Identification of GGC Repeat Expansions in *ZFHX3* Among Chilean Movement Disorder Patients

Paula Saffie-Awad[1], Abraham Moller[2], Kensuke Daida[3], Pilar Alvarez Jerez[2,4], Zhongbo Chen[4,5,6], Zachary B. Anderson[7], Mariam Isayan[8], Kimberly Paquette[2], Sophia B Gibson[7,9], Madison Fulcher[3], Abigail Miano-Burkhardt[3], Laksh Malik[2], Breeana Baker[2], Paige Jarreau[2], Henry Houlden[10], Mina Ryten[11,12,13,14], Bida Gu[15], Mark JP Chaisson[15], Danny E. Miller[9,16,17], Pedro Chaná-Cuevas[18], Cornelis Blauwendraat [2,3+], Andrew B. Singleton[2,3], Kimberley J. Billingsley[2].

1. Clínica Santa María, Santiago, Chile.
2. Center for Alzheimer's and Related Dementias, National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
3. Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
4. Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, University College London, London, UK
5. Department of Genetics and Genomic Medicine, Great Ormond Street Institute of Child Health, University College London, London WC1N 1EH, UK 3
6. NIHR Great Ormond Street Hospital Biomedical Research Centre, University College London, London WC1N 1EH, UK
7. Division of Genetic Medicine, Department of Pediatrics, University of Washington, Seattle, Washington, USA
8. Department of Neurology and Neurosurgery, National Institute of Health, Yerevan, Armenia
9. Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.
10. Department of Neuromuscular Disease, Queen Square Institute of Neurology, UCL, London, UK
11. Department of Genetics and Genomic Medicine, Great Ormond Street Institute of Child Health, UCL, London, UK
12. NIHR Great Ormond Street Hospital Biomedical Research Centre, UCL, London, UK
13. UK Dementia Research Institute at the University of Cambridge, Cambridge, UK
14. Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge, UK
15. Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA
16. Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington 98195, USA.
17. Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, Washington 98195, USA.
18. Centro de Trastornos del Movimiento, Facultad de Ciencias Médicas, Universidad de Santiago de Chile, Santiago, Chile

Corresponding author
Kimberley J. Billingsley
Staff Scientist, Head, Applied Neurogenomics Group
Center for Alzheimer's Disease and Related Dementias (CARD)
NIH, NIA, 9000 Rockville Pike, NIH Building T44, Bethesda, MD 20892.
Email: kimberley.billingsley@nih.gov

Word count: 2682

Runnning title: *ZFHX3* Repeat Expansions in Chilean Patients.

The authors declare no financial or non-financial conflicts of interest related to this manuscript.

This study was supported in part by the Intramural Research Program of the National Institute on Aging (NIA) and the Center for Alzheimer's and Related Dementias (CARD), within the Intramural Research Program of the NIA and the National Institute of Neurological Disorders and Stroke (ZIANS003154, ZIAAG000538), National Institutes of Health (AG000538).

Additional support was provided by the National Institutes of Health (DP5OD033357, R01HG011649, 5T32HG000035-29), as well as grants from the Michael J. Fox Foundation, the Medical Research Council, the Wellcome Trust, and the National Institute for Health and Care Research (UCL/UCLH Biomedical Research Centre).

Abstract

Background

Hereditary ataxias are genetically diverse, yet up to 75% remain undiagnosed due to technological and financial barriers. A pathogenic *ZFHX3* GGC repeat expansion was recently linked to spinocerebellar ataxia type 4 (SCA4), characterized by progressive ataxia and sensory neuropathy, with all reported cases in individuals of Northern European ancestry.

Methods

We performed Oxford Nanopore Technologies (ONT) genome long-read sequencing (>115 GB per sample) on a total of 15 individuals from Chile; 14 patients with suspected hereditary movement disorders and one unrelated family member. Variants were identified using PEPPER-Margin-DeepVariant 0.8 (SNVs), Sniffles 2.4 (SVs), and Vamos 2.1.3 (STRs). Ancestry was inferred using GenoTools with reference data from the 1000 Genomes Project, Human Genome Diversity Project, and an Ashkenazi Jewish panel. Haplotype analysis was conducted by phasing SNVs within *ZFHX3*, and methylation profiling was performed with modbamtools.

Results

We identified *ZFHX3* GGC repeat expansions (47–55 repeats) in four individuals with progressive ataxia, polyneuropathy, and vermis atrophy. One case presented parkinsonism–ataxia, expanding the phenotype. Longer expansions correlated with earlier onset and greater severity. Hypermethylation was detected on the expanded allele, and haplotype analysis linked ultra-rare *ZFHX3* variants to distant Swedish ancestry.

Conclusion

This is the first report of SCA4 outside Northern Europe, confirming a shared founder haplotype and expansion instability. The presence of parkinsonism broadens the clinical spectrum. Comprehensive genetic testing across diverse populations is crucial, and long-read sequencing enhances diagnostic yield by detecting repeat expansions and SNVs in a single assay.

Introduction

Hereditary ataxias, like many other rare neurological disorders, are challenging to diagnose due to their extensive genetic and clinical heterogeneity. These conditions can result from a wide range of genetic variant types, including short tandem repeats (STRs), single nucleotide variants (SNVs), and structural variants (SVs) such as large insertions, deletions, and other complex rearrangements, making comprehensive molecular diagnosis complex. Despite the availability of targeted gene panels, diagnostic yields remain around 75% [1] and can be even lower in resource-limited settings. For example, a specialized center in Chile reported a diagnostic rate of only 23% in patients presenting with ataxia[2]. These gaps underscore the need for continued refinements in both clinical and molecular evaluation strategies.

Among the more than 50 recognized spinocerebellar ataxias (SCAs), SCA4 was first mapped to chromosome 16q22.1 over 25 years ago in a large Utah-based family of Swedish ancestry [3]. Its underlying genetic etiology was recently identified, as an uninterrupted GGC repeat expansion in the final exon of *ZFHX3* gene, possibly linked to a single founder Northern European haplotype [4,5]. Notably, this expansion has not been identified in the non-European population [6,7] despite screening in Japanese [8] and Brazilian cohorts [9]. Expanded alleles typically range from 42 to 74 repeats, while normal alleles are generally shorter (14–31 repeats) and often contain interruptions by other sequences. Alleles in the 32–41 range have uncertain clinical significance [10]. Clinically, SCA4 is characterized by a combination of cerebellar ataxia and sensory axonal neuropathy, frequently accompanied by dysautonomia and oculomotor abnormalities [11]. A notable anticipation phenomenon correlated with repeat size, with a typical age of onset between 30 and 50 years[4–6]. Unlike more frequent ataxias—such as SCA2 and SCA3, SCA4's full phenotypic spectrum, prevalence in non-European populations, and molecular mechanisms are still being defined.

In this study, we performed Oxford Nanopore Technologies (ONT) genome long-read sequencing on 15 individuals, including two multiplex families and eleven index cases, with suspected hereditary movement disorders—ataxia, ataxia-parkinsonism, parkinsonism, or atypical parkinsonism. We identified SCA4-related heterozygous *ZFHX3* expansions in three unrelated Chilean families, suggesting that SCA4 may be more geographically and ethnically diverse than previously recognized. Here, we present these findings and emphasize the advantages of long-read sequencing for accurately diagnosing SCA4 and other neurodegenerative diseases.

Methods

Patients and Participants

This study was approved by the local ethics committee in Chile, ensuring compliance with ethical research standards. Written informed consent was obtained from all participants. Peripheral venous blood samples were collected and stored at −80 °C to preserve the integrity of genetic material for subsequent analyses. A total of 15 samples were collected, from two multiplex families and eleven index cases, with suspected hereditary movement disorders—ataxia, ataxia-parkinsonism, parkinsonism, or atypical parkinsonism **(Supplementary Table 1)**.

Data generation

The DNA protocols are publicly available on protocols.io (DOI: 10.17504/protocols.io.n92ldmx3ol5b/v1) [12]. In brief, DNA was extracted from whole blood using the QIAamp Blood Midi Kit (Qiagen), following the spin-column protocol. Whole blood samples, collected in EDTA tubes, were equilibrated to room temperature prior to processing. DNA extraction was performed according to the manufacturer's protocol, with minor modifications to optimize yield and purity. The extracted DNA was quantified using the Qubit Fluorometer (Thermo Fisher Scientific) with the Qubit dsDNA Broad Range Assay Kit, ensuring high-quality DNA suitable for downstream applications. On average, the protocol yielded 20–30 µg of DNA per sample The samples were then size selected with Sage Science's Blue Pippin following the 0.75 Agarose Dye-Free 10 kb High Pass Plus Marker U1 protocol. An average of 3.56 µg of DNA was loaded into size selection for each sample. After size selection, the average peak size was 27.8kb. Libraries were constructed using an SQK-LSK114 ONT kit (https://www.protocols.io/view/processing-frozen-archival-human-dna-samples-for-l-5jyl82morl2w/v1) using the same modifications as above however where 1.3-2.5 µg of DNA was used as starting input. PromethION sequencing (MinKNOW version 24.02.10) was performed as per manufacturer's guidelines with minor adjustments, 20 fmol of the library was loaded onto each primed R10.4.1 flow cell. Each sample required at least two or three loads to achieve > 115GB total data output over 72 hours.

Data processing

All samples were basecalled and aligned on the NIH HPC Biowulf cluster. Basecalling was performed using dorado v0.7.1 on HPC nodes with 2 NVIDIA V100X GPUs (super-accurate basecalling model - dna_r10.4.1_e8.2_400bps_sup@v5.0.0, 5mCG/5hCG methylation calling) and samples were mapped to the GRCh38 human genome using Minimap2 v2.28 with the map-ont preset. We called SNVs, SVs, and STRs from alignments against GRCh38 with a set of variant detection tools. SNVs were called with Clair3 v1.0.10V and identified with PEPPER-Margin-DeepVariant 0.8 [13]). SVs were called and merged amongst individuals with Sniffles 2.4[14] and then split into harmonized per individual VCFs using bcftools 1.21 plugin +split. STRs were called with Vamos 2.1.3[15]. For ancestry analysis, SNVs were assessed using reference data from the 1000 Genomes Project, the Human Genome Diversity Project, and an Ashkenazi Jewish panel, analyzed with GenoTools[16,17].

In order to compare haplotypes, we took the SNV calls for our samples and looked for the six ultra-rare SNVs coming from the Swedish ancestry founder effect as reported in Chen et al[18]. We then plotted the haplotypes for visual comparison using the following script: https://github.com/zanderson82/SNP-Haplotype-Plotting/tree/main.

Methylation analysis

To look at methylation patterns, we used modbamtools v0.4.8 (https://www.biorxiv.org/content/10.1101/2022.07.07.499188v1) to generate plots for our expansion carriers and a control. We provided the haplotagged bams from PEPPER-Margin-DeepVariant as input, and the Gencode v38GRCh38 gtf for the gene tracks. Additionally, we added the –hap option to split out the methylation frequency by haplotype.

Data sharing and code availability

The data supporting the findings of this study are available upon reasonable request due to ethical and legal restrictions related to patient confidentiality. Summary statistics and relevant scripts used for data analysis are publicly available in the https://github.com/molleraj/CARDlongread-chile-data-processing associated with this study. Due to privacy concerns, individual-level genetic and clinical data cannot be shared publicly but can be accessed through data-sharing agreements with the corresponding author upon reasonable request and approval from the relevant ethics committees.

Results

Data generation and analysis

We faced significant challenges implementing high-molecular-weight DNA protocols in Chile. These protocols typically require advanced expertise, costly equipment, and complex logistics, including dry ice shipping. To address these obstacles, we developed and publicly shared a cost-effective wet-lab DNA extraction protocol tailored for resource-limited settings, available on the protocols.io platform (https://www.protocols.io/view/protocol-purification-of-dna-from-whole-blood-usin-c7ypzpvn)[12]. DNA was extracted from frozen blood in Chile, with subsequent processing and sequencing performed at the NIH. This workflow generated high-quality ONT whole-genome long-read sequencing data (average N50: 21 kb; 130 GB per flow cell) for under $1,000 per sample **(Supplementary Table 2)**. This dataset facilitated comprehensive analyses of SNVs, SVs, and repeat expansions. A detailed overview of the workflow used for these analyses is presented in **Figure 1a.** Ancestry analysis predicted all the Chilean samples to be Latino/admixed American by Genotools (**Supplementary Figure 1**).

Identifying four carriers of the GGC *ZFHX3* expansion

Repeat expansions, such as those in *HTT* (CAG repeats in Huntington's disease), *ATXN1* and *ATXN2* (CAG repeats in spinocerebellar ataxias), and *FXN* (GAA repeats in Friedreich's ataxia), are well-established causes of neurodegenerative disorders [19,20]. Using the *vamos* tool, we assessed the lengths of a catalog of known pathogenic STRs. Among the cohort, we identified four patients with pathogenic-length expansions of the recently described *ZFHX3* GGC repeat, associated with SCA4 (**Figure 1.b)**. No other pathogenic-length STRs were detected in the remaining patients. **Supplementary Table 3** provides a summary of the STR lengths determined from long-read sequencing for all tested pathogenic loci. Additionally, no known pathogenic SNVs or SVs were identified in the four *ZFHX3* carriers. An inverse relationship was observed between repeat length and age at onset, with longer expansions associated with earlier disease manifestation. While this trend was not statistically significant ($R^2 = 0.54$, $p = 0.26$) **(Supplementary Figure 2**), it aligns with previous reports [10].

To date, the SCA4 expansion has only been identified in individuals who carry ultra-rare SNVs linked to a distant common founder event in Sweden. Here, we analyzed phased SNVs surrounding the repeat expansion and identified four of the six ultra-rare SNVs previously reported in individuals with SCA4 by Figueroa et al.[4] and Chen et al[18] **.**Notably, these SNVs were all absent in non-carriers **(Figure 2, Supplementary Table 4)**.

Methylation analysis

Chen et al. previously reported hypermethylation associated with the SCA4 expansion, prompting us to investigate methylation patterns in our cohort[18]. Consistent with their findings, haplotype-specific methylation calling around the *ZFHX3* GGC repeat expansion showed hypermethylation around the STRs compared with the non-expanded allele overlapping with the last exon of *ZFHX3*. In contrast, samples without the expansion showed hypomethylation

in this region **(Figure 3)**. This differential methylation pattern present at the expanded allele suggests that the presence of the GGC repeat could have downstream effects on *ZFHX3* expression through epigenetic regulation.

Clinical characteristics and phenotype of *ZFHX3* carriers

All four *ZFHX3* carriers initially presented with generalized ataxia without vertigo or cerebellar atrophy, accompanied by various neurological and non-neurological features, including chronic cough. Notably, one patient (CL_OC_II-1_A1) progressed to a rapidly evolving parkinsonism–ataxia syndrome. A detailed summary of each patient's clinical presentation is provided in **Table 1.**

CL_PLO_III-1_A2, a female with a 10-year disease duration, carries 55 CAG repeats. She exhibited a classic ataxic phenotype (SARA 10) without dysarthria, motor neuron involvement, or cognitive impairment. Electromyography (EMG) confirmed polyneuropathy, and her only non-neurological symptom was chronic cough. Magnetic Resonance Imaging (MRI) showed mild cerebellar atrophy. CL_PLO_II-3_A1, a male with a 9-year disease duration, carries 49 CAG repeats. His presentation included severe ataxia (SARA 28), dysarthria, and upper motor neuron signs, such as spasticity and a positive Babinski sign. He demonstrated mild cognitive decline (MoCA 21) and polyneuropathy with hypopalesthesia. Additional features included bladder incontinence and REM sleep behavior disorder (RBD). Neuroimaging revealed cerebellar vermis atrophy. CL_PAV_II-1_A1, a female with an 8-year disease duration, carries 47 CAG repeats. She presented with gait ataxia (SARA 19), dysarthria and sensory polyneuropathy. No vertigo, upper motor neuron involvement, or cognitive decline was observed. Her only additional complaint was chronic cough. MRI demonstrated upper vermis and upper spinal cord atrophy. There was no reported family history of neurological disorders. CL_OC_II-1_A1, a male with a 5-year disease duration, carries 49 CAG repeats. He presented with a rapidly progressive parkinsonism–ataxia phenotype characterized by severe ataxia, spasticity, a positive Babinski sign, and moderate-to-severe polyneuropathy confirmed by EMG. Significant cognitive decline was observed alongside other features, including gaze palsy, severe bladder incontinence, constipation, and myoclonus. MRI revealed cerebellar vermis atrophy.

This analysis highlights the overlap of clinical features associated with *ZFHX3* expansions with more common ataxias, such as *RFC1*-related disorders and Friedreich's ataxia making diagnosis challenging when relying solely on the clinical phenotype and testing for a single expansion.

***ZFHX3* Repeat Length Distribution Across Diverse Cohorts**

Although no large-scale long-read sequencing datasets are currently available from individuals of Chilean ancestry, ancestry analysis of our samples aligned with Admixed American (AMR) populations. Chileans exhibit a complex genetic background primarily composed of Native American and European ancestry, with the Native American component closely related to Andean indigenous groups such as the Mapuche. The European ancestry in Chileans predominantly derives from Southern Europe, with some Northern European influence, while a smaller but detectable contribution of African ancestry reflects historical migration patterns in Latin America [21].

To contextualize our findings, we leveraged data from two large long-read sequencing resources: the CARD Long-Read Initiative and the 1000 Genomes (1000G) Long-Read Project. From the CARD Long-Read Initiative, we analyzed 205 control samples of European ancestry from the North American Brain Expression Consortium (NABEC) cohort and 133 control samples of African and African-admixed ancestry from the Human Brain Collection Core (HBCC) cohort[22]. Additionally, we screened 100 samples of mixed ancestry from the 1000G Long-Read Project[23]. To ensure methodological consistency with the Chilean samples, all repeat lengths were analyzed using *vamos*. In the CARD dataset, *ZFHX3* GGC repeat lengths ranged from 11 to 24 in the NABEC cohort and from 18 to 23 in the HBCC cohort. Similarly, in the 1000G dataset, repeat lengths ranged from 18 to 29 **(Figure 4).** To date, this represents the most comprehensive analysis of *ZFHX3* repeat length variability using long-read sequencing. These results establish a robust reference for *ZFHX3* repeat length variation across ancestrally diverse control cohorts, providing a critical framework for comparison with patient samples. Further, our findings support the repeat size threshold of ≥42 proposed by Wallenius et al., reinforcing its potential role in SCA4 pathogenicity[5].

Discussion

Our study represents the largest analysis of the SCA4 STR using long-read sequencing across diverse ancestries, incorporating NABEC (n=205), HBCC (n=133), and a subset from the 1000 Genomes Project (n=100). Despite limited South American datasets, this work expands the known geographic distribution of SCA4, reporting the first confirmed cases outside Northern Europe and reinforcing the need for broader population studies[24]

A key finding is the intergenerational expansion of the *ZFHX3* repeat, with an increase from 49 to 55 units in an affected family. This supports the anticipation phenomenon, a well-documented feature in repeat expansion disorders. Notably, a recent study demonstrated a strong inverse correlation between repeat length and age at onset, with longer repeats associated with earlier disease presentation (CL_PLO_III-1_A2) and greater severity (CL_OC_II-1_A1). Our findings align with this, suggesting that repeat instability may influence disease progression in SCA4[10].

Clinically, all affected individuals exhibited ataxia, polyneuropathy, upper motor neuron signs, and dysautonomia, consistent with SCA4 [25]. However, we identified parkinsonism in one patient, a feature not previously linked to this disorder. Given its occurrence in other SCAs (SCA2, SCA3, SCA6, SCA8, SCA17)[26,27], and the the overlapping neurodegenerative processes affecting multiple brain regions [28] this suggests that SCA4 may have a broader clinical spectrum rather than this being a coincidental finding. Further studies, including neuroimaging and functional assays, are needed to clarify this association.

At the molecular level, we confirmed hypermethylation of the expanded *ZFHX3* allele, suggesting an epigenetic role in disease pathogenesis. Similar methylation changes in *C9orf72*-related ALS/FTD [29,30] and Fragile X syndrome[31] have been associated with transcriptional dysregulation, highlighting the need for further investigation into the functional consequences of methylation in SCA4. Additional follow-up studies are required to determine how the expansion affects *ZFHX3* expression and function. Consistent with previous reports, we identified a common Swedish founder haplotype, estimated to have originated approximately 2,200 years ago. The detection of this haplotype outside Northern Europe, including in our Chilean cohort, suggests a wider historical dispersal of the pathogenic expansion[6].

Despite these findings, limitations remain. The small sample size underscores the need for larger, multi-ethnic studies, and the lack of genomic data from diverse populations limits full assessment of *ZFHX3* variability. Expanding sequencing efforts in underrepresented populations is critical for refining our understanding of SCA4. While long-read sequencing is needed to capture the full spectrum of structural variants, repeat expansions, and complex regions, it can be resource-intensive. An efficient approach would be to implement adaptive sampling[6], enabling targeted enrichment of regions associated with ataxias to comprehensively detect all variant types including SNVs, STRs, and SVs, while optimizing sequencing efficiency and reducing costs. This targeted sequencing strategy could effectively address the current limitations of ataxia panels, providing a comprehensive solution for genetic diagnosis.

In summary, this study expands the knowledge of SCA4 by confirming a shared founder haplotype, documenting repeat expansion instability, and identifying a potential link to parkinsonism. Addressing current limitations through broader genetic studies and improved diagnostic access will be key to advancing understanding and clinical management of SCA4.

Acknowledgements

We sincerely thank the patients and their families for their participation in this study. Their invaluable contributions make this research possible.

We also acknowledge the support of Oxford Nanopore Technologies staff in generating this dataset, in particular A. Markham, J. Anderson, and C. Vacher.

This work was supported in part by the Intramural Research Program of the National Institute on Aging (NIA) and the Center for Alzheimer's and Related Dementias (CARD), within the Intramural Research Program of the NIA and the National Institute of Neurological Disorders and Stroke (ZIANS003154, ZIAAG000538), National Institutes of Health (AG000538). Computational resources were provided by the NIH HPC Biowulf cluster (https://hpc.nih.gov).

Additional support was provided by the National Institutes of Health (DP5OD033357, R01HG011649, 5T32HG000035-29), as well as grants from the Michael J. Fox Foundation, the Medical Research Council, the Wellcome Trust, and the National Institute for Health and Care Research (UCL/UCLH Biomedical Research Centre).

Authors' Roles

P.S.A., A.M., K.D., P.A.J., C.B., Z.C.: Design, execution, analysis, writing, editing of final version of the manuscript.

K.P., L.M., B.B., M.F., M.I., A.M.B., A.B.S., Z.B.A., D.E.M: Execution, analysis, editing of final version of the manuscript.

H.H., M.R., B.G., M.J.P., P.J. S.B.G: : Analysis, editing of final version of the manuscript.

K.J.B.: Supervision, design, execution, analysis, writing, editing of final version of the manuscript.

Financial Disclosure of all authors (for the preceding 12 months)

This work was supported in part by the Intramural Research Program of the National Institute on Aging (NIA) and the Center for Alzheimer's and Related Dementias (CARD), within the Intramural Research Program of the NIA and the National Institute of Neurological Disorders and Stroke (ZIANS003154, ZIAAG000538, AG000538), National Institutes of Health (NIH). Computational resources were provided by the NIH HPC Biowulf cluster (*https://hpc.nih.gov*).
Note: Individuals affiliated with CARD or the Laboratory of Neurogenetics at NIH are marked with an asterisk (*) below, as this is their funding source. None of the authors have received consultancies, honoraria, hold intellectual property rights, or receive royalties

- *Paula Saffie-Awad* – Employment: Clínica Santa María, Santiago, Chile. Grant: Michael J. Fox Foundation. Advisory Boards: Biogen.
- *Pilar Alvarez Jerez** – Employment: Laboratory of Neurogenetics, National Institute on Aging, NIH, Bethesda, MD, USA; Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, University College London, UK.
- *Cornelis Blauwendraat, Andrew B. Singleton** – Employment: Center for Alzheimer's and Related Dementias, National Institute on Aging and National Institute of Neurological Disorders and Stroke, NIH, Bethesda, MD, USA; Laboratory of Neurogenetics, National Institute on Aging, NIH, Bethesda, MD, USA.
- *Abraham Moller**, *Kimberly Paquette**, Laksh Malik*, Breeana Baker*, Paige Jarreau*, Kimberley J. Billingsley* – Employment: Center for Alzheimer's and Related Dementias, NIH, USA.
- *Kensuke Daida**, *Madison Fulcher**, Abigail Miano-Burkhardt* – Employment: Laboratory of Neurogenetics, NIH, Bethesda, MD, USA.
- *Zhongbo Chen* – Employment: UCL Queen Square Institute of Neurology, UK; Great Ormond Street Institute of Child Health, UCL, UK.
- *Zachary B. Anderson* – Employment: University of Washington, Seattle, WA, USA.
- *Sophia B. Gibson* – Employment: Division of Genetic Medicine, Department of Pediatrics, University of Washington, Seattle, WA, USA; Department of Genome Sciences, University of Washington. **Grant Support:** NIH grant 5T32HG000035-29.
- *Danny E. Miller* – Employment: Division of Genetic Medicine, Department of Pediatrics, University of Washington, Seattle, WA, USA; Department of Laboratory Medicine and Pathology and the Brotman Baty Institute for Precision Medicine, University of Washington. **Stock Ownership:** MyOme. **Advisory Boards:** Oxford Nanopore Technologies (ONT), Basis Genetics. **Partnerships:** ONT, Pacific Biosciences, Basis Genetics. **Grant Support:** NIH DP5OD033357.
- *Mariam Isayan* – Employment: National Institute of Health, Yerevan, Armenia.
- *Henry Houlden* – Employment: UCL Queen Square Institute of Neurology, UK. **Grants:** Michael J. Fox Foundation, MRC, Wellcome Trust, NIHR UCL/UCLH BRC.
- *Mina Ryten* – Employment: UCL, UK; University of Cambridge, UK.
- *Bida Gu, Mark JP* – Employment: University of Southern California, USA. **Grant Support:** NIH R01HG011649.

References

1. Chen Z, Tucci A, Cipriani V, Gustavsson EK, Ibañez K, Reynolds RH, et al. Functional genomics provide key insights to improve the diagnostic yield of hereditary ataxia. Brain. 2023 Jul 3;146(7):2869–84.

2. P. Saffie, A. Schuh, D. Muñoz, J. Fernández, F. Canals, P. Chaná-Cuevas. LBA-24: Diagnostic yield of Next Generation Sequencing techniques in a movement disorders center in Chile. In International Congress of Parkinson's Disease and Movement Disorders; Available from: https://www.mdscongress.org/Congress-Files/2022-Madrid-Congress_Late-Breaking-Abstracts-LBA.pdf

3. Flanigan K, Gardner K, Alderson K, Galster B, Otterud B, Leppert MF, et al. Autosomal dominant spinocerebellar ataxia with sensory axonal neuropathy (SCA4): clinical description and genetic localization to chromosome 16q22.1. Am J Hum Genet. 1996 Aug;59(2):392–9.

4. 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. Nat Genet. 2024 Jun;56(6):1080–9.

5. Wallenius J, Kafantari E, Jhaveri E, Gorcenco S, Ameur A, Karremo C, et al. Exonic trinucleotide repeat expansions in ZFHX3 cause spinocerebellar ataxia type 4: A poly-glycine disease. Am J Hum Genet. 2024 Jan 4;111(1):82–95.

6. Chen Z, Gustavsson EK, Macpherson H, Anderson C, Clarkson C, Rocca C, et al. Adaptive long-read sequencing reveals GGC repeat expansion in ZFHX3 associated with spinocerebellar ataxia type 4. Mov Disord. 2024 Mar;39(3):486–97.

7. Matsushima M, Yaguchi H, Koshimizu E, Kudo A, Shirai S, Matsuoka T, et al. FGF14 GAA repeat expansion and ZFHX3 GGC repeat expansion in clinically diagnosed multiple system atrophy patients. J Neurol. 2024 Jun;271(6):3643–7.

8. Shirai S, Mizushima K, Shibata Y, Matsushima M, Iwata I, Yaguchi H, et al. Spinocerebellar ataxia type 4 is not detected in a cohort from Hokkaido, the northernmost island of Japan. J Neurol Sci. 2024 May 15;460(122974):122974.

9. Novis LE, Alavi S, Pellerin D, Della Coleta MV, Raskin S, Spitz M, et al. Unraveling the genetic landscape of undiagnosed cerebellar ataxia in Brazilian patients. Parkinsonism Relat Disord. 2023 Dec 20;119:105961.

10. Dalski A, Pauly MG, Hanssen H, Hagenah J, Hellenbroich Y, Schmidt C, et al. Repeat length in spinocerebellar ataxia type 4 (SCA4) predicts age at onset and disease severity. J Neurol. 2024 Sep 2;271(9):6289–300.

11. Rudaks LI, Yeow D, Kumar KR. SCA4 unravelled after more than 25 years using advanced genomic technologies. Mov Disord. 2024 Mar;39(3):457–61.

12. Saffie P, Baker B, Billingsley KJ. Protocol: Purification of DNA from Whole Blood using the QIAamp Blood Midi Kit (Spin Protocol) + QC with Qu. 2024 Mar 6 [cited 2025 Feb 3]; Available from: https://www.protocols.io/view/protocol-purification-of-dna-from-whole-blood-usin-c7ypzpvn

13. Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021 Nov;18(11):1322–32.

14. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024 Oct;42(10):1571–80.

15. Ren J, Gu B, Chaisson MJP. Vamos: Variable-number tandem repeats annotation using efficient motif sets. Genome Biol. 2023 Jul 27;24(1):175.

16. Vitale D, Koretsky MJ, Kuznetsov N, Hong S, Martin J, James M, et al. GenoTools: an open-source Python package for efficient genotype data quality control and analysis. G3 (Bethesda) [Internet]. 2025 Jan 8;15(1).

Available from: http://dx.doi.org/10.1093/g3journal/jkae268

17. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST. Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. Proc Natl Acad Sci U S A. 2010 Sep 14;107(37):16222–7.

18. Chen Z, Jerez PA, Anderson C, Paucar M, Lee J, Nilsson D, et al. The ZFHX3 GGC Repeat Expansion Underlying Spinocerebellar Ataxia Type 4 has a Common Ancestral Founder. Movement Disorders [Internet]. 2024 Dec 5 [cited 2025 Feb 4]; Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.30077

19. Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. Acta Neuropathol Commun. 2021 May 25;9(1):98.

20. Chen Z, Morris HR, Polke J, Wood NW, Gandhi S, Ryten M, et al. Repeat expansion disorders. Pract Neurol. 2024 Sep 30;n-2023-003938.

21. Adhikari K, Mendoza-Revilla J, Chacón-Duque JC, Fuentes-Guajardo M, Ruiz-Linares A. Admixture in Latin America. Curr Opin Genet Dev. 2016 Dec;41:106–14.

22. Billingsley KJ, Meredith M, Daida K, Jerez PA, Negi S, Malik L, et al. Long-read sequencing of hundreds of diverse brains provides insight into the impact of structural variation on gene expression and DNA methylation [Internet]. bioRxivorg. 2024 [cited 2025 Feb 6]. p. 2024.12.16.628723. Available from: https://www.biorxiv.org/content/10.1101/2024.12.16.628723v1.abstract

23. Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D, et al. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. Genome Res. 2024 Nov 20;34(11):2061–73.

24. Pellerin D, Iruzubieta P, Xu IRL, Danzi MC, Cortese A, Synofzik M, et al. Recent advances in the genetics of ataxias: An update on novel autosomal dominant repeat expansions. Curr Neurol Neurosci Rep. 2025 Jan 16;25(1):16.

25. 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. J Intern Med. 2024 Sep;296(3):234–48.

26. Synofzik M. Parkinsonism in neurodegenerative diseases predominantly presenting with ataxia. Int Rev Neurobiol. 2019 Nov 21;149:277–98.

27. Park H, Kim HJ, Jeon BS. Parkinsonism in spinocerebellar ataxia. Biomed Res Int. 2015 Mar 19;2015:125273.

28. Hellenbroich Y, Gierga K, Reusche E, Schwinger E, Deller T, de Vos RAI, et al. Spinocerebellar ataxia type 4 (SCA4): Initial pathoanatomical study reveals widespread cerebellar and brainstem degeneration. J Neural Transm (Vienna). 2006 Jul;113(7):829–43.

29. Udine E, Finch NA, DeJesus-Hernandez M, Jackson JL, Baker MC, Saravanaperumal SA, et al. Targeted long-read sequencing to quantify methylation of the C9orf72 repeat expansion. Mol Neurodegener. 2024 Dec 21;19(1):99.

30. Russ J, Liu EY, Wu K, Neal D, Suh E, Irwin DJ, et al. Hypermethylation of repeat expanded C9orf72 is a clinical and molecular disease modifier. Acta Neuropathol. 2015 Jan;129(1):39–52.

31. Sutcliffe JS, Nelson DL, Zhang F, Pieretti M, Caskey CT, Saxe D, et al. DNA methylation represses FMR-1 transcription in fragile X syndrome. Hum Mol Genet. 1992 Sep;1(6):397–400.

Tables and figures Legends

Figure 1. a) Schematic overview of the study design. **b)** Waterfall plots displaying ONT long-read sequencing data for the four predicted *ZFHX3* GGC repeat carriers. The red dotted line marks the pathogenic threshold. Created using BioRender.com.

Figure 2: Haplotype analysis of *ZFHX3* GGC repeat expansion carriers. Out of the six rare SNVs reported as part of the distance common founder event, four were found in our samples within the repeat region (Highlighted by the pink box) and compared to the Utah index patient from Chen et al[18]. These SNVs are missing in the unaffected Chilean individual.

Figure 3. Haplotype specific Differential methylation around the expansion. a-d) Modbamtools plots of methylation frequency for the four heterozygous expansion carriers. Methylation frequency is plotted at the top where haplotype 1 (blue) represents the non-expanded allele and haplotype 2 (orange) represents the expanded allele. The *ZFHX3* gene track is overlaid at the top, showing the last two exons. Haplotype-specific reads are shown at the bottom with blue sections of the reads denoting hypomethylation and red sections of the reads denoting hypermethylation. Purple box indicates repeat region. For all four carriers, the expanded allele is hypermethylated compared to the non-expanded allele. **e)** Modbamtools plot of an unaffected related individual, homozygous non-repeat carrier, showing that hypomethylation in this region is expected under normal conditions.

Figure 4. a) Distribution of ZFHX3 GGC repeat lengths in the Chilean samples (n=15), read indicates pathogenic length carrier. **b)** Distribution of ZFHX3 GGC repeat lengths in the 1000G control cohort (n=100) comprising individuals of mixed ancestry. **c)** Distribution of ZFHX3 GGC repeat lengths in the NABEC/HBCC control cohort (n=338) comprising individuals of European and African and African-admixed ancestry. The red dotted line marks the pathogenic threshold.

Table 1. Clinical characteristics of the four *ZFHX3* GGC repeat carriers. Abbreviations: SARA, Scale for the Assessment and Rating of Ataxia; MRI, magnetic resonance imaging; RBD, Rapid Eye Movement sleep behavior disorder, PKN, for parkinsonism, EMG, Electromyography.

Supplementary Figure 1. Scatter plot of principal components 1 and 2 from the Ancestry analysis Cluster plot of PC1 and PC2 of the ancestry analysis showed the Chilean samples clustering close to Latino/admixed American (AMR).

AAC; African American/Afro-Caribbean, AFR; African, AJ; Ashkenazi Jewish, AMR; Admixed American, CAS; Central Asian, EAS; Eastern Asian, EUR; European, FIN; Finnish, MDE; Middle Eastern, SAS; South Asian.

Supplementary figure 2. Inverse Correlation Between *ZFHX3* GGC Repeat Length and Age at Onset in SCA4 Patients. A negative trend is observed, but it is not statistically significant ($R^2 = 0.54$, $p = 0.26$).

Supplementary Tables

Supplementary Table 1. Clinical phenotypes of the individuals studied. Sample ID based on their country (CL for Chile), family ID, generation within the pedigree (Roman numerals: I, II, III, etc.), individual order within that generation (numbered from left to right in the pedigree), and affected status (A for affected, U for unaffected).

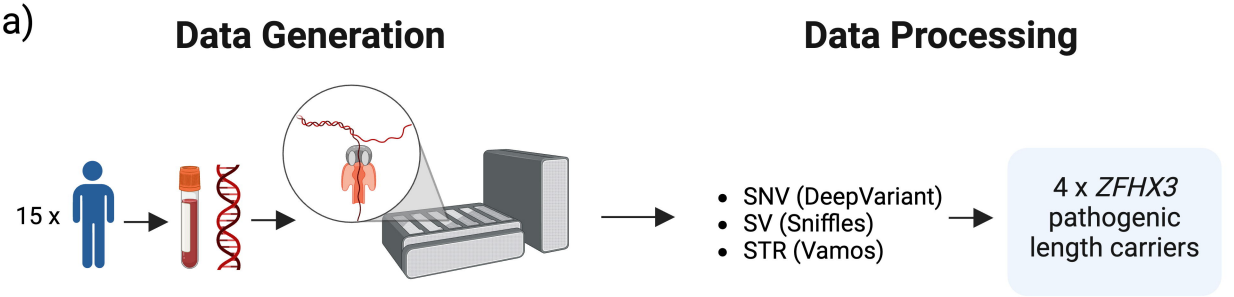
Supplementary Table 2. Overview of long-read sequencing data for the studied.N50 represents the read length at which 50% of total bases are in reads of that length or longer. Yield (Gb) indicates total bases sequenced. Mean coverage refers to the average sequencing depth.

Supplementary Table 3. Overview of lengths per allele of known pathogenic expansion loci. The table includes chromosome (#chr) position, affected gene, associated disease, pathogenic repeat number, and repeat lengths for each individual.

Supplementary Table 4. Summary results of haplotype analysis of six ultra-rare SNVs associated with the Swedish ancestry founder effect[4,18]. Genotypes for the *ZFHX3* carriers at six SNV on chromosome 16 (hg38). Shared haplotypes are highlighted. *Abbreviations:* SNV, single nucleotide variant; hg38, human genome assembly GRCh38.

Figure 1

a)


Data Generation
Data Processing
15 x
SNV (DeepVariant)
SV (Sniffles)
STR (Vamos)
4 x ZFHX3 pathogenic length carriers

b)

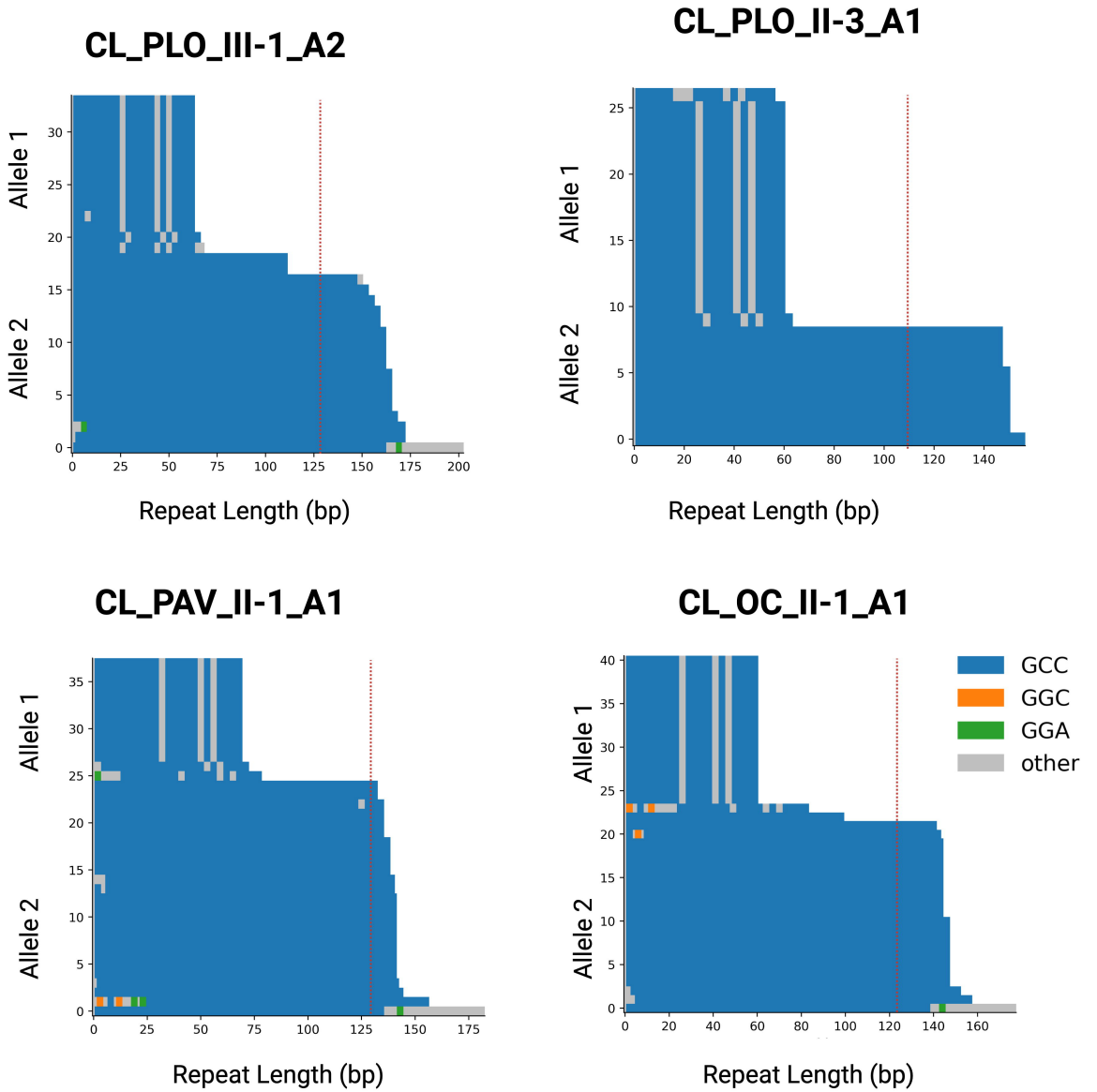

CL_PLO_III-1_A2
CL_PLO_II-3_A1
CL_PAV_II-1_A1
CL_OC_II-1_A1
Allele 1
Allele 2
Repeat Length (bp)
GCC
GGC
GGA
other

Figure 2

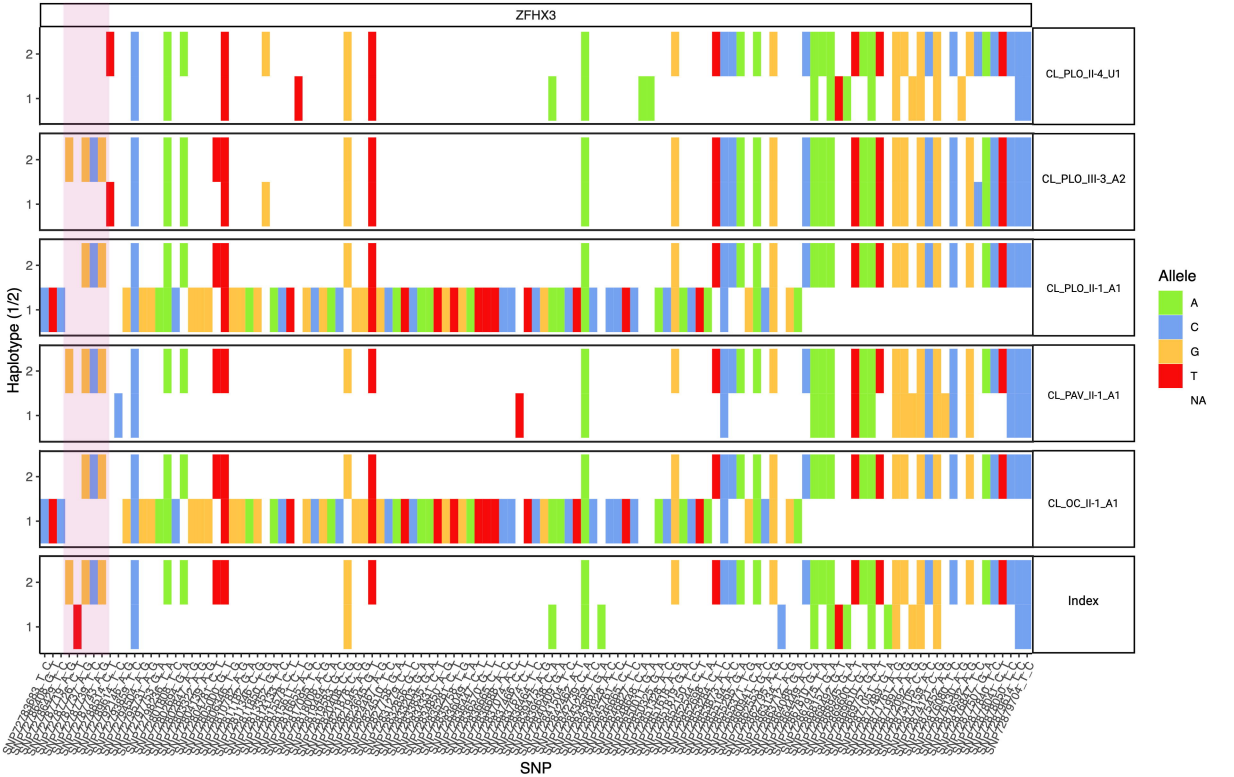

ZFHX3
CL_PLO_II-4_U1
CL_PLO_III-3_A2
CL_PLO_II-1_A1
CL_PAV_II-1_A1
CL_OC_II-1_A1
Index
Haplotype (1/2)
2
1
Allele
A
C
G
T
NA
SNP

Figure 3

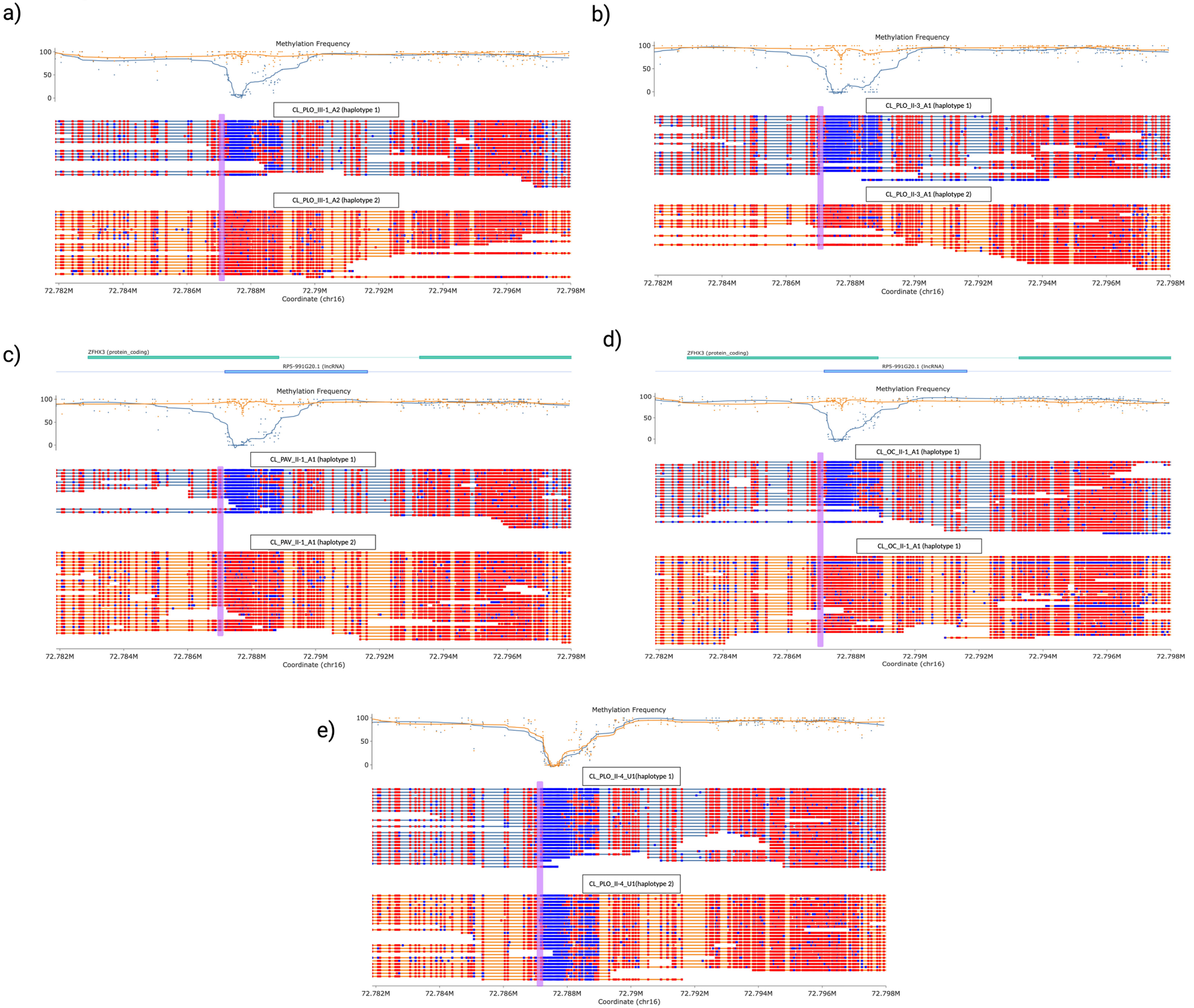
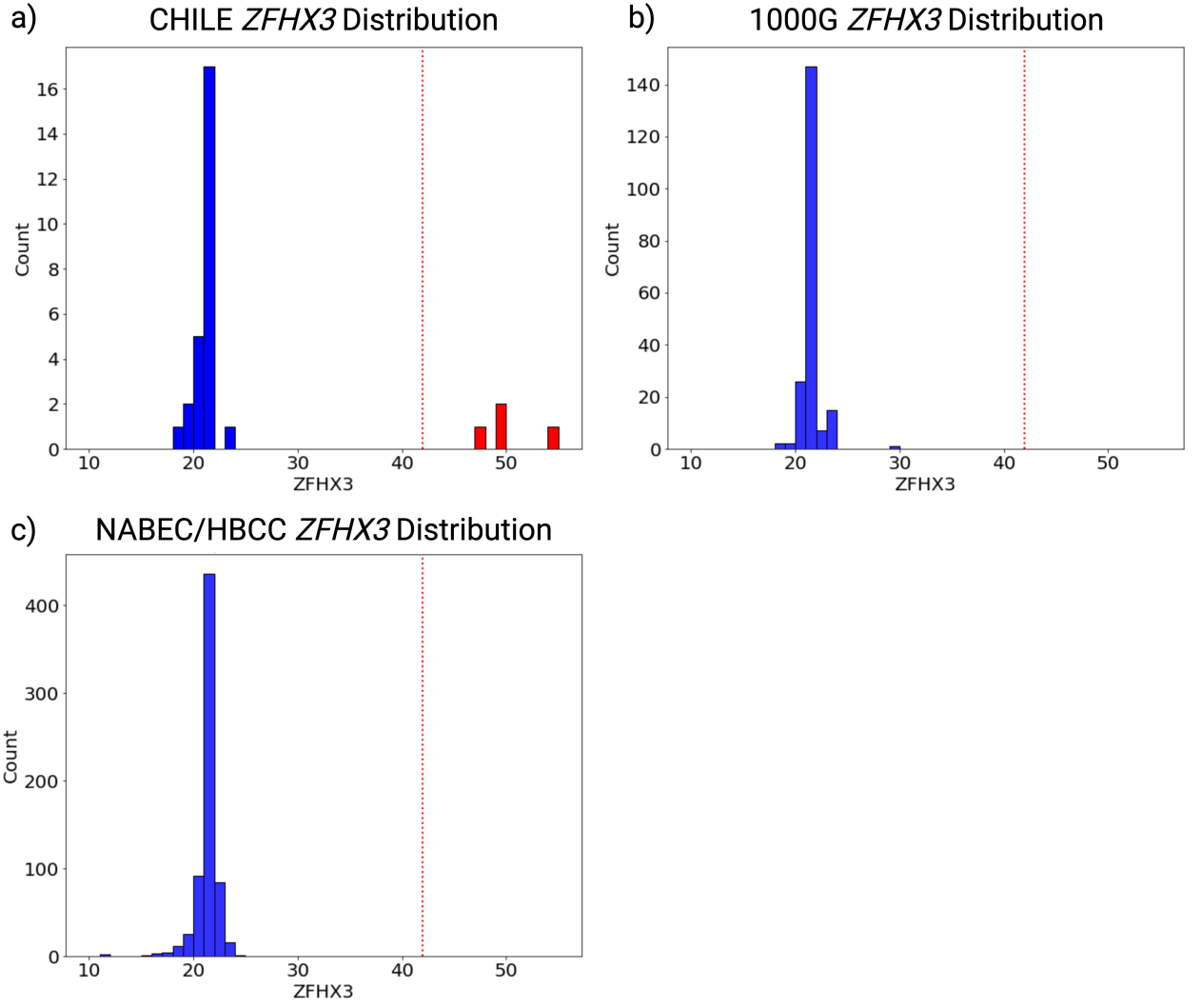

a)
Methylation Frequency
100
50
0
CL_PLO_III-1_A2 (haplotype 1)
CL_PLO_III-1_A2 (haplotype 2)
72.782M
72.784M
72.786M
72.788M
72.79M
72.792M
72.794M
72.796M
72.798M
Coordinate (chr16)
b)
Methylation Frequency
CL_PLO_II-3_A1 (haplotype 1)
CL_PLO_II-3_A1 (haplotype 2)
Coordinate (chr16)
c)
ZFHX3 (protein_coding)
RP5-991G20.1 (lncRNA)
Methylation Frequency
CL_PAV_II-1_A1 (haplotype 1)
CL_PAV_II-1_A1 (haplotype 2)
Coordinate (chr16)
d)
ZFHX3 (protein_coding)
RP5-991G20.1 (lncRNA)
Methylation Frequency
CL_OC_II-1_A1 (haplotype 1)
CL_OC_II-1_A1 (haplotype 1)
Coordinate (chr16)
e)
Methylation Frequency
CL_PLO_II-4_U1(haplotype 1)
CL_PLO_II-4_U1(haplotype 2)
Coordinate (chr16)

Figure 4


a)
CHILE ZFHX3 Distribution
Count
ZFHX3
b)
1000G ZFHX3 Distribution
Count
ZFHX3
c)
NABEC/HBCC ZFHX3 Distribution
Count
ZFHX3