

8. Elden, A. C., Kim, H. J., Hart, M. P., Chen-Plotkin, A. S., Johnson, B. S., Fang, X., Armakola, M., Geser, F., Greene, R., Lu, M. M., Padmanabhan, A., Clay-Falcone, D., McCluskey, L., Elman, L., Juhr, D., Gruber, P. J., Rüb, U., Auburger, G., Trojanowski, J. Q., … Gitler, A. D. (2010). Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature, 466(7310), 1069–1075. https://doi.org/10.1038/nature09320

9. Figueroa, K. P., Gross, C., Buena-Atienza, E., Paul, S., Gandelman, M., Kakar, N., Sturm, M., Casadei, N., Admard, J., Park, J., Zühlke, C., Hellenbroich, Y., Pozojevic, J., Balachandran, S., Händler, K., Zittel, S., Timmann, D., Erdlenbruch, F., Herrmann, L., … Pulst, S. M. (2024). A GGC-repeat expansion in ZFHX3 encoding polyglycine causes spinocerebellar ataxia type 4 and impairs autophagy. In Nature Genetics (Vol. 56, Issue 6, pp. 1080–1089). Nature Research. https://doi.org/10.1038/s41588-024-01719-5

10. Paucar, M., Nilsson, D., Engvall, M., Laffita-Mesa, J., Söderhäll, C., Skorpil, M., Halldin, C., Fazio, P., Lagerstedt-Robinson, K., Solders, G., Angeria, M., Varrone, A., Risling, M., Jiao, H., Nennesmo, I., Wedell, A., & Svenningsson, P. (2024). 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, 296(3), 234–248. https://doi.org/10.1111/joim.13815

11. Wallenius, J., Kafantari, E., Jhaveri, E., Gorcenco, S., Ameur, A., Karremo, C., Dobloug, S., Karrman, K., de Koning, T., Ilinca, A., Landqvist Waldö, M., Arvidsson, A., Persson, S., Englund, E., Ehrencrona, H., & Puschmann, A. (2024). Exonic trinucleotide repeat expansions in ZFHX3 cause spinocerebellar ataxia type 4: A poly-glycine disease. American Journal of Human Genetics, 111(1), 82–95. https://doi.org/10.1016/j.ajhg.2023.11.008

12. Pérez Baca, M. del R., Jacobs, E. Z., Vantomme, L., Leblanc, P., Bogaert, E., Dheedene, A., de Cock, L., Haghshenas, S., Foroutan, A., Levy, M. A., Kerkhof, J., McConkey, H., Chen, C. A., Batzir, N. A., Wang, X., Palomares, M., Carels, M., Agrawal, P., Armstrong Scott, D., … Callewaert, B. (2024). Haploinsufficiency of ZFHX3, encoding a key player in neuronal development, causes syndromic intellectual disability. American Journal of Human Genetics, 111(3), 509–528. https://doi.org/10.1016/j.ajhg.2024.01.013

13. Rajan-Babu, I. S., Dolzhenko, E., Eberle, M. A., & Friedman, J. M. (2024). Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. In Nature Reviews Genetics (Vol. 25, Issue 7, pp. 476–499). Nature Research. https://doi.org/10.1038/s41576-024-00696-z

14. Wright, G. E. B., Findlay Black Bsc, H., Collins Ba, J. A., Caron, N. S., Mbchb, H., Gall-Duncan Bsc, T., Pearson, C. E., Hayden, M. R., Wright, G. E. B., Black, H. F., Collins, J. A., Gall-Duncan, T., Caron, N. S., & Pearson, C. E. (2020). Interrupting sequence variants and age of onset in Huntington's disease: clinical implications and emerging therapies. In Rapid Review Lancet Neurol (Vol. 19). www.thelancet.com/neurology

15. Aston, A. N., & Dion, V. (2025). Interruptions impact clinical features of repeat expansion diseases, but how are they gained and lost? Trends in Genetics, 41(12), 1056–1067. https://doi.org/10.1016/j.tig.2025.07.005

16. Huang, X. R., Tang, B. S., Jin, P., & Guo, J. F. (2022). The Phenotypes and Mechanisms of NOTCH2NLC-Related GGC Repeat Expansion Disorders: a Comprehensive Review. In Molecular Neurobiology (Vol. 59, Issue 1, pp. 523–534). Springer. https://doi.org/10.1007/s12035-021-02616-2

17. Wang, L., Milton, M., Fearnley, L. G., Bhalala, O. G., Bahlo, M., & Rafehi, H. (2025). Identification of expanded and interrupted ATXN2 repeat expansions in Parkinson's disease and Lewy Body Dementia cohorts. Npj Parkinson's Disease, 11(1). https://doi.org/10.1038/s41531-025-01188-5

18. Wu, S. S. L., McCann, E. P., Chan Moi Fat, S., Grima, N., Henden, L., Fearnley, L. G., Chiu, P., Williams, K. L., Rowe, D. B., Nicholson, G. A., Kiernan, M. C., Blair, I. P., Yang, S., & Fifita, J. A. (2026). Genetic characterisation of ATXN2 in Australian amyotrophic lateral sclerosis. Brain Disorders, 21. https://doi.org/10.1016/j.dscb.2025.100301

19. Kozlowski, P., Sobczak, K., & Krzyzosiak, W. J. (2010). Genomic repeats and human diseases. http://genomemedicine.com/content/2/4/29

20. Brooks, B. R., Miller, R. G., Swash, M., & Munsat, T. L. (2000). El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotrophic Lateral Sclerosis, 1(5), 293–299. https://doi.org/10.1080/146608200300079536

21. van Rheenen, W., Pulit, S. L., Dekker, A. M., al Khleifat, A., Brands, W. J., Iacoangeli, A., Kenna, K. P., Kavak, E., Kooyman, M., McLaughlin, R. L., Middelkoop, B., Moisse, M., Schellevis, R. D., Shatunov, A., Sproviero, W., Tazelaar, G. H. P., van der Spek, R. A. A., van Doormaal, P. T. C., van Eijk, K. R., … Veldink, J. H. (2018). Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. European Journal of Human Genetics, 26(10), 1537–1546. https://doi.org/10.1038/s41431-018-0177-4

22. The International HapMap Project The International HapMap Consortium*. (2003). http://www.hapmap.

23. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81(3), 559–575. https://doi.org/10.1086/519795

24. Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., Scheffler, K., van Vugt, J. J. F. A., French, C., Sanchis-Juan, A., Ibáñez, K., Tucci, A., Lajoie, B. R., Veldink, J. H., Raymond, F. L., … Birol, I.

(2019). ExpansionHunter: A sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics, 35(22), 4754–4756. https://doi.org/10.1093/bioinformatics/btz431

25. Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., & Gymrek, M. (2021). TRTools: A toolkit for genome-wide analysis of tandem repeats. Bioinformatics, 37(5), 731–733. https://doi.org/10.1093/bioinformatics/btaa736

26. Weisburd, B., Tiao, G., & Rehm, H. L. (2023). Insights from a genome-wide truth set of tandem repeat variation. https://doi.org/10.1101/2023.05.05.539588

27. Jam, H. Z., Li, Y., DeVito, R., Mousavi, N., Ma, N., Lujumba, I., Adam, Y., Maksimov, M., Huang, B., Dolzhenko, E., Qiu, Y., Kakembo, F. E., Joseph, H., Onyido, B., Adeyemi, J., Bakhtiari, M., Park, J., Javadzadeh, S., Jjingo, D., … Gymrek, M. (2023). A deep population reference panel of tandem repeat variation. https://doi.org/10.1101/2023.03.09.531600

28. Dolzhenko, E., Weisburd, B., Ibañez, K., Rajan-Babu, I. S., Anyansi, C., Bennett, M. F., Billingsley, K., Carroll, A., Clamons, S., Danzi, M. C., Deshpande, V., Ding, J., Fazal, S., Halman, A., Jadhav, B., Qiu, Y., Richmond, P. A., Saunders, C. T., Scheffler, K., … Eberle, M. A. (2022). REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. Genome Medicine, 14(1). https://doi.org/10.1186/s13073-022-01085-z

29. Thiele, C., & Hirschfeld, G. (2021). Cutpointr: Improved estimation and validation of optimal cutpoints in r. Journal of Statistical Software, 98. https://doi.org/10.18637/jss.v098.i11

30. Gianferrari, G., Martinelli, I., Zucchi, E., Simonini, C., Fini, N., Vinceti, M., Ferro, S., Gessani, A., Canali, E., Valzania, F., Sette, E., Pugliatti, M., Tugnoli, V., Zinno, L., Stano, S., Santangelo, M., de Pasqua, S., Terlizzi, E., Guidetti, D., … Mandrioli, J. (2022). Epidemiological, Clinical and Genetic Features of ALS in the Last Decade: A Prospective Population-Based Study in the Emilia Romagna Region of Italy. Biomedicines, 10(4). https://doi.org/10.3390/biomedicines10040819

31. Nolin, S. L., Glicksman, A., Ersalesi, N., Dobkin, C., Brown, W. T., Cao, R., Blatt, E., Sah, S., Latham, G. J., & Hadd, A. G. (2015). Fragile X full mutation expansions are inhibited by one or more AGG interruptions in premutation carriers. Genetics in Medicine, 17(5), 358–364. https://doi.org/10.1038/gim.2014.106

32. Menon, R. P., Nethisinghe, S., Faggiano, S., Vannocci, T., Rezaei, H., Pemble, S., Sweeney, M. G., Wood, N. W., Davis, M. B., Pastore, A., & Giunti, P. (2013). The Role of Interruptions in polyQ in the Pathology of SCA1. PLoS Genetics, 9(7). https://doi.org/10.1371/journal.pgen.1003648

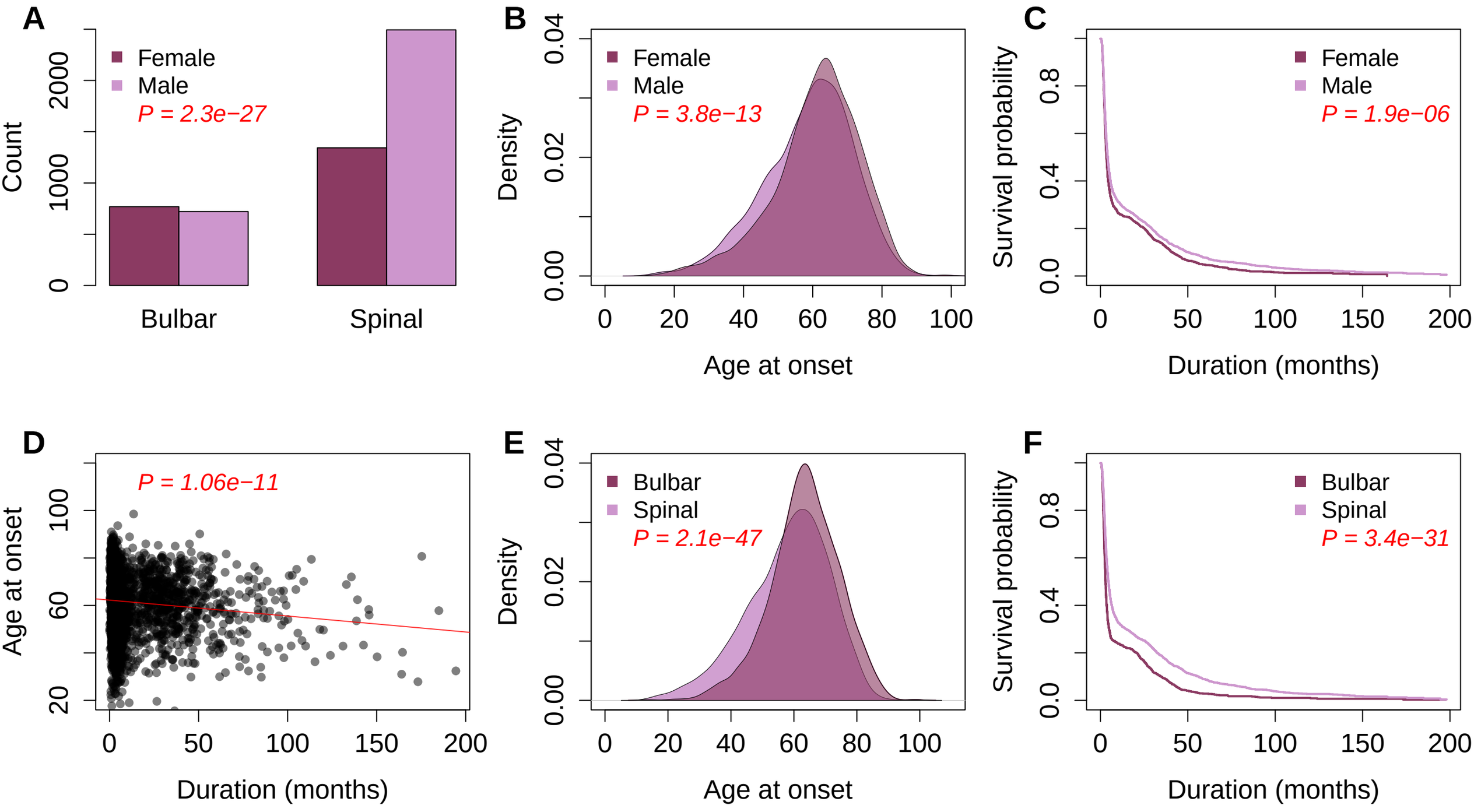
33. Mitina, A., Engchuan, W., Trost, B., Pellecchia, G., Scherer, S. W., & Yuen, R. K. C. (2025). Diverse short tandem repeat sequences influence gene regulation in human populations. Genome Biology, 26(1). https://doi.org/10.1186/s13059-025-03754-9

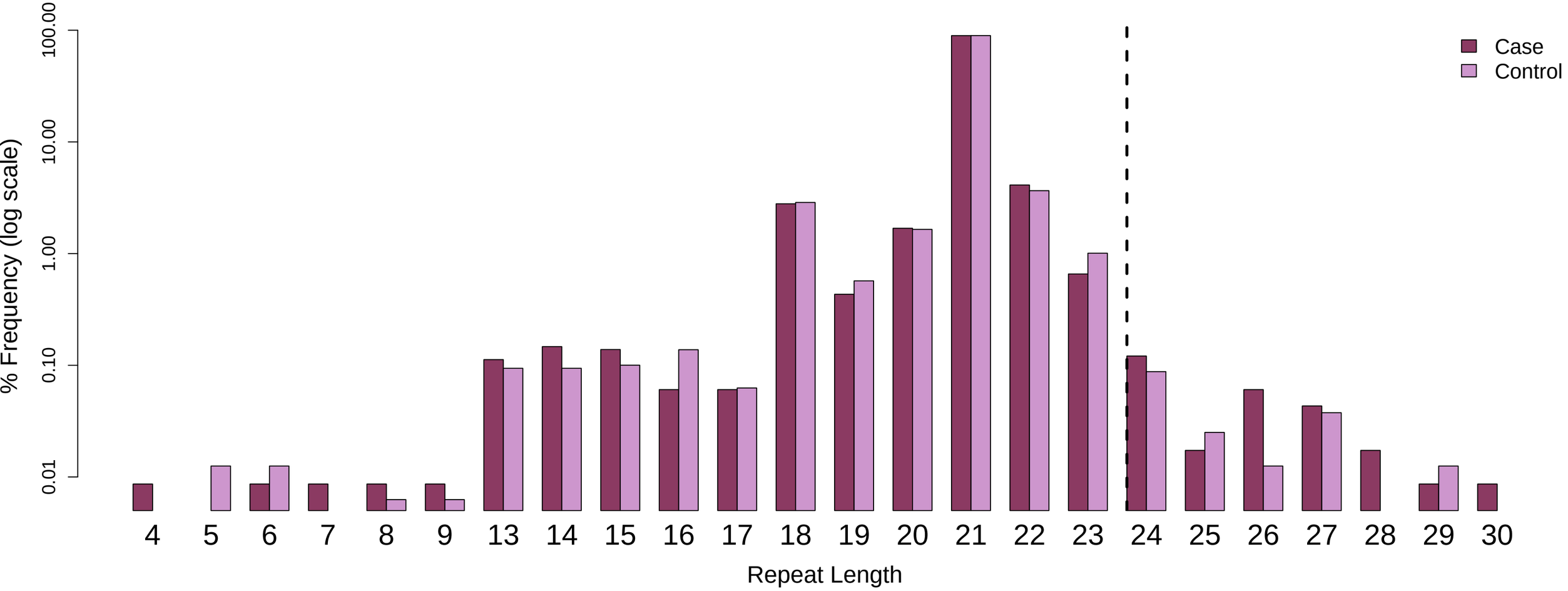
34. Matuszek, Z., Arbab, M., Kesavan, M., Hsu, A., Roy, J. C. L., Zhao, J., Yu, T., Weisburd, B., Newby, G. A., Doherty, N. J., Wu, M., Shibata, S., Cristian, A., Tao, Y. A., Fearnley, L. G., Bahlo, M., Rehm, H. L., Xie, J., Gao, G., … Liu, D. R. (2025). Base editing of trinucleotide repeats that cause Huntington's disease and Friedreich's ataxia reduces somatic repeat expansions in patient cells and in mice. Nature Genetics, 57(6), 1437–1451. https://doi.org/10.1038/s41588-025-02172-8

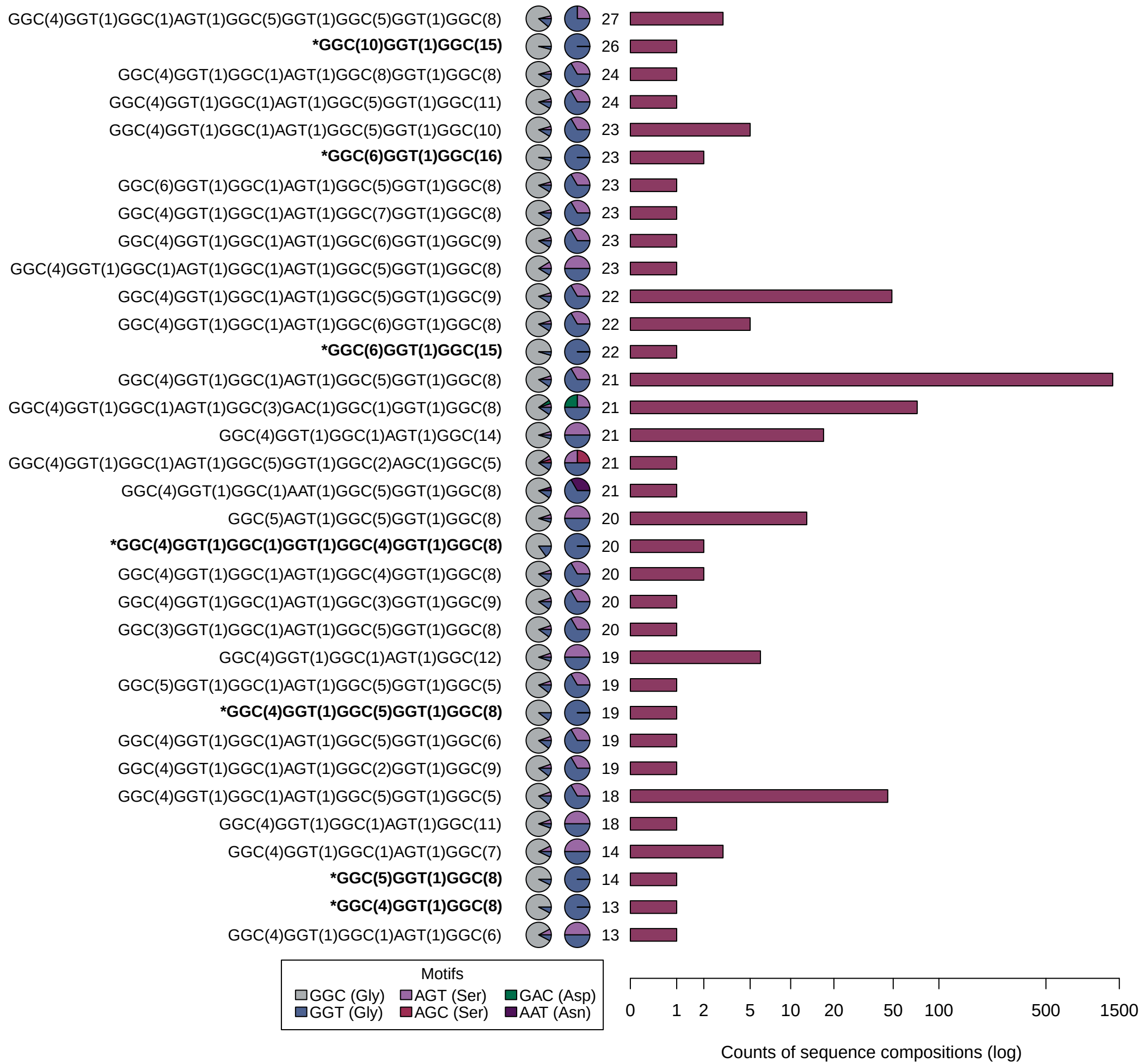
35. Wei, Q., Zhou, Q., Chen, Y., Ou, R., Cao, B., Xu, Y., Yang, J., & Shang, H. F. (2017). Analysis of SOD1 mutations in a Chinese population with amyotrophic lateral sclerosis: A case-control study and literature review. Scientific Reports, 7. https://doi.org/10.1038/srep44606

36. Edgar, S., Ellis, M., Abdul-Aziz, N. A., Goh, K. J., Shahrizaila, N., Kennerson, M. L., & Ahmad-Annuar, A. (2021). Mutation analysis of SOD1, C9orf72, TARDBP and FUS genes in ethnically-diverse Malaysian patients with amyotrophic lateral sclerosis (ALS). Neurobiology of Aging, 108, 200–206. https://doi.org/10.1016/j.neurobiolaging.2021.07.008

37. Chen, Z., Gustavsson, E. K., Macpherson, H., Anderson, C., Clarkson, C., Rocca, C., Self, E., Alvarez Jerez, P., Scardamaglia, A., Pellerin, D., Montgomery, K., Lee, J., Gagliardi, D., Luo, H., Hardy, J., Polke, J., Singleton, A. B., Blauwendraat, C., Mathews, K. D., … Ptáček, L. J. (2024). Adaptive Long-Read Sequencing Reveals GGC Repeat Expansion in ZFHX3 Associated with Spinocerebellar Ataxia Type 4. Movement Disorders, 39(3), 486–497. https://doi.org/10.1002/mds.29704


A
Count
0
1000
2000
Female
Male
P = 2.3e−27
Bulbar
Spinal
B
Density
0.00
0.02
0.04
Female
Male
P = 3.8e−13
0
20
40
60
80
100
Age at onset
C
Survival probability
0.0
0.4
0.8
Female
Male
P = 1.9e−06
0
50
100
150
200
Duration (months)
D
Age at onset
20
60
100
P = 1.06e−11
0
50
100
150
200
Duration (months)
E
Density
0.00
0.02
0.04
Bulbar
Spinal
P = 2.1e−47
0
20
40
60
80
100
Age at onset
F
Survival probability
0.0
0.4
0.8
Bulbar
Spinal
P = 3.4e−31
0
50
100
150
200
Duration (months)


% Frequency (log scale)
100.00
10.00
1.00
0.10
0.01
Case
Control
4
5
6
7
8
9
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
Repeat Length


GGC(4)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(5)GGT(1)GGC(8) 27
*GGC(10)GGT(1)GGC(15) 26
GGC(4)GGT(1)GGC(1)AGT(1)GGC(8)GGT(1)GGC(8) 24
GGC(4)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(11) 24
GGC(4)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(10) 23
*GGC(6)GGT(1)GGC(16) 23
GGC(6)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(8) 23
GGC(4)GGT(1)GGC(1)AGT(1)GGC(7)GGT(1)GGC(8) 23
GGC(4)GGT(1)GGC(1)AGT(1)GGC(6)GGT(1)GGC(9) 23
GGC(4)GGT(1)GGC(1)AGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(8) 23
GGC(4)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(9) 22
GGC(4)GGT(1)GGC(1)AGT(1)GGC(6)GGT(1)GGC(8) 22
*GGC(6)GGT(1)GGC(15) 22
GGC(4)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(8) 21
GGC(4)GGT(1)GGC(1)AGT(1)GGC(3)GAC(1)GGC(1)GGT(1)GGC(8) 21
GGC(4)GGT(1)GGC(1)AGT(1)GGC(14) 21
GGC(4)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(2)AGC(1)GGC(5) 21
GGC(4)GGT(1)GGC(1)AAT(1)GGC(5)GGT(1)GGC(8) 21
GGC(5)AGT(1)GGC(5)GGT(1)GGC(8) 20
*GGC(4)GGT(1)GGC(1)GGT(1)GGC(4)GGT(1)GGC(8) 20
GGC(4)GGT(1)GGC(1)AGT(1)GGC(4)GGT(1)GGC(8) 20
GGC(4)GGT(1)GGC(1)AGT(1)GGC(3)GGT(1)GGC(9) 20
GGC(3)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(8) 20
GGC(4)GGT(1)GGC(1)AGT(1)GGC(12) 19
GGC(5)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(5) 19
*GGC(4)GGT(1)GGC(5)GGT(1)GGC(8) 19
GGC(4)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(6) 19
GGC(4)GGT(1)GGC(1)AGT(1)GGC(2)GGT(1)GGC(9) 19
GGC(4)GGT(1)GGC(1)AGT(1)GGC(5)GGT(1)GGC(5) 18
GGC(4)GGT(1)GGC(1)AGT(1)GGC(11) 18
GGC(4)GGT(1)GGC(1)AGT(1)GGC(7) 14
*GGC(5)GGT(1)GGC(8) 14
*GGC(4)GGT(1)GGC(8) 13
GGC(4)GGT(1)GGC(1)AGT(1)GGC(6) 13
Motifs
GGC (Gly)
AGT (Ser)
GAC (Asp)
GGT (Gly)
AGC (Ser)
AAT (Asn)
0
1
2
5
10
20
50
100
500
1500
Counts of sequence compositions (log)