

https://dspace.org/

Analysis of Whole Exome Sequence Data from African Patients with HD-Like Features and No Known HDPhenocopy Syndrome

Naicker, Racilya; Naicker, Racilya

2023

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/43123

Krause, Amanda . (2024). Analysis of Whole Exome Sequence Data from African Patients with HD-Like Features and No Known HDPhenocopy Syndrome [Master's dissertation, University of the Witwatersrand, Johannesburg]. WireDSpace.

Downloaded from a DSpace repository. https://dspace.org/

ANALYSIS OF WHOLE EXOME SEQUENCE DATA FROM AFRICAN PATIENTS WITH HD-LIKE FEATURES AND NO KNOWN HD PHENOCOPY SYNDROME

Racilya Naicker

Student number: 1717757

Supervisor: Dr Fiona Baine-Savanhu

Co-supervisor: Prof Amanda Krause



A research report (in the format of a “submissible” paper) submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Medicine (Genomic Medicine) by Research and Coursework

Johannesburg, 2023

NATIONAL HEALTH LABORATORY SERVICE
School of Pathology, University of the Witwatersrand

DIVISION OF HUMAN GENETICS

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Hospital Street, Johannesburg, 2001 I PO Box 1038, Johannesburg, 2000
[T] +27 11 489 9223 I [M] +27 78 080 8841 I [F] +27 11 489 9226 I [E] human.genetics@nhls.ac.za

29 November 2023

Re: Research report submitted to the Faculty of Health Sciences, University of the Witwatersrand, by Ms Racilya Naicker in partial fulfilment of the requirements for the degree Master of Science in Medicine (Genomic Medicine)

Dear Examiner,

Thank you for agreeing to examine the research report submitted by Ms Racilya Naicker (student number: 1717757) entitled "Analysis of Whole Exome Sequence Data from African Patients with HD-Like Features and No Known HD Phenocopy Syndrome". This research report is being submitted by the candidate in the format of a "submissible" paper as per the Style Guide for Thesis, Dissertation and Research Reports (updated March 2016) by the University of the Witwatersrand Faculty of Health Sciences. This research report contributes 33% towards the final mark of the degree for which the candidate is enrolled.

The candidate and her supervisors have chosen to submit the paper to the journal Neuroepidemiology. This is an international peer-reviewed journal published by Karger Publishers. The submissible paper is therefore written in concordance with the author guidelines of the stipulated journal. These guidelines have been attached for your reference in the appendices section of this document. The only deviation from these guidelines is that the manuscript has been presented as one complete document (including all tables and figures) for ease of marking, rather than being submitted separately.

The research protocol has also been attached as an appendix for the purpose of providing an extended literature review and further contextualise the study. The protocol was assessed and approved by the Faculty of Health Sciences Post-Graduate Assessors committee and is therefore not for examination.

This report is intended solely to record the observations and/or opinion of the writer. It does not constitute a medico-legal report

Chairperson: Prof Eric Buch CEO: Dr Karmani Chetty
Physical Address: 1 Modderfontein Road, Sandringham, Johannesburg, South Africa Postal Address: Private Bag X8, Sandringham, 2131, South Africa
Tel: +27 (0) 11 386 6000/ 0860 00 NHLS(6457) www.nhls.ac.za
Practice number: 5200296

NATIONAL HEALTH LABORATORY SERVICE
School of Pathology, University of the Witwatersrand

DIVISION OF HUMAN GENETICS

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Hospital Street, Johannesburg, 2001 I PO Box 1038, Johannesburg, 2000 [T] +27 11 489 9223 I [M] +27 78 080 8841 I [F] +27 11 489 9226 I [E] human.genetics@nhls.ac.za

Please refer to the Table of Contents for a complete outline of the contents of the research report.

Yours sincerely,

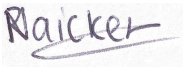
Racilya Naicker (Candidate)

Dr Fiona Baine-Savanhu (Supervisor)

Prof Amanda Krause (Co-supervisor)

Declaration

I, Racilya Naicker, declare that this Research Report (in the format of a "submissible" paper) is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine (Genomic Medicine) by Research and Coursework at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



Racilya Naicker

Student number: 1717757

25th day of February 2024 in Johannesburg

Contribution of Candidate

Student's contribution to the article(s) and agreement of co-author(s)

I, Racilya Naicker, student number 1717757, declare that this Research Report, entitled "Analysis of Whole Exome Sequence Data from African Patients with HD-Like Features and No Known HD Phenocopy Syndrome", is my own work and that I contributed significantly towards the research findings presented in the paper intended for publication below.



27/11/2023

Signature of Student Date

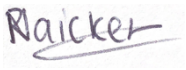


27/11/2023

Signature of Primary Supervisor Date

Agreement by co-authors:

By signing this declaration, the co-authors listed below agree to the use of the research findings intended for a published article by the student as part of their Research Report.

Authors	Name	Signature	Date
1st	Ms Racilya Naicker		27/11/2023
2nd	Dr Fiona Baine-Savanhu		27/11/2023
3rd	Prof Amanda Krause		27/11/2023

Abstract

Huntington disease (HD) is a rare progressive neurodegenerative disorder that results from a CAG repeat expansion within the huntingtin gene (*HTT*). Several syndromes present with HD-like features in the absence of the *HTT* expansion and are termed HD phenocopies. Huntington disease-like 2 (HDL2), a known phenocopy, is most commonly observed in individuals with African, or probable African, ancestry. Therefore, previous diagnostic testing in the Division of Human Genetics, National Health Laboratory Service (Johannesburg, South Africa) screened for both HD and HDL2 in patients with HD-like phenotypes and an indication of African ancestry. Patients who tested negative for both syndromes remain undiagnosed, highlighting the need for further testing strategies. This study aimed to identify variants implicated in the HD-like phenotype of these patients. In a group of nine undiagnosed patients with Black African ancestry, a virtual gene panel was analysed through a whole exome sequencing (WES) approach. The data was filtered, and candidate variants were prioritised according to the frequency, type, and location of the variants as well as *in-silico* pathogenicity prediction scores. A total of 20 candidate variants in 15 genes were shortlisted and classified according to ACMG/AMP guidelines. Notably, variants in the mitochondrial DNA polymerase subunit gamma (*POLG*; c.2246T>C; p.Phe749Ser) and the glutaryl-CoA dehydrogenase (*GCDH*; c.877G>A; p.Ala293Thr) genes were classified as likely pathogenic and pathogenic, respectively. Genotype-phenotype correlation indicated a potential diagnosis of autosomal dominant progressive external ophthalmoplegia in the patient carrying the *POLG* variant, whereas the *GCDH* variant was considered an incidental finding due to a lack of correlation with the characteristics of glutaric aciduria type 1. These findings highlight the diagnostic challenges faced in the African context for patients with HD-like clinical features and call for further validation studies and re-analysis of the WES data using alternative gene panels for the patients for whom no putative pathogenic variants were identified.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr Fiona Baine-Savanhu, for her guidance, invaluable insights, and continuous encouragement throughout the course of this research. Her expertise and thoughtful advice have been fundamental in shaping both the direction and outcome of this work. Equally, I extend my sincere thanks to my co-supervisor, Prof Amanda Krause, whose expertise, and perspective have been instrumental in refining my research approach and analysis.

I am also thankful for the supportive environment provided by everyone at the Division of Human Genetics. Their collective knowledge and willingness to help have contributed significantly to my personal and academic growth.

I would also like to thank my classmates for their constant support. The journey through this Masters course has been a collaborative effort, and the shared experiences and encouragement from my peers have been a source of motivation during challenging times.

My deepest thanks goes to my family and friends, whose love and unwavering support have been my cornerstone. Their belief in my abilities has been a constant source of inspiration and motivation. This journey would not have been possible without them and for that, I am eternally grateful.

Table of Contents

Declaration ..... **iv**

Contribution of Candidate ..... **v**

Abstract ..... **vi**

Acknowledgements ..... **vii**

List of Figures ..... **x**

List of Tables ..... **xi**

List of Abbreviations ..... **xii**

Research report in the format of a "submissible" paper ..... **xiv**

Abstract for Journal Submission **1**

1. Introduction **2**

1.1. Huntington Disease 2

1.2. Huntington Disease Phenocopies 3

1.3. Challenges in the Diagnosis of Patients with HD-like Clinical Features 3

1.4. Next-Generation Sequencing 4

1.5. Relevance of this Study 4

2. Methods **5**

2.1. Workflow 5

2.2. Sample Selection 5

2.3. Virtual Gene Panel Development 7

2.4. Whole Exome Sequencing 7

2.5. Sequence Alignment and Variant Calling 7

2.6. Variant Annotation and Filtering 7

2.7. Variant Prioritisation 8

2.8. Variant Classification 8

2.9. Genotype-Phenotype Correlation 9

3. Results **9**

3.1. Patient Characteristics and Whole Exome Sequencing Summary Statistics 9

3.2. Generation of the Virtual Gene Panel 11

3.3. Identification and Classification of Candidate Variants 11

3.3.1. Pathogenic and Likely Pathogenic Variants 14

3.3.2. Benign, Likely Benign and Uncertain Significance Variants 17

4. Discussion 17
4.1. Pathogenic and Likely Pathogenic Variants 17
4.2. The Whole Exome Sequencing Approach for Patients with HD-like Features 19
4.3. Limitations of the Study 19
4.4. Future Work 20
5. Conclusion 20
Acknowledgement 21
Statement of Ethics 21
Conflict of Interest Statement 21
Funding Sources 21
Author Contributions 21
Data Availability Statement 21
References 22
Appendices 27
Appendix A: Supplementary Tables and Figures 28
Appendix B: Ethics Clearance Certificate 39
Appendix C: Approved Research Protocol 41
Appendix D: Turnitin Report and Plagiarism Declaration 61
Appendix E: Author Guidelines for *Neuroepidemiology* 66

List of Figures

Figure 1. Relative proportions of HD and HD phenocopies. 2
Figure 2. Methodology flow chart for this study. 5
Figure 3. Studies performed in the Division of Human Genetics on this patient group. 6
Figure 4. Proportion of coding sequence variant types across the nine patients following annotation by Ensembl VEP. 11
Figure 5. Inclusion and exclusion criteria for the analysis of the WES data for the patient group, with the number of variants shown at each step. 12
Figure 6. Genomic location of the POLG NM_002693.2: c.2246T>C (p.Phe749Ser) variant. . 15
Figure 7. Genomic location of the GCDH NM_000159.4: c.877G>A (p.Ala293Thr) variant. .. 16

Supplementary Figure 1. Code for Filtering Variants in RStudio 29
Supplementary Figure 2. Candidate Variant Visualisation using IGV 33

List of Tables

Table 1. Clinical Characteristics of the Patient Group (n = 9) 10

Table 2. Details of the Likely Pathogenic and Pathogenic Candidate Variants (n = 2)..... 13

Supplementary Table 1. Virtual Gene Panel (n = 447)..... 28

Supplementary Table 2. Details of the Benign, Likely Benign and Uncertain Significance Candidate Variants (n = 18) 30

List of Abbreviations

ACMG/AMP	American College of Medical Genetics and Genomics/Association for Molecular Pathology
ADCY5	Adenylate Cyclase 5 gene
ATM	Ataxia Telangiectasia Mutated gene
ATN1	Atrophin 1 gene
ATP7B	Atpase Copper Transporting Beta gene
ATXN	Ataxin gene
BAM	Binary Alignment Map
C9orf72	Chromosome 9 Open Reading Frame 72 gene
CADD	Combined Annotation-Dependent Depletion
CP	Ceruloplasmin gene
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
DRPLA	Dentatorubral-Pallidoluysian Atrophy
FATHMM	Functional Analysis Through Hidden Markov Models
FRRS1L	Ferric Chelate Reductase 1 Like gene
FTL	Ferritin Light Chain gene
GA1	Glutaric Aciduria Type 1
GCDH	Glutaryl-CoA Dehydrogenase gene
gnomAD	Genome Aggregation Database
HD	Huntington Disease
HDL	Huntington Disease-Like
HDL2	Huntington Disease-Like 2
HTT	Huntingtin gene
Indel	Insertions and Deletions
JPH3	Junctophilin 3 gene
MAF	Minor Allele Frequency
NBIA	Neurodegeneration with Brain Iron Accumulation
NCBI	National Centre for Biotechnology Information
NGS	Next-Generation Sequencing
PDGFB	Platelet-Derived Growth Factor BB gene
PDGFRB	Platelet-Derived Growth Factor Receptor Beta gene
PEO	Progressive External Ophthalmoplegia
POLG	Mitochondrial DNA Polymerase Gamma gene
PolyPhen-2	Polymorphism Phenotyping Version 2
PRNP	Prion Protein gene
REVEL	Rare Exome Variant Ensemble Learner
RNG216	Ring Finger Protein 216 gene
SCA	Spinocerebellar Ataxia

SIFT	Sorting Intolerant From Tolerant
SLC20A2	Solute Carrier Family 20 Member 2 gene
SNV	Single Nucleotide Variant
TBP	TATA-Box Binding Protein gene
VCF	Variant Call Format
VEP	Variant Effect Predictor
VPS13A	Vacuolar Protein Sorting-Associated Protein 13A gene
VUS	Variant of Uncertain Significance
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
XK	X-Linked Kx Blood Group Antigen gene
XPR1	Xenotropic and Polytropic Retrovirus Receptor 1 gene

Research report in the format of a "submissible" paper

To be submitted to the Neuroepidemiology journal.

Research Article
Analysis of Whole Exome Sequence Data from African Patients with HD-Like Features and No Known HD Phenocopy Syndrome

Racilya Naicker[a], Fiona Baine-Savanhu[a], Amanda Krause[a]

[a] Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, National Health Laboratory Services, Johannesburg, South Africa

Short Title: Analysis of Whole Exome Sequence Data from African Patients With HD-Like Features

Corresponding Author: Racilya Naicker

E-mail address: racilya@gmail.com / 1717757@students.witz.ac.za

Keywords: Huntington Disease, Huntington Disease Phenocopies, Whole Exome Sequencing, Virtual Gene Panel, South Africa

Abstract for Journal Submission

Introduction: Huntington disease (HD) is a rare neurodegenerative disorder characterised by a triad of psychiatric, cognitive, and motor impairments. HD is caused by a CAG repeat expansion in the huntingtin (*HTT*) gene. Approximately 1% of cases worldwide with HD-like features lack this mutation and are termed HD phenocopies. Commonly observed in South Africa, Huntington disease-like 2 (HDL2) is clinically nearly indistinguishable from HD but results from a CTG/CAG repeat expansion in the junctophilin 3 (*JPH3*) gene. Several other HD phenocopy syndromes have been described, with various causative mutations ranging from single nucleotide variants to repeat expansions, thus complicating the diagnosis of HD and its phenocopies. This study aimed to investigate the genetic variants associated with HD-like clinical features in undiagnosed patients of Black African ancestry with HD-like features using a virtual panel to analyse whole exome sequencing (WES) data.

Methods: WES was performed on nine African patients who had previously tested negative for known pathogenic expansions underlying HD and some of its phenocopies. A broad virtual panel of 447 genes was created to filter and prioritise variants, focusing on the identification of rare, disease-causing mutations. Prioritised candidate variants, which were identified by *in-silico* pathogenicity prediction tools, low allele frequency data and published literature, were classified according to ACMG/AMP guidelines prior to genotype-phenotype correlation.

Results: From the 397,269 variants detected, rigorous filtering and prioritisation steps yielded 20 candidate variants in 15 genes. Following classification, a likely pathogenic variant (NM_002693.2: c.2246T>C; NP_002684.1: p.Phe749Ser) in the mitochondrial DNA polymerase subunit gamma (*POLG*) gene was identified in a young patient with chorea; this variant may be associated with autosomal dominant progressive external ophthalmoplegia. A pathogenic variant (NM_000159.4: c.877G>A; NP_000150.1: p.Ala293Thr) in the glutaryl-CoA dehydrogenase (*GCDH*) gene was found in an older patient with a progressive movement disorder. However, this was deemed an incidental finding due to the lack of correlation of the associated disorder, glutaric aciduria type 1, with the patient's clinical features and family history.

Conclusion: This study used a broad virtual panel to identify potential causative variants in WES data from Black African patients with HD-like features. A likely pathogenic and pathogenic variant was detected in *POLG* and *GCDH*, respectively, in two patients. Further studies are required to validate these findings, and to re-analyse the WES data of the patients where no pathogenic variants were found using a different virtual gene panel. This study highlights the complexity of diagnosing patients with HD-like features and the need for broader genetic testing strategies, particularly in the African context where diverse HD-like conditions may be more prevalent than previously recognised.

1. Introduction

1.1. Huntington Disease

Huntington disease (HD) is a rare, progressive neurodegenerative disorder that typically manifests in mid-adulthood. Its global prevalence is estimated to be 4.88 per 100,000 individuals, although this rate is higher in European and North American populations compared to that in Asia and Africa [1], [2]. South African individuals of African ancestry demonstrate a lower prevalence than the Caucasian subpopulation [1], [3]. HD, which is inherited in an autosomal dominant pattern, results from a trinucleotide (CAG) repeat expansion, encoding a glutamine amino acid tract, in the huntingtin (*HTT*) gene. Consequently, this polyglutamine expansion leads to a mutant HTT protein that imparts pathogenicity through the disruption of various cellular functions [4]. This results in the triad of characteristic symptoms associated with HD which are psychiatric manifestations, cognitive impairment, and a movement disorder (chorea) [5]. Around 1% of suspected HD cases present with overlapping clinical features in the absence of the *HTT* expansion and are thus termed HD phenocopies (Figure 1) [6]. There are numerous genetically heterogenous phenocopy syndromes and their relative worldwide frequencies are shown in Figure 1.

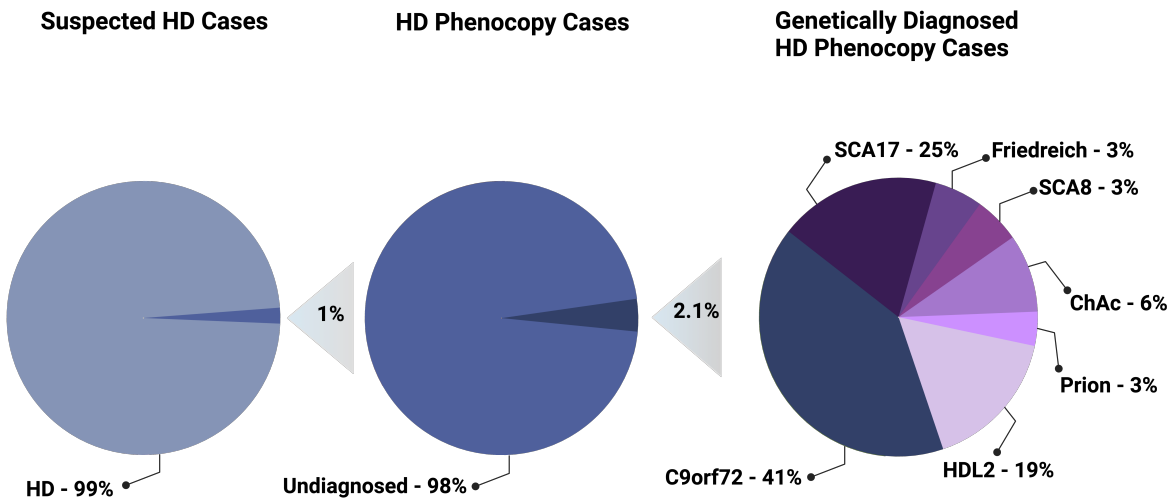


Figure 1. Relative proportions of HD and HD phenocopies.

Relative proportions of HD and HD phenocopies cases (left), successful genetic diagnosis of HD phenocopies (middle), and the distribution of the individual HD phenocopy syndromes (right). Values are based on several global genetic screening studies. HD, Huntington disease; SCA17, spinocerebellar ataxia 17; Friedreich, Friedreich's ataxia; ChAc, chorea-acanthocytosis. SCA8, spinocerebellar ataxia 8; Prion, Prion disease; HDL2, Huntington disease-like 2; C9orf72, chromosome 9 open reading frame 72 mutations. Figure adapted from [7].

1.2. Huntington Disease Phenocopies

Huntington disease-like 2 (HDL2) is phenotypically, nearly indistinguishable from HD throughout its disease progression, however its causative mutation is a CTG/CAG repeat expansion in the junctophilin 3 gene (*JPH3*) [8], [9]. Individuals with African, or probable African, ancestry tend to display a higher relative frequency of HDL2, with South Africa having the highest number of reported HDL2 cases worldwide [10], [11]. A study by Krause et al demonstrated that roughly one-third of South African patients with African ancestry and HD-like features, who had a positive molecular diagnosis, were found to have HDL2 [10].

There are several other HD phenocopy syndromes that present with overlapping clinical features, but with various causative mutations; these are shown in Figure 1. Phenocopies that result from repeat expansions include spinocerebellar ataxias (SCAs), SCA1, SCA2, SCA3, SCA7 and SCA8; HD-like syndromes, HDL1 and HDL4 (SCA17); dentatorubral-pallidoluysian atrophy (DRPLA); and chromosome 9 open reading frame 72 (*C9orf72*) repeat expansions [6], [12]–[14]. Certain HD phenocopies are caused by other types of mutations (loss-of-function, missense, and frameshift mutations) such as chorea-acanthocytosis and neurodegeneration with brain iron accumulation (NBIA) syndromes [6], [15]. Importantly, the underlying genetic mutations responsible for each HD phenocopy syndrome are distinct, ranging from single nucleotide variants (SNVs) to repeat expansions, which occur in different genes.

1.3. Challenges in the Diagnosis of Patients with HD-like Clinical Features

Diagnosing HD phenocopies presents significant challenges, as these conditions are complex and there is a limited understanding of the genetic factors that contribute to the overlapping clinical phenotypes. This challenge is further compounded by the increasing number of conditions as well as other neurological disorders with similar clinical presentations [16]. Therefore, poor, or inadequate clinical phenotyping may also result in misdiagnosis, an issue that is particularly prevalent in patients of African ancestry who are underrepresented in HD and HD phenocopy studies worldwide.

A study by Koriath et al illustrated that within a patient group of 50 individuals where an HD-like phenotype was apparent in the absence of the *HTT* CAG expansion, a diagnosis could only be made for one individual who had an ataxin 1 (*ATXN1*) gene expansion, underlying SCA1 [17]. This emphasises the difficulty in diagnosing patients who present with HD-like features without the *HTT* expansion, as only about 2% of these patients receive a molecular diagnosis (Figure 1) [17]. The Division of Human Genetics, National Health Laboratory Service (NHLS), Johannesburg, South Africa has been involved in

HD diagnosis since the mid-1980s [18]. In South Africa, individuals of African ancestry who display HD-like clinical features, undergo molecular diagnostic testing for expansions in both the *HTT* and *JPH3* genes [10]. Some patients carry neither of these expansions, thereby excluding HD and HDL2 from their diagnosis. A previous study performed in the Division investigated five relevant genes (atrophin 1 (*ATN1*), TATA-box binding protein (*TBP*), *C9orf72*, ataxin 2 (*ATXN2*) and ataxin 7 (*ATXN7*)) in the group of patients for whom HD and HDL2 had been excluded [11]. Of the 105 samples analysed, one individual had an expansion in the *ATXN2* gene, confirming a diagnosis of SCA2 in this patient [11]. In a subsequent study in the Division, based on the work of Mariani et al., 16 undiagnosed patient samples were screened for mutations using a targeted custom next-generation sequencing (NGS) panel comprising 69 genes relevant to neurodegenerative disorders [19]. From this study, three individuals were found to have potential pathogenic mutations in *JPH3* and ATPase copper transporting beta (*ATP7B*) [unpublished data]. Importantly, these studies conducted within the Division demonstrated different outcomes among patients of African ancestry in contrast to the global studies performed on non-African populations. Thus, illustrating the need for broader testing strategies to be employed for this undiagnosed African ancestry patient group.

1.4. Next-Generation Sequencing

Next-generation sequencing is a high-throughput massively parallel technique utilised to obtain the sequence of a specific DNA sample, which can subsequently be compared to the reference genome to investigate the presence of variants [20]. NGS can be used to sequence specific genes or regions (targeted panel sequencing), the coding regions (whole exome sequencing (WES)) or the entire genome (whole genome sequencing (WGS) [20]. While WGS can provide information regarding structural, copy number and SNVs in both non-coding and coding DNA sequences, it is significantly more time-consuming, costly, and requires expertise and computing capacity [20]. On the other hand, WES is cost-effective, provides fewer variants of uncertain significance (VUSs) than WGS, and is able to detect disease-causing mutations, which are mostly present in exons [21]. Importantly, as previously mentioned, the patients in this study have tested negative for certain HD phenocopies caused by repeat expansions. This makes WES the crucial next step to investigate SNVs and insertions and deletions (indels) in specific regions underlying other HD phenocopy syndromes.

1.5. Relevance of this Study

While the diagnosis and research of HD and HD-like syndromes has been carried out in the Division for many years, it has presented several challenges, leaving a number of patients without a molecular diagnosis. This study therefore aimed to identify variants implicated in the HD-like phenotype of these

African patients by utilising a WES-based approach. This was achieved by (i) the development of a virtual gene panel, for the analysis of the WES data; (ii) the identification and classification of pathogenic and likely pathogenic variants that could be contributing to the HD-like features of this patient group; and (iii) performing genotype-phenotype correlations, where possible, for pathogenic and likely pathogenic variants using de-identified patient clinical features.

2. Methods

2.1. Workflow

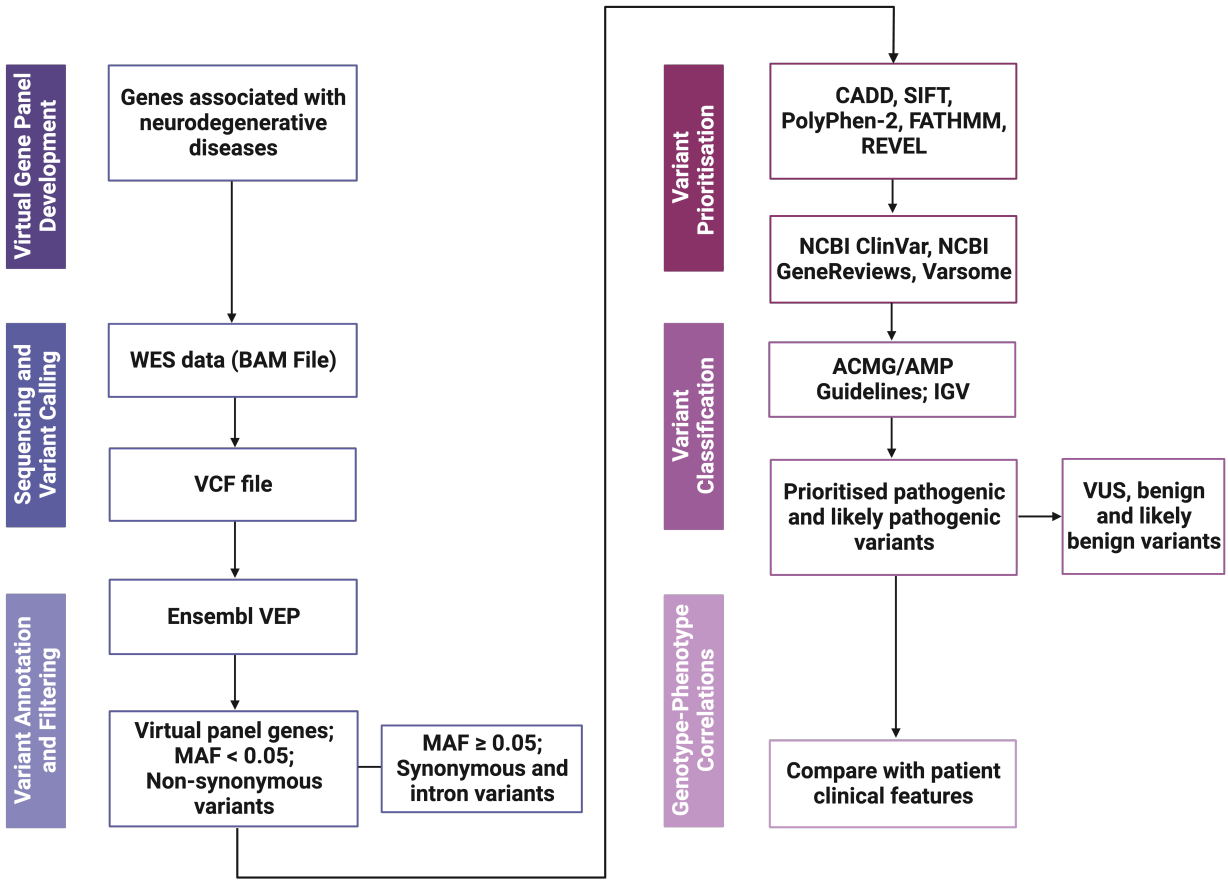


Figure 2. Methodology flow chart for this study.

2.2. Sample Selection

This study utilised de-identified DNA samples from nine unrelated undiagnosed patients of Black African ancestry. These patients were carefully selected to include those who were most likely to have a genetic diagnosis. While there was limited data available, patients with a young age, positive family history and suggestive phenotype were selected for this study. The previous testing strategies

performed on this patient group are shown in Figure 3, where they had previously tested negative for pathogenic expansions in *HTT*, *JPH3*, *ATN1*, *TBP*, *C9orf72*, *ATXN2* and *ATXN7*.

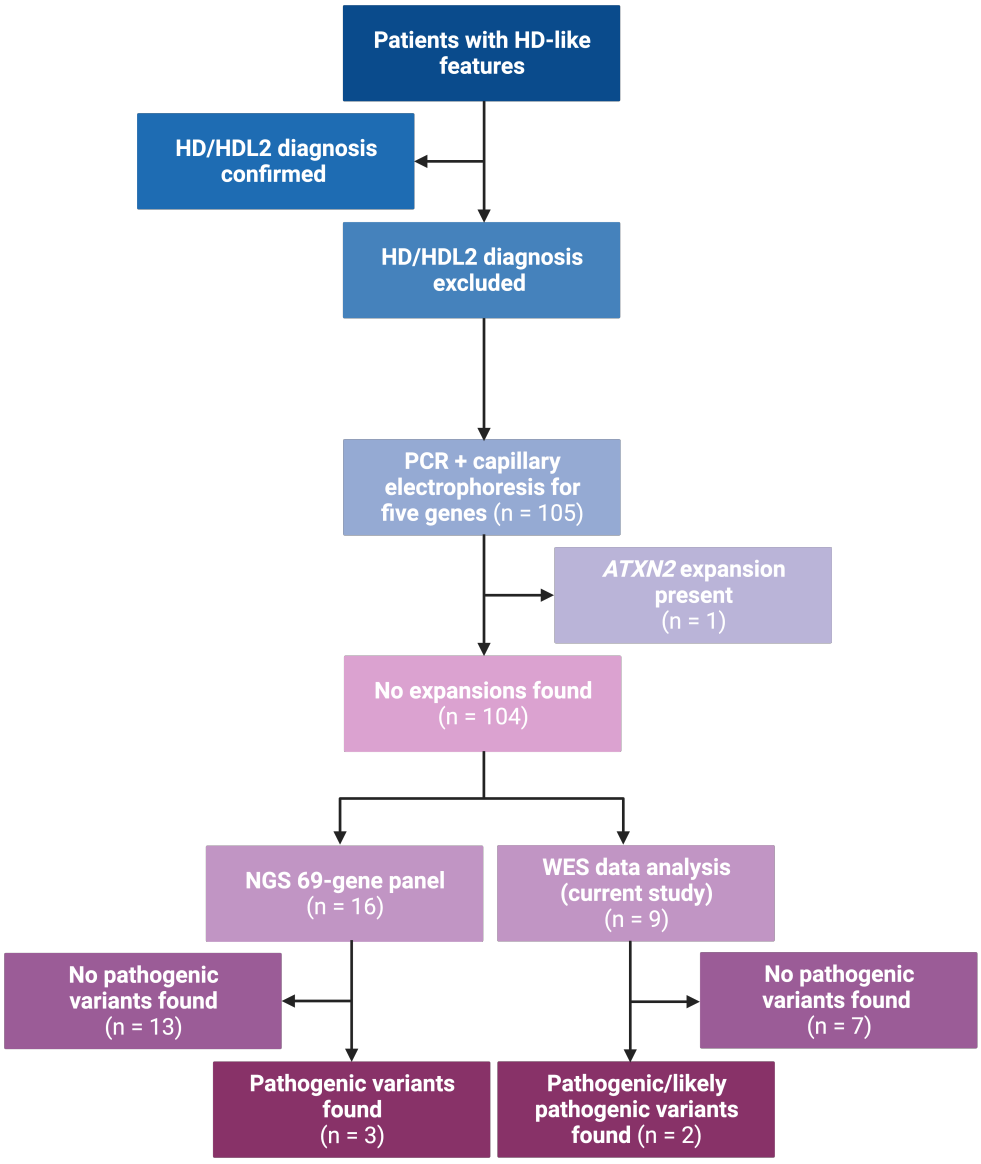


Figure 3. Studies performed in the Division of Human Genetics on this patient group.

The summary of the studies done for the undiagnosed patients with HD-like clinical manifestations. The sample sizes for each study are included. HD/HDL2 diagnosis was achieved by HTT and JPH3 repeat expansion detection. PCR and capillary electrophoresis were performed for ATN1, TBP, C9orf72, ATXN2 and ATXN7 gene expansions. The NGS 69-gene panel study found potentially pathogenic variants in JPH3 and ATP7B [unpublished data]. The current study encompassed the analysis of a virtual gene panel from WES data from the undiagnosed patients, where a likely pathogenic and pathogenic variant were identified in two patients.

2.3. Virtual Gene Panel Development

The generation of the virtual gene panel for this study involved the investigation of genes that are implicated in neurodegenerative disorders with features overlapping those of HD through a review of relevant literature. The literature search was conducted in the National Centre for Biotechnology Information (NCBI) PubMed (https://pubmed.ncbi.nlm.nih.gov/) and was limited to studies from the past ten years. Search terms included "HD phenocopies", "HD-like", "neurodegenerative disorders" and "genes", which were combined with "and" to broaden the search. Further, curated gene-disease panels were selected from Genomics England PanelApp (https://panelapp.genomicsengland.co.uk/). The 69 genes analysed as part of a previous targeted NGS panel study in the Division on a different patient group were also included in the virtual panel developed for this study (Figure 3). Those genes had been selected based on the custom 63-gene panel utilised by Mariani et al and other relevant neurodegenerative disorders [19]. The list encompassing all relevant genes was compiled into the virtual panel, and details are given in Supplementary Table 1 (Appendix A).

2.4. Whole Exome Sequencing

To identify potential disease-causing variants, WES was performed for each patient. The DNA was quantified with the Qubit™ Fluorometer using the Qubit™ dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific, USA). Thereafter, library preparation was performed manually using the Ion AmpliSeq™ Exome RDY Library Preparation Kit (Thermo Fisher Scientific, USA) according to the manufacturer's instructions. The libraries were then templated using the Ion 540™ Kit – Chef (Thermo Fisher Scientific, USA) before being sequenced by the Ion GeneStudio™ S5 Prime Sequencer (Thermo Fisher Scientific, USA).

2.5. Sequence Alignment and Variant Calling

The Ion Torrent instrument software, Torrent Suite™ (Thermo Fisher Scientific, USA), was used to generate the binary alignment map (BAM) files aligned to the reference genome sequence (GRCh37). Variant calling was subsequently performed using the variantCaller plug-in (Thermo Fisher Scientific, USA) and variant call format (VCF) files were produced from each of the BAM files.

2.6. Variant Annotation and Filtering

The VCF files were annotated using the Ensembl Variant Effect Predictor (VEP) tool (https://www.ensembl.org/Tools/VEP) [22]. The variants were then filtered using a custom bioinformatics pipeline in RStudio (version 4.2.2) according to several parameters (see Supplementary

Figure 1; Appendix A). The virtual gene panel (Supplementary Table 1; Appendix A) was applied to select a subset of genes for first-level analysis. Intronic, synonymous and variants with a genome Aggregation Database (gnomAD v4.0.0) African/African American population minor allele frequency (MAF) greater than 0.05 were removed. These filtering steps generated a list of candidate variants that were investigated further.

2.7. Variant Prioritisation

The list of candidate variants was prioritised based on scores from various *in-silico* pathogenicity prediction tools. These tools included: Combined Annotation-Dependent Depletion (CADD), which provides variants (SNVs and indels) with deleteriousness scores [23]; Sorting Intolerant from Tolerant (SIFT) and Polymorphism Phenotyping version 2 (PolyPhen-2), which are used to predict the effect of amino acid substitutions on the structure and functionality of the encoded protein [24], [25]; Functional Analysis Through Hidden Markov Models (FATHMM), which predicts the functional impact of protein missense variants [26]; and the Rare Exome Variant Ensemble Learner (REVEL) tool, which integrates results from 13 pathogenicity prediction tools to generate one pathogenicity prediction score for missense variants [27]. Databases containing information regarding specific genes and/or variants were used to aid variant prioritisation. These included NCBI ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), gnomAD (https://gnomad.broadinstitute.org/), VarSome (https://varsome.com/), NCBI GeneReviews (https://www.ncbi.nlm.nih.gov/books/NBK1116/), UniProt (https://www.uniprot.org/), Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER; https://www.deciphergenomics.org/) and relevant published literature.

2.8. Variant Classification

Based on collected evidence, the prioritised candidate variants were then classified according to the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines [28]. There are five different classes to which variants can be assigned: (1) pathogenic, (2) likely pathogenic, (3) uncertain significance, (4) likely benign, or (5) benign. The quality of pathogenic and likely pathogenic variants was verified using the Integrative Genomics Viewer (IGV) (https://software.broadinstitute.org/software/igv/) to remove the likelihood of false positives.

2.9. Genotype-Phenotype Correlation

The clinical phenotypes of the de-identified patients were made available. Although, the information was limited, this enabled the correlation of the genotype (variant) to the phenotype (clinical feature), where possible.

3. Results

3.1. Patient Characteristics and Whole Exome Sequencing Summary Statistics

Nine undiagnosed patients with Black African ancestry were included in this study, and their clinical characteristics are summarised in Table 1. The group encompassed five female and four male patients, aged between 19 and 69. No age was given for patient 9, while patients 1 and 5 exhibited early-onset conditions (under 18 years), and the remainder manifested adult-onset of symptoms. The age at which each patient underwent diagnostic testing for *HTT* and *JPH3* expansions ranged from 15 to 65 years, and these results are shown in Table 1. The clinical features and family history of each patient are also summarised in Table 1.

All sequencing runs were successful, yielding a mean depth of 127X. A total of 397,269 variants were detected across the nine patients, with an average of 44,141 variants per patient. Of the coding sequence variants, the majority were synonymous, followed by missense, then frameshift and in-frame deletions were least commonly detected, as per Ensembl VEP variant annotations (Figure 4).

234 **Table 1. Clinical Characteristics of the Patient Group (n = 9)**

Patient	Sex	*HTT* Repeat Number	*JPH3* Repeat Number	Age of Diagnostic Testing (Years)	Clinical Features	Family History
1	F	15/16	14/*	17	Chorea.	Abnormal movements observed in mother.
2	F	14/17	6/*	64	Three-year progressive movement disorder, involuntary mood and sleep disturbances. At present cognitive outbursts, generalised choreiform movements and global atrophy.	Two siblings (sister and second brother) and 35-year-old nephew said to present similarly.
3	F	18/20	11/14	65	Choreiform movements, dyskinesia, and neuro-cognitive decline.	ND
4	M	15/16	12/12	61	Cognitive decline since 2011 followed by abnormal behaviour and choreiform movements and psychiatric manifestations.	ND
5	M	16/*	6/10	15	Choreiform movements, cognitive impairment and ocular saccades.	ND
6	M	15/17	11/17	51	Choreiform-athetosis and saccadic eye movements.	ND
7	M	16/21	12/12	33	Cognitive decline, rigidity, and abnormal movements.	ND
8	F	18/*	12/22	45	Chorea, cognitive decline, and involuntary movement disorder.	ND
9	F	15/15	16/17	ND	Choreiform movements, oromandibular dystonia, cognitive impairment ,and chorea.	ND

235 *HTT, huntingtin gene; JPH3, junctophilin 3 gene; F, female; M, male; ND, no data; *, presumed homozygous (no second allele identified).*

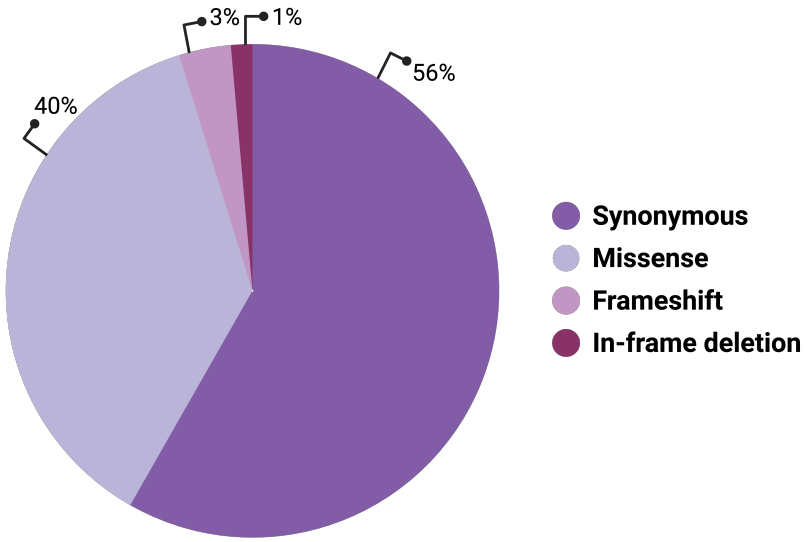


Figure 4. Proportion of coding sequence variant types across the nine patients following annotation by Ensembl VEP.

3.2. Generation of the Virtual Gene Panel

The resulting articles from the extensive literature search in NCBI PubMed were manually inspected for the identification of genes underlying HD phenocopies and neurodegenerative conditions. Genes that were identified from this search were summarised by Schneider and Bird in 2016, and included *HTT*, *JPH3*, prion protein (*PRNP*), *ATXN1*, *ATXN2*, *ATXN3*, *ATXN7*, *ATXN8*, *C9orf72*, *TBP*, *ATN1*, vacuolar protein sorting-associated protein 13A (*VPS13A*), ferritin light chain (*FTL*), adenylate cyclase 5 (*ADCY5*), solute carrier family 20 member 2 (*SLC20A2*), platelet-derived growth factor receptor beta (*PDGFRB*), platelet-derived growth factor BB (*PDGFB*), xenotropic and polytropic retrovirus receptor 1 (*XPR1*), X-Linked Kx blood group antigen (*XK*), ring finger protein 216 (*RNF216*), ferric chelate reductase 1 like (*FRRS1L*), ataxia telangiectasia mutated (*ATM*), and ceruloplasmin (*CP*) [7]. *ATP7B* was also included in this panel due to the previous identification of a potentially pathogenic variant in this gene [unpublished data]. The final virtual panel of 447 genes, shown in Supplementary Table 1 (Appendix A), was compiled using the relevant genes identified from literature, the previous 69-gene panel study and the curated gene-disease panels from Genomics England PanelApp (adult-onset neurodegenerative disorder (Version 4.41) and early onset dementia (encompassing fronto-temporal dementia and prion disease panels) (Version 1.83)) [19].

3.3. Identification and Classification of Candidate Variants

A shortlist of 20 candidate variants in 15 genes was generated and subsequently classified according to the ACMG/AMP guidelines. Figure 5 shows the stepwise process of filtering and prioritisation of the

extensive total number of variants detected to the candidate variants shortlist. A comprehensive search for supporting evidence was conducted using gnomAD, NCBI ClinVar, DECIPHER, VarSome, UniProt, and NCBI GeneReviews to aid in classification. Additionally, IGV was used to inspect each variant to confirm its presence and determine zygosity (Supplementary Figure 1; Appendix A). While five, four and nine candidate variants were classified as benign, likely benign and uncertain significance, respectively (see Supplementary Table 2; Appendix A), two variants were classified as likely pathogenic and pathogenic (Table 2).

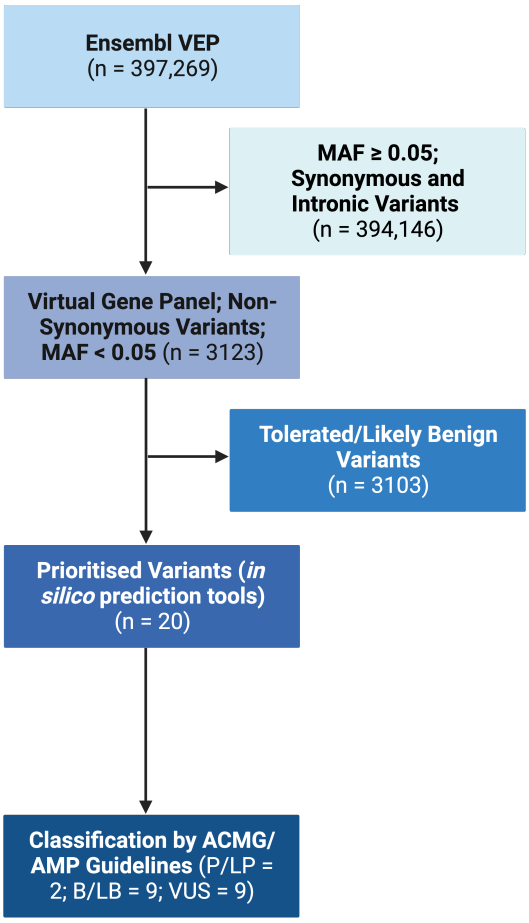


Figure 5. Inclusion and exclusion criteria for the analysis of the WES data for the patient group, with the number of variants shown at each step.

The stepwise process of variant analysis started from the filtering of the total variants identified to only investigate variants in genes included in the virtual gene panel, with MAF less than 5% and non-synonymous variants. Further prioritisation was based on in-silico pathogenicity prediction scores. The shortlist of 20 prioritised candidate variants were classified according to the ACMG/AMP guidelines. VEP, variant effect predictor; MAF, minor allele frequency; ACMG/AMP, American College of Medical Genetics and Genomics and the Association for Molecular Pathology; P/LP, pathogenic/likely pathogenic; B/LB, benign/likely benign; VUS, variant of uncertain significance.

Table 2. Details of the Likely Pathogenic and Pathogenic Candidate Variants (n = 2)

Patient ID	Gene	OMIM ID	RefSeq	Location	Coding mutation	Protein Change	ClinVar ID	Zygosity	SIFT/PolyPhen/CADD/ FATHMM/REVEL	ACMG/AMP Classification	Criteria Met for Classification and Evidence	Disease Associated with Gene
1	*POLG*	174763	NM_002693.3	15:89866654	c.2246T>C	p.Phe749Ser	206520	Het	Deleterious (0)/Probably Damaging (0.999)/29.2/ -4.94/0.959	LP	PM3_Moderate: detected in a compound heterozygous state in Alpers syndrome patients [29], [30]; PM2_Supporting: gnomAD AFR/AA AF = 0.0006811; PP3_Strong: multiple *in-silico* prediction tools suggest pathogenicity; PP5_Supporting: multiple ClinVar submissions suggesting pathogenicity; PP2_Supporting: 12.1% P/LP missense variants vs 1.49% B/LB missense variants	*POLG*-Related Disorders; Mitochondrial Disease
2	*GCDH*	608801	NM_000159.4	19:13007748	c.877G>A	p.Ala293Thr	2083	Het	Deleterious Low Confidence (0)/Probably Damaging (1)/31/ -3.89/0.950	P	PM3_Moderate: homozygous mutation found in individuals with disease [31], [31]; PM2_Supporting: gnomAD AFR/AA AF = 0.0004013; PP3_Moderate: multiple *in-silico* prediction tools suggest pathogenicity; PM1_Strong: variant is present in the binding site of the gene product (functional domain); PP5_Supporting: multiple ClinVar submissions suggesting pathogenicity; PP2_Supporting: 59.4% P/LP missense variants vs 2.9% B/LB missense variants	Glutaric Aciduria Type 1

OMIM, Online Mendelian Inheritance in Man; RefSeq, reference sequence; ACMG/AMP, American College of Medical Genetics and Genomics and the Association for Molecular Pathology; POLG, DNA polymerase subunit gamma; GCDH, glutaryl-CoA dehydrogenase; Het, heterozygous; LP, likely pathogenic; P, pathogenic; B, benign; LB, likely benign; AFR/AA AF, African/African American allele frequency.

3.3.1. Pathogenic and Likely Pathogenic Variants

Patient 1 carried a heterozygous missense variant in the mitochondrial DNA polymerase subunit gamma (*POLG*) gene (OMIM 174763) (Table 2). This variant NM_002693.2: c.2246T>C (NP_002684.1: p.Phe749Ser), shown in Figure 6, was classified as likely pathogenic according to ACMG/AMP guidelines based on the following evidence: the c.2246T>C variant has been identified in a compound heterozygous state in *POLG*-related disorders affected cases (PM3_Moderate) [29], [30]; it was present at a low African/African American allele frequency of 0.0006811 in gnomAD (PM2_Supporting); all five computational tools (SIFT, PolyPhen, CADD, FATHMM and REVEL) predicted a deleterious effect on the encoded protein (PP3_Strong); reputable sources have identified the variant as likely pathogenic or pathogenic as shown by 11 ClinVar submissions (PP5_Supporting); and pathogenic missense mutations were found to be more common than those that are benign in VarSome (PP2_Supporting).

The *POLG* gene encodes mitochondrial DNA polymerase gamma, a crucial component in the replication of mitochondrial DNA [32]. Mutations in *POLG* underlie several disorders including: Alpers-Huttenlocher syndrome, which has a severe phenotype; childhood myocerebrohepatopathy spectrum, which manifests in the first three years of life; myoclonic epilepsy myopathy sensory ataxia; ataxia neuropathy spectrum; autosomal recessive progressive external ophthalmoplegia (PEO); and autosomal dominant PEO [32]. These disorders are part of the spectrum of *POLG*-related conditions [32]. The clinical manifestations observed in affected individuals are broad but can include severe infantile encephalopathy and liver failure to symptoms that appear later in life such as external ophthalmoplegia, ataxia, myopathy, and axonal sensorimotor neuropathy [32]. Epilepsy and movement disorders are also commonly observed [32].

In this study, patient 1 was a 17-year-old female who presented with chorea, a hyperkinetic movement disorder characterised by involuntary and irregular muscle movements [33]. While chorea is typically associated with HD, it is also observed in several other neurological conditions [33]. Patient 1 also presented with a family history of symptoms as abnormal movements were observed in her mother, possibly demonstrating an autosomal dominant inheritance pattern. Since the phenotypic spectrum of *POLG*-related conditions is broad and includes movement disorders, the clinical features noted in patient 1 could be associated with the likely pathogenic variant identified. Autosomal dominant PEO could therefore be a likely diagnosis, but confirmation is required.

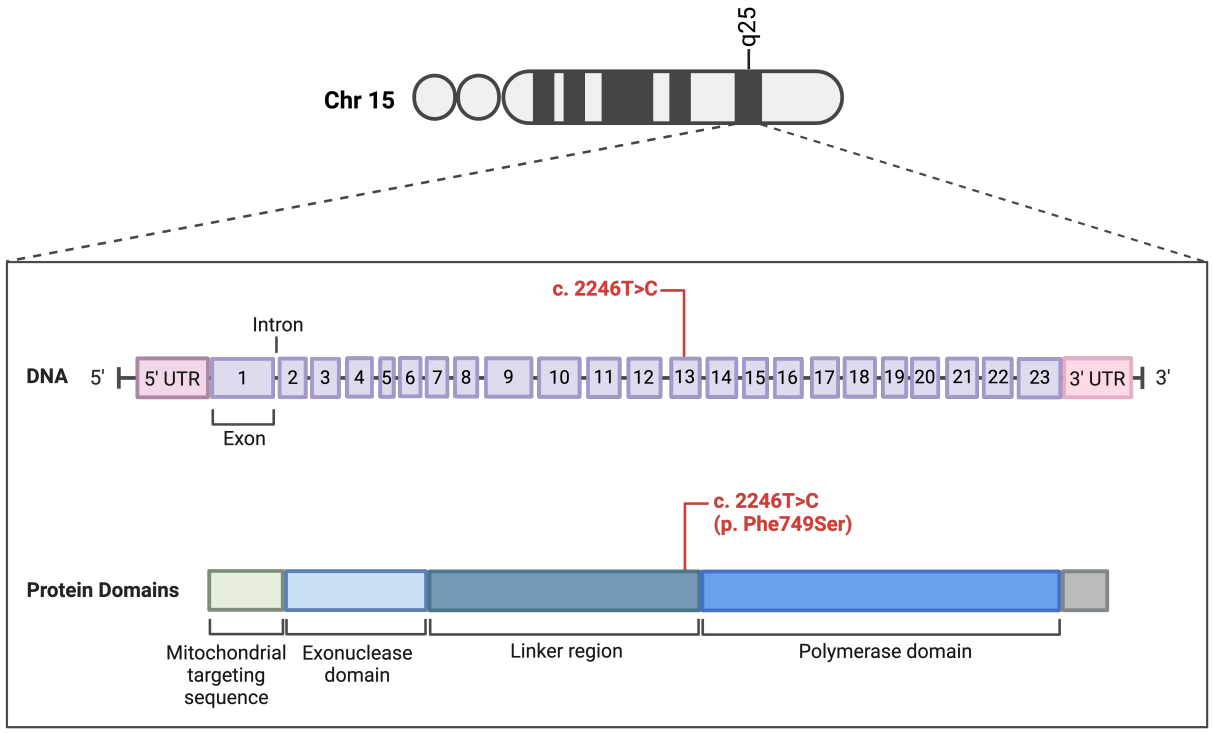


Figure 6. Genomic location of the *POLG* NM_002693.2: c.2246T>C (p.Phe749Ser) variant.

The POLG gene is located on chromosome 15q25, and the NM_002693.2: c.2246T>C (p.Phe749Ser) variant (shown in red) is present in exon 13 of 23. This variant is found in the linker region of POLG, which is between the N-terminal exonuclease and C-terminal polymerase domains. Chr, chromosome; UTR, untranslated region. (Figure created using BioRender https://www.biorender.com/).

Patient 2 had a heterozygous missense mutation (NM_000159.4: c.877G>A; NP_000150.1: p.Ala293Thr) in the glutaryl-CoA dehydrogenase (*GCDH*) gene (OMIM 608801) shown in Figure 7, which was classified as pathogenic (Table 2). Evidence for this classification included that its homozygous state has been detected in glutaric aciduria type 1 (GA1) affected cases (PM3_Moderate) [31], [34]; the variant was located within the binding site of the gene product as determined in UniProt (PM1_Strong); the variant was found to have a low African/African American allele frequency in gnomAD of 0.0004013 (PM2_Supporting); missense variants in this gene were more commonly pathogenic than benign as shown in VarSome (PP2_Supporting); all five *in-silico* prediction tools inferred pathogenicity of this variant (PP3_Moderate); and 10 submissions on ClinVar suggest that the variant is pathogenic (PP5_Supporting).

Mutations in the *GCDH* gene underlie GA1 (OMIM 231670), an autosomal recessive neurometabolic condition with clinical manifestations encompassing macrocephaly, dystonia, frontotemporal atrophy, acute encephalopathic crises during the first six years of life, which can result in striatal injury and consequently, severe dystonic movement disorders [32] . In some patients, neurological conditions

may also develop without clinically apparent crises at any age [35], [36]. In this study, the c.877G>A pathogenic variant was identified in a 64-year-old female patient who presented with a three-year progressive movement disorder (choreiform movements), global atrophy, sleep disturbances and cognitive outbursts. Her sister, second brother and 35-year-old nephew all presented with similar clinical manifestations. Since GA1 is autosomal recessive, either homozygosity or compound heterozygosity would infer the presence of the condition [36]. However, the variant was found to be heterozygous in this patient and no other pathogenic variants in the *GCDH* gene were found; therefore, a direct clinical diagnosis of GA1 in this patient could not be made. While GA1 is common in the South African population with a frequency of 1.7 per 1000 individuals, it typically manifests within the first three years of life, so the late presentation of symptoms in this patient does not correlate with this phenotype [36], [37]. Given the prominent family history of clinical manifestations, an autosomal dominant condition would be the likely cause of disease, which also does not correlate with the GA1 clinical characteristics.

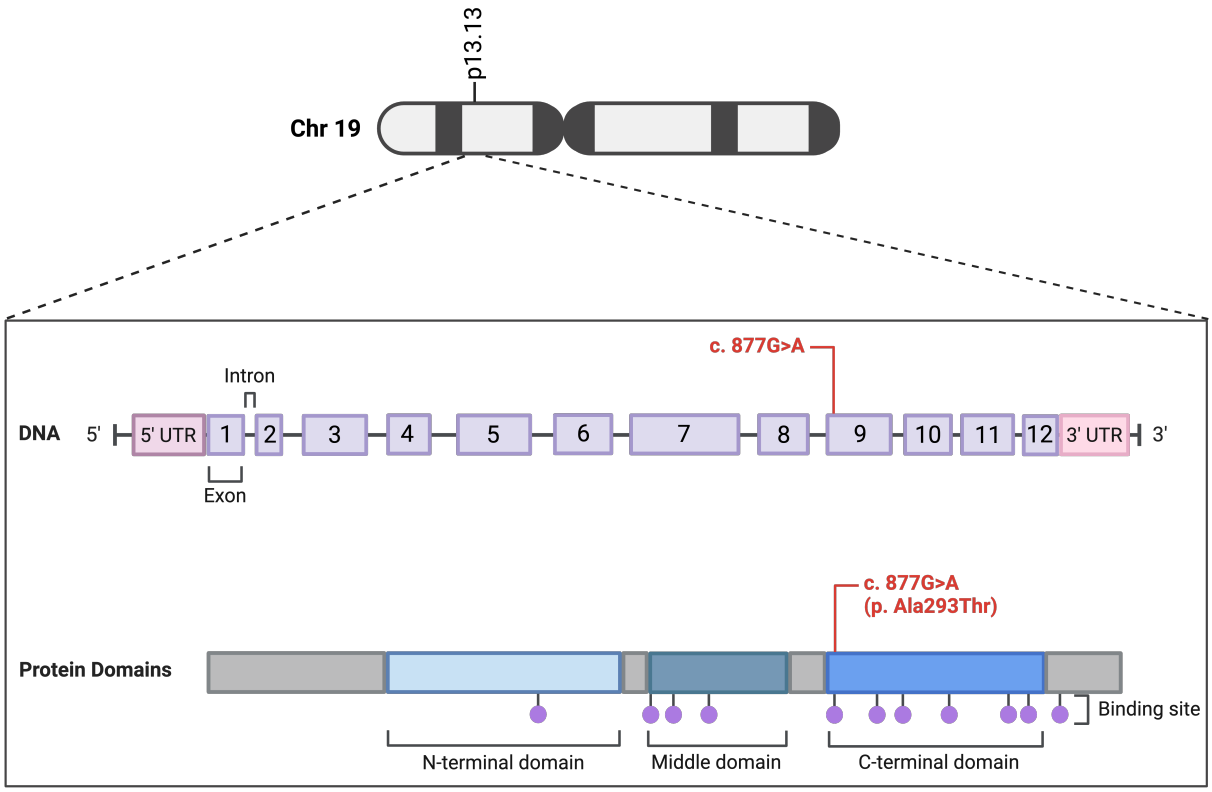


Figure 7. Genomic location of the *GCDH* NM_000159.4: c.877G>A (p.Ala293Thr) variant.

The NM_000159.4: c.877G>A (p.Ala293Thr) variant (shown in red) is located in exon 9 of 12 in the GCDH gene, which is found on chromosome 19p13.13. It is present within a binding site in the C-terminal protein domain. Chr, chromosome; UTR, untranslated region. (Figure created using BioRender https://www.biorender.com/).

3.3.2. Benign, Likely Benign and Uncertain Significance Variants

The details of the prioritised candidate variants classified as benign, likely benign and uncertain significance are shown in Supplementary Table 2 (Appendix A). To determine if patients 1 and 2, who had likely pathogenic and pathogenic variants, were compound heterozygotes, a re-analysis of the *POLG* and *GCDH* genes was conducted. This was done to identify any additional variants that might have been previously overlooked. From this investigation, two heterozygous variants were found in *POLG* (NM_002693.2: c.3708G>T; NP_002684.1: p.Gln1236His and NM_002693.2: c.128A>G; NP_002684.1: p.Gln43Arg), and one homozygous variant was found in *GCDH* (NM_000159.4: c.*165A>G). However, these variants were classified as benign and considered not to be associated with the patients' clinical symptoms.

The classification of VUSs in this study resulted either from conflicting criteria for classification or by insufficient evidence being available. No VUSs identified in this study were determined to be 'hot'. 'Hot' VUSs are those that have substantial supporting evidence and with further information (such as if they were confirmed to be *de novo*) may be reclassified as pathogenic.

4. Discussion

The complexity associated with the overlapping clinical presentations of HD and its phenocopies necessitates a thorough investigative approach to evaluate the underlying genetic causes. This study utilised a virtual gene panel-based strategy to identify potential pathogenic and likely pathogenic variants from WES data generated from undiagnosed African patients with HD-like clinical symptoms. The inclusion of genes relating to both early- and late-onset neurodegenerative conditions in the virtual gene panel allowed for a wide spectrum of genes, that were possibly associated with neurodegenerative phenotypes, to be investigated simultaneously. This approach yielded a considerable number of variants, which, upon filtering, narrowed down to 20 candidate variants across 15 distinct genes. Two variants in two different genes were classified as pathogenic and likely pathogenic in two patients, respectively.

4.1. Pathogenic and Likely Pathogenic Variants

The likely pathogenic variant (c.2246T>C) found in patient 1 within the *POLG* gene, is associated with *POLG*-related disorders. The wide range of mutations (missense and null variants) in various regions (mitochondrial targeting sequence, exonuclease, linker, and polymerase domains) of the POLG protein contribute to the diverse clinical symptoms seen in patients with *POLG* mutations [38]. Autosomal

388 dominant PEO is characterised by parkinsonism, hypogonadism, ataxia, sensorineural hearing loss (of
389 variable degrees), generalised myopathy, depression, cataracts and axonal neuropathy [32]. In this
390 study, the likely diagnosis was autosomal dominant PEO due to the clinical feature (chorea), the family
391 history that displayed a possible autosomal dominant pattern, and a heterozygous likely pathogenic
392 variant detected with no other pathogenic variants found in this gene. However, confirmation of this
393 diagnosis is required by re-examination of the clinical features of the patient and the affected family
394 members to identify any other symptoms consistent with this condition. This would allow for the
395 family history and segregation of the variant to better understand the underlying genetic mechanism
396 of disease. This method of confirmation has proved beneficial in another study, whereby WGS data
397 was analysed to uncover a heterozygous variant (c.3436C>T; p.Arg1146Cys) variant in *POLG*
398 underlying autosomal dominant PEO [39]. The clinical phenotype of the 22-year-old female patient
399 was well-described and the patient's mother, with similar clinical features, was also found to have the
400 same variant [39]. Along with Sanger sequencing, this aided in the confirmation of diagnosis of this
401 condition [39]. A different study also utilised a WES strategy coupled with mitochondrial DNA
402 sequencing, deletion and copy number analysis to successfully confirm a diagnosis of autosomal
403 dominant PEO in a 57-year-old female patient with a loss-of-function frameshift variant
404 (p.Gly23Serfs*236) in *POLG* and determine its impact on mitochondrial DNA depletion [40]. This
405 approach should be considered for the patient in our study to fully characterise the impact of the
406 identified c.2246T>C variant.
407
408 The heterozygous nature of the pathogenic variant (c.877G>A) in the C-terminal domain of the *GCDH*
409 gene detected in patient 2, did not allow for a direct molecular diagnosis of the autosomal recessive
410 condition, GA1. The lack of identification of a second pathogenic variant in *GCDH* could have been due
411 to the lack of a second pathogenic variant in this gene, the presence of the variant in the intronic
412 regions, which were not analysed using this WES approach, or the variant could have been filtered out
413 during analysis. While pathogenic variants are spread across all domains of the GCDH protein, there
414 is an increased density of pathogenic variants in the C-terminal domain, which is linked to a milder
415 form of disease [41]. The c.877G>A variant is clinically relevant, particularly in the Black South African
416 population group. In a study by van der Watt et al., 14 unrelated Black South African GA1 patients
417 were screened for their genetic, clinical, and biochemical features [31]. A homozygous c.877G>A
418 pathogenic mutation in *GCDH* was observed in 11 of the 14 patients. The c.877G>A variant is known
419 to be a severe mutation that, if present on both alleles, results in no residual GCDH enzymatic activity
420 as well as other features associated with GA1 [31], [42]. In contrast, the heterozygous form of the
421 c.877G>A mutation is not known to cause the GA1 phenotype, and the authors found a 1 in 36

heterozygosity rate for this variant in unrelated Black South African patients, which is deemed a founder mutation in this population group [31]. This suggests that the detection of the c.877G>A variant in our patient could have been an incidental finding as the carrier rate is high in this population group, there is no evidence of manifesting heterozygotes, and the phenotypic presentation would be atypical [32]. Nonetheless, to gain a complete diagnosis of GA1 in this patient, further molecular testing of the *GCDH* gene would be required as well as biochemical analysis of the enzymatic activity of GCDH from cultured fibroblasts or leukocytes [36]. This would ascertain whether the variant is in trans with a hypomorphic variant and thus, metabolically confirm its impact on GCDH enzyme activity.

4.2. The Whole Exome Sequencing Approach for Patients with HD-like Features

This study made use of WES data, which was analysed through a broad virtual gene panel, to investigate the presence of pathogenic variants in undiagnosed patients with HD-like symptoms. This method allowed for the efficient analysis of the coding regions, where most disease-causing variants are located. Both known and novel variants that may be associated with the clinical manifestations of the patient could be identified, facilitating a more accurate diagnosis. WES has the potential benefit of identifying mutations in genes not previously associated with HD-like phenotypes, broadening the understanding of the genetic contributors to the disease. In this study, the patients had previously tested negative for repeat expansions in five HD phenocopy syndromes. Thus, a short-read WES approach was selected, as it offers higher base-calling accuracy, thereby enabling a more thorough investigation of SNVs and small indels, which had not previously been explored in these patients. Furthermore, in the resource-limited African setting, WES would be beneficial as it is more cost-effective than WGS and has a wider gene coverage than targeted panels, which are limited to the genes chosen for analysis, resulting in a lower yield.

4.3. Limitations of the Study

There were a number of limitations associated with this study. Firstly, the clinical features of the patients were not completely detailed, which reduced the ability for correlation of the genotype to the phenotype. Secondly, given that this patient group was African, and there is low representation of African genomic data in public databases, there was a larger number of VUSs identified. Another reason behind the identification of VUSs was that certain classification codes could not be applied due to lack of evidence. Since the patients in this study were singletons, *de novo* data (PM6 or PM2) and segregation data (PP1) were not available, conclusions regarding the inheritance and co-segregation of the variants detected could not be made. Lastly, many HD phenocopy syndromes are caused by

pathogenic repeat expansions, which cannot be accurately detected using this short-read WES approach.

4.4. Future Work

Future studies include the validation of the pathogenic and likely pathogenic variants identified in this study using PCR and Sanger sequencing. An advantage of the WES approach is that the data that was generated in this study can be re-analysed. Different genes may be investigated in the patients where no plausible variants were identified. Another approach could employ a long-read sequencing methodology, using the stored patient DNA samples, to combat the issue of short-read sequencing detection of repeat expansions. Importantly, the parents and/or affected family members of the patient should also be included in future work to gain insights into the segregation patterns and heritability of the identified variants. This would also allow for definitive conclusions to be drawn regarding pathogenicity of the identified variants.

5. Conclusion

This study utilised WES data analysed by a virtual gene panel, to investigate the presence of disease-causing variants in undiagnosed patients of Black African ancestry with HD-like clinical features. A likely pathogenic variant (c.2246T>C; p.Phe749Ser) was identified in *POLG*, which possibly underlies autosomal dominant PEO in this patient; however, diagnostic confirmation is required. A pathogenic variant (c.877G>A; p.Ala293Thr) in the *GCDH* gene was also uncovered; however, its presence was thought to be incidental due to the lack of genotype-phenotype correlation for the variant and GA1. Nonetheless, this study highlighted the genetic heterogeneity underlying HD-like conditions as patients with similar features could have different causative mutations. Further studies, encompassing a larger patient group, including the affected and unaffected family members, and in-depth genotype-phenotype correlations, are required to validate these findings and to gain more insights into the complex nature of HD phenocopy syndromes. This could potentially aid in molecular diagnoses being achieved for some patients, thereby alleviating the diagnostic uncertainty they may face. It is also essential for the generation of more knowledge regarding the genetic epidemiology of HD phenocopy syndromes in South Africa.

Acknowledgement

Drs T Ngcungcu, L Lamola and M Mudau assisted R Naicker with WES.

Statement of Ethics

This sub-study protocol was reviewed and approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (approval number M230689), under the parent study (approval number M150368). The ethical clearance certificate is provided in Appendix B.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This study was supported by the National Health Laboratory Service Research Trust Grant. The funder had no role in the design, data collection, data analysis, and reporting of this study.

Author Contributions

R Naicker, F Baine-Savanhu and A Krause designed the study. R Naicker, F Baine-Savanhu and A Krause contributed to acquisition and interpretation of the data. R Naicker analysed the data. R Naicker drafted the manuscript. R Naicker, F Baine-Savanhu and A Krause revised the manuscript for important intellectual content. F Baine-Savanhu and A Krause supervised the research. All authors approved the final version of the manuscript.

Data Availability Statement

Data available upon request.

References

[1] Hayden, M. R., MacGregor, J. M., & Beighton, P. H. (1980). The prevalence of Huntington's chorea in South Africa. *South African Medical Journal*, *58*(5), 193–196.

[2] Medina, A., Mahjoub, Y., Shaver, L., & Pringsheim, T. (2022). Prevalence and Incidence of Huntington's Disease: An Updated Systematic Review and Meta-Analysis. *Movement Disorders*, *37*(12), 2327–2335.

[3] Baine, F. K., Krause, A., & Greenberg, L. J. (2016). The Frequency of Huntington Disease and Huntington Disease-Like 2 in the South African Population. *Neuroepidemiology*, *46*(3), 198–202.

[4] Bates, G. P., Dorsey, R., Gusella, J. F., Hayden, M. R., Kay, C., Leavitt, B. R., Nance, M., Ross, C. A., Scahill, R. I., & Wetzel, R. (2015). Huntington disease. *Nature Reviews Disease Primers*, *1*(1), 1–21.

[5] Podvin, S., Reardon, H. T., Yin, K., Mosier, C., & Hook, V. (2019). Multiple clinical features of Huntington's disease correlate with mutant HTT gene CAG repeat lengths and neurodegeneration. *Journal of Neurology*, *266*(3), 551–564.

[6] Wild, E. J., & Tabrizi, S. J. (2007). Huntington's disease phenocopy syndromes. *Current Opinion in Neurology*, *20*(6), 681–687.

[7] Schneider, S. A., & Bird, T. (2016). Huntington's Disease, Huntington's Disease Look-Alikes, and Benign Hereditary Chorea: What's New? *Movement Disorders Clinical Practice*, *3*(4), 342–354.

[8] Anderson, D. G., Ferreira-Correia, A., Rodrigues, F. B., Aziz, N. A., Carr, J., Wild, E. J., Margolis, R. L., & Krause, A. (2019). Comparison of the Huntington's Disease like 2 and Huntington's Disease Clinical Phenotypes. *Movement Disorders Clinical Practice*, *6*(4), 302–311.

[9] Holmes, S. E., O'hearn, E., Rosenblatt, A., Callahan, C., Hwang, H. S., Ingersoll-Ashworth, R. G., Fleisher, A., Stevanin, G., Brice, A., Potter, N. T., Ross, C. A., & Margolis, R. L. (2001). A repeat expansion in the gene encoding junctophilin-3 is associated with Huntington disease-like 2. *Nature Genetics*, *29*(4), 377–378.

[10] Krause, A., Mitchell, C., Essop, F., Tager, S., Temlett, J., Stevanin, G., Ross, C., Rudnicki, D., & Margolis, R. (2015). Junctophilin 3 (JPH3) Expansion Mutations Causing Huntington Disease Like 2 (HDL2) Are Common in South African Patients with African Ancestry and a Huntington Disease Phenotype. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics : The Official Publication of the International Society of Psychiatric Genetics*, *168*(7), 573–585.

[11] Baine, F. K., Peerbhai, N., & Krause, A. (2018). A study of Huntington disease-like syndromes in black South African patients reveals a single SCA2 mutation and a unique distribution of normal alleles across five repeat loci. *Journal of the Neurological Sciences*, *390*(2018), 200–204.

[12] Schöls, L., Bauer, P., Schmidt, T., Schulte, T., & Riess, O. (2004). Autosomal dominant cerebellar

ataxias: Clinical features, genetics, and pathogenesis. *Lancet Neurology*, *3*(5), 291–304.

[13] Moore, R. C., Xiang, F., Monaghan, J., Han, D., Zhang, Z., Edström, L., Anvret, M., & Prusiner, S. B. (2001). Huntington Disease Phenocopy Is a Familial Prion Disease. *American Journal of Human Genetics*, *69*(6), 1385–1388.

[14] Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F., & Tsuji, S. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral–pallidoluysian atrophy (DRPLA). *Nature Genetics*, *6*(1), 9–13.

[15] Ueno, S. I., Maruki, Y., Nakamura, M., Tomemori, Y., Kamae, K., Tanabe, H., Yamashita, Y., Matsuda, S., Kaneko, S., & Sano, A. (2001). The gene encoding a newly discovered protein, chorein, is mutated in chorea-acanthocytosis. *Nature Genetics*, *28*(2), 121–122.

[16] Nguyen, Q. T. R., Ortigoza Escobar, J. D., Burgunder, J. M., Mariotti, C., Saft, C., Hjermind, L. E., Youssov, K., Landwehrmeyer, G. B., & Bachoud-Lévi, A. C. (2022). Combining Literature Review With a Ground Truth Approach for Diagnosing Huntington's Disease Phenocopy. *Frontiers in Neurology*, *13*, 817753.

[17] Koriath, C. A. M., Guntotoi, F., Norseworthy, P., Dolzhenko, E., Eberle, M., Moss, D. H., Flower, M., Hummerich, H., Rosser, A., Tabrizi, S. J., Mead, S., & Wild, E. J. (2022). Huntington's disease phenocopy syndromes revisited: a clinical comparison and next-generation sequencing exploration. *BioRxiv*, 1–23.

[18] Baine, F. K., Kay, C., Ketelaar, M. E., Collins, J. A., Semaka, A., Doty, C. N., Krause, A., Greenberg, J., & Hayden, M. R. (2013). Huntington disease in the South African population occurs on diverse and ethnically distinct genetic haplotypes. *European Journal of Human Genetics*, *21*, 1120–1127.

[19] Mariani, L. L., Tesson, C., Charles, P., Cazeneuve, C., Hahn, V., Youssov, K., Freeman, L., Grabli, D., Roze, E., No, S., Peuvion, J. N., Bachoud-Levi, A. C., Brice, A., Stevanin, G., & Durr, A. (2016). Expanding the Spectrum of Genes Involved in Huntington Disease Using a Combined Clinical and Genetic Approach. *JAMA Neurology*, *73*(9), 1105–1114.

[20] Bertier, G., Hétu, M., & Joly, Y. (2016). Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views. *BMC Medical Genomics*, *9*(1), 52.

[21] Klein, H. G., Bauer, P., & Hambuch, T. (2014). Whole genome sequencing (WGS), whole exome sequencing (WES) and clinical exome sequencing (CES) in patient care. *LaboratoriumsMedizin*, *38*(4), 221–230.

[22] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 1–14.

[23] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894.

[24] Ng, P. C., & Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814.

[25] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249.

[26] Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, *34*(1), 57.

[27] Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., & Karyadi, D. (2016). REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, *99*(4), 877–885.

[28] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *17*(5), 405–424.

[29] Nguyen, K. V., Sharief, F. S., Chan, S. S. L., Copeland, W. C., & Naviaux, R. K. (2006). Molecular diagnosis of Alpers syndrome. *Journal of Hepatology*, *45*(1), 108–116.

[30] Algahtani, H., Shirah, B., Alsaggaf, K., Al-Qahtani, M. H., Abdulkareem, A. A., Naseer, M. I., & Abuzinadah, A. R. (2021). Chronic progressive external ophthalmoplegia in a Saudi patient with a mutation in the *POLG* gene successfully managed with bilateral frontalis sling. *Journal of Genetic Medicine*, *18*(2), 121–126.

[31] van der Watt, G., Owen, E. P., Berman, P., Meldau, S., Watermeyer, N., Olpin, S. E., Manning, N. J., Baumgarten, I., Leisegang, F., & Henderson, H. (2010). Glutaric aciduria type 1 in South Africa-high incidence of glutaryl-CoA dehydrogenase deficiency in black South Africans. *Molecular Genetics and Metabolism*, *101*(2–3), 178–182.

[32] Rahman, S., & Copeland, W. C. (2019). POLG-related disorders and their neurological manifestations. *Nature Reviews Neurology*, *15*(1), 40–52.

[33] Ehrlich, D. J., & Walker, R. H. (2017). Functional neuroimaging and chorea: a systematic review.

Journal of Clinical Movement Disorders, *4*(1), 1–15.

[34] Busquets, C., Merinero, B., Christensen, E., Gelpí, J. L., Campistol, J., Pineda, M., Fernández-Alvarez, E., Prats, J. M., Sans, A., Arteaga, R., Martí, M., Campos, J., Martínez-Pardo, M., Martínez-Bermejo, A., Vaquerizo, J., Orozco, M., Ugarte, M., Coll, M. J., & Ribes, A. (2000). Glutaryl-CoA Dehydrogenase Deficiency in Spain: Evidence of Two Groups of Patients, Genetically, and Biochemically Distinct. *Pediatric Research*, *48*(3), 315–322.

[35] Boy, N., Mühlhausen, C., Maier, E. M., Heringer, J., Assmann, B., Burgard, P., Dixon, M., Fleissner, S., Greenberg, C. R., Harting, I., Hoffmann, G. F., Karall, D., Koeller, D. M., Krawinkel, M. B., Okun, J. G., Opladen, T., Posset, R., Sahm, K., Zschocke, J., … vom Dahl, S. (2017). Proposed recommendations for diagnosing and managing individuals with glutaric aciduria type I: second revision. *Journal of Inherited Metabolic Disease*, *40*(1), 75–101.

[36] Healy, L., O'Shea, M., McNulty, J., King, G., Twomey, E., Treacy, E., Crushell, E., Hughes, J., Knerr, I., & Monavari, A. A. (2022). Glutaric aciduria type 1: Diagnosis, clinical features and long-term outcome in a large cohort of 34 Irish patients. *JIMD Reports*, *63*(4), 379.

[37] Govender, R., Mitha, A., & Mubaiwa, L. (2017). A review of patients with glutaric aciduria type 1 at Inkosi Albert Luthuli Central Hospital, Durban, South Africa. *SAMJ: South African Medical Journal*, *107*(3), 201–204.

[38] Wong, L. J. C., Naviaux, R. K., Brunetti-Pierri, N., Zhang, Q., Schmitt, E. S., Truong, C., Milone, M., Cohen, B. H., Wical, B., Ganesh, J., Basinger, A. A., Burton, B. K., Swoboda, K., Gilbert, D. L., Vanderver, A., Saneto, R. P., Maranda, B., Arnold, G., Abdenur, J. E., … Copeland, W. C. (2008). Molecular and clinical genetics of mitochondrial diseases due to POLG mutations. *Human Mutation*, *29*(9), E150–E172.

[39] Rossi, M., Medina Escobar, A., Radrizzani, M., Tenembaum, S., Perandones, C., & Merello, M. (2017). Dystonia in a Patient with Autosomal-Dominant Progressive External Ophthalmoplegia Type 1 Caused by Mutation in the POLG Gene. *Movement Disorders Clinical Practice*, *4*(2), 266–269.

[40] Kurtz, J., Fernandes, J. A., Mansukhani, M., Copeland, W. C., & Naini, A. B. (2021). Whole-Exome Sequencing Identifies a Novel POLG Frameshift Variant in an Adult Patient Presenting with Progressive External Ophthalmoplegia and Mitochondrial DNA Depletion. *Case Reports in Genetics*, *2021*, 1–7.

[41] Schuurmans, I. M. E., Dimitrov, B., Schröter, J., Ribes, A., de la Fuente, R. P., Zamora, B., van Karnebeek, C. D. M., Kölker, S., & Garanto, A. (2023). Exploring genotype–phenotype correlations in glutaric aciduria type 1. *Journal of Inherited Metabolic Disease*, *46*(3), 371–390.

[42] Christensen, E., Ribes, A., Merinero, B., & Zschocke, J. (2004). Correlation of genotype and

phenotype in glutaryl-CoA dehydrogenase deficiency. *Journal of Inherited Metabolic Disease*, *27*(6), 861–868.

Appendices

Appendix A: Supplementary Tables and Figures

Supplementary Table 1. Virtual Gene Panel (n = 447)

Supplementary Table 2. Details of the Benign, Likely Benign and Uncertain Significance Candidate Variants (n = 18)

Supplementary Figure 1. Code for Filtering Variants in RStudio

Supplementary Figure 2. Candidate Variant Visualisation using IGV

Appendix B: Ethics Clearance Certificate

Appendix C: Approved Research Protocol

Appendix D: Turnitin Report and Plagiarism Declaration

Appendix E: Author Guidelines for *Neuroepidemiology*

Appendix A: Supplementary Tables and Figures

Supplementary Table 1. Virtual Gene Panel (n = 447)

*AAAS, AARS, ABCB7, ABCD1, ABHD12, ACTB, ADAR, ADCY5, AFG3L2, AIMP1, ALAS2, ALDH18A1, ALDH4A1, ALS2, AMPD2, ANG, ANO10, ANO3, ANXA11, AP1S2, AP4B1, AP4E1, AP4M1, AP4S1, AP5Z1, APP, APTX, AR, ARG1, ARHGEF28, ARL6IP1, ARSA, ARSI, ARX, ATCAY, ATL1, ATM, ATN1, ATP13A2, ATP1A2, ATP1A3, ATP2B3, ATP5A1A, ATP6AP2, ATP7B, ATP8A2, ATXN1, ATXN10, ATXN2, ATXN3, ATXN7, ATXN8, AUH, B4GALNT1, BCAP31, BEAN1, BSCL2, C12orf65, C19orf12, C9orf72, CA8, CACNA1A, CACNA1G, CACNB4, CAMTA1, CAPN1, CASK, CCDC88C, CCNF, CCT5, CDK16, CHCHD10, CHCHD2, CHMP1A, CHMP2B, CHRM4, CIZ1, CLCN2, CLN6, CLP1, COASY, COG5, COL4A1, COL4A2, COL6A3, COQ2, COQ8A, COX20, CP, CSF1R, CST3, CSTB, CTCF, CTSA, CTSF, CWF19L1, CYP27A1, CYP2U1, CYP7B1, DAB1, DAO, DARS, DARS2, DCAF17, DCTN1, DDC, DDHD1, DDHD2, DLAT, DMXL2, DNAH10, DNAH11, DNAJB2, DNAJC13, DNAJC19, DNAJC5, DNAJC6, DNAJC7, DNMT1, DRD2, DRD5, DSTYK, DYNC1H1, DYRK3, EARS2, ECI1, EIF2B1, EIF2B2, EIF2B3, EIF2B4, EIF2B5, EIF4G1, ELOVL4, ELOVL5, ENTPD1, EPM2A, ERBB4, ERCC6, ERLIN1, ERLIN2, EWSR1, EXOSC3, FA2H, FAM154B, FARS2, FASTKD2, FBXO7, FGF14, FIG4, FLVCR1, FMR1, FOLR1, FOXG1, FOXRED1, FRRS1L, FTL, FUS, FUT8, FXN, GAD1, GAMT, GBA, GBA2, GBE1, **GCDH**, GCH1, GFAP, GIGYF2, GJC2, GLA, GLRA1, GLRB, GLT8D1, GNAL, GNAO1, GNS, GOSR2, GPAA1, GRID2, GRM1, GRN, GSN, GTF2F2, HACE1, HEXA, HEXB, HFE, HNRNPA1, HNRNPA2B1, HPCA, HPRT1, HSPA1L, HSPD1, HTRA1, HTRA2, HTT, IBA57, IGF2R, IPPK, ITM2B, ITPR1, JPH3, JPH4, KCNA1, KCNC3, KCND3, KCNJ10, KCNK18, KCNQ2, KCNQ3, KDM5C, KIAA1161, KIDINS220, KIF1A, KIF1C, KIF5A, KLC4, KMT2B, L1CAM, L2HGDH, LAMA3, LAMA4, LAMB1, LRP1B, LRRK2, LYST, MAG, MAP3K10, MAPT, MARS, MARS2, MAT1A, MATR3, MCOLN1, MECR, MMACHC, MMADHC, MPV17, MR1, MRE11, MT-ATP6, MT-ND6, MTPAP, MTTP, MVK, MYH10, NAGLU, NCOR1, NCOR2, NDUFA12, NEDD4, NEFH, NEK1, NEK9, NHLRC1, NIPA1, NKX2-1, NKX6-2, NOP56, NOTCH3, NPC1, NPC2, NR4A2, NT5C2, OPA3, OPHN1, OPTN, PANK2, PARK7, PAX2, PAX6, PCDH12, PCLO, PDGFB, PDGFRA, PDGFRB, PDHX, PDYN, PEX16, PFN1, PGAP1, PIK3R5, PINK1, PLA2G6, PLP1, PMPCA, PNKD, PNKP, PNPLA6, PNPT1, **POLG**, POLR3A, PPP2R2B, PRICKLE1, PRKCG, PRKN, PRKRA, PRNP, PRPH, PRRT2, PSAP, PSEN1, PSEN2, PTEN, PTS, PXDN, QDPR, RAB39B, RAB3GAP2, RARS2, REEP1, REEP2, RELN, REPS1, RNASEH2A, RNASEH2B, RNASEH2C, RNF170, RNF216, RTN2, RUBCN, SACS, SAMHD1, SAR1B, SCN1A, SCN8A, SCN9A, SCP2, SDHAF1, SDK1, SEPSECS, SERAC1, SETX, SGCE, SIGMAR1, SIL1, SLC16A2, SLC19A3, SLC1A3, SLC1A4, SLC20A2, SLC24A2, SLC25A46, SLC2A1, SLC30A10, SLC33A1, SLC39A14, SLC41A1, SLC46A1, SLC52A1, SLC52A2, SLC52A3, SLC6A3, SLC6A5, SLC9A6, SLIT2, SNCA, SNCAIP, SNCB, SNX14, SOD1, SORL1, SPART, SPAST, SPG11, SPG21, SPG7, SPR, SPTBN2, SPTLC1, SQSTM1, SRD5A3, SS18L1, STUB1, SUCLA2, SUOX, SYNE1, SYNJ1, SYT14, TAF1, TAF15, TANC1, TARDBP, TBK1, TBP, TDP1, TECPR2, TET2, TFG, TGM6, TH, THAP1, TIA1, TIMM8A, TMEM240, TOR1A, TPK1, TPP1, TRAP1, TREM2, TREX1, TSEN2, TSEN34, TSEN54, TTBK2, TTC19, TTPA, TTR, TUBA4A, TUBB4A, TWNK, TYROBP, UBQLN2, UBR4, UCHL1, UNC13A, UQCRC1, USP8, VAC14, VAMP1, VAPB, VCP, VEGFA, VLDLR, VPS13A, VPS13C, VPS13D, VPS35, VPS37A, VPS53, VRK1, WASHC5, WDR45, WDR45B, WDR48, WDR73, WDR81, WFS1, WWOX, XK, XPR1, XRCC1, YY1, ZEB2, ZFYVE26, ZFYVE27, ZNF592*

Supplementary Figure 1. Code for Filtering Variants in RStudio

```
# Install package
#install.packages("dplyr")

# Load packages
library(dplyr)
library(writexl)

# Generate the list of files to be filtered
vep_analysis_file_paths <- list.files(path = "/Users/racilya/Desktop/Racil
ya/2023/Research Project/Results/VEP Analysis/vep_files_directory", patter
n = ".txt", full.names = TRUE)

# Read virtual gene panel file and trim white space
gene_panel <- trimws(readLines("/Users/racilya/Desktop/virtual_gene_panel.
txt"))

# Filtering each Ensembl VEP analysis file simultaneously
for (file_path in vep_analysis_file_paths) {
  ## Read the variant annotation file
  variant_data <- read.csv(file_path, sep = "\t")

  ## Filter the data based on the virtual gene panel
  filtered_data <- variant_data %>%
    filter(SYMBOL %in% gene_panel)

  ## Filter the data based on MAF and type of variant
  filtered_data2 <- filtered_data %>%
    filter(`gnomADe_AFR_AF` <= 0.05,
           !(`Consequence` %in% c("synonymous_variant", "intron_variant,no
n_coding_transcript_variant", "intron_variant")))

  ## Add ' to exon number column to keep correct format in Excel
  filtered_data2$EXON <- paste0("'",filtered_data2$EXON)

  ## Get the file name (without path)
  file_name <- basename(file_path)

  ## Define the output file path with .xlsx extension
  output_file <- paste0("filtered_", tools::file_path_sans_ext(file_name),
 ".xlsx")

  ## Save the filtered data as an .xlsx file
  write_xlsx(filtered_data2, path = output_file)

  cat("Filtered data saved to:", output_file, "\n")
}
```

Supplementary Table 2. Details of the Benign, Likely Benign and Uncertain Significance Candidate Variants (n = 18)

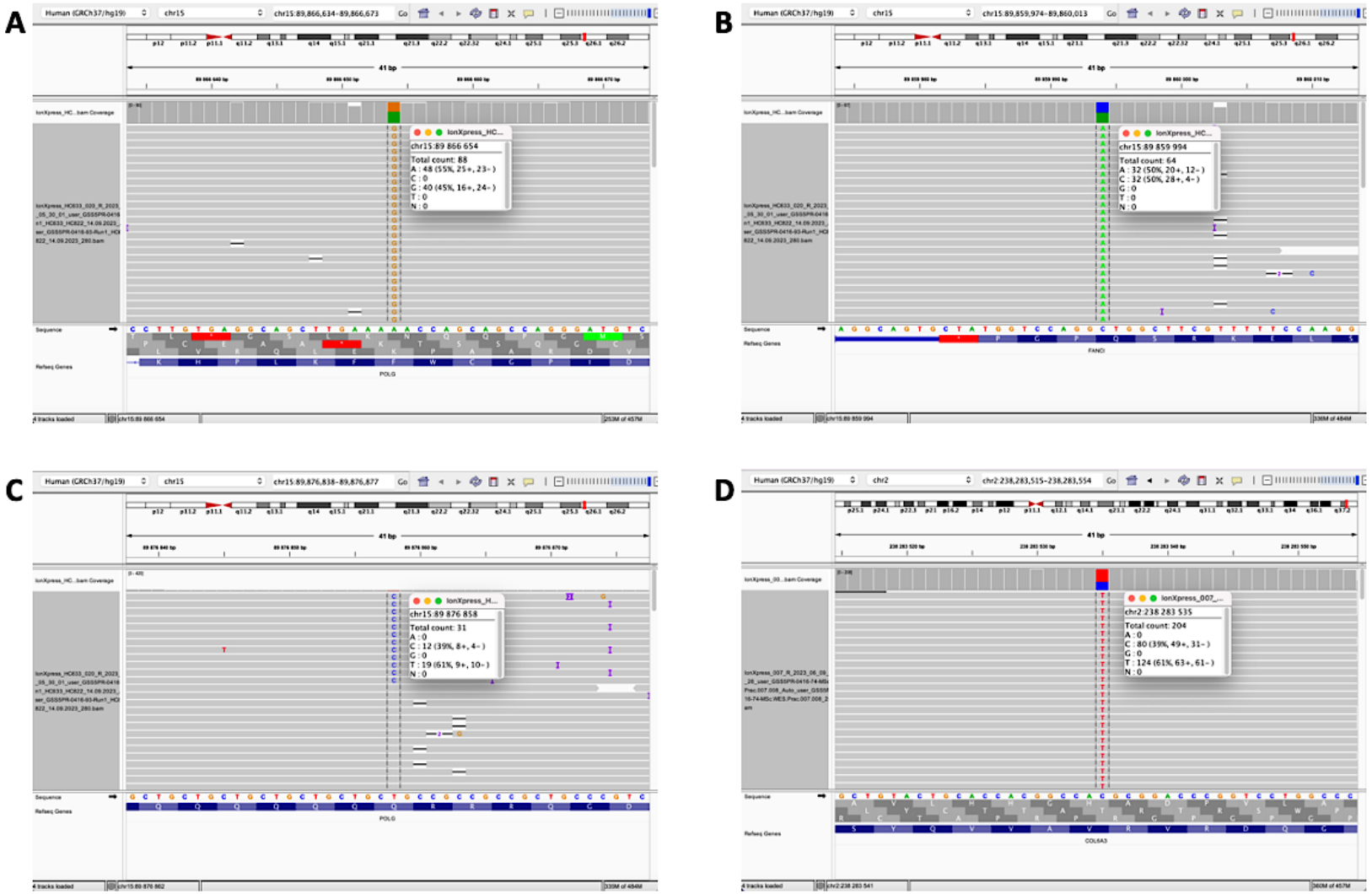
Patient ID	Gene	OMIM ID	RefSeq	Location	Coding mutation	Protein Change	ClinVar ID	Zygosity	SIFT/PolyPhen/CADD/ FATHMM/REVEL	ACMG/AMP Classification	Criteria Met for Classification and Evidence	Disease Associated with Gene
1	*POLG*	174763	NM_002693.3	15:89866654	c.3708G>T	p.Gln1236His	21316	Het	Tolerated Low Confidence (0.06)/Benign (0.365)/22.6/-4.0/0.358	B	PM2_Supporting: gnomAD AFR/AA AF = 0.01274; PP2_Supporting: 12.1% P/LP missense variants vs 1.49% B/LB missense variants; BS2_Strong: variant was observed in a homozygous state in gnomAD more than expected for disease (814 individuals); BP6_Strong: multiple ClinVar submissions suggesting variant is B/LB	*POLG*-Related Disorders; Mitochondrial Disease
1	*POLG*	174763	NM_002693.3	15:89866654	c.128A>G	p.Gln43Arg	129989	Het	Tolerated (0.51)/Benign (0)/0.001/-3.92/0.305	B	PM2_Supporting: gnomAD AFR/AA AF = 0.03691; PP2_Supporting: 12.1% P/LP missense variants vs 1.49% B/LB missense variants; BS2_Strong: variant was observed in a homozygous state in gnomAD more than expected for disease (21 individuals); BP6_Strong: multiple ClinVar submissions suggesting variant is B	*POLG*-Related Disorders; Mitochondrial Disease
2	*COL6A3*	120250	NM_004369.4	2:238283535	c.3199G>A	p.Val1067Met	284121	Het	Deleterious (0)/Probably Damaging (0.993)/26.2/-2.28/0.601	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.0008024; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity	Collagen VI-Related Dystrophies
2	*GCDH*	608801	NM_000159.4	19:13007748	c.*165A>G	N/A	208504	Hom	Deleterious Low Confidence (0)/Benign (0)/0.771/-/-	B	BA1_Very Strong: allele frequency is common for disease in gnomAD (0.8270); PP2_Supporting: 59.4% P/LP missense variants vs 2.9% B/LB missense variants; BS2_Strong: variant was observed in a homozygous state in gnomAD more than expected for disease (54766 individuals); BP6_Strong: multiple ClinVar submissions suggesting variant is B	Glutaric Aciduria Type 1
2	*FARS2*	611592	NM_006567.5	6:554547 9	c.971A>G	p.Tyr324Cys	214333	Het	Deleterious (0)/Probably Damaging (0.994)/26.3/-0.3/0.684	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.002003; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity	*FARS2* Deficiency; Leigh Syndrome
3	*SPG11*	610844	NM_025137.4	15:44878032	c.4923G>C	p.Lys1641Asn	376990	Het	Deleterious (0)/Probably Damaging (0.998)/24.0/-1.17/0.517	B	PM2_Supporting: gnomAD AFR/AA AF = 0.01430; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity; BP6_Strong: multiple ClinVar submissions suggesting variant is B/LB; BS2_Strong: variant was observed in a homozygous state in gnomAD more than expected for disease (7 individuals)	Hereditary Spastic Paraplegia 11
4	*LAMA3*	600805	NM_198129.4	18:21530071	c.9590A>G	p.His3197Arg	714745	Het	Deleterious (0)/Probably Damaging (0.998)/23.0/-1.09/0.718	LB	PM2_Supporting: gnomAD AFR/AA AF = 0.01146; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity; BP6_Supporting: multiple ClinVar submissions suggest variant is B; BS2_Strong: variant was observed in a homozygous state in gnomAD more than expected for disease (9 individuals)	Junctional Epidermolysis Bullosa

Patient ID	Gene	OMIM ID	RefSeq	Location	Coding mutation	Protein Change	ClinVar ID	Zygosity	SIFT/PolyPhen/CADD/ FATHMM/REVEL	ACMG/AMP Classification	Criteria Met for Classification and Evidence	Disease Associated with Gene
4	*COL6A3*	120250	NM_004369.4	2:238280477	c.4183C>T	p.Arg1395Trp	235302	Het	Deleterious (0)/Probably Damaging (0.996)/22.8/-1.75/0.531	LB	PM2_Supporting: gnomAD AFR/AA AF = 0.01873; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity; BP6_Supporting: multiple ClinVar submissions suggesting variant is B/LB; BS2_Supporting: variant was observed in a homozygous state in gnomAD more than expected for disease (1 individual)	Collagen VI-Related Dystrophies
5	*LRP1B*	608766	NM_018557.3	2:141473651	c.5914C>T	p.His1972Tyr	N/A	Het	Deleterious (0.02)/Probably Damaging (0.996)/27.5/-1.81/0.685	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.0001090; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity	NSCLC, melanoma, oesophageal, stomach, head and neck, uterine and bladder cancers
5	*SCN1A*	182389	NM_001165963.4	2:166847834	c.5951C>A	p.Pro1984His	93661	Het	Deleterious (0)/Possibly Damaging (0.819)/27.0/-4.14/0.801	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.002365; PP3_Moderade: multiple in-silico prediction tools suggest pathogenicity; PP2_Supporting: 48.1% P/LP missense variants vs 2.4% B/LB missense variants; BP6_Supporting: multiple ClinVar submissions suggesting variant is B/LB	*SCN1A* Seizure Disorders
6	*COL6A3*	120250	NM_004369.4	2:238287746	c.2030G>A	p.Arg677His	198167	Het	Deleterious (0)/Probably Damaging (0.993)/28.8/-2.08/0.630	B	PM2_Supporting: gnomAD AFR/AA AF = 0.01318; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity; BP6_Strong: multiple ClinVar submissions suggesting variant is B/LB; BS2_Strong: variant was observed in a homozygous state in gnomAD more than expected for disease (5 individuals)	Collagen VI-Related Dystrophies
6	*DNAH11*	603339	NM_001277115.2	7:21813586	c.9305G>A	p.Gly3102Asp	835960	Het	Deleterious (0)/Probably Damaging (1)/25.1/-1.68/0.776	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.0002564; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity	Primary Ciliary Dyskinesia
7	*WFS1*	606201	NM_006005.3	4:6303248	c.1726G>A	p.Gly576Ser	45439	Het	Deleterious (0.04)/Benign (0.404)/24.4/-1.52/0.537	LB	PM2_Supporting: gnomAD AFR/AA AF = 0.02361; PP2_Supporting: 22.81% P/LP missense variants vs 3.8% B/LB missense variants; BP6_Supporting: multiple ClinVar submissions suggest variant is B/LB; BS2_Strong: variant was observed in a homozygous state in gnomAD more than expected for disease (62 individuals)	*WFS1* Spectrum Disorder; Wolfram Syndrome
7	*ATXN1*	601556	NM_000332.3	6:16327335	c.1207C>T	p.Arg403Cys	N/A	Het	Deleterious (0.01)/Probably Damaging (0.993)/23.4/-1.26/0.307	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.009973; BS2_Strong: the variant was observed in a homozygous state in gnomAD more than expected for the disease (2 individuals)	Spinocerebellar Ataxia Type 1
8	*ALDH4A1*	606811	NM_003748.4	1:19211987	c.433G>A	p.Ala145Thr	294389	Het	Deleterious (0)/Probably Damaging (0.99)/26.5/-1.34/0.766	LB	PM2_Supporting: gnomAD AFR/AA AF = 0.02760; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity; BP6_Supporting: multiple ClinVar submissions suggesting variant is B/LB; BS2_Strong: variant was observed in a homozygous state in gnomAD more than expected for disease (8 individuals)	Hyperprolinemia, Type II

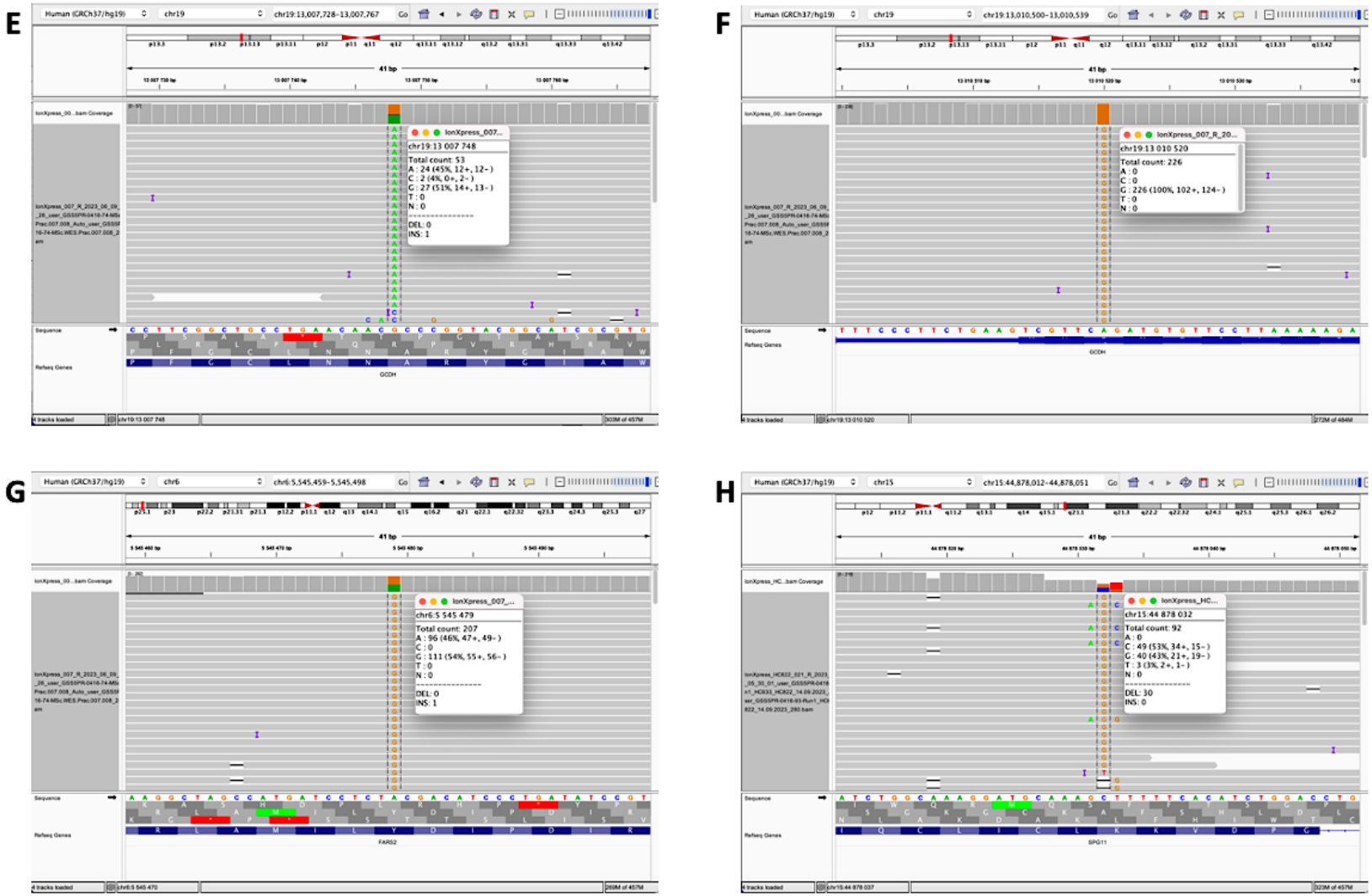
Patient ID	Gene	OMIM ID	RefSeq	Location	Coding mutation	Protein Change	ClinVar ID	Zygosity	SIFT/PolyPhen/CADD/ FATHMM/REVEL	ACMG/AMP Classification	Criteria Met for Classification and Evidence	Disease Associated with Gene
8	*DSTYK*	612666	NM_015375.3	1:205131207	c.1775G>A	p.Arg592Gln	224356	Het	Deleterious (0.01)/Probably Damaging (0.943)/32/-1.39/0.609	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.004248; PP3_Supporting: multiple in-silico prediction tools suggest pathogenicity	Complex hereditary spastic paraplegia
9	*RNASEH2B*	610326	NM_024570.4	13:51503677	c.203A>G	p.Glu68Gly	706377	Het	Deleterious (0)/Possibly Damaging (0.878)/28.3/-3.96/0.834	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.003286; PP3_Moderate: multiple in-silico prediction tools suggest pathogenicity; PP2_Supporting: 13.8% P/LP missense variants vs 1.4% B/LB missense variants	Aicardi-Goutières Syndrome
9	*TGM6*	613900	NM_198994.3	20:2384241	c.1108G>A	p.Gly370Ser	727483	Het	Deleterious (0)/Probably Damaging (0.989)/25.2/-3.98/0.925	VUS	PM2_Supporting: gnomAD AFR/AA AF = 0.006293; PP3_Moderade: multiple in-silico prediction tools suggest pathogenicity; BP6_Supporting: multiple ClinVar submissions suggesting variant is B	Spinocerebellar Ataxia Type 35

OMIM, Online Mendelian Inheritance in Man; RefSeq, reference sequence; ACMG/AMP, American College of Medical Genetics and Genomics and the Association for Molecular Pathology; POLG, DNA polymerase subunit gamma; COL6A3, collagen type VI alpha 3; GCDH, glutaryl-CoA dehydrogenase; FARS2, phenylalanyl-tRNA synthetase 2; SPG11, spastic paraplegia 11; LAMA3, laminin subunit alpha 3; LRP1B, low-density lipoprotein receptor-related protein 1B; SCN1A, sodium voltage-gated channel alpha subunit 1; DNAH11, dynein axonemal heavy chain 11; WFS1, wolframin endoplasmic reticulum transmembrane glycoprotein; ATXN1; ataxin 1; ALDH4A1, aldehyde dehydrogenase 4 family member A1; DSTYK, dual serine/threonine and tyrosine protein kinase ; RNASEH2B, ribonuclease H2 subunit B ; TGM6, transglutaminase 6; Het, heterozygous; Hom, homozygous; LB, likely benign; B, benign; VUS, variant of uncertain significance; AFR/AA AF, African/African American allele frequency.

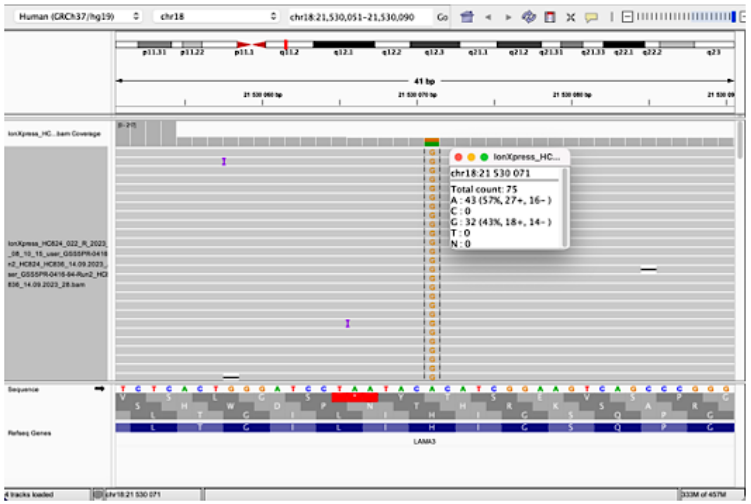
Supplementary Figure 2. Candidate Variant Visualisation using IGV



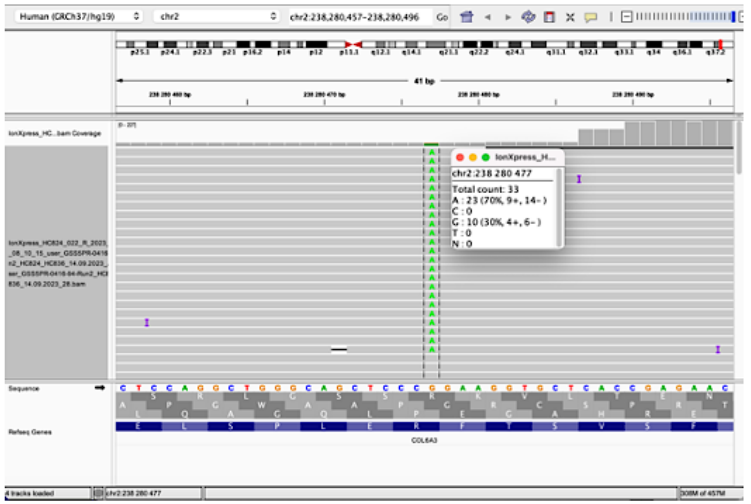
Supplementary Figure 1. Candidate Variant Visualisation using IGV Continued



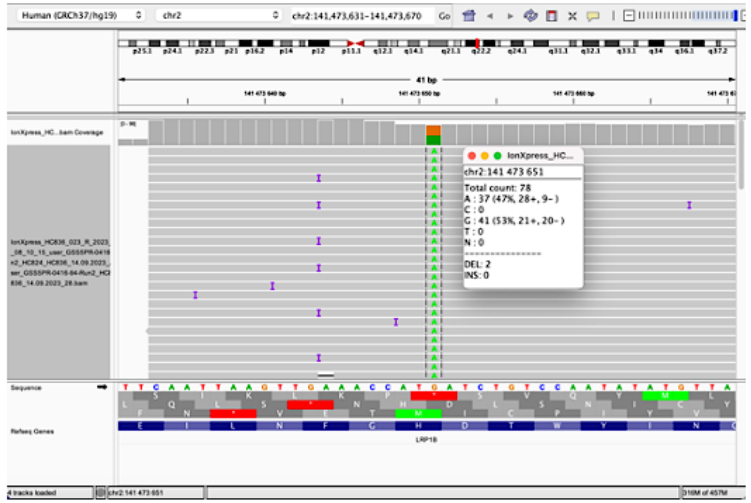
I


Human (GRCh37/hg19)
chr18
chr18:21,530,051-21,530,090
Go
41 bp
IonXpress_HC...
chr18:21 530 071
Total count: 75
A : 43 (57%, 27+, 16-)
C : 0
G : 32 (43%, 18+, 14-)
T : 0
N : 0
Sequence
Refseq Genes
LAMA3

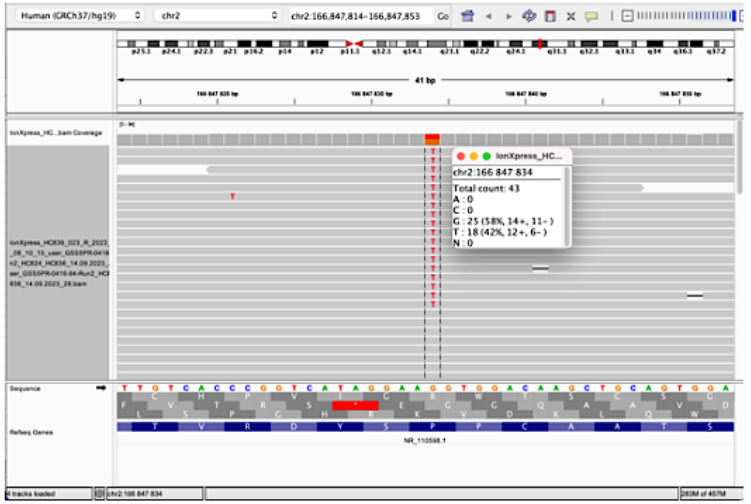
J


Human (GRCh37/hg19)
chr2
chr2:238,280,457-238,280,496
Go
41 bp
IonXpress_H...
chr2:238 280 477
Total count: 33
A : 23 (70%, 9+, 14-)
C : 0
G : 10 (30%, 4+, 6-)
T : 0
N : 0
Sequence
Refseq Genes
COL6A3

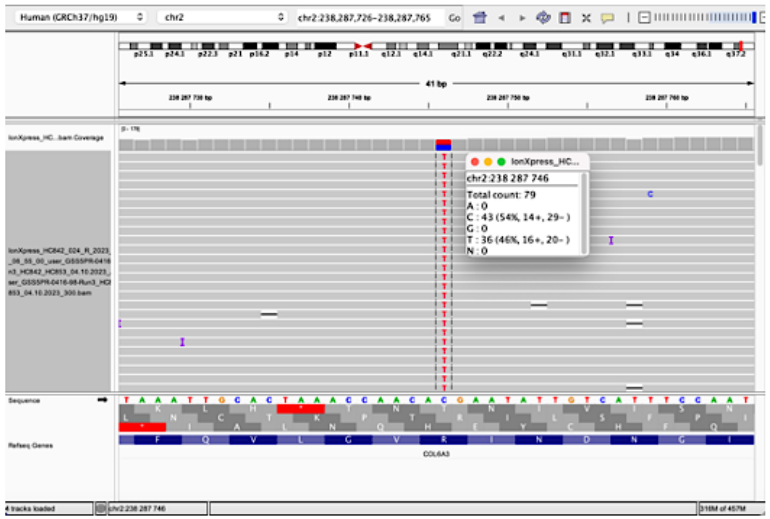
K


Human (GRCh37/hg19)
chr2
chr2:141,473,631-141,473,670
Go
41 bp
IonXpress_HC...
chr2:141 473 651
Total count: 78
A : 37 (47%, 28+, 9-)
C : 0
G : 41 (53%, 21+, 20-)
T : 0
N : 0
DEL: 2
INS: 0
Sequence
Refseq Genes
LRP1B

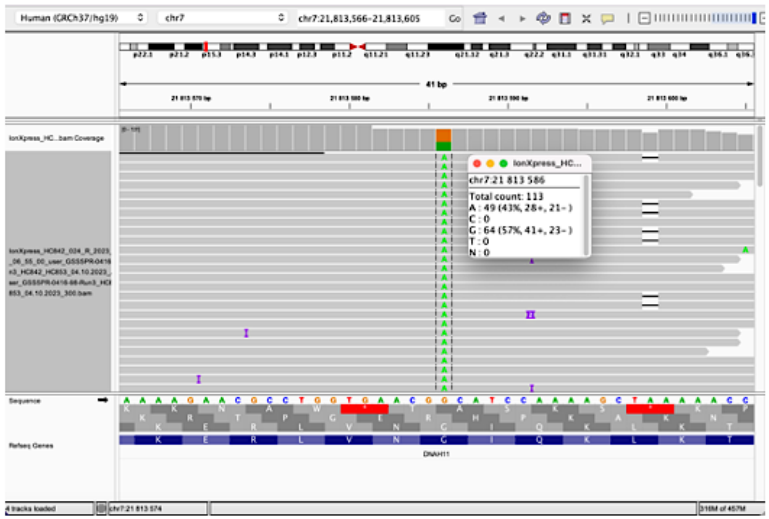
L


Human (GRCh37/hg19)
chr2
chr2:166,847,814-166,847,853
Go
41 bp
IonXpress_HC...
chr2:166 847 834
Total count: 43
A : 0
C : 0
G : 25 (58%, 14+, 11-)
T : 18 (42%, 12+, 6-)
N : 0
Sequence
Refseq Genes
NR_110598.1

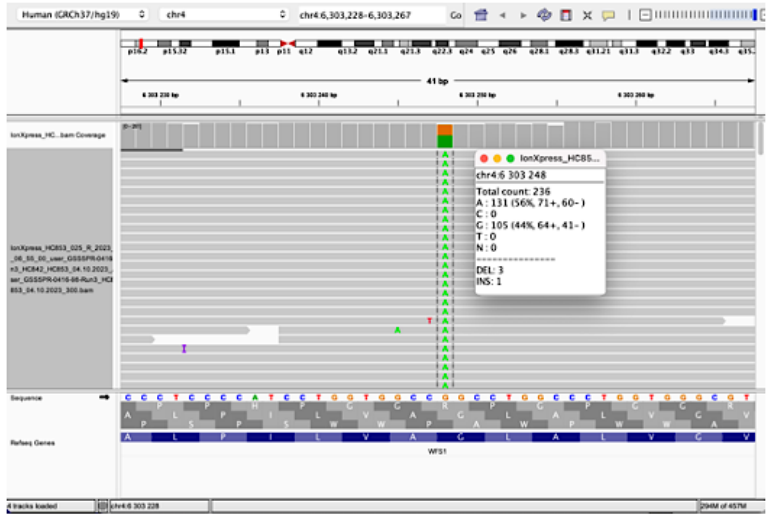
M


Human (GRCh37/hg19)
chr2
chr2:238,287,726-238,287,765
41 bp
IonXpress_HC...
chr2:238 287 746
Total count: 79
A : 0
C : 43 (54%, 14+, 29-)
G : 0
T : 36 (46%, 16+, 20-)
N : 0
COL6A3

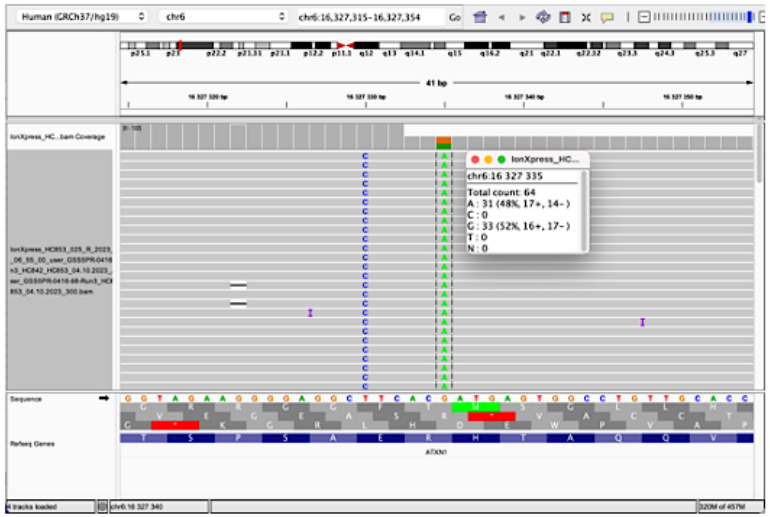
N


Human (GRCh37/hg19)
chr7
chr7:21,813,566-21,813,605
41 bp
IonXpress_HC...
chr7:21 813 586
Total count: 113
A : 49 (43%, 28+, 21-)
C : 0
G : 64 (57%, 41+, 23-)
T : 0
N : 0
DNAH11

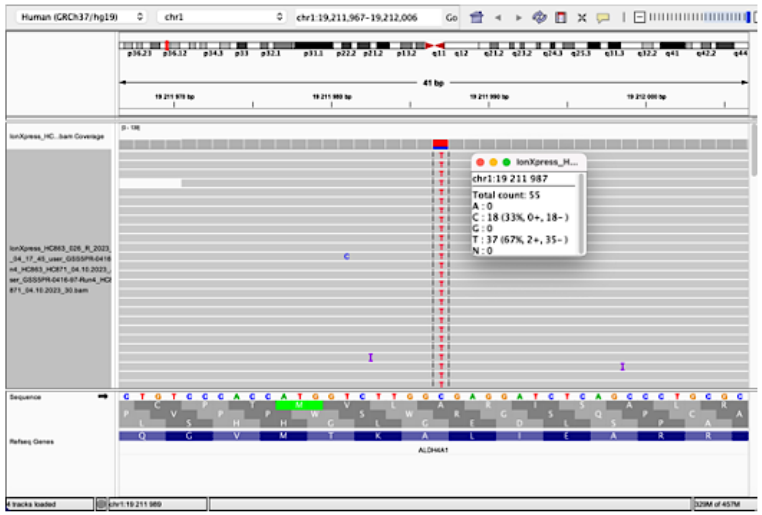
O


Human (GRCh37/hg19)
chr4
chr4:6,303,228-6,303,267
41 bp
IonXpress_HC85...
chr4:6 303 248
Total count: 236
A : 131 (56%, 71+, 60-)
C : 0
G : 105 (44%, 64+, 41-)
T : 0
N : 0
DEL: 3
INS: 1
WFS1

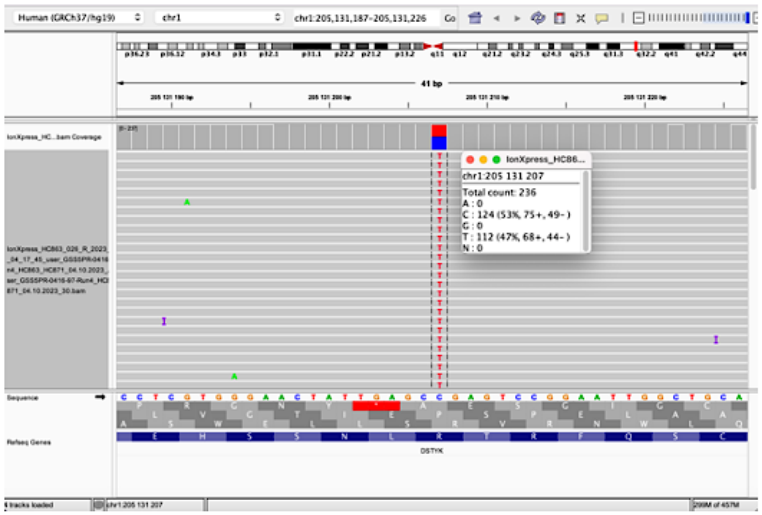
P


Human (GRCh37/hg19)
chr6
chr6:16,327,315-16,327,354
41 bp
IonXpress_HC...
chr6:16 327 335
Total count: 64
A : 31 (48%, 17+, 14-)
C : 0
G : 33 (52%, 16+, 17-)
T : 0
N : 0
ATXN1

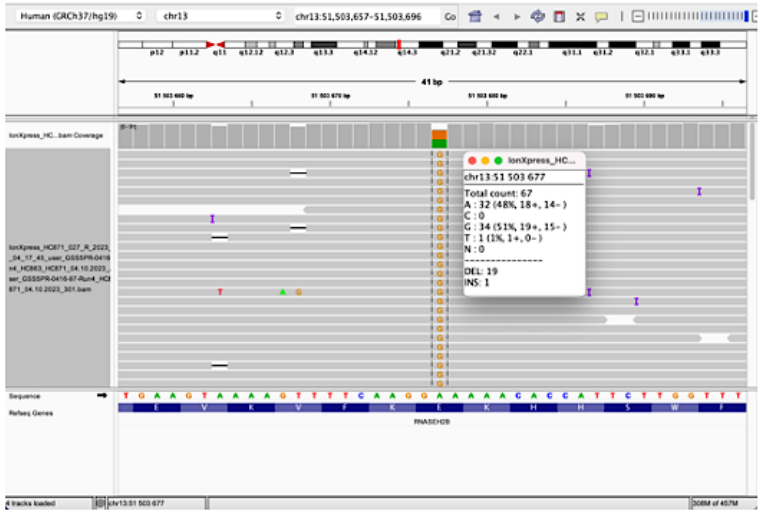
Q


Human (GRCh37/hg19)
chr1
chr1:19,211,967-19,212,006
41 bp
IonXpress_H...
chr1:19 211 987
Total count: 55
A : 0
C : 18 (33%, 0+, 18-)
G : 0
T : 37 (67%, 2+, 35-)
N : 0
ALDH4A1

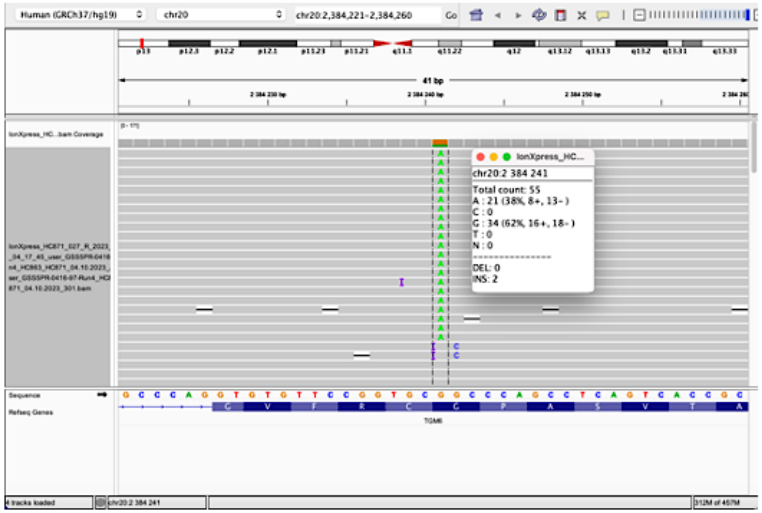
R


Human (GRCh37/hg19)
chr1
chr1:205,131,187-205,131,226
41 bp
IonXpress_HC86...
chr1:205 131 207
Total count: 236
A : 0
C : 124 (53%, 75+, 49-)
G : 0
T : 112 (47%, 68+, 44-)
N : 0
DSTYK

S


Human (GRCh37/hg19)
chr13
chr13:51,503,657-51,503,696
41 bp
IonXpress_HC...
chr13:51 503 677
Total count: 67
A : 32 (48%, 18+, 14-)
C : 0
G : 34 (51%, 19+, 15-)
T : 1 (1%, 1+, 0-)
N : 0
DEL: 19
INS: 1
RNASEH2B

T


Human (GRCh37/hg19)
chr20
chr20:2,384,221-2,384,260
41 bp
IonXpress_HC...
chr20:2 384 241
Total count: 55
A : 21 (38%, 8+, 13-)
C : 0
G : 34 (62%, 16+, 18-)
T : 0
N : 0
DEL: 0
INS: 2
TGM6

Visualisation of each candidate variant for the determination of zygosity and coverage assessment using Integrative Genomics Viewer (IGV). Patient 1: NM_002693.3:c.2246T>C:p.Phe749Ser ***(A)****, NM_002693.3:c.3708G>T p.Gln1236His* ***(B)*** *and NM_002693.3:c.128A>G p.Gln43Arg* ***(C)****. Patient 2: NM_004369.4:c.3199G>A:p.Val1067Met* ***(D)****, NM_000159.4:c.877G>A:p.Ala293Thr* ***(E)****, NM_000159.4:c.*165A>G* ***(F)*** *and NM_006567.5:c.971A>G:p.Tyr324Cys* ***(G)****. Patient 3: NM_025137.4:c.4923G>C:p.Lys1641Asn* ***(H)****. Patient 4: NM_198129.4:c.9590A>G:p.His3197Arg* ***(I)*** *and NM_004369.4:c.4183C>T:p.Arg1395Trp* ***(J)****. Patient 5: NM_018557.3:c.5914C>T:p.His1972Tyr* ***(K)*** *and NM_001165963.4:c.5951C>A:p.Pro1984His* ***(L)****. Patient 6: NM_004369.4:c.2030G>A:p.Arg677His* ***(M)*** *and NM_001277115.2:c.9305G>A:p.Gly3102Asp* ***(N)****. Patient 7: NM_006005.3:c.1726G>A:p.Gly576Ser* ***(O)*** *and NM_000332.3:c.1207C>T:p.Arg403Cys* ***(P)****. Patient 8: NM_003748.4:c.433G>A:p.Ala145Thr* ***(Q)*** *and NM_015375.3:c.1775G>A:p.Arg592Gln* ***(R)****. Patient 9: NM_024570.4:c.203A>G:p.Glu68Gly* ***(S)*** *and NM_198994.3:c.1108G>A:p.Gly370S* ***(T)****.*

Appendix B: Ethics Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO:	Ms R Naicker School of Pathology Division of Human Genetics Medical School University E-mail: racilya@gmail.com
CC:	Supervisor: Dr F Baine-Savanhu and Professor A Krause <Fiona.Baine-Savanhu@wits.ac.za> and <HREC-Medical Research Office@wits.ac.za>
FROM:	Mr Iain Burns Human Research Ethics Committee (Medical) Tel: 011 717 1252 E-mail: Iain.Burns@wits.ac.za
DATE:	2023/07/24
REF:	R14/49
PROTOCOL NO:	**M230689** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)
PROJECT TITLE:	*Analysis of whole exome sequence data from African patients with HD-like features and no known HD phenocopy syndrome*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.



MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

R49 Ms R Naicker

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M230689

NAME: (Principal Investigator)	Ms R Naicker
DEPARTMENT:	School of Pathology Division of Human Genetics Medical School University
PROJECT TITLE:	*Analysis of whole exome sequence data from African patients with HD-like features and no known HD phenocopy syndrome*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M15/03/68
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Dr F Baine-Savanhu and Professor A Krause
APPROVED BY:	Dr M Vorster, Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	2023/07/24 **EXPIRY DATE:** 2028/07/23

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in **June** and therefore reports and re-certification will be due in the month of **June** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator Date

Appendix C: Approved Research Protocol

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Private Bag 3 Wits, 2050
Fax: 027117172119
Tel: 02711 7172076

Reference: Mrs Sandra Benn
E-mail: sandra.benn@wits.ac.za

11 July 2023
Person No: 1717757
PAG

Miss R Naicker
11 Morningview Estate
8 Fir Road
Morningside
2196
South Africa

Dear Miss Racilya Naicker

Master of Science in Medicine: Approval of Title

We have pleasure in advising that your proposal entitled *Analysis of Whole Exome Sequence Data from African Patients with HD-Like Features and No Known HD Phenocopy Syndrome* has been approved. Please note that any amendments to this title have to be endorsed by the Faculty's higher degrees committee and formally approved.

Yours sincerely

Mrs Sandra Benn
Faculty Registrar
Faculty of Health Sciences

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

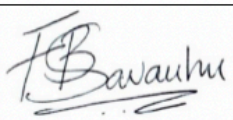
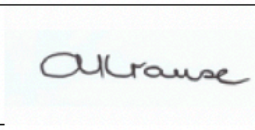
FACULTY OF HEALTH SCIENCES

CANDIDATE'S SURNAME: [Please print] NAICKER	FIRST NAME/S: RACILYA	STUDENT NUMBER: 1717757
CURRENT QUALIFICATIONS: BSc BIOLOGICAL SCIENCES & BSc (Hons) BIOCHEMISTRY AND CELL BIOLOGY		
TEL: N/A	CELL: 083 642 7116	E-MAIL: 1717757@students.wits.ac.za / FAX: N/A
DEGREE FOR WHICH PROTOCOL IS BEING SUBMITTED: MSc (MED) GENOMIC MEDICINE		
PART-TIME OR FULL-TIME: FULL-TIME		
FIRST REGISTERED FOR THIS DEGREE:	TERM: 2023	YEAR: 1
DEPARTMENT: HUMAN GENETICS		
TITLE OF PROPOSED RESEARCH: ANALYSIS OF WHOLE EXOME SEQUENCE DATA FROM AFRICAN PATIENTS WITH HD-LIKE FEATURES AND NO KNOWN HD PHENOCOPY SYNDROME		
CANDIDATE'S SIGNATURE: Racicker		DATE: 25/04/2023
SUPERVISOR 1 (NAME & SURNAME): DR FIONA BAINE-SAVANHU		% Supervision: 70%
SUPERVISOR'S QUALIFICATIONS: PhD		
SUPERVISOR'S DEPARTMENT: HUMAN GENETICS		
SUPERVISOR'S ADDRESS / TEL / E-MAIL: NHLS BRAAMFONTEIN, 1 HOSPITAL STREET, BRAAMFONTEIN 011 489 9293 fiona.baine-savanhu@wits.ac.za		
SUPERVISOR 2 (NAME & SURNAME): PROF AMANDA KRAUSE		% Supervision: 30%
SUPERVISOR'S QUALIFICATIONS: MBBCh, PhD		
SUPERVISOR'S ADDRESS / TEL / E-MAIL: NHLS BRAAMFONTEIN, 1 HOSPITAK STREET, BRAAMFONTEIN 011 489 9211 amanda.krause@nhls.ac.za		

SYNOPSIS OF RESEARCH:(Brief summary of proposed research project; between 200-300 words only; with sub-headings: an introduction and justification for study, aim/s, proposed methodology and expected outcome/s)
[Use reverse side of this page if more space is required]

Introduction and Justification of Study

Huntington disease (HD) is a rare progressive neurodegenerative disorder that results from a CAG repeat expansion within the huntingtin gene (*HTT*). Several syndromes present with HD-like features in the absence of the *HTT* expansion, and are termed HD phenocopies. Huntington disease-like 2 (HDL2) is the phenocopy most commonly observed in individuals with African, or probable African ancestry. Therefore, the Division of Human Genetics, National Health Laboratory Service screens for both HD and HDL2 in patients with HD-like phenotypes and African ancestry. However, many patients test negative for both syndromes and remain without a molecular diagnosis. This highlights the urgent need for alternative testing strategies to be employed to aid in patient diagnosis and to understand the genetic basis underlying the HD-like clinical presentation.

WITS ETHICS NOT REQUIRED: Yes No WITS ETHICS PENDING: (Yes) No WITS ETHICS APPROVED: Yes No (circle appropriate symbol)* ***Please note the final human ethics clearance certificate or animal ethics certificate must be available prior to starting research**		IF Y SUPPLY ETHICS CLEARANCE CERTIFICATE AS ATTACHMENT AND INCLUDE ETHICS NUMBER HERE:
As supervisor/s, I/we confirm that I have read the protocol which has been submitted for assessment.		
SIGNATURE OF SUPERVISOR/S:		
SIGNATURE PG OFFICE STAFF ……	REGISTERED YES….. NO…..	STAMP

SYNOPSIS OF RESEARCH CONTINUED

Aim
To identify variants and/or genes implicated in the HD-like phenotype of South African patients without a molecular diagnosis.

Proposed Methodology
Whole exome sequencing (WES) data derived from eight patients with HD-like features will be analysed in a stepwise bioinformatics process to identify potential disease-causing variants. A virtual gene panel will be created by compiling a list of genes associated with neurodegenerative disorders that present with overlapping features. Variants in these genes will be filtered and prioritised using *in-silico* pathogenicity prediction tools and mutation databases. Lastly, the variants will be classified according to ACMG/AMP guidelines and where possible, genotype-phenotype correlations will be performed for pathogenic and likely pathogenic variants.

Expected Outcomes
Variants determined to be associated with the HD-like features in this group of patients may provide a molecular diagnosis. This study could therefore alleviate the diagnostic uncertainty these patients face and generate more knowledge regarding the genetic epidemiology of these HD phenocopy syndromes.

11 March 2019/MP

Table of Contents

Abstract *1*

1. Literature Review *2*

1.1. Huntington Disease **2**

1.2. Huntington Disease Phenocopies **2**

1.2.1. Huntington Disease-like 2 2

1.2.2. Other Relevant HD Phenocopies 3

1.3. Challenges in the Diagnosis of Patients with HD-like Clinical Features **4**

1.4. Next-Generation Sequencing **5**

1.5. Relevance of this Study **6**

2. Aim and Objectives *6*

2.1. Aim **6**

2.2. Objectives **6**

3. Methods *7*

3.1. Workflow **7**

3.2. Sample Selection **7**

3.3. Virtual Gene Panel Development **8**

3.4. Sequencing, Alignment and Variant Calling **9**

3.5. Variant Annotation and Filtering **9**

3.6. Variant Prioritisation **10**

3.7. Variant Classification **10**

3.8. Genotype-Phenotype Correlations **11**

4. Time Frame *11*

5. Budget and Funding Source *11*

6. Ethics *11*

7. Limitations of this Study *12*

8. References *13*

Abstract

Huntington disease (HD) is a rare progressive neurodegenerative disorder that results from a CAG repeat expansion within the huntingtin gene (*HTT*). Several syndromes present with HD-like features in the absence of the *HTT* expansion, and are termed HD phenocopies. Huntington disease-like 2 (HDL2), a known phenocopy, is most commonly observed in individuals with African, or probable African, ancestry. Therefore, the Division of Human Genetics, National Health Laboratory Service screens for both HD and HDL2 in patients with HD-like phenotypes and an indication of African ancestry. Some patients test negative for both syndromes and are therefore undiagnosed. This highlights the urgent need for further testing strategies to be employed to aid in patient diagnosis. Therefore, this study aims to identify variants and/or genes implicated in the HD-like phenotype of these undiagnosed patients. In a selection of eight South African patients, for whom all previous tests yielded negative results, whole exome sequencing data will be analysed using a stepwise bioinformatic process. A virtual gene panel will be generated to analyse variants present in genes associated with neurodegenerative disorders with overlapping features. Variants will then be filtered to create a prioritised list of candidate pathogenic and likely pathogenic variants. Where possible, genotype-phenotype correlations will be done to aid in the identification of the pathogenic variants that could be implicated in the HD-like features of the patients. This study could therefore alleviate the diagnostic uncertainty these patients face and generate more knowledge regarding the genetic epidemiology of these HD phenocopy syndromes.

1. Literature Review

1.1. Huntington Disease

Huntington disease (HD) is a rare progressive neurodegenerative disorder with onset typically occurring in mid-adulthood. The estimated worldwide prevalence is 4.88 individuals per 100 000, however, it generally differs depending on the geographical region (Medina et al., 2022). The HD prevalence is significantly greater in European and North American populations compared to Asian and African populations (Medina et al., 2022). While the HD prevalence estimates in South Africa have increased since the first figures determined by Hayden et al in 1980, a lower HD frequency is observed in the black South African population group in comparison with the mixed ancestry and Caucasian population groups (Baine et al., 2016; Hayden et al., 1980).

Huntington disease is inherited in an autosomal dominant pattern and is caused by a trinucleotide (CAG) repeat expansion within the Huntingtin gene (*HTT*). This polyglutamine expansion promotes pathogenicity by imparting a toxic gain of function to the resultant mutant HTT protein (Bates et al., 2015). Normally, the HTT protein is expressed ubiquitously throughout the body but exhibits its highest levels in the brain and testes (Schulte & Littleton, 2011). Within the brain, HTT is found in all neurons and glial cells. While the exact function of wildtype HTT is not fully elucidated, mutant HTT is known to disturb the basal ganglia, leading to the associated symptoms (Schulte & Littleton, 2011). The clinical manifestations of HD include a triad of psychiatric symptoms (anxiety, depression), cognitive impairment (executive dysfunction) and movement disorder (chorea) (Podvin et al., 2019). Interestingly, certain clinically relevant syndromes share overlapping features with HD in the absence of the *HTT* expansion. Therefore, they are termed HD phenocopies and account for approximately 1% of suspected HD cases (Wild & Tabrizi, 2007).

1.2. Huntington Disease Phenocopies

1.2.1. Huntington Disease-like 2

Apart from a later onset of chorea, Huntington disease-like 2 (HDL2) is phenotypically nearly indistinguishable from HD (Anderson et al., 2019). While the HDL2 phenotype most closely resembles HD throughout its disease progression, its causative mutation is a CTG/CAG repeat expansion in the junctophilin 3 gene (*JPH3*) (Holmes et al., 2001). Pathogenicity is thought to arise from neurotoxicity of the expanded transcript

or by loss of expression of the protein (Rudnicki et al., 2007; Seixas et al., 2012). This is due to the disturbance of the normal JPH3 protein function, which is to regulate the excitability of neurons and the intracellular calcium signalling pathways in the endoplasmic reticulum (Bourinaris et al., 2021). Interestingly, individuals with African, or probable African, ancestry tend to display a higher relative frequency of HDL2, with South Africa having the highest number of reported cases of HDL2 worldwide (Baine et al., 2018; Krause et al., 2015). Therefore, standard South African HD diagnostic laboratory practice offers simultaneous screening for both the *HTT* and *JPH3* expansions when individuals possess African ancestry (mixed ancestry and African patients).

1.2.2. Other Relevant HD Phenocopies

There are several other HD phenocopy syndromes that present with overlapping clinical features, but with various causative mutations. Phenocopies that result from repeat expansions include spinocerebellar ataxias (SCAs), which are a diverse group of disorders characterised by cerebellar ataxia, cognitive impairment, and chorea (Martino et al., 2013). The CAG repeat expansion is the underlying genetic cause for many of the most severe and prevalent SCAs (Paulson et al., 2017). Dentatorubral-pallidoluysian atrophy (DRPLA) presents with similar features to HD, such as chorea, ataxia, and myoclonus and is caused by an expansion in the atrophin 1 gene (*ATN1*) (Le Ber et al., 2003; Martino et al., 2013). Chromosome 9 open reding frame 72 (*C9orf72*) repeat expansions present with clinical features including chorea, behavioural issues, psychiatric symptoms, and tremor (Hensman et al., 2014). Huntington disease-like 1 (HDL1) is caused by a repeat expansion in the prion protein gene (*PRNP*), with the resemblance to HD being through the clinical features such as motor impairment with chorea, cognitive decline, prominent personality changes and psychiatric symptoms (Moore et al., 2001). The overlapping clinical features of these HD phenocopy syndromes may be explained by the mechanism by which polyglutamine repeat expansions induce pathogenicity, as it is the underlying cause for many of these disorders (Saudou & Humbert, 2016). Also, the proteins associated with each of these phenocopies have all been noted to be expressed in the brain, thereby providing a further possible explanation for the similarity of the phenotypes to HD. However, the underlying functions of the genes involved in the different

phenocopy syndromes are still poorly understood, as are the factors contributing to the similar clinical phenotypes that are seen.

Some HD phenocopies are caused by mutations other than repeat expansions, and these include neuroferritinopathy, which can result from various mutations in the ferritin light chain gene (*FTL1*). Such mutations may be missense variants and small intragenic duplications, which lead to clinical features such as increased serum ferritin levels, dementia and asymmetric dystonia or chorea (Hartig, 2022; Kumar et al., 2016). Chorea-acanthocytosis is caused by loss of function mutations in the Vacuolar Protein Sorting 13 Homolog A gene (*VPS13A*), which encodes chorein (Sokolov et al., 2012). Clinical presentations include psychiatric and cognitive dysfunction that resembles HD, as well as movement disorders (Sokolov et al., 2012). While these HD phenocopies result from a wide range of mutations in various genes, which are all expressed in the brain, they all present with overlapping clinical features that resemble HD at some point during disease progression. Therefore, inaccurate clinical phenotyping may lead to difficulties in patient diagnosis.

1.3. Challenges in the Diagnosis of Patients with HD-like Clinical Features

Only a few HD phenocopy syndromes are discussed but there are many other conditions not included in this literature review that share overlapping features with classic HD, for at least part of their disease course. The diagnosis of HD phenocopies is therefore challenging due to the complexity and lack of knowledge surrounding these numerous disorders, as well as poor clinical phenotyping, which results in misdiagnosis. A 2022 study by Koriath et al illustrated that within a patient group of 50 individuals where the HD-like phenotype was apparent in the absence of the *HTT* CAG expansion, a diagnosis could only be made for one individual due to the presence of the ataxin 1 (*ATXN1*) expansion, underlying spinocerebellar ataxia 1 (SCA1) (Koriath et al., 2022). Thus, emphasising the difficulty in diagnosing patients who present with the HD-like features without the *HTT* expansion as only about 2% of HD phenocopies receive a molecular diagnosis, while the rest remain undiagnosed (Koriath et al., 2022).

The Division of Human Genetics, National Health Laboratory Service (NHLS), Johannesburg, South Africa has been involved in HD diagnosis since the mid-1980s (Baine et al., 2013). As previously mentioned, in South Africa individuals of African ancestry, who display HD clinical features, undergo molecular diagnostic testing for

expansions in both the *HTT* and *JPH3* genes (Krause et al., 2015). While this has proven beneficial for some patients, many others carry neither of these expansions, therefore excluding HD and HDL2 from their diagnosis. A previous study performed in the Division investigated five relevant genes (namely, *ATN1*, *TBP*, *C9orf72*, *ATXN2* and *ATXN7*) in the patient cohort where HD and HDL2 had been excluded (Baine et al., 2018). These genes were selected as their pathogenic repeat expansions are known to underlie HD phenocopy syndromes. Of the 105 samples analysed using PCR and capillary electrophoresis, one individual had an expansion in the *ATXN2* gene which underlies SCA2, thereby confirming a diagnosis in this patient (Baine et al., 2018). Thereafter, a further study was performed whereby 16 of these patient samples, for whom previous tests yielded negative results, were screened for mutations using a targeted custom next-generation sequencing (NGS) gene panel comprising 69 genes relevant to neurodegenerative disorders (unpublished data). These genes were chosen based on their role in the pathogenesis of HD phenocopies and to investigate a broader range of genes that could be underlying their HD-like clinical features. From this study, three individuals were found to have potential pathogenic mutations, thus illustrating the need for further testing on this patient cohort to aid in diagnosis of the remaining undiagnosed patients. However, in this proposed study, patients who were previously assessed using this NGS testing panel will be excluded in order to further investigate more of the undiagnosed patient cohort.

1.4. Next-Generation Sequencing

Next-generation sequencing is a high-throughput massively parallel technique utilised to obtain the sequence of a specific DNA sample, which can subsequently be compared to the reference genome to investigate the presence of variants. NGS can be used to sequence specific genes (targeted gene sequencing panels), exons (whole exome sequencing (WES)) or the entire genome (whole genome sequencing (WGS)). While WGS can provide information regarding structural, copy number and single nucleotide variants (SNVs) in both non-coding and coding DNA sequences, it is significantly more time consuming and costly. Analysis of the large dataset produced also requires expertise. On the other hand, WES is more cost-effective and provides fewer variants of uncertain significance (VUS) than WGS. While WES only analyses about 1 to 2% of the total genetic sequence, most mutations are present in the exons, making it critical to study these regions to investigate the presence of small insertions

and deletions, as well as SNVs. One limitation of WES using NGS technology is that the detection of repeat expansions is difficult due to the process of read alignment mapping. Nonetheless, WES is currently a more efficient and cost-effective option as it has a wider genetic coverage than the previously used NGS targeted gene panel. Thus, this study elected to utilise WES.

1.5. Relevance of this Study

The diagnosis of HD and phenocopies is often challenging stemming from the numerous overlapping clinical features and the lack of knowledge regarding all phenocopy syndromes. An ongoing study in the Division has screened a cohort of patients, for whom HD and HDL2 had previously been excluded, for several genes implicated in neurodegenerative disorders. Most of these patients remain without a molecular diagnosis; therefore, this study will investigate whether potential disease-causing mutations are present in their exomes. This will allow for the understanding of the genetic basis underlying disease, which is of utmost importance as it can help provide patients with an accurate diagnosis and potential therapeutic options. This study could therefore alleviate the diagnostic uncertainty these patients face and generate more knowledge regarding the genetic epidemiology of HD phenocopy syndromes in South Africa.

2. Aim and Objectives

2.1. Aim

To identify variants and/or genes implicated in the HD-like phenotype of South African patients without a molecular diagnosis using WES data.

2.2. Objectives

To achieve the aforementioned aim, the following objectives will be set:

1. To develop a virtual gene panel, for the analysis of the WES data, using literature to identify genes associated with neurodegenerative disorders with features overlapping HD.
2. To identify and classify pathogenic and likely pathogenic variants that could be contributing to the HD-like features of this patient cohort (n=8).
3. To perform genotype-phenotype correlations, where possible, for pathogenic and likely pathogenic variants using de-identified patient clinical features.

3. Methods

3.1. Workflow

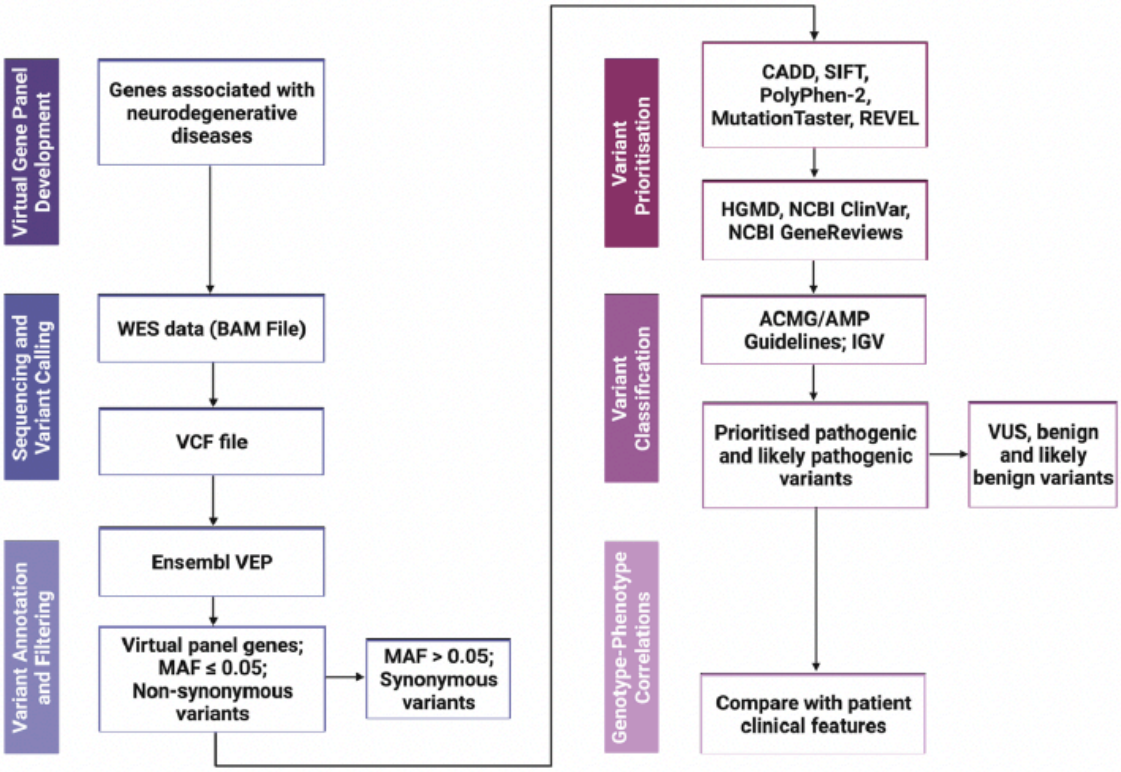


Figure 1. Methodology flow chart for the current study.

3.2. Sample Selection

This study will make use of de-identified DNA samples from eight undiagnosed South African patients who have presented with HD-like clinical features. These patients have previously tested negative for pathogenic expansions in *HTT*, *JPH3*, *ATN1*, *TBP*, *C9orf72*, *ATXN2* and *ATXN7*, thereby largely excluding the associated disorders from their diagnosis (Figure 2). The current study will utilise WES data from these patients to investigate the presence of any likely disease-causing variants. The sixteen patients who were screened using targeted NGS will be excluded as the virtual panel may have a lot of overlap with the panel previously used.

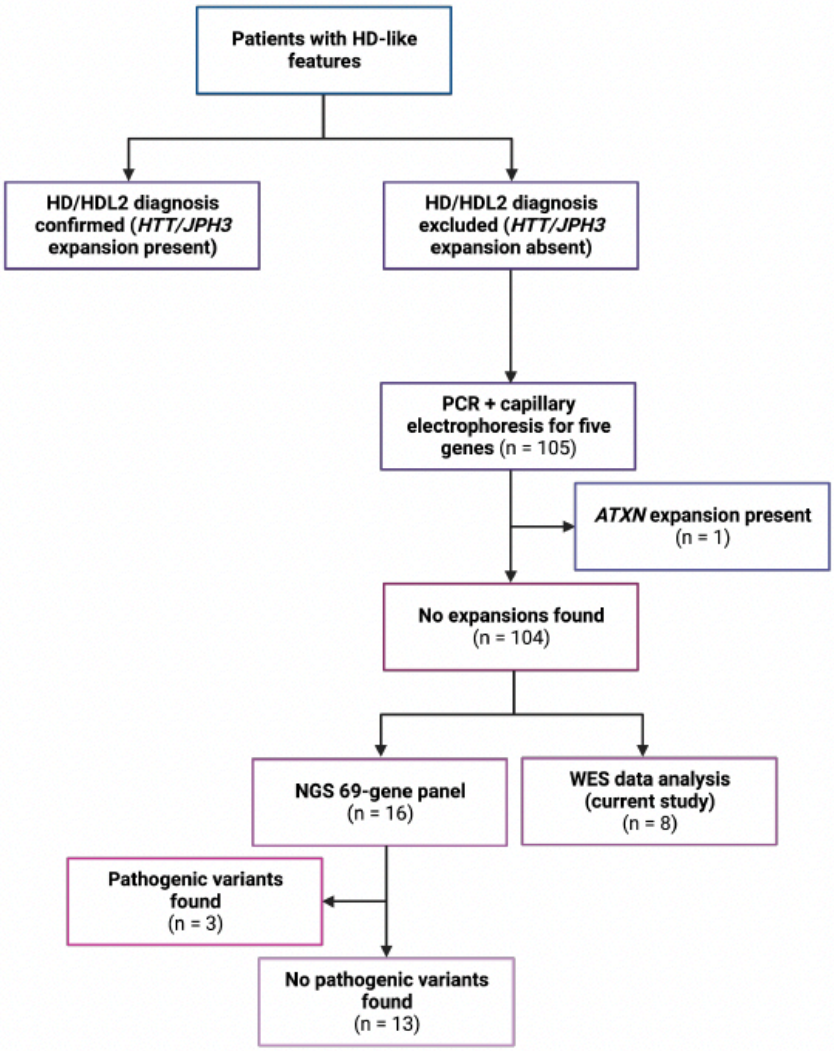


Figure 2. Previous studies performed in the Division of Human Genetics on this patient cohort.

3.3. Virtual Gene Panel Development

Genes that have been implicated in neurodegenerative disorders with features overlapping those of HD will be investigated through the review of relevant literature. Genes that are associated with disorders that present overlapping clinical features with HD and phenocopies will be compiled into a list to be used as a virtual gene panel for the analysis of the WES data. Also, the 69 genes from the previous targeted NGS panel study in the Division will be included in the virtual panel in this study. Thus, allowing for the other patients who were not previously investigated using NGS to be assessed for a wider range of causative genes and/or variants that could be underlying their HD-like phenotypes.

3.4. Sequencing, Alignment and Variant Calling

De-identified stored DNA samples from the Division were previously sequenced using the Ion GeneStudio™ S5 system (Thermo Fisher Scientific, USA) to generate the WES data. The Ion Torrent instrument software, Torrent Suite™ (Thermo Fisher Scientific, USA), was used to generate the binary alignment map (BAM) files for each sequence. BAM files are binary versions of Sequence Alignment Map (SAM) files and comprise sequencing data aligned to the reference genome sequence (GRCh38). Thereafter, variant calling was performed using the variantCaller plug-in (Thermo Fisher Scientific, USA) and variant call format (VCF) files were produced from each of the BAM files. VCF files are used to determine the variation of the sequence from the reference genome and document SNVs, copy number variants (CNVs), structural variants (SVs) as well as insertions and deletions (indels).

3.5. Variant Annotation and Filtering

The VCF files will be annotated using the Ensembl Variant Effect Predictor (VEP) tool (McLaren et al., 2016). Ensembl VEP can predict the molecular consequences associated with the variant by comparison to the reference transcripts Ensembl/GENCODE and the National Centre for Biotechnology Information (NCBI) Reference Sequence (RefSeq) database (Frankish et al., 2021; O'Leary et al., 2016). Sequence variant consequences on the protein produced that are determined by Ensembl VEP include missense, frameshift, stop loss and gain, as well as in-frame indels (McLaren et al., 2016).

The variants will then be filtered using Ensembl VEP according to several parameters. Firstly, a virtual gene panel will be applied to select a subset of genes for first-level analysis. Genes that will be included will be identified through a manual investigation of potentially relevant genes as part of Objective 1. Secondly, the minor allele frequency (MAF) will be used to remove any variants with a frequency greater than 0.05, thus allowing only low-frequency and rare variants to be selected for further investigation (Goswami et al., 2021). Variants with a MAF between 0.01 and 0.05 are deemed low frequency, while MAFs less than 0.01 are termed rare. In this study, the focus will be on these rare variants as these tend to display a greater impact on the phenotype than more common variants (MAF larger than 0.05) (Goswami et al., 2021). Lastly, the variants will be filtered according to the type to remove any synonymous

variants as these do not alter the encoded protein and thus, are less likely to be disease-causing. These filtering steps will generate a list of variants that will be subjected to further analysis.

3.6. Variant Prioritisation

The list of candidate variants will be prioritised further based on various *in-silico* pathogenicity prediction tools. These tools include Combined Annotation-Dependent Depletion (CADD), which provides variants (SNVs and indels) with deleteriousness scores (Rentzsch et al., 2019). The Sorting Intolerant from Tolerant (SIFT) and the Polymorphism Phenotyping version 2 (PolyPhen-2) tools are used to predict the effect of amino acid substitutions on the structure and functionality of the encoded protein (Adzhubei et al., 2010; Ng & Henikoff, 2003). Another tool that will be used is MutationTaster, which predicts whether a variant is disease-causing (Schwarz et al., 2010). Lastly, the Rare Exome Variant Ensemble Learner (REVEL) tool, which integrates results from 13 pathogenicity prediction tools to generate one pathogenicity prediction score for missense variants, will be used and compared to the individual tools (Ioannidis et al., 2016). Certain databases contain information regarding specific genes and/or variants and highlights their impact in disease. These will be used in this study to aid in variant prioritisation. Such databases will include NCBI ClinVar, the Human Gene Mutation Database (HGMD) and NCBI GeneReviews. Information obtained about these variants will be used to support the prioritisation of candidate variants to be taken forward for classification.

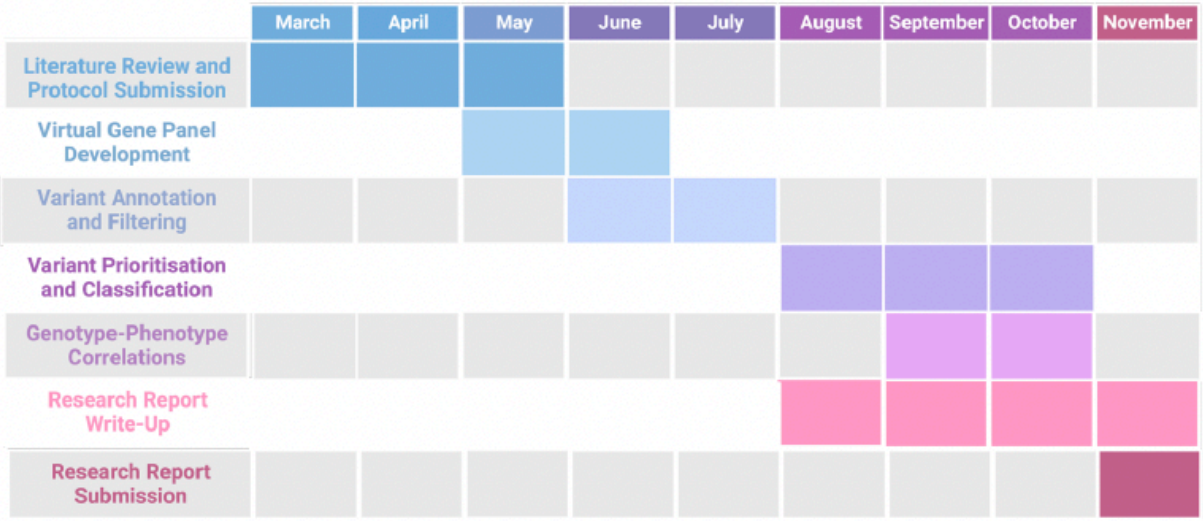
3.7. Variant Classification

Prioritised candidate variants will then be classified according to the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines (Richards et al., 2015). There are five different classes to which variants can be assigned: (1) pathogenic, (2) likely pathogenic, (3) uncertain significance, (4) likely benign, or (5) benign. Of interest to this study will be the pathogenic and likely pathogenic variants. The quality of these variants will be verified using the Integrative Genomics Viewer (IGV) to remove any false positives, and thereafter these variants will be taken forward for genotype-phenotype correlations.

3.8. Genotype-Phenotype Correlations

The clinical phenotypes of the de-identified patients with one or more variants classified as pathogenic or likely pathogenic will be made available. This will enable correlation of the genotype (variant) to the phenotype (clinical feature), where possible.

4. Time Frame



The project will begin with the literature review and submission of the protocol. Genes to include in the virtual gene panel will be researched and following approval of the protocol, the analysis of the sequence data will commence with annotation of the VCF files and filtering of the variants. Thereafter, the variants will be prioritised and classified to obtain candidate variants which will be taken forward for genotype-phenotype correlations. The research report will be written up throughout the duration of the project and will be submitted in November 2023.

5. Budget and Funding Source

Since this study involves the bioinformatic analysis of whole exome sequencing data, no funding will be required.

6. Ethics

The ethics application for this study has been submitted to the Human Research Ethics Committee and is pending approval of the protocol. It has been submitted as a sub-study under the parent study (Ethics clearance certificate number M150368).

7. Limitations of this Study

This study will perform a stepwise bioinformatic analysis of WES data from eight patient samples, for whom all previous testing strategies yielded negative results. This sample size is small, and therefore there is a chance that no disease-causing variants will be identified. Additionally, WES using NGS technology is not an ideal approach to assess repeat expansions (which have been reported to underlie several HD phenocopy syndromes). However, these patients have previously been screened for repeat expansions in five genes associated with neurodegenerative disease, with negative results. Therefore, repeat expansions in these genes are not expected to be identified in this group of patients. Another limitation is that the clinical phenotypes of these patients may not be fully described, which could hinder the genotype-phenotype correlations. Finally, given that these patients are African, and there is a low representation of African genomic data in public databases, it is expected that a number of VUSs may be identified. These will not be analysed in the current study but will be documented for future research.

8. References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249.

Anderson, D. G., Ferreira-Correia, A., Rodrigues, F. B., Aziz, N. A., Carr, J., Wild, E. J., Margolis, R. L., & Krause, A. (2019). Comparison of the Huntington's Disease like 2 and Huntington's Disease Clinical Phenotypes. *Movement Disorders Clinical Practice*, *6*(4), 302–311.

Baine, F. K., Kay, C., Ketelaar, M. E., Collins, J. A., Semaka, A., Doty, C. N., Krause, A., Greenberg, J., & Hayden, M. R. (2013). Huntington disease in the South African population occurs on diverse and ethnically distinct genetic haplotypes. *European Journal of Human Genetics*, *21*, 1120–1127.

Baine, F. K., Krause, A., & Greenberg, L. J. (2016). The Frequency of Huntington Disease and Huntington Disease-Like 2 in the South African Population. *Neuroepidemiology*, *46*(3), 198–202.

Baine, F. K., Peerbhai, N., & Krause, A. (2018). A study of Huntington disease-like syndromes in black South African patients reveals a single SCA2 mutation and a unique distribution of normal alleles across five repeat loci. *Journal of the Neurological Sciences*, *390*(2018), 200–204.

Bates, G. P., Dorsey, R., Gusella, J. F., Hayden, M. R., Kay, C., Leavitt, B. R., Nance, M., Ross, C. A., Scahill, R. I., & Wetzel, R. (2015). Huntington disease. *Nature Reviews Disease Primers*, *1*(1), 1–21.

Bourinaris, T., Athanasiou, A., Efthymiou, S., Wiethoff, S., Salpietro, V., & Houlden, H. (2021). Allelic and phenotypic heterogeneity in Junctophillin-3 related neurodevelopmental and movement disorders. *European Journal of Human Genetics*, *29*(6), 1027–1031.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Sala, S. C., Cunningham, F., Domenico, T. Di, Donaldson, S., Fiddes, I. T., Girón, C. G., & Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, *49*(D1), D916–D923.

Goswami, C., Chattopadhyay, A., & Chuang, E. Y. (2021). Rare variants: data types and analysis strategies. *Annals of Translational Medicine*, *9*(12), 961.

Hartig, M. B. (2022). Neuroferritinopathy. *Encyclopedia of Movement Disorders*, 262–

265.

Hayden, M. R., MacGregor, J. M., & Beighton, P. H. (1980). The prevalence of Huntington's chorea in South Africa. *South African Medical Journal*, *58*(5), 193–196.

Hensman, D. J., Poulter, M., Beck, J., Hehir, J., Polke, J. M., Campbell, T., Adamson, G., Mudanohwo, E., McColgan, P., Haworth, A., Wild, E. J., Sweeney, M. G., Houlden, H., Mead, S., & Tabrizi, S. J. (2014). C9orf72 expansions are the most common genetic cause of Huntington disease phenocopies. *Neurology*, *82*(4), 292–299.

Holmes, S. E., O'hearn, E., Rosenblatt, A., Callahan, C., Hwang, H. S., Ingersoll-Ashworth, R. G., Fleisher, A., Stevanin, G., Brice, A., Potter, N. T., Ross, C. A., & Margolis, R. L. (2001). A repeat expansion in the gene encoding junctophilin-3 is associated with Huntington disease-like 2. *Nature Genetics*, *29*(4), 377–378.

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., & Karyadi, D. (2016). REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, *99*(4), 877–885.

Koriath, C. A. M., Guntotoi, F., Norseworthy, P., Dolzhenko, E., Eberle, M., Moss, D. H., Flower, M., Hummerich, H., Rosser, A., Tabrizi, S. J., Mead, S., & Wild, E. J. (2022). Huntington's disease phenocopy syndromes revisited: a clinical comparison and next-generation sequencing exploration. *BioRxiv*, 1–23.

Krause, A., Mitchell, C., Essop, F., Tager, S., Temlett, J., Stevanin, G., Ross, C., Rudnicki, D., & Margolis, R. (2015). Junctophilin 3 (JPH3) Expansion Mutations Causing Huntington Disease Like 2 (HDL2) Are Common in South African Patients with African Ancestry and a Huntington Disease Phenotype. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics : The Official Publication of the International Society of Psychiatric Genetics*, *168*(7), 573–585.

Kumar, N., Rizek, P., & Jog, M. (2016). Neuroferritinopathy: Pathophysiology, Presentation, Differential Diagnoses and Management. *Tremor and Other Hyperkinetic Movements*, *6*, 355.

Le Ber, I., Camuzat, A., Castelnovo, G., Azulay, J. P., Genton, P., Gastaut, J. L., Broglin, D., Labauge, P., Brice, A., & Dürr, A. (2003). Prevalence of Dentatorubral-Pallidoluysian Atrophy in a Large Series of White Patients With Cerebellar Ataxia. *Archives of Neurology*, *60*(8), 1097–1099.

Martino, D., Stamelou, M., & Bhatia, K. P. (2013). The differential diagnosis of Huntington's disease-like syndromes: "red flags" for the clinician. *Journal of Neurology, Neurosurgery, and Psychiatry*, *84*, 650–656.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 1–14.

Medina, A., Mahjoub, Y., Shaver, L., & Pringsheim, T. (2022). Prevalence and Incidence of Huntington's Disease: An Updated Systematic Review and Meta-Analysis. *Movement Disorders*, *37*(12), 2327–2335.

Ng, P. C., & Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(Database issue), D733–D745.

Paulson, H. L., Shakkottai, V. G., Clark, H. B., & Orr, H. T. (2017). Polyglutamine spinocerebellar ataxias — from genes to potential treatments. *Nature Reviews Neuroscience*, *18*(10), 613–626.

Podvin, S., Reardon, H. T., Yin, K., Mosier, C., & Hook, V. (2019). Multiple clinical features of Huntington's disease correlate with mutant HTT gene CAG repeat lengths and neurodegeneration. *Journal of Neurology*, *266*(3), 551–564.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *17*(5), 405–424.

Rudnicki, D. D., Holmes, S. E., Lin, M. W., Thornton, C. A., Ross, C. A., & Margolis, R. L. (2007). Huntington's disease–like 2 is associated with CUG repeat-

containing RNA foci. *Annals of Neurology*, *61*(3), 272–282.

Saudou, F., & Humbert, S. (2016). The Biology of Huntingtin. *Neuron*, *89*(5), 910–926.

Schulte, J., & Littleton, J. T. (2011). The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology. *Current Trends in Neurology*, *5*, 65–78.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, *7*(8), 575–576.

Seixas, A. I., Holmes, S. E., Takeshima, H., Pavlovich, A., Sachs, N., Pruitt, J. L., Silveira, I., Ross, C. A., Margolis, R. L., & Rudnicki, D. D. (2012). Loss of junctophilin-3 contributes to huntington disease-like 2 pathogenesis. *Annals of Neurology*, *71*(2), 245–257.

Sokolov, E., Schneider, S. A., & Bain, P. G. (2012). Chorea-acanthocytosis. *Practical Neurology*, *12*(1), 40–43.

Wild, E. J., & Tabrizi, S. J. (2007). Huntington's disease phenocopy syndromes. *Current Opinion in Neurology*, *20*(6), 681–687.

Appendix D: Turnitin Report and Plagiarism Declaration



PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Racilya Naicker (Student number: 1717757) am a student registered for the degree of MSc (Med) Genomic Medicine in the academic year 2023.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: Naicker Date: 27/11/2023

Research_Report_Turnitin_Copy_Racilya_Naicker-1.docx

ORIGINALITY REPORT

15%	10%	13%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Witwatersrand Student Paper	2%
2	www.mdpi.com Internet Source	1%
3	Fiona K. Baine, Nabeelah Peerbhai, Amanda Krause. "A study of Huntington disease-like syndromes in black South African patients reveals a single SCA2 mutation and a unique distribution of normal alleles across five repeat loci", Journal of the Neurological Sciences, 2018 Publication	1%
4	www.disgenet.org Internet Source	1%
5	www.ncbi.nlm.nih.gov Internet Source	1%
6	www.frontiersin.org Internet Source	1%
7	reg.clinicalgenome.org Internet Source	1%
8	"Abstracts from the 50th European Society of Human Genetics Conference: Posters", European Journal of Human Genetics, 2018 Publication	1%
9	Submitted to University of Cape Town Student Paper	<1%
10	Jess Nithianantharajah. "Dynamic mutations as digital genetic modulators of brain development, function and dysfunction", BioEssays, 06/2007 Publication	<1%
11	G. Koutsis, G. Karadima, A. Pandraud, M. G. Sweeney, R. Paudel, H. Houlden, N. W. Wood, M. Panas. "Genetic screening of Greek patients with Huntington's disease phenocopies identifies an SCA8 expansion", Journal of Neurology, 2012 Publication	<1%
12	Klaus Seppi, Michael F.H. Schocke, Katherina J. Mair, Regina Esterhammer et al. "Diffusion-weighted imaging in Huntington's disease", Movement Disorders, 2006 Publication	<1%
13	genomebiology.biomedcentral.com Internet Source	<1%
14	Maria Kousi. "Transcriptomics in rare diseases", Elsevier BV, 2021	<1%

	Publication	
15	Robert B. Hufnagel, Wendi Liang, Jacque L. Duncan, Carmen C. Brewer et al. "Tissue-specific genotype-phenotype correlations among USH2A-related disorders in the RUSH2A study", Human Mutation, 2022 Publication	<1 %
16	cris.maastrichtuniversity.nl Internet Source	<1 %
17	Jiandong Shen, Dianyun Qu, Yan Gao, Fangxi Sun et al. "Genetic etiologic analysis in 74 Chinese Han women with idiopathic premature ovarian insufficiency by combined molecular genetic testing", Journal of Assisted Reproduction and Genetics, 2021 Publication	<1 %
18	Magdalena Podlacha, Karolina Pierzynowska, Lidia Gaffke, Grażyna Jerzemowska, Ewa Piotrowska, Grzegorz Węgrzyn. "Behavioral- and blood-based biomarkers for Huntington's disease: Studies on the R6/1 mouse model with prospects for early diagnosis and monitoring of the disease", Brain, Behavior, & Immunity - Health, 2022 Publication	<1 %
19	pure.manchester.ac.uk Internet Source	<1 %
20	www.science.gov Internet Source	<1 %
21	M. Manto. "Spinocerebellar Atrophy", Encyclopedia of Neuroscience, 2009 Publication	<1 %
22	medworm.com Internet Source	<1 %
23	pure.eur.nl Internet Source	<1 %
24	"Poster Presentations", Movement Disorders, 2012 Publication	<1 %
25	Joseph Hofman, Katrin Wudy. "In situ process monitoring in laser-based powder bed fusion of polyamide 12 using thermal imaging", The International Journal of Advanced Manufacturing Technology, 2022 Publication	<1 %
26	Margarita Chumarina, Kaspar Russ, Carla Azevedo, Andreas Heuer et al. "Cellular alterations identified in pluripotent stem cell-derived midbrain spheroids generated from a female patient with progressive external ophthalmoplegia and parkinsonism who carries a novel variation (p.Q811R) in the POLG1 gene", Acta Neuropathologica Communications, 2019	<1 %

	Publication	
27	Nikhil Shri Sahajpal, Ashis K. Mondal, Timothy Fee, Benjamin Hilton et al. "Clinical Validation and Diagnostic Utility of Optical Genome Mapping in Prenatal Diagnostic Testing", The Journal of Molecular Diagnostics, 2023 Publication	<1 %
28	Zunyan Dai, Zachary Whitt, Lindsey C. Mighion, Alessandro Pontoglio, Lora J.H. Bean, Roberto Colombo, Madhuri Hegde. "Caution in interpretation of disease causality for heterozygous loss-of-function variants in the MYH8 gene associated with autosomal dominant disorder", European Journal of Medical Genetics, 2017 Publication	<1 %
29	core.ac.uk Internet Source	<1 %
30	d.docksci.com Internet Source	<1 %
31	humgenomics.biomedcentral.com Internet Source	<1 %
32	pure.rug.nl Internet Source	<1 %
33	Fontana, Jason. "CRISPR-Cas Activation for Regulating Multi-gene Expression in Bacteria.", University of Washington, 2020 Publication	<1 %
34	Molecular Pathology in Clinical Practice, 2016. Publication	<1 %
35	Russell P. Saneto, Robert K. Naviaux. "Polymerase gamma disease through the ages", Developmental Disabilities Research Reviews, 2010 Publication	<1 %
36	Yang Kunfang, Meng Linyi, Zhang Yuanfeng, Yang Yongchen, Cheng Hongyi, Jiang Zhihu, Zhang Hong, Chen Yucai. "*POLG* Mutations Are Probably Rare in the Han Chinese Population", Chinese Medical Sciences Journal, 2020 Publication	<1 %
37	doaj.org Internet Source	<1 %
38	mg.co.za Internet Source	<1 %
39	molecularcytogenetics.biomedcentral.com Internet Source	<1 %
40	onlinelibrary.wiley.com Internet Source	<1 %
41	www.angis.org.au Internet Source	<1 %
	www.biorxiv.org	

42	Internet Source	<1%
43	www.medrxiv.org Internet Source	<1%
44	www.nature.com Internet Source	<1%
45	Baudhuin, Linnea M, Katrina E Kotzer, and Susan A Lagerstedt. "Decreased frequency of FBN1 missense variants in Ghent criteria-positive Marfan syndrome and characterization of novel FBN1 variants", Journal of Human Genetics, 2015. Publication	<1%
46	Felipe Antonio de Oliveira Garcia, Edilene Santos de Andrade, Edenir Inez Palmero. "Insights on variant analysis in silico tools for pathogenicity prediction", Frontiers in Genetics, 2022 Publication	<1%
47	Gillian M. Blue, Mauk Mekel, Debjani Das, Michael Troup et al. "Whole genome sequencing in transposition of the great arteries and associations with clinically relevant heart, brain and laterality genes.", American Heart Journal, 2021 Publication	<1%
48	L. Campbell, J. Fredericks, K. Mathivha, P. Moshesh et al. "The implementation and utility of clinical exome sequencing in a South African infant cohort", Frontiers in Genetics, 2023 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Appendix E: Author Guidelines for *Neuroepidemiology*

Research Article

Manuscript Title

First Name(s) Surname[a], First Name(s) Surname[a], First Name(s) Surname[b], First Name(s) Surname[c], First Name(s) Surname[a]

[a] Department, Institute/University/Hospital, City, (State,) Country

[b] Department, Institute/University/Hospital, City, (State,) Country

[c] Department, Institute/University/Hospital, City, (State,) Country

Short Title: to be used as running head

Corresponding Author:

First and last name

E-mail address:

Keywords: Please provide 3–5 keywords highlighting the most important points of your paper.

Abstract

A short Abstract should summarize the main points and reflect the content of the article. It should be written in a clear and concise way and be structured using the following subheadings: Introduction, Methods, Results, and Conclusion. Abbreviations used in the main text may be introduced and used. Use neither bibliographic references nor references to figures or tables in the Abstract.

Introduction

The Introduction should provide a summary of the background to the relevant field of research and the specific problems addressed and should state the hypotheses being explored as well as the main goal(s) of the study. Conclusions or findings should not appear in the Introduction.

Methods

The Methods section should clearly list all inclusion and exclusion criteria, methods of research, and variables evaluated and should state how outcomes were assessed. All terms should be adequately defined and statistical information should be sufficiently detailed so that a study can be repeated. If your manuscript is a clinical trial, please provide the clinical trial number.

Results

The Results section should describe the most important findings of the study, analysis, or experiment. The most important results should be indicated, and relevant trends and patterns should be described.

Discussion

The Discussion should provide an evaluation of the results. There should be a clear discussion of the implications, significance, and novelty of the results presented and whether the data support or contradict previous studies. A final conclusion may be added.

Statements

All papers must contain the following statements after the main body of the text and before the reference list. More detailed information can be found at Publication Ethics and Editorial Policies | Karger Publishers.

Acknowledgement (optional)

In the Acknowledgement section, authors may include individuals, who are not listed as authors, and organizations that have made substantive contributions to the research or the manuscript. An exception is where funding was provided, which should be included in Funding Sources.

Statement of Ethics

Please address the following aspects in your Statement of Ethics.

Study approval statement: Provide name and affiliation of the committee who approved the study and the decision reference number. An example statement can be found here: "*This study protocol was reviewed and approved by [committee name and affiliation], approval number [XXX]*." If ethics approval was not required, or if the study has been granted an exemption from requiring ethics approval, this should also be stated, including the name of the ethics committee who made that decision.

Consent to participate statement: For studies using human participants, state whether written informed consent was obtained from participants (or their parent/legal guardian/next of kin) to participate in the study. If written informed consent was not required, or if the study has been granted an exemption from requiring written informed consent, this should also be stated, including the name of the ethics committee who made that decision.

Conflict of Interest Statement

Authors are required to disclose any possible conflicts of interest. All forms of support and financial involvement (e.g. employment, consultancies, honoraria, stock ownership and options, expert testimony, grants or patents received or pending, royalties) which took place in the previous three years should be listed, regardless of their potential relevance to the paper. Also the nonfinancial relationships (personal, political, or professional) that may potentially influence the writing of the

manuscript should be declared. If there is no conflict of interest, please state: “The authors have no conflicts of interest to declare.”

Funding Sources

Authors must give full details about the funding of any research relevant to their study, including sponsor names and explanations of the roles of these sources in the study design, execution and analysis, and manuscript conception, planning, writing and decision to publish. If the sponsor or funder had no role in any of the above, please use the following statement: "The funder had no role in the design, data collection, data analysis, and reporting of this study." Please ensure to include any support that could be perceived as a potential conflict of interest in the Conflict of Interest Statement. It is strongly advised to write out the funding body in full and add the grant number in brackets. Multiple grant numbers should be separated by commas and spaces. If no funding was received for the study, please use the following statement: "This study was not supported by any sponsor or funder."

Author Contributions

In the Author Contributions section, a short statement detailing the contributions of each person named as an author should be included. Please include the initials of each author and their contribution. Contributors to the paper who do not fulfil the ICMJE Criteria for Authorship should be credited in the Acknowledgement section.

Data Availability Statement

Authors are required to provide a Data Availability Statement in their article that details whether data are available and where they can be found. The journal’s data sharing policy strongly encourages authors to make all datasets on which the conclusions of the paper rely available to editors, reviewers and readers without unnecessary restriction wherever possible. In cases where research data are not publicly available on legal or ethical grounds, this should be clearly stated in the Data Availability Statement along with any conditions for accessing the data. Examples of Data Availability Statements and additional information on the data sharing policy can be found at Publication Ethics and Editorial Policies | Karger Publishers.

References [Numerical]

References in the text should be identified sequentially using Arabic numerals [in square brackets].

The reference list should include only those publications which are cited in the text, arranged numerically in the order in which they are cited. More information and examples can be found on the Author Guidelines of the respective journal.

Figure Legends

Fig. 1. Legend text.

Fig. 2. Legend text.

Figures should be mentioned in the manuscript text as follows:
Without round brackets:

"…shown in Figure 1…" or "…shown in Figures 1 and 4…" or "…shown in Figures 2–6…" always with capital letters and written out.

With round brackets:

"(shown in Fig. 1)" or "(shown in Fig. 1, 4)" or "(shown in Fig. 2–6)", always abbreviated as "Fig." followed by the number or numbers after a full stop and a space.

In the Legend:

"Fig. 1." or "Fig. 1. a", always abbreviated as "Fig." followed by the number after a full stop and a space

Please note that the actual figures and all tables should be uploaded as separate items in their original file format.