Investigating genetic causes of neurological movement disorders

Kafantari, Efthymia

2026

Document Version:
Publisher's PDF, also known as Version of record

Investigating genetic causes of neurological movement disorders

EFTHYMIA KAFANTARI
DEPARTMENT OF CLINICAL SCIENCES, LUND | FACULTY OF MEDICINE | LUND UNIVERSITY

Investigating genetic causes of neurological movement disorders

Investigating genetic causes of neurological movement disorders

Efthymia Kafantari

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Medicine, Lund University, Sweden. To be defended at Segerfalksalen on 2 June 2026 at 13:00.

Faculty opponent:
Enza Maria Valente, Professor of Medical Genetics, University of Pavia, Italy

Organization: LUND UNIVERSITY

Document name: Doctoral Thesis **Date of issue:** June 2, 2026

Author(s): Efthymia Kafantari **Sponsoring organization:** Lund University

Title and subtitle: Investigating genetic causes of neurological movement disorders

Abstract: Neurological movement disorders such as dystonia, hereditary ataxia, and Parkinson's disease (PD) are characterized by genetic and clinical heterogeneity, making diagnosis challenging. Next-generation sequencing (NGS) has enabled comprehensive genetic analysis allowing both known and novel pathogenic variants to be identified. In this study, multiple Swedish cases and families with movement disorders were analyzed for uncovering their genetic causes. The development of advanced bioinformatics pipelines enabled the processing and prioritization of variants from large NGS datasets, for the identification of pathogenic variants and the detection of complex variations.

In dystonia, in two families, rare variants in TOR1AIP2 encoding LULL1 (an activator of TorsinA involved in nuclear envelope function), were identified. Functional studies showed that TorsinA-LULL1 interaction is disrupted, similar to the known TOR1A mutation causing early-onset isolated dystonia (DYT1). The patients manifested dystonia with hemichorea and hemiballism. TOR1AIP2 is proposed as a novel candidate gene for hereditary nuclear envelope-related movement disorders. For hereditary ataxia, 87 patients from 76 families were genetically screened. A diagnostic yield of 39.5% was obtained. The study highlights the importance of integrating clinical and genetic data. Remarkably, GGC trinucleotide repeat expansions in ZFHX3 were found in affected members of five Swedish families, establishing these expansions as the genetic cause of spinocerebellar ataxia type 4 (SCA4). Affected individuals showed progressive balance and gait issues, sensory neuropathy and autonomic dysfunction. Whole-exome sequencing (WES) data from 285 PD cases were analyzed for monogenic causes, including early-onset and familial forms. Pathogenic variants were identified in several known PD associated genes, VPS35, SNCA, LRRK2, GBA1. A novel CHCHD2 frameshift variant was found in two unrelated patients. However, Mendelian PD forms accounted for only 2.1% of cases, emphasizing PD's complex genetics.

By integrating NGS, bioinformatics, and clinical data, diagnostic precision was improved and our understanding of the genetic background of movement disorders was broadened.

Key words: Movement disorders; dystonia; hereditary ataxia; Parkinson's disease; Next-generation sequencing; bioinformatics; genetic investigation; diagnostic yield

Classification system and/or index terms (if any) Supplementary bibliographical information

Language: English **Number of pages**: 102

ISSN and key title: 1652-8220

ISBN: 978-91-8021-889-4

Recipient's notes Price Security classification

Signature Date 2026-03-25

Investigating genetic causes of neurological movement disorders

Efthymia Kafantari

LUND
UNIVERSITY

Published by:

Department of Clinical Sciences, Lund

Faculty of Medicine

Lund University

Lund 2026

ISBN 978-91-8021-889-4

Series title: Lund University, Faculty of Medicine Doctoral Dissertation Series 2026:91

ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University,

Lund, 2026

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

ἓν οἶδα ὅτι οὐδὲν οἶδα
Socrates

Table of Contents

List of papers

Paper I:

Kafantari E, Hernandez VJ, Necpál J, Leonidou M, Baureder R, Hedberg-Oldfors C, Jech R, Zech M, Schwartz TU, Puschmann A.

TOR1AIP2 as a candidate gene for dystonia-hemichorea/hemiballism.

Parkinsonism Relat Disord. 2025 May;134:107781. doi: 10.1016/j.parkreldis.2025.107781.

Paper II:

Gorcenco S[#], **Kafantari E**[#], Wallenius J, Karremo C, Alinder E, Dobloug S, Landqvist Waldö M, Englund E, Ehrencrona H, Wictorin K, Karrman K, Puschmann A.

Clinical and genetic analyses of a Swedish patient series diagnosed with ataxia.

J Neurol. 2024 Jan;271(1):526-542. doi: 10.1007/s00415-023-11990-x.

# Equal contributions

Paper III:

Wallenius J, **Kafantari E**, Jhaveri E, Gorcenco S, Ameur A, Karremo C, Dobloug S, Karrman K, de Koning T, Ilinca A, Landqvist Waldö M, Arvidsson A, Persson S, Englund E, Ehrencrona H, Puschmann A.

Exonic trinucleotide repeat expansions in ZFHX3 cause spinocerebellar ataxia type 4: A poly-glycine disease.

Am J Hum Genet. 2024 Jan 4;111(1):82-95. doi: 10.1016/j.ajhg.2023.11.008.

Paper IV:

Kafantari E, Atterling Brolin K, Wallenius J, Swanberg M, Puschmann A.

WES-based screening of a Swedish patient series with Parkinson's disease.

Genes. 2025 Dec;16(12):1482. doi: 10.3390/genes16121482

Abstract

Neurological movement disorders such as dystonia, hereditary ataxia, and Parkinson's disease (PD) are characterized by genetic and clinical heterogeneity, making diagnosis challenging. Next-generation sequencing (NGS) has enabled comprehensive genetic analysis allowing both known and novel pathogenic variants to be identified. In this study, multiple Swedish cases and families with movement disorders were analyzed for uncovering their genetic causes.

The development of advanced bioinformatics pipelines enabled the processing and prioritization of variants from large NGS datasets, for the identification of pathogenic variants and the detection of complex variations.

In dystonia, in two families, rare variants in *TOR1AIP2* encoding LULL1 (an activator of TorsinA involved in nuclear envelope function), were identified. Functional studies showed that TorsinA-LULL1 interaction is disrupted, similar to the known *TOR1A* mutation causing early-onset isolated dystonia (DYT1). The patients manifested dystonia with hemichorea and hemiballism. *TOR1AIP2* is proposed as a novel candidate gene for hereditary nuclear envelope-related movement disorders.

For hereditary ataxia, 87 patients from 76 families were genetically screened. A diagnostic yield of 39.5% was obtained. The study highlights the importance of integrating clinical and genetic data. Remarkably, GGC trinucleotide repeat expansions in *ZFHX3* were found in affected members of five Swedish families, establishing these expansions as the genetic cause of spinocerebellar ataxia type 4 (SCA4). Affected individuals showed progressive balance and gait issues, sensory neuropathy and autonomic dysfunction.

Whole-exome sequencing (WES) data from 285 PD cases were analyzed for monogenic causes, including early-onset and familial forms. Pathogenic variants were identified in several known PD-associated genes*, VPS35, SNCA, LRRK2***,** *GBA1***.** A novel *CHCHD2* frameshift variant was found in two unrelated patients. However, Mendelian PD forms accounted for only 2.1% of cases, emphasizing PD's complex genetics.

By integrating NGS, bioinformatics, and clinical data, diagnostic precision was improved and our understanding of the genetic background of movement disorders was broadened.

Popular Science Summary

Movement disorders such as dystonia**,** ataxia**,** and Parkinson's disease (PD) are neurological conditions which disrupt the brain's ability to control movement, coordination and balance. They can cause abnormal involuntary movements, tremor, unsteady gait, or slowness of movement. They may occur sporadically but in many cases there is an underlying genetic cause. Identifying the pathogenic variant or responsible gene is often challenging, as clinical symptoms resulting from different genetic causes may be similar.

This thesis investigates the genetic causes of dystonia, hereditary ataxias, and PD, in Swedish patients, using whole-exome sequencing (WES) and whole-genome sequencing (WGS). These techniques enable comprehensive screening of the protein-coding genes, allowing the identification of genetic variants. Advanced bioinformatics methods were developed for the analysis of the large amount of raw sequencing data which are transformed into interpretable results through several computational steps. The combination of computational predictions with clinical data and experimental techniques enabled the identification of both known and novel genetic causes.

Dystonia – identifying a new candidate gene:

Dystonia is a movement disorder characterized by involuntary muscle contractions and abnormal movements and postures. In one Swedish family with dystonia, hemichorea, and hemiballism, a rare variant in the *TOR1AIP2* gene was identified. This gene encodes LULL1, a nuclear envelope protein which interacts with TorsinA, the product of *TOR1A*, a well-established dystonia gene. Functional studies confirmed that the *TOR1AIP2* p.(Arg412Gly) variant weakens the TorsinA–LULL1 interaction, mimicking the early-onset isolated dystonia DYT1-causing *TOR1A* mutation. A second family with a different *TOR1AIP2* variant showed a milder, related movement disorder. These findings propose *TOR1AIP2* as a novel candidate gene for hereditary dystonia with overlapping hyperkinetic symptoms, a hereditary nuclear envelope-related movement disorder.

Hereditary ataxia – confirmed genetic diagnosis in 39.5% of cases:

Ataxia refers to a group of disorders causing impaired coordination and unsteady movements often due to cerebellar dysfunction. In a cohort of 87 Swedish patients from 76 families with progressive ataxia, genetic testing using WES or WGS identified a clear genetic diagnosis in nearly 40% of families, though several variants remained of uncertain significance. This study highlights the importance of combining genomic data analysis with detailed clinical phenotyping and close

collaboration between clinicians, bioinformaticians and geneticists to achieve accurate diagnoses.

Autosomal-dominant ataxia with sensory and autonomic neuropathy – addressing a long-unresolved genetic enigma:

In two Swedish kindreds with a distinctive form of ataxia combined with sensory and autonomic neuropathy, WGS and bioinformatic repeat expansion analysis identified exonic GGC trinucleotide repeat expansions in the *ZFHX3* gene. These expansions encode abnormally long poly-glycine tracts that presumably lead to toxic protein accumulation in nerve cells. Anticipation (earlier and more severe symptoms in successive generations) was observed. Three more Swedish families were found to also carry repeat expansions in *ZFHX3* gene. This discovery demonstrated that spinocerebellar ataxia type 4 (SCA4) is caused by *ZFHX3* repeat expansions, resolving a 25-year-old genetic enigma first described in a family of Swedish origin.

Parkinson's disease – rare monogenic forms in a Swedish cohort:

PD is a common neurodegenerative disorder, characterized by tremor, rigidity, and slowness of movement. Screening of WES data from 285 PD patients from southern Sweden, focusing on 44 PD-related genes identified pathogenic variants in *SNCA, LRRK2* including the first Swedish case with the *VPS35* p.(Asp620Asn) variant and a novel *CHCHD2* frameshift variant which was found in two unrelated patients. *GBA1* variants were the most frequent, identified in 15.1% of patients. However, monogenic (single-gene) PD forms explained only 2.1% of cases, underling the complex aetiology of PD. This study provides a genetic overview of PD in southern Sweden and further confirms *CHCHD2* as a PD gene.

Together, these studies reveal the genetic complexity of movement disorders in the Swedish population. The identification of *TOR1AIP2* as a dystonia-associated gene, the discovery of *ZFHX3* repeat expansions as the cause of SCA4, and the finding of a novel *CHCHD2* PD variant illustrate how NGS and bioinformatics analysis can uncover novel genetic causes successfully. This thesis demonstrates that combining genomic technology**,** bioinformatic analysis, and clinical expertise greatly enhances our understanding of neurological movement disorders. Such interdisciplinary approaches not only improve diagnostic precision but also broaden the genetic spectrum of movement disorders.

Populärvetenskaplig sammanfattning

Rörelsestörningar såsom dystoni, ataxi och Parkinsons sjukdom (PD) är neurologiska tillstånd som påverkar hjärnans förmåga att kontrollera rörelse, koordination och balans. De kan orsaka abnorma ofrivilliga rörelser, tremor, ostadig gång eller långsamma rörelser. Störningarna kan uppstå sporadiskt, men i många fall finns en underliggande genetisk orsak. Att identifiera den sjukdomsorsakande varianten eller genen är ofta utmanande, eftersom kliniska symtom från olika genetiska orsaker kan likna varandra.

Denna avhandling undersöker de genetiska orsakerna till dystoni, hereditär ataxi och PD hos svenska patienter, med hjälp av hel-exomsekvensering (WES) och hel-genomsekvensering (WGS). Dessa tekniker möjliggör en omfattande kartläggning av de protein-kodande generna, vilket gör det möjligt att identifiera genetiska varianter. Avancerade bioinformatiska metoder utvecklades för analys av den stora mängden rådata från sekvenseringen, som omvandlas till tolkningsbara resultat genom flera beräkningssteg. Kombinationen av datorbaserade prediktioner med kliniska data och experimentella tekniker identifierande både kända och nya genetiska orsaker.

Dystoni – identifiering av en ny kandidatgen:

Dystoni är en rörelsestörning som kännetecknas av ofrivilliga muskelkontraktioner samt onormala rörelser och kroppsställningar. I en svensk familj med dystoni, hemikorea och hemiballism identifierades en sällsynt variant i *TOR1AIP2*-genen**.** Denna gen kodar för LULL1**,** ett protein i kärnmembranet som interagerar med TorsinA**,** produkten av den välkända dystonigenen *TOR1A***.** Funktionella studier visade att *TOR1AIP2*-varianten p.(Arg412Gly) försvagar TorsinA–LULL1-interaktionen, vilket liknar den *TOR1A*-mutation som orsakar tidig debut av isolerad dystoni (DYT1). En andra familj med en annan *TOR1AIP2*-variant visade en mildare, relaterad rörelsestörning. Dessa fynd föreslår *TOR1AIP2* som en ny kandidatgen för ärftlig dystoni med överlappande hyperkinetiska symtom, en ny ärftlig rörelsestörning relaterad till kärnmembranet.

Ärftlig ataxi – genetisk diagnos i 39,5 % av fallen:

Ataxi avser en grupp störningar som orsakar nedsatt koordination och ostadiga rörelser, ofta på grund av cerebellär dysfunktion. I en kohort av 87 svenska patienter från 76 familjer med kroniskt framskridande ataxi identifierade genetisk testning med WES eller WGS en tydlig genetisk diagnos i nästan 40 % av familjerna, även om flera varianter fortfarande var av osäker betydelse. Studien understryker vikten av att kombinera analys av genomiska data med noggrann klinisk fenotypning och ett nära samarbete mellan kliniker, bioinformatiker och genetiker för att uppnå korrekta diagnoser.

Autosomal dominant ataxi med sensorisk och autonom neuropati – att ta itu med ett sedan länge olöst genetisk gåta:

I två svenska familjer med en särpräglad form av ataxi kombinerad med sensorisk och autonom neuropati identifierade WGS och bioinformatisk analys av upprepningsexpansioner exoniska GGC-trinukleotidupprepningar som expanderats i *ZFHX3*-genen. Dessa upprepningar kodar för onormalt långa poly-glycin-sekvenser som antas leda till toxisk proteinansamling i nervceller. Anticipation (tidigare och allvarligare symtom i efterföljande generationer) observerades. Tre ytterligare svenska familjer visade sig också bära upprepningsexpansioner i *ZFHX3*-genen. Denna upptäckt visade att spinocerebellär ataxi typ 4 (SCA4) orsakas av trinukleotidupprepningsexpansioner i *ZFHX3* och löste ett 25 år gammalt genetisk gåta som först beskrevs en släkt av svenskt urprungi.

Parkinsons sjukdom – sällsynta monogena former i en svensk kohort:

PD är en vanlig neurodegenerativ sjukdom som kännetecknas av tremor, rigiditet och långsamma rörelser. Analys av WES-data från 285 PD-patienter från södra Sverige, med fokus på 44 PD-relaterade gener, identifierade sjukdomsorsakande varianter i *SNCA, LRRK2*, samt det första svenska fallet med *VPS35* p.(Asp620Asn) och en ny läsramsförskjutningsmutation i *CHCHD2* som hittades hos två obesläktade patienter. *GBA1*-varianter var de vanligaste och identifierades hos 15,1 % av patienterna. Monogena former av PD (variation i en enda gen) förklarade dock endast 2,1 % av fallen, vilket understryker sjukdomens komplexa etiologi. Studien ger en genetisk översikt över PD i södra Sverige och bekräftar *CHCHD2* som en PD-gen.

Tillsammans visar dessa studier den genetiska komplexiteten hos rörelsestörningar i den svenska populationen. Identifieringen av *TOR1AIP2* som en dystoni-associerad gen, upptäckten av upprepningsexpansioner i *ZFHX3* som orsaken till SCA4 och upptäckten av en ny PD-variant i *CHCHD2* illustrerar hur NGS och bioinformatisk analys framgångsrikt kan avslöja nya genetiska orsaker. Denna avhandling visar att kombinationen av genomisk teknik, bioinformatisk analys och klinisk expertis förbättrar vår förståelse av neurologiska rörelsestörningar. Sådana tvärvetenskapliga metoder förbättrar inte bara diagnostisk precision utan breddar också det genetiska spektrumet av rörelsestörningar.

Περίληψη

Οι κινητικές διαταραχές όπως η δυστονία, η αταξία και η νόσος Πάρκινσον (PD) είναι νευρολογικές παθήσεις που επηρεάζουν την ικανότητα του εγκεφάλου να ελέγχει τις κινήσεις, τον συντονισμό και την ισορροπία. Μπορούν να προκαλέσουν μη φυσιολογικές ακούσιες κινήσεις, τρέμουλο, ασταθή βάδιση ή επιβράδυνση των κινήσεων. Μπορεί να εμφανίζονται σποραδικά, αλλά σε πολλές περιπτώσεις υπάρχει υποκείμενη γενετική αιτία. Ο εντοπισμός της παθογόνου μετάλλαξης ή του υπεύθυνου γονιδίου είναι συχνά δύσκολη, καθώς κλινικά συμπτώματα που προκύπτουν από διαφορετικούς γενετικούς παράγοντες μπορεί να είναι παρόμοια.

Η παρούσα διατριβή διερευνά τις γενετικές αιτίες της δυστονίας, των κληρονομικών αταξιών και της νόσου Πάρκινσον, σε Σουηδούς ασθενείς, χρησιμοποιώντας ολική αλληλούχιση εξωνίων (WES) και ολική αλληλούχιση γονιδιώματος (WGS). Αυτές οι τεχνικές επιτρέπουν την εκτενή ανάλυση των γονιδίων που κωδικοποιούν τις πρωτεΐνες και καθιστούν δυνατή την εύρεση γενετικών μεταλλάξεων. Αναπτύχθηκαν προηγμένες βιοπληροφορικές μέθοδοι για την ανάλυση του μεγάλου όγκου των ακατέργαστων δεδομένων αλληλούχισης, τα οποία μετατρέπονται σε ερμηνεύσιμα αποτελέσματα μέσω πολλών υπολογιστικών βημάτων. Ο συνδυασμός υπολογιστικών προβλέψεων με κλινικά δεδομένα και πειραματικές τεχνικές επέτρεψε την εύρεση τόσο γνωστών όσο και νέων γενετικών αιτιών.

Δυστονία – εύρεση νέου υποψήφιου γονιδίου:

Η δυστονία χαρακτηρίζεται από ακούσιες μυϊκές συσπάσεις και μη φυσιολογικές κινήσεις και στάσεις σώματος. Σε μια σουηδική οικογένεια με δυστονία σε συνδυασμό με ημικορεία και ημιβαλισμό, ανακαλύφθηκε μια σπάνια μετάλλαξη στο γονίδιο *TOR1AIP2*. Το γονίδιο αυτό κωδικοποιεί την πρωτεΐνη LULL1, μια πρωτεΐνη της πυρηνικής μεμβράνης που αλληλεπιδρά με την TorsinA, το προϊόν του γνωστού γονιδίου δυστονίας, *TOR1A*. Πειραματικές μελέτες έδειξαν ότι η μετάλλαξη *TOR1AIP2* p.(Arg412Gly) μειώνει τη σύνδεση TorsinA–LULL1, με τρόπο παρόμοιο με την *TOR1A* μετάλλαξη που προκαλεί τη κλασική πρόωρης έναρξης δυστονία (DYT1). Μια δεύτερη οικογένεια με διαφορετική μετάλλαξη στο *TOR1AIP2* γονίδιο παρουσίασε ηπιότερη αλλά παρόμοια κινητική διαταραχή. Αυτά τα ευρήματα προτείνουν το *TOR1AIP2* ως νέο υποψήφιο γονίδιο για κληρονομική δυστονία με υπερκινητικά συμπτώματα, μια νέα κληρονομική διαταραχή κίνησης του πυρηνικού φακέλου.

Κληρονομική αταξία – γενετική διάγνωση στο 39,5% των περιπτώσεων:

Η αταξία αναφέρεται σε μια ομάδα διαταραχών που προκαλούν μειωμένο συντονισμό και ασταθείς κινήσεις, συχνά λόγω δυσλειτουργίας της

παρεγκεφαλίδας (cerebellum). Σε μια ομάδα 87 Σουηδών ασθενών από 76 οικογένειες με προοδευτική αταξία, οι αναλύσεις WES και WGS επέτρεψαν την ταυτοποίηση σαφούς γενετικής διάγνωσης σε σχεδόν 40% των εξεταζόμενων οικογενειών αν και πολλές μεταλλάξεις παρέμειναν αβέβαιες. Η μελέτη δείχνει τη σημασία της συνδυαστικής χρήσης γενετικών αναλύσεων με λεπτομερή κλινική φαινοτυπική εκτίμηση και της στενής συνεργασίας κλινικών γιατρών, βιοπληροφορικών και γενετιστών για ακριβείς διαγνώσεις.

Αυτοσωματική επικρατής αταξία με αισθητική και αυτόνομη νευροπάθεια – εξετάζοντας ένα μακροχρόνιο και άλυτο γενετικό αίνιγμα:

Σε δύο σουηδικές οικογένειες με ιδιαίτερη μορφή αταξίας σε συνδυασμό με αισθητική και αυτόνομη νευροπάθεια, μέσω WGS και βιοπληροφορικής ανάλυσης επαναλήψεων, εντοπίστηκαν GGC τρινοκλεοτιδικές επαναλήψεις στο γονίδιο *ZFHX3*. Αυτές οι επαναλήψεις κωδικοποιούν ασυνήθιστα μεγάλες αλυσίδες πολυγλυκίνης που πιθανώς οδηγούν σε τοξική συσσώρευση πρωτεϊνών στα νευρικά κύτταρα. Παρατηρήθηκε **'**επίσπευση', δηλαδή εκδήλωση σοβαρότερων συμπτωμάτων νωρίτερα στις επόμενες γενιές. Τρεις ακόμα οικογένειες με αταξία από τη νότια Σουηδία βρέθηκαν με τις ίδιες επεκτάσεις στο *ZFHX3* γονίδιο. Η ανακάλυψη αυτή έδειξε ότι η **ν**ωτιαιοπαρεγκεφαλιδική αταξία τύπου 4 (SCA4) προκαλείται από τις επαναλήψεις στο *ZFHX3* γονίδιο, λύνοντας ένα 25ετές γενετικό αίνιγμα που είχε πρωτοπεριγραφεί σε μία οικογένεια σουηδικής καταγωγής.

Νόσος Πάρκινσον – σπάνιες μονογονιδιακές μορφές σε σουηδική ομάδα μελέτης:

Η νόσος Πάρκινσον (PD) είναι μια νευροεκφυλιστική πάθηση η οποία χαρακτηρίζεται από τρέμουλο, ακαμψία και επιβράδυνση των κινήσεων. Η ανάλυση δεδομένων WES από 285 ασθενείς με PD από τη νότια Σουηδία, επικεντρωμένη σε 44 γονίδια σχετιζόμενα με την νόσο, εντόπισε παθογόνες μεταλλάξεις στα γονίδια *SNCA, LRRK2* περιλαμβανομένου του πρώτου σουηδικού περιστατικού με *VPS35* p.(Asp620Asn) και μιας νέας *CHCHD2* μετάλλαξης μετατόπισης πλαισίου ανάγνωσης σε δύο μη σχετιζόμενους ασθενείς. Οι μεταλλαγές στο γονίδιο *GBA1* ήταν οι πιο συχνές, εντοπισμένες σε περίπου 15% των ασθενών. Ωστόσο, οι μονογονιδιακές μορφές PD εξήγησαν μόνο το 2,1% των περιπτώσεων, υπογραμμίζοντας την πολύπλοκη αιτιολογία της νόσου. Η μελέτη παρέχει μια γενετική επισκόπηση της PD στη νότια Σουηδία και επιβεβαιώνει περαιτέρω το *CHCHD2* ως γονίδιο σχετιζόμενο με την νόσο.

Συνολικά, αυτές οι μελέτες δείχνουν την γενετική πολυπλοκότητα των κινητικών διαταραχών στον σουηδικό πληθυσμό. Η ταυτοποίηση του *TOR1AIP2* ως γονιδίου σχετιζόμενο με δυστονία, η ανακάλυψη των *ZFHX3* τρινοκλεοτιδικών επαναλήψεων ως αιτία της SCA4 και η ανακάλυψη της νέας *CHCHD2* σχετιζόμενη με PD μετάλλαξης δείχνουν πώς οι σύγχρονες τεχνολογίες αλληλούχισης και η

βιοπληροφορική μπορούν να ανιχνεύσουν νέες γενετικές αιτίες επιτυχώς. Η παρούσα διατριβή καταδεικνύει ότι ο συνδυασμός γονιδιωματικής τεχνολογίας, βιοπληροφορικής ανάλυσης και κλινικής αξιολόγησης ενισχύουν σημαντικά την κατανόηση των νευρολογικών κινητικών διαταραχών. Τέτοιες διεπιστημονικές προσεγγίσεις όχι μόνο βελτιώνουν την ακρίβεια της διάγνωσης αλλά και συμβάλλουν στη διεύρυνση του γενετικού φάσματος των κινητικών διαταραχών.

Abbreviations

ACMG	American College of Medical Genetics and Genomics
ADCA	Autosomal dominant cerebellar ataxias
ADCADN	Autosomal dominant cerebellar ataxia, deafness, and narcolepsy
ARCA	Autosomal recessive cerebellar ataxias
ARSA	Arylsulfatase A
AT	Ataxia–telangiectasia
ATP7B	ATPase copper transporting beta
ATXPC	Ataxia-pancytopenia syndrome
BAM	Binary alignment map
bcftools	Binary call format tools
BF	Bayes factor
BG	Basal ganglia
BLAT	BLAST-like alignment
BQSR	Base quality score recalibration
BVVLS2	Brown–Vialetto–Van Laere syndrome-2
BWA-MEM	Burrows–Wheeler aligner-maximum exact matches
CADD	Combined annotation dependent depletion
CANVAS	Cerebellar ataxia, neuropathy and vestibular areflexia syndrome
CAPN1	Calpain 1
CCDS	Consensus coding DNA sequence
CDS	Coding DNA sequence
CHCHD2	Coiled-coil-helix-coiled-coil-helix domain containing 2
CNS	Central nervous system
CNV	Copy number variation
dbNSFP	Database for nonsynonymous SNPs' functional predictions
dbSNP	Database for single nucleotide polymorphisms
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
DNA	Deoxyribonucleic acid
DP	Depth of coverage
DYT1	Early-onset isolated dystonia type 1
EA	Episodic ataxias
ELOVL4	ELOVL fatty acid elongase 4
FRDA	Friedreich ataxia
FXTAS	Fragile X tremor ataxia syndrome
GATK	Genome Analysis Toolkit
GBA1	Glucosylceramidase beta 1

GERP	Genomic Evolutionary Rate Profiling
GNAL	G protein subunit alpha L
gnomAD	Genome Aggregation Database
GQ	Genotype quality
GRCh37	Genome Reference Consortium Human build 37
GRCh38	Genome Reference Consortium Human build 38
gVCF	Genomic variant call format
GWAVA	Genome Wide Annotation of Variants
hap-ibd	Haplotype-based identity-by-descent
HD	Huntington's disease
HGNC	HUGO Gene Nomenclature Committee
HGVS	Human Genome Variation Society
hg19	Human genome version 19
hg38	Human genome version 38
HPO	Human Phenotype Ontology
HTT	Huntingtin
HWE	Hardy-Weinberg Equilibrium
IBD	Identity-by-descent
IBS	Identity-by-state
ICARS	International Cooperative Ataxia Rating Scale
ICD-10	International Statistical Classification of Diseases and Related Health Problems 10th Revision
IGV	Integrative Genomic Viewer
Indels	Insertions and deletions
IRRs	In-repeat reads
KMT2B	Lysine methyltransferase 2B
LAP1	Lamina-associated polypeptide 1
LRRK2	Leucine-rich repeat kinase 2
LULL1	Luminal domain-like LAP1
MAF	Minor allele frequency
MD	Movement disorders
MDCS	Malmö Diet and Cancer Study
MPBC	MultiPark Biobank Sample Collection
MRI	Magnetic Resonance Imaging
mRNA	Messenger ribonucleic acid
mtDNA	Mitochondrial DNA
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NMD	Nonsense-mediated mRNA decay
OMIM	Online Mendelian Inheritance in Man
PacBio	Pacific Biosciences
PARK2	Parkin RBR E3 ubiquitin protein ligase
PARK7	Parkinsonism associated deglycase
PARLU	PARkinsonLUnd study

PCR	Polymerase Chain Reaction
PD	Parkinson's disease
PINK1	PTEN induced kinase 1
POLR3B	RNA polymerase III subunit B
PolyPhen-2	Polymorphism Phenotyping v2
pre-mRNA	Premature messenger ribonucleic acid
PRKN	Parkin RBR E3 ubiquitin protein ligase
PROVEAN	Protein Variation Effect Analyzer
REVEL	Rare Exome Variant Ensemble Learner
RFC1	Replication factor C subunit 1
rRNA	Ribosomal ribonucleic acid
SAM	Sequence alignment map
SAMD9L	Sterile alpha motif domain containing 9 like
SARA	Scale for the Assessment and Rating of Ataxia
SCA	Spinocerebellar ataxias
SCAR16	Spinocerebellar ataxia autosomal recessive 16
SIFT	Sorting Intolerant From Tolerant
SLC52A2	Solute carrier family 52 member 2
SNCA	Synuclein alpha
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
STR	Short tandem repeat
STUB1	STIP1 homology and U-box containing protein 1
SPAST	Spastin
SPG	Spastic paraplegia
SweFreq	The Swedish Frequency resource for genomics
THAP1	THAP domain containing 1
TOR1A	Torsin family 1 member A
TOR1AIP1	Torsin 1A interacting protein 1
TOR1AIP2	Torsin 1A interacting protein 2
TraP	Transcript-inferred Pathogenicity
tRNA	Transfer ribonucleic acid
UTR	Untranslated region
VCF	Variant calling format
VEP	Variant Effect Predictor
VPS13C	Vacuolar protein sorting 13 homolog C
VPS16	Vacuolar protein sorting-associated protein 16
VPS35	Vacuolar protein sorting-associated protein 35
VQSLOD	Variant quality score log-odds
VQSR	Variant quality score recalibration
VUS	Variants of uncertain significance
WES	Whole exome sequencing
WGS	Whole genome sequencing
ZFHX3	Zinc finger homeobox 3

Introduction

Genetics

The nuclear human genome

The human genetic material (the human genome) comprises approximately 6.2 billion base pairs (diploid organism) of deoxyribonucleic acid (DNA) organized into 23 pairs of chromosomes: 22 autosomes and one pair of sex chromosomes (X and Y). The DNA is located in the cells' nucleus. It contains approximately 20,000 protein-coding genes.[1] The protein-coding parts of these genes represent only about 1–2% of the genome, while the remaining 98–99% consists of non-coding regions, including regulatory sequences, introns, and intergenic DNA.[2]

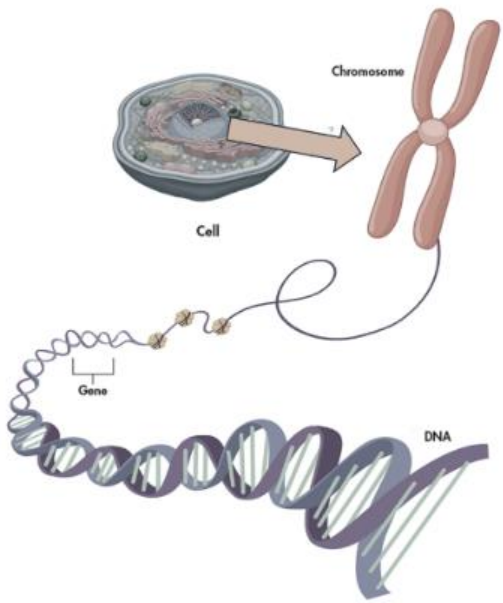

Figure 1:The human genome. The chromosomes located in the cell nucleus, and the DNA contains genes and intergenic regions organized in chromosomes. 'Chromosome and dna structure' by brgfx on Freepik. Image modified from the original.

Those regions that contain the information for producing proteins are called exons and located within the genes. Exons represent less than 2% of the genome and all exons combined are referred to as the exome. The gene contains both coding (exons) and non-coding (introns) regions as well as extra sequences that are transcribed but not translated which border the coding DNA sequence (CDS), known as the 5′ and 3′ untranslated regions. They contain regulatory regions which play a crucial role for the proper expression of the gene. The CDS is the actual region of DNA that is translated to form proteins.[3]

DNA consists of nucleotides (bases) which are abbreviated with the letters A (Adenine), T (Thymine), G (Guanine) and C (Cytosine). Protein biosynthesis is performed in two phases, transcription and translation. During transcription a gene, is "read" and transcribed into a premature messenger ribonucleic acid (pre-mRNA) molecule. Then a mature mRNA is formed which contains only exons as the introns are spliced out[4] (Figure 2). For translation, the mRNA travels to a ribosome in the cytoplasm. Each set of three nucleotides, called a codon, specifies a particular amino acid. Transfer RNA (tRNA) molecules bring the corresponding amino acids, which are then linked together to form a polypeptide chain that folds into a protein. Different trinucleotides of RNA may encode the same amino acid. There are 64 different codons: 61 specify amino acids and 3 are used as stop signals. The CDS runs from the first AUG (start codon) to the first in-frame stop codon.[4]

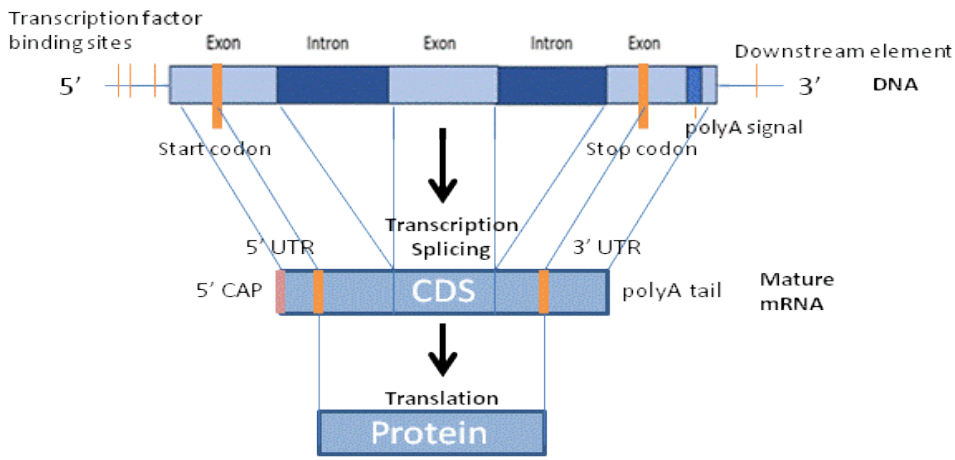

Figure 2: Protein biosynthesis. DNA is copied into messenger RNA (mRNA). This occurs in the nucleus. Then splicing occurs whereby the introns are removed, and the remaining exons joined together to form the mature mRNA. 5' cap and a poly-A tail at the ends of the mRNA molecule are added. They help the molecule to bind to the ribosomes and protect it from degradation. This processed mRNA is then ready to be translated into protein in the cytoplasm. *DNA*: Deoxyribonucleic acid, *CDS*: coding DNA sequence, *mRNA*: messenger ribonucleic acid, *UTR*: Untranslated region.

Genetic variation

Genomic variation refers to DNA sequence differences among individuals or populations. Some variants influence biological function (such as pathogenic variants that cause a genetic disease), while others have no biological effects. The DNA sequences of any two human genomes are estimated to be approximately 99.6% identical. The remaining 0.4%, primarily consisting of single nucleotide variants (SNVs), is thought to account for the phenotypic differences among individuals, such as variations in physical traits, susceptibility to diseases, and responses to medications.[5] The human genome of each individual contains roughly 4 million genetic variants.[6]

Hereditary human diseases, such as hereditary movement disorders are caused by variants in the genome. There are different types of variants. The types of variants that were detected by investigating the genome or the exome of our patients are described here:

Single nucleotide variants

A single nucleotide variant (SNV) is a substitution of a single nucleotide at a specific position in the genome. Each genome contains about 3.5 million SNVs.[6] They can be located in coding as well as in non-coding regions. There are two types of coding SNVs. *Synonymous* variants do not change the amino acid sequence of the protein and are generally considered non-deleterious. However, synonymous SNVs can influence transcription and translation efficiency, alter splicing, or impact mRNA stability.[6] In contrast, *non-synonymous* variants alter the amino acid sequence either by substituting one amino acid for another (*missense* variants) (Figure 3), which may be harmful or benign depending on chemical similarity, or by introducing a premature stop codon (*nonsense* variants). *Non-synonymous* variants may result in the expression of a non-functional or abnormal protein with an entirely different function (loss- or gain-of-function SNV, respectively).[6,7] Transcripts with premature stop codons are recognized by cellular surveillance mechanisms and eliminated through the nonsense-mediated mRNA decay (NMD) pathway.[7] Other types of SNVs can lead to the loss of a start or stop codon.

Small insertions and deletions

Small insertions and deletions (indels) are insertions or deletions of a number of nucleotides in a DNA sequence, affecting 1-50 nucleotides, resulting in a completely different translation from the original. Short indels involving 1 to 3 nucleotides are the most frequent, accounting for approximately 70% of all indels in the genome. Additionally, deletions occur about twice as often as insertions.[6] If

the number of nucleotides that an indel consists of is divisible by three the variation is called *in-frame*, otherwise the indel leads to a *frameshift* variant. *In-frame* indels can lead to a pathomechanism but in most cases they are not pathogenic as they maintain the open reading frame of the protein. A *frameshift* variant leads to a different reading frame from the normal and alters the stop codon resulting in abnormally short or long dysfunctional protein, usually with different amino acids (Figure 3). NMD may occur leading to no protein or to a significantly reduced amount of protein product.

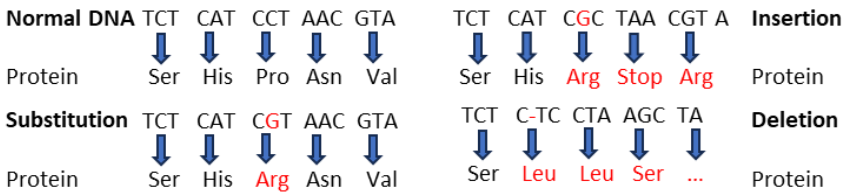

Figure 3: Different mutation types. The normal DNA contains codons which encode for amino acids of the normal protein product. In case of substitution of one base by another, different amino acid from the normal is encoded (missense mutation). In case of insertion or deletion of a nucleotide, the reading frame is changed, and it is possible to cause the creation of a premature stop codon.

Splice-site variants

Alternative splicing produces multiple transcripts from a single gene, enriching the diversity of proteins. However, in some cases splice-site alterations which occur at the boundaries between exons and introns during transcription can be deleterious. These changes can disrupt mRNA splicing resulting in the loss of exons or the inclusion of introns and in altered protein-coding sequence (Figure 4). Around 10% of all pathogenic variants linked to human genetic diseases affect the canonical splice sites.[6,7]

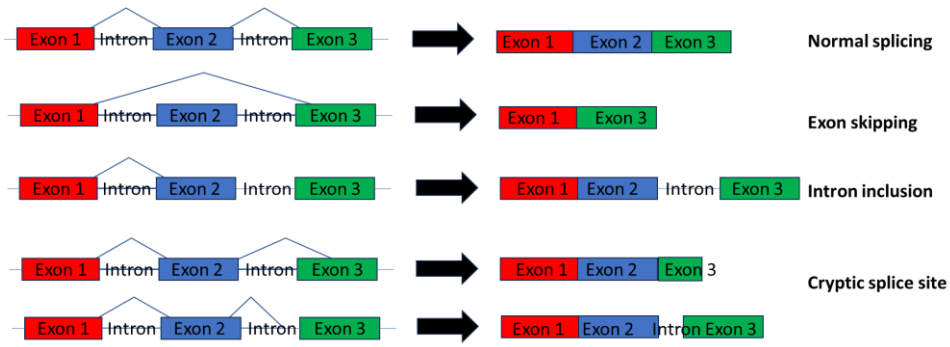

Figure 4: Effects of splicing disruption on mRNA. The different potential effects are shown. Exon skipping or intron inclusion occur when the splice donor sequence or splice acceptor sequence is altered due to a variation. Cryptic splice sites can be generated as a consequence from a SNV or indel or to be activated by variants in canonical splice sites or splicing regulatory elements. Activation of these sites leads to abnormal proteins.

Copy number variation

Copy number variants (CNVs) refer to an intermediate-scale genetic change. They affect segments greater than 1,000 base pairs in length but typically less than 5 megabases. CNVs include both losses of genetic material (deletions) and additional copies of sequence (multiplications such as duplications, triplications, etc.) (Figure 5). Many CNVs span multiple genes and influence the expression of both the affected genes and nearby ones. Generally, an increased gene copy number leads to higher expression levels, while a reduced copy number results in decreased expression.[6]

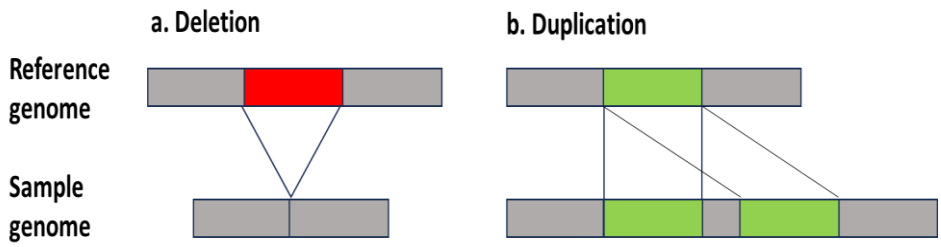

Figure 5: CNVs. a. Deletion and b. Tandem duplication of a genomic region.

Short tandem repeats

Short tandem repeats (STRs) are short sequences of DNA, usually a unit of 1-6 bases, that are repeated many times in a row. When the number of repeats exceeds a specific threshold (number of unit repeats in normal individuals) the generated

repeat expansions, may lead to disease (repeat disorders) (Figure 6). In the majority of cases, the larger the expansion the more severe the symptoms and the younger the age at disease onset are.[8] Earlier age at onset and more severe symptoms are observed as these expansions are transmitted from generation to generation further expanded. This phenomenon is called anticipation.[8] However, in some cases the expansions, which are very unstable, may become shorter as they are passed on to the offspring. Loss-of-function by gene silencing due to DNA hypermethylation, mRNA molecules with expanded repeats which accumulate to form inclusions in the nuclei leading to RNA toxicity and neuronal intranuclear protein misfolding and aggregation due to longer segments with the same amino acids are some of the pathophysiological mechanisms associated with STRs.[8]

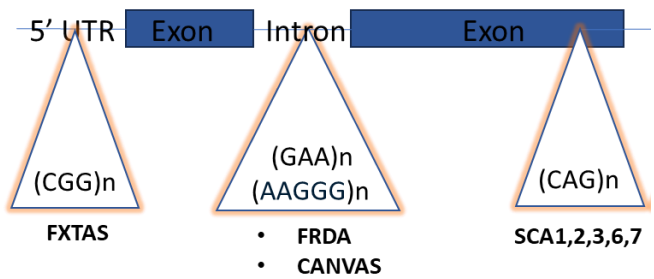

Figure 6: STRs causing diseases. Genomic locations and expression products of STRs causing human disorders. *FXTAS*: Fragile X tremor ataxia syndrome, *FRDA*: Friedreich ataxia, *CANVAS*: Cerebellar ataxia, neuropathy and vestibular areflexia syndrome, *SCA*: Spinocerebellar ataxias

Intronic variants

Intronic variants typically do not have biological effects unless they create novel or activate cryptic splice sites, potentially causing intron inclusion, or disrupt regulatory elements such as enhancers or silencers.[6]

The mitochondrial DNA

The mitochondrial genome is circular, double-stranded, and relatively small, measuring about 16.6 kilobases. The two strands vary in their base composition and are distinguished as the heavy strand (h-strand) and the light strand (l-strand). It has no introns or intergenic regions. The heavy strand carries most of the genetic information, encoding 28 of the 37 genes.[9] These include genes for 12 polypeptides, two rRNAs (12S and 16S), and 14 tRNAs (Figure 7). In contrast, the light strand encodes just one protein and eight tRNAs (Figure 7). All 13 protein-coding genes

are vital components of the oxidative phosphorylation complexes (OXPHOS), which are essential for energy production.[9]

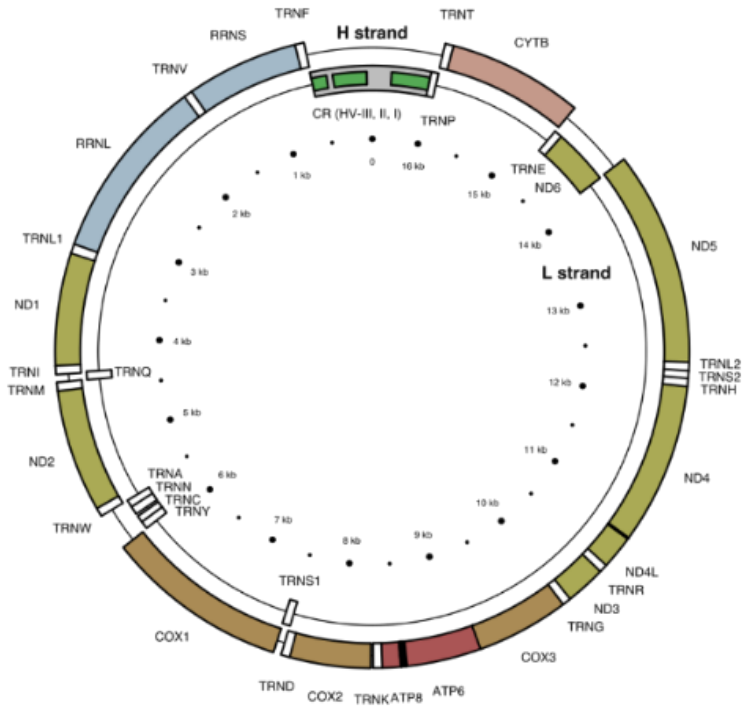

Figure 7: The circular, double-stranded, mitochondrial DNA and the genes that contains. "Map of the human mitochondrial genome" by Emmanuel Douzery, licensed under CC BY-SA 4.0. Source: Wikimedia. Commons. Image modified from the original.

Mitochondrial DNA (mtDNA) is inherited exclusively from the mother. This pattern of maternal inheritance accounts for the genetic transmission of diseases associated with mitochondrial genome abnormalities.[9] Within a cell, multiple populations of mtDNA can coexist. For example, mutated mtDNA molecules may be present alongside normal, or wild-type, mtDNA. This is known as heteroplasmy. In contrast, when only one form of mtDNA is present, this is described as homoplasmy.[9]

Movement disorders

Neurological conditions which affect ease, fluency and/or speed of movement and lead to abnormal motor functions are described as movement disorders. They are classified as *hyperkinetic* when there are excessive, involuntary movements and as *hypokinetic* when bradykinesia, hypokinesia, akinesia or rigidity is observed. In both types, the ability to perform intended movements can be compromised. In the first category ataxia, dystonia, chorea, ballism and myoclonus are included among other conditions while in the second, Parkinson's disease (PD) is the primary disorder[10] (Figure 8).

Hypokinesias	
Akinesia/bradykinesia (parkinsonism)	Rigidity
Apraxia	Stiff muscles
Blocking (holding) tics	Freezing phenomenon
Catatonia	Hesitant gaits
Hyperkinesias	
Ataxia/asynergia/dysmetria	Myoclonus
Dystonia	Paroxysmal dyskinesias
Chorea	Tremor
Ballism	Tics
Athetosis	Restless legs
Akathitic movements	Stereotypy
Hemifacial spasm	REM sleep behavior disorder

Figure 8: Movement disorders' classification. Categorization of movement disorders into two main groups, hypokinesias and hyperkinesias. From Fahn, 2011,[10] modified.

The main brain areas involved in the production and coordination of movement are the motor cortex, the basal ganglia (BG), the thalamus, the brain stem and the cerebellum which are structural areas of the central nervous system (CNS).[11] These regions are interconnected, and the full network is essential for generating proper movements. Not all movement disorders are linked to dysfunction or damage in a specific area of the CNS. However, as the primary motor cortex is responsible for the initiation of voluntary movements and the thalamus contains motor and sensory information, damage of these areas can cause myoclonus or tremor. Decreased or involuntary movements resulting in PD or dystonia, can be caused by disruptions between brain nuclei connections of the BG which is responsible for voluntary

muscle movements and coordination of postures.[11] On the other hand, lack of coordination which is caused by dysfunction of the cerebellum can lead to ataxia.[12]

In principle, there are a number of reasons and causes for developing a movement disorder. Infections and toxins as well as lesions of the brain or spinal cord can lead to acquired movement disorders. Such lesions can be caused by mechanical trauma, stroke or vascular diseases. Movement disorders can be the side effect of specific medications (such as anti-psychotic drugs) or occur as symptoms of inherited metabolic disorders. A relatively large proportion of movement disorders have a genetic cause and hundreds different genes are known to cause various movement disorders.[13]

Some of the movement disorders are caused by mutations located on a single gene (monogenic) such as forms of ataxia and dystonia, while others may be associated with additive effects of multiple genes that interact with each other and with the environment (multifactorial). Many cases of PD are sporadic with a late age at onset and a negative family history, while other entities of movement disorders may be mostly familial, often those with earlier age at disease onset.

In case of monogenic diseases, there are more than one causative gene for each disorder, and these may have different modes of inheritance. Thus, there often are different subtypes of the same movement disorder. Moreover, abnormal movements may be the primary manifestation of a disease or not. By taking all these into account, neurological movement disorders can be characterized as complex and their clinical and genetic diagnosis remaining a challenging process.

Dystonia

Dystonia is characterized by sustained or intermittent muscle contractions resulting in abnormal postures and/or movements. Common dystonic movements are twisting, patterned, tremulous and repetitive movements.[14] A hallmark of the disease is simultaneous activation of an agonist and an antagonist muscle over the same joint. Voluntary action usually initiates or worsens dystonia. The degree of severity varies and in some cases activities of daily life cannot be performed, while other patients live a relatively normal lifestyle.

Dystonia is characterized by heterogeneity due to its diverse etiology and clinical presentation. Based on the etiology, dystonia has been classified into primary and secondary. Primary dystonia includes idiopathic or genetic disorders in which dystonia is the only sign. Secondary dystonia means that dystonia has external causes (lesions within the brain, drug side effects, others).[14] In 2013, a new classification was proposed with two-axis structure.[15] Axis I summarizes clinical characteristics of dystonia. Based on the associated features, dystonia entities are

divided to isolated (dystonia is the sole clinical feature with no other neurological signs and in most cases are caused by genetic alterations) or combined (whereby dystonia is combined with other neurological signs like parkinsonism, myoclonus or paroxysmal dystonia plus other dyskinesia). Dystonia can also be complex and in that case is not the prominent disease manifestation. Axis I also includes the classification of dystonia based on the regions of body affected, age at onset and temporal pattern. Regarding the affected body part, it is subdivided to focal (one region of the body is affected), multifocal (more than one non-contiguous body parts are affected), segmental (adjacent regions of the body are affected) and generalized (several regions of the body are affected). The age at onset ranges from infancy (neonatal- 2 years) to late adulthood (>40 years) and the temporal pattern can be persistent, action specific, with diurnal fluctuations or paroxysmal.[16] Axis II addresses etiology and includes two categories, whether there is evidence of degeneration or not and whether dystonia is inherited or acquired. When caused by neurodegeneration, dystonia progressively worsens over time. Recently, a new revised classification system was proposed. This system continues to use a two-axis framework, though with some updates.[17] Axis I covers additional clinical characteristics such as family history and phenomenology while axis II focuses on the underlying causes, organizing them into categories like genetic factors, acquired conditions, anatomical changes, and common disease mechanisms. The revision aims to provide a clearer structure that supports consistent application in clinical practice.[17]

Epidemiology

Different types of dystonia have different prevalence with isolated idiopathic and genetic forms to be relatively more common.[14] Prevalence increases with age therefore, adult-onset focal dystonia is more frequent than early-onset dystonia. For adult-onset focal dystonia the estimated prevalence is ~16/100,000 based on clinical studies while early-onset has a prevalence of ~7.6/100,000.[14] About 4,300 dystonia patients are included in the Swedish national patient register and the prevalence for primary dystonia in Sweden was found to be 35.1/100,000.[18]

Pathophysiology

Dysfunction of the BG has been thought to be the principal cause of dystonia as this region integrates motor control. It has been observed that abnormal connectivity and alterations in plasticity might also play a role in the pathophysiology of the disorder. Moreover, there is evidence of dysfunctions in numerous regions of the CNS involved in sensory and motor control like cerebellum, spinal cord, cortex and brainstem.[19]

Genetics

The first genes implicated in dystonia, were *TOR1A* which encodes TorsinA, a member of the AAA+ family of adenosine triphosphatases (ATPases) that is expressed in several brain regions, and *THAP1* which encodes a DNA-binding transcription regulator. A recurrent three base-pair inframe deletion in exon 5 of *TOR1A* (c.907_909delGAG), p.(delGlu303), (ΔE303) is the single most common monogenic cause for dystonia. More than 100 distinct pathogenic variants in *THAP1* have been detected in sporadic and familial dystonia cases.[14] Variants in these genes cause isolated dystonia and are inherited in an autosomal dominant manner. Variants in genes such as *GNAL, ANO3, KMT2B* and *VPS16* are also well validated causes of monogenic dystonia.[20] In the last two decades more genes were incriminated to cause autosomal dominant or recessive isolated dystonia. Incomplete penetrance is common in the known forms of monogenic dystonia and index cases may have negative family history. For instance in case of individuals with *TOR1A* p.(delGlu303), penetrance has been estimated to be 20-30%.[14] The remaining large number of genes included in dystonia gene panels is associated with combined forms of the condition and with complex dystonias as more than a hundred genetic diseases present with dystonia as one of their neurological symptoms.

Diagnosis

Clinical presentation of patients constitutes the cornerstone of assessment. When an individual is diagnosed with dystonia, further evaluation is performed in order the specific cause to be determined. Physical and neurologic examination, neuroimaging, obtaining records of detailed medical and family history are included in the approach that is routinely followed by clinicians.[21] If environmental factors that affect the brain and manifestation of more neurologic abnormalities (such as dementia, seizures, spasticity) are reported, secondary dystonia is the most suitable diagnosis.[19] In case of primary dystonia, features like age and site of onset, ethnicity, presence or absence of other neurologic or non-neurologic defects as well as the clinical course are examined. Clinical, non-DNA based tests are performed for further evaluation.[21] Molecular genetic testing may be used when a monogenic cause is suspected. Due to the genetic heterogeneity of dystonia, multigene panels or comprehensive genomic testing are frequently used. Multigene panels are commercially available but may include different genes of interest as the design of targeted panels for a genetically heterogeneous condition as dystonia is challenging. Comprehensive genomic testing refers to genomic data derived by Next-generation sequencing technologies (NGS). Nowadays, Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) are usually used when genetic testing of a person with dystonia is indicated.

Ataxia

Ataxia can both be a disease and a physical finding. It is defined as an inability to coordinate voluntary movements resulting in defects of gait, gaze, arm movements and hand coordination and/or speech. It can be classified, according to the location of the underlying cause, as cerebellar, sensory, or vestibular.[22] Depending on the etiology, the onset can be insidious with a slow progression, acute with a rapid progression, or subacute. Subacute onset can be the result of deficiency syndromes or inflammation. In case of acute onset, common etiologies include hemorrhagic strokes or infections. Chronic progression is mainly associated with hereditary ataxias.[12] Ataxia can also be subdivided into acquired, sporadic (patients have negative family history) and hereditary.[22] In case of acquired ataxia, symptoms may worsen over time and life expectancy may be reduced. For people with hereditary ataxia life expectancy is generally shorter. Incoordination of gait is mostly observed early in the disease course. However, cerebellar atrophy, peripheral neuropathy, poor coordination of eye movements, hands and speech, dysartria, nystagmus or intellectual disability can also be observed.[23]

Monogenic ataxias are subdivided by the genetic locus on which the pathogenic variant is located or by the mode of inheritance. Spinocerebellar ataxias (SCA) and episodic ataxias (EA) belong to the category of autosomal dominant cerebellar ataxias (ADCA).[24] Adult-onset is frequent for some ADCAs. Autosomal recessive cerebellar ataxias (ARCA) such as cerebellar ataxia with neuropathy and vestibular areflexia syndrome (CANVAS) and ataxia–telangiectasia (AT) are transmitted in a recessive pattern. The most common recessive hereditary ataxia is Friedreich ataxia (FRDA).[24] In contrast with ADCA, the age of onset in recessive ataxias is more often in childhood or early adulthood. The most well-known X-linked cerebellar ataxia is fragile X tremor ataxia syndrome (FXTAS).[24] Progressive ataxia can also be caused from abnormalities occurring on mitochondrial DNA and in that case, ataxia constitutes one of several symptoms of a mitochondrial disorder.[23]

Epidemiology

The prevalence of ADCAs in the general population is 2.7/100,000 with the most common form in many European populations being SCA3.[25] Autosomal recessive ataxias (ARCA) can be characterized as relatively frequent. The prevalence in total is 3.3/100,000[25] and the most common among hereditary ataxias is Friedreich ataxia (FRDA) with a prevalence 1/40,000.[26] X-linked cerebellar ataxias seem to be uncommon in the general population except from FXTAS. In Sweden, the prevalence of progressive ataxia was found to be 6.0/100,000. [18]

Pathophysiology

The proprioceptive sensory pathway, the vestibular system and the cerebellum and its afferent and efferent connections may all be involved in ataxia. Lesions to midline cerebellar structures usually leads to gait ataxia, while damage of the spinal cord causes sensory ataxia and when the vestibular system is impaired, disequilibrium is observed.[12]

Genetics

Hereditary ataxias are characterized by both genetic and phenotypic heterogeneity. The same clinical features can result from variants in various genes, and a single gene can lead to different types of ataxia depending on the specific variation.[27] Dynamic mutations, which are not static and can change dynamically over time or through generations, are the cause of several SCAs and of other types of ataxias. Specifically, they are caused by nucleotide repeat expansions in coding or non-coding regions within genes.[28] Slippage events may occur during DNA replication or repair, resulting in these repeat expansions that may lead to toxic abnormal translational products.[29] These products contain a large number of a specific residue, most of the times glutamine (encoded by CAG codons) and therefore, some of the most common SCAs are polyglutamine diseases (e.g. SCA 1,2,3,6,7,17).[8] The normal alleles contain normal number of nucleotide repeats and when the number of glutamine residues in the abnormal product is expanded beyond a threshold, SCA may develop. Mutated genes containing repeat expansions are highly unstable and may be transmitted to the next generation with further expansion of the repeats. As a result, the age of onset is mainly decreased while the severity of the symptoms may be increased in the next generations. Therefore, families affected with ataxias are frequently subjected to the phenomenon of anticipation.[8] Other forms of ataxias, including EAs and AT, are caused mostly by SNVs and indels but deletions and duplications may also occur.[27] The number of known causative genes is large in case of ARCAs. However, some causative genes may be 'private' and be found in only one family worldwide due to consanguinity or founder effect.[27]

Diagnosis

When the underling etiology of a phenotype of a proband is suspected to be one of the types of hereditary ataxias, physical and neurologic examination, recording of medical and family history followed by neuroimaging studies and molecular genetic testing is the approach that is routinely followed for the determination of the exact subtype of the disease.[28] Single-gene testing or multigene panel can be used if the clinician, based on the findings, hypothesizes which gene or group of genes to test.[28] For this decision, clinicians rely on features such as distinctive clinical features, age

at onset, mode of inheritance, ethnicity, and apparent anticipation among others. At our hospital, patients used to be tested for the most common spinocerebellar ataxias (SCA1,2,3,6,7) first. Gene panels for autosomal dominant and recessive ataxias are now widely available. When the phenotype under investigation is difficult to be associated with a specific type of ataxia or if a diagnosis cannot be established after single or multiple gene testing, more comprehensive tests should be performed. NGS technologies can be used as an alternative method for the genetic investigation of ataxias, and WGS has become a standard method today.

Parkinson's disease

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's. It is a disorder of the CNS which causes involuntary and uncontrollable movements like resting tremor in some patients, as well as, bradykinesia, rigidity, and postural instability.[30] There is a decreased ability to perform intended movements. PD is progressive and worsens over time. Non-motor symptoms in PD may be cognitive impairment, mood or sleep disorders, constipation, anosmia. Dementia is very common in advanced stages with a prevalence of 30%-40% while among PD patients with a disease duration of 20 years, the prevalence of dementia increases to as high as 83%.[30,31] They are labelled as idiopathic or sporadic when the index case has negative family history. Life expectancy of persons with PD is near normal. However, many patients experience negative effects on their functioning and their quality of life after 7 to 10 years. The major risk factor for developing PD is age but also exposure to toxins (herbicides and pesticides) increases the risk.[31] On the other hand, environmental factors that decrease the PD risk found to be the tobacco smoking and coffee drinking among others,[31] although this association is disputed and may represent reverse causation. PD can be either monogenic or arise from a complex interaction between genetic and environmental factors that impact many essential cellular processes.

Epidemiology

The prevalence of PD rises in older individuals from 0.5–1% among those 65–69 years of age, to 1–3% among persons 80 years of age and older.[32] Today more than 6 million people worldwide are affected and by 2040 the number of affected individuals is predicted to increase to about 12-17 million people.[33] The prevalence in Europe is estimated to be 108-257/100,000.[34] In Sweden 15-20,000 individuals and approximately 1% of all Swedish over the age of 65 have PD (https://www.neuroreg.se/en/parkinsons-sjukdom/).

Pathophysiology

In PD the loss of neurons in substantia nigra pars compacta leads to dopamine deficit resulting in the motor symptoms.[31] The pathologic hallmark of PD is misfolded proteins, especially alpha-synuclein, which are aggregated into filaments forming the so-called Lewy bodies in substantia nigra pars compacta and other brain areas. Lewy pathology is not limited to the brain; it can also be detected in the spinal cord and peripheral nervous system.[31] The substantia nigra pars compacta is a part of a group of midbrain neurons that project to the BG, which integrate motor control.

Genetics

PD is generally considered multifactorial, resulting from complex interactions between genetic and environmental factors, which is why most cases are classified as sporadic. However, about 10-15% of PD cases have a positive family history, indicating a familial form. Mendelian forms of PD, caused by rare variants in a limited number of genes, reportedly account for roughly 5-10% of cases.[35] In these cases, a mutation in a single gene is sufficient to cause the disease (monogenic), and these variants tend to have a strong effect. Some of these genes follow an autosomal dominant inheritance pattern, such as *SNCA, LRRK2*, and *VPS35*, while others are inherited recessively, like *PRKN, PARK7*, and *PINK1*. However, some very common variants such as *LRRK2* p.(Gly2019Ser), exhibit incomplete penetrance. Only 20-70% of those carrying this variant develop PD by the age of 80. Variants in the *GBA1* gene do not cause Mendelian PD but represent the most significant genetic risk factor for developing the disease, with about 5-15% of PD patients carrying *GBA1* variants.[36]

Although PD can be monogenic, in many cases, single genetic variation cannot explain familial aggregation or heritability. The missing heritability refers to the discrepancy between the heritability estimates of PD derived from clinical and epidemiological observations including family and twin studies and the proportion of that heritability explained by identified genetic variants. Previous findings show that most occurrences of familial aggregation are not caused by monogenic inheritance. In two genetic analyses that were performed in Southern Sweden, familial PD was studied. In the first, more than 2,000 PD patients were recruited and 21.6 % reported first or second-degree relatives with PD. Only 0.6% were found to carry a one of eight genetic variants causing dominant PD.[37] Eight families from Southern Sweden were diagnosed 80 years ago with familial PD, yet none of the presently living members have developed the condition. In another study from Greece, two large families from Crete with PD where each one had one 'founder' parent were followed up. Five of 11 (45%) of the children but only 3 of 44 (7%) of grandchildren over 60 years developed PD.[38] The authors postulated digenic

inheritance. As the proportion of patients with family history of PD, is larger than the identified genetic causes, there may be monogenic forms that haven't been discovered yet as many rare variants with small effect sizes remain undiscovered. Another explanation may be the impact of complex factors (combination of genetic and environmental factors) which are difficult to detect. Moreover, structural variants and non-coding genetic regions are underexplored. Another reason may be the epigenetic modifications that may contribute to inherited risk. The sporadic cases as well as the familial cases that cannot be explained by monogenic inheritance, may have an inheritance pattern that have not been identified yet (such as digenic inheritance, when the interaction of two genes is required for expressing a phenotype).

Diagnosis

The diagnosis of PD is mainly clinical, despite advances in neuroimaging and genetics. Misdiagnosis is common and definite diagnosis can be obtained only pathologically.[39] Specific Magnetic resonance imaging (MRI) findings can help differentiate PD from other parkinsonisms,[40] and there are additional tests that can distinguish PD from other disorders. Family history is very important and most commonly a three-generation family history is obtained as a first- or second-degree relative with PD increases the likelihood that the patient has one of the known monogenic forms. Age at onset will help to determine if the condition is inherited in autosomal dominant or recessive manner. Age will also help to distinguish unaffected individuals from family members who have not reached the age at onset and who may still develop PD in the future. Ethnicity is another important factor as some populations are more likely to have certain causative pathogenic variants.[41]

Genetic testing for recessive PD is currently considered in Swedish health care for people with unusually early symptom onset, up to 30–35 years of age. Genetic testing should include both sequence analysis of *PRKN, PINK1* or *PARK7*, and analysis for CNVs of these genes. For testing for dominant PD genes, *LRRK2*, *VPS35* and *SNCA* need to be analyzed for SNVs and indels, for example by targeted gene panel or WES; in Swedish health care this is generally considered if the patient has at least 2 first- or second-degree family members with PD, or shows a symptomatology suggestive of SNCA-related PD. In case of negative findings, CNVs for *SNCA* needs to be performed. With a multigene panel is highly likely to identify a genetic cause of PD. However, comprehensive genomic testing which does not require the clinician to determine which gene(s) are likely involved can also be used by neurological clinics as the cost of NGS methods is constantly dropping.

Next generation sequencing

The first sequencing of the human genome with Sanger sequencing was performed in 1977 and Sanger sequencing remained the only sequencing method for more than 20 years.[42] Next generation sequencing (NGS, Second-generation sequencing) or massively parallel sequencing is a high-throughput technique which emerged in the 1990's and became commercially available in 2005.[42] Sanger technology is a chain-termination method that uses labeled dideoxynucleotides and has low throughput as it sequences one or a few DNA fragments at a time. Sanger sequencing offers high accuracy and long read lengths for targeted regions, while NGS enables massively parallel analysis of large genomic regions, making it more efficient and cost-effective for high-throughput applications. NGS has become a fundamental tool for scientists and researchers and it is today widely applied in clinical diagnostics.

Ion Torrent and Illumina sequencing platforms were used to obtain the NGS data that was analyzed in this project (Table 1). The two platforms differ in the amplification method that they use, their chemistry and each technology has its own limitations. They both generate short-read sequencing data.

In recent years, a newer sequencing technology, third-generation sequencing, has been developed and is now widely used. It provides long-read sequencing enabling the sequencing of much larger DNA fragments and overcome limitations of previous sequencing generations. Short-read sequencing can generate reads up to 600–700 bases, while long-read sequencing can generate reads up to 25–30 kilobases.[42] Long-read sequencing is polymerase chain reaction (PCR) free, therefore, it eliminates amplification bias (uneven or preferential amplification of certain DNA sequences) and the mapping of the generated reads on the reference genome is performed more accurately. It provides uniform genome coverage, even in complex regions, and complex structural variants, such as large deletions and insertions and repeat expansions, can be more precisely detected. However, the cost remains relatively high compared with short-read sequencing techniques. Pacific Bioscience (PacBio) sequencing is an example of these technologies. The PacBio Revio instrument was used in this project for sequencing two patients' genomes for the identification of the exact number of triplet repeats in *ZFHX3*.

Table 1: Different technologies of NGS platforms used for the analyses in this thesis. From Satam *et.al* 2023,[42] modified.

Platform	Use	Amplification type	Read length (bp)	Principle	Limitations
Ion Torrent	Short read sequencing	Emulsion PCR	200–400	Fragments of DNA are attached to beads. The chip is then flooded with each one of the four nucleotides. The release of hydrogen ions or pH change during DNA synthesis is detected and the sequence is determined	Prone to homopolymer errors.* Sequencing of homopolymer regions may lead to loss in signal strength
Illumina	Short read sequencing	Bridge PCR	36–300	Single-stranded DNA binds on immobilized surface, the flow cell, for amplification and clusters generation, followed by the addition of single fluorescent-labelled deoxynucleoside triphosphates to sequence the amplified DNA template. Fluorescent signals are detected identifying which nucleotide was incorporated at each cycle	In case of sample overloading, the sequencing may lead to overlapping signals, raising the error rate up to 1%
PacBio Single-molecule real-time sequencing technology (SMRT)	Long read sequencing	PCR free	average 10,000–25,000	Individual DNA molecules are immobilized within zero-mode waveguides (ZMWs), emitting light as the polymerase incorporates each nucleotide, allowing real-time measurement of nucleotide incorporation	Higher cost compared to other sequencing platforms

*The technology has difficulty accurately counting the number of repeated identical bases in a row, leading to insertion or deletion mistakes.

Whole exome sequencing vs Whole genome sequencing

With WES only the exome is captured and sequenced. Enrichment of exonic regions using target specific amplification methods or hybrid capture is performed prior to sequencing.[42] Although the exome contains only 1-2% of the genome, 85% of the disease-causing variants are located on these regions.[43] By using appropriate tools, SNVs, indels, CNVs and STRs of the protein-coding regions can be identified. WES has long been preferred for clinical diagnostics because of its lower cost, and at the time before WGS became the more established method. However, specific types of variations such as intronic variants and STRs of non-coding genomic areas cannot

be detected by WES. WGS is a comprehensive technique as almost the entire genome is sequenced, providing information of the complete genetic background of an individual's genome including genes, non-coding elements and regulatory regions. As the cost of WGS is constantly dropping, it is now widely preferred over WES for clinical diagnostics. However, its basic limitation is the problem that arises regarding the storage and analysis of the generated data as it is significantly larger.

Aims

The major aim of this project was to investigate the genetic background of patients with neurological movement disorders (ataxia, dystonia, PD) by utilizing algorithms and appropriate *in-silico* tools for analyzing NGS data and to develop bioinformatic strategies and pipelines for the identification of known and novel candidate disease-causing variations and genes. The diagnostic yield of NGS data analysis in neurological movement disorders was also estimated.

Paper I: The aims were to investigate the genetic cause of a Swedish family with dystonia hemichorea/hemiballism, and to find evidence for the causality of the identified variant and gene (*TOR1AIP2*) for this particular type of dystonia by functionally investigating the encoded protein and searching for additional families with similar phenotype and genotype.

Paper II: The aims were the clinical and genetical examination of a case series with ataxia from Southern Sweden with known or unknown genetic cause, the development of bioinformatic strategies and collaborative approaches for the investigation of the molecular etiology of previously undiagnosed cases, the clinical and genetical description of the findings, and the determination of the diagnostic yield by using up-to-date genetic testing methods.

Paper III: The aims were to investigate the genetic cause of two large families from the Skåne region in Sweden with balance and gait disturbances, sensory neuropathy and slow saccades, to find the shared haplotypes among the affected family members, to determine the exact number and architecture of the identified repeats in *ZFHX3*, to examine our case series with ataxia for the identified variation, to screen larger datasets for the determination of the architecture of the normal locus, and to interpret the findings of postmortem examination.

Paper IV: The aims were to establish a genetic diagnosis of a selected cohort of PD patients, including also patients with early age at disease-onset and/or positive family history, factors that increase the likelihood of a genetic cause, by investigating their genome using WES data that were obtained from the Multipark NGS database, and to determine the diagnostic yield.

Methods

The principal study designs that were used for the papers of this thesis were family-based sequencing studies for the papers I and III and sequencing studies of case-series for the papers II and IV.

Patients and families

Paper I: All studied individuals in a family from Sweden with dystonia combined with hemichorea and hemiballism, were included in a research study on rare neurological diseases. They provided written informed consent. The Swedish index patient and her mildly affected son were repeatedly referred to the movement disorders clinic at our tertiary center. Her unaffected daughter as well as her two unaffected half-siblings provided blood samples and were interviewed, observed and examined via video connection (Figure 13). Members of a second family from Slovakia were part of a genomic sequencing project focusing on heterogeneous dystonia syndromes;[44] written informed consent was obtained. DNA was analyzed from 4 family members as shown in Figure 13. The index patient and her son were re-examined clinically within the present study.

Paper II: This paper described a case series of patients with ataxia from Southern Sweden. Patients were recruited by searching in the diagnosis register of the department of Neurology at Skåne University Hospital between the years 2011-2020, by using the ICD-10 Version:2019 codes for hereditary ataxia. Additional patients were recruited from among those referred to our center from other neurologists than the authors or identified through family members or through the SCA-Network, a Swedish patient organization. Finally, a consecutive single-center series of 87 patients from 76 families with ataxia, were investigated clinically and genetically.

Paper III: Two kindred with a particular form of ataxia with balance and gait disturbances, sensory neuropathy, slow saccades, and severe autonomic dysfunction were described in 2014 by some of the authors[45] (Figure 18) and all surviving affected family members were re-contacted in 2022. All were interviewed and re-examined in a standardized manner in the context of the present study. In the authors' clinics, additional affected members were identified via the families and/or

from among persons with ataxia found to carry *ZFHX3* repeat expansions who were re-approached for the present study (Figure 18).

Paper IV: Out of a total of 1,150 Parkinson's patients who had enrolled in either the PARkinsonLUnd (PARLU) study[46] or the MultiPark biobank sample collection (MPBC),[47] 285 individuals were selected for WES. Many had young age at disease onset and/or reported affected family members. In the PARLU study, PD diagnoses were confirmed through clinical examinations conducted by one of the authors, along with longitudinal follow-up within the research framework. For MPBC participants, diagnoses were verified using data from the national Parkinson registry,[48] which compiles information from treating neurologists. All participants provided informed written consent for genetic research related to PD. Detailed information regarding disease history and family background was collected.

Sequencing analysis

In Paper I, genomic DNA was extracted from peripheral blood samples of the family members under investigation. The two affected individuals were sequenced using Ion Torrent technology at Clinical Genetics in Lund. The three unaffected individuals were sequenced using Illumina technology, following the manufacturers' protocols at Centogene in Rostock, Germany. Ataxia patients of Paper II were sequenced with Illumina platform. The majority of WES and WGS analyses were carried out at the Center for Translational Genomics at Lund University. Additional sequencing and genotyping were conducted at Blueprint Genetics in Helsinki, Finland, and Centogene in Rostock, Germany. In paper III, short-read WGS of the individuals of the families under investigation was performed at the Center for Translational Genomics at Lund University or at Centogene, Rostock, Germany. Moreover, long-read sequencing for two family members was also conducted by using the PacBio Revio platform. For the PD patients of Paper IV, blood was drawn and DNA was extracted following standard procedures. WES was performed within the MultiPark NGS database (https://www.multipark.lu.se/infrastructures/ngs-database-infrastructure) infrastructure on an Illumina platform at the Center for Translational Genomics at Lund University.

Bioinformatic analysis after sequencing

The image below illustrates the bioinformatic analysis workflow used in this project. Initially, in some of the studies, raw sequencing reads were aligned to the reference genome, followed by variant calling to generate variant calling format

(VCF) files containing SNVs and indels. In some genomic regions, multiple variant calls might occur at the same position; these regions were decomposed so that each line in the VCF file represented a single variant call. The VCF files were then annotated with relevant information, and a filtering strategy was applied to select variants that meet specific criteria for further evaluation. Additionally, CNVs and STRs were investigated using specialized *in-silico* tools. In certain cases, supplementary analyses were performed to address specific questions. See the next sections for explanations.

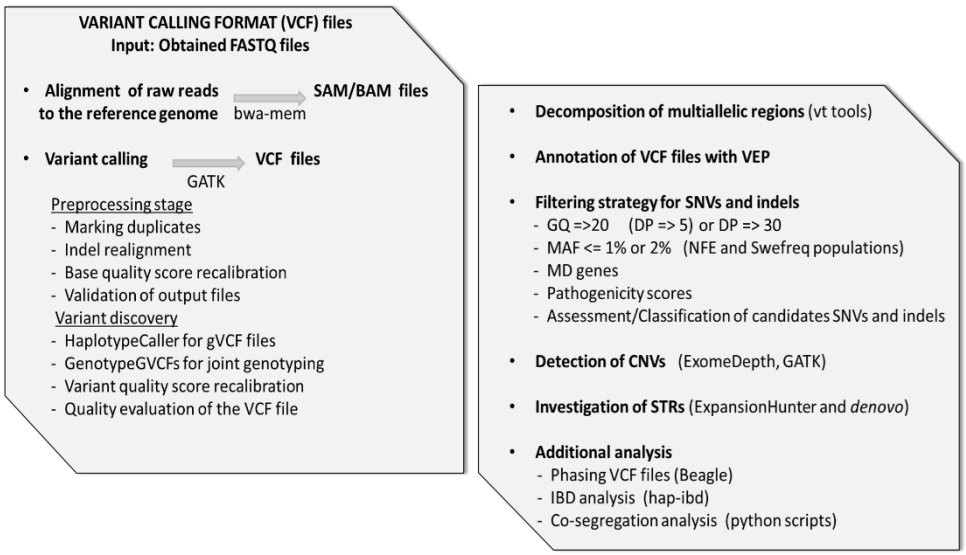

Figure 9: The bioinformatic workflow of this project.
bwa-mem: Burrows–Wheeler aligner-maximum exact matches, *GATK*: Genome analysis toolkit, *VEP*: Variant effect predictor, *GQ*: Genotype quality, *MAF*: Minor allele frequency, *NFE*: Non-Finnish Europeans, *MD*: Movement disorders, *CNVs*: Copy number variants, *STRs*: Short tandem repeats, *IBD*: Identity-by-descent

Reference genome

The entire human genome was sequenced and first published in 2004. Different assemblies have been published since then based on new discoveries and advancements of previous versions. The most widely used version was published in 2009 the GRCh37/hg19 although the most resent one is the GRCh38/hg38 published in 2013. Both these reference assemblies were used for analysis in this thesis. They are both publicly available and can be found at http://genome.ucsc.edu.

Bioinformatics tools

After alignment of raw reads on the reference genome with Burrows–Wheeler aligner-maximum exact matches (bwa-mem) v.0.7.17,[49] by using Genome Analysis Toolkit (GATK) v.4.0.4.0,[50,51] on the L-SENS or COSMOS-SENS cluster on LUNARC, VCF files can be produced. GATK was used for generating VCF files for downstream analysis of the Parkinson's patients' data.

The final VCF files were annotated with Variant Effect Predictor v.102 or v.109 (VEP) by Ensembl.[52] The plugin modules dbNSFP (Database for nonsynonymous SNPs' functional predictions) v3.0[53,54] or v.4.1[53,55] and CADD (Combined Annotation Dependent Depletion) v.1.7[56] were used in order to annotate the output further.

CNV analysis of WES data was performed by using the R package, ExomeDepth v.1.1.15[57] while CNVs from WGS data were detected by using GATK run in 'cohort' mode with default settings as it is described in the official guide: https://gatk.broadinstitute.org/hc/en-us/articles/360035531152--How-to-Call-common-and-rare-germline-copy-number-variants. CNVs are then curated with Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER).[58]

Candidate disease-causing SNVs, indels and CNVs were visually inspected with Integrative Genomic Viewers (IGV) v.2.16.2[59] and they were classified based on the criteria that were proposed by the American Collage of Medical Genetics and Genomics (ACMG) in 2015.[60]

STRs were detected with ExpansionHunter v.5.0.0,[61,62] using the bundled hg19 short tandem repeat catalog and default settings. The outputs of ExpansionHunter were visualized with REViewer v.0.2.7.[63] STR expansion loci not previously associated with disease were discovered with ExpansionHunter Denovo v.0.9.0.[64]

VCF files were phased with Beagle v5.4[65] and then the genomic segments shared by affected individuals were detected by identity-by-descent (IBD) analysis with hap-ibd v1.0.[66]

Other tools and freely available software that were used in this project are bcftools v.1.9,[67] vt variant tool set (https://genome.sph.umich.edu/wiki/Vt), spliceAI v.1.3,[68] TraP-score v.3,[69] PredictSNP2 v.1.0,[70] GWAVA v.1.0.[71]

Assessment of mitochondrial variants was performed with MITOMAP[72] and MitImpact 3D 3.1.2.[73] These databases provide, among other tools, APOGEE2 scores and classes which is a mitochondrially-centered ensemble method.[74]

Some of the pipelines were developed by using the Snakemake v.9.2.0 workflow management system[75] which was installed with a Miniconda Python3 distribution.

Databases and datasets that were visited or used during this project are ClinVar of NCBI,[76] Online Mendelian Inheritance in Man (OMIM),[77] Human Phenotype Ontology (HPO),[78] SweGen Variant Frequency Dataset of SweFreq,[79] Malacards,[80] GeneReviews,[81] VarSome[82] and Franklin by Genoox available at https://franklin.genoox.com/clinical-db/home.

Co-segregation analysis that was performed in familial cases of Papers I, II and III, retrieving the allele frequencies from the SweGen Variant Frequency Dataset of SweFreq and selecting variants located in genes that have been associated with the diseases under investigation were achieved by using appropriate Python v.3.12.10 scripts. The final results including SNVs and indels were obtained in excel files by running *in-house* Python v.3.12.10 script.

Genome Analysis Toolkit

As it was mentioned above, first the generated raw reads were aligned to the reference genome and then variant calling, the process of pinpointing genetic differences in sequencing data, which is crucial for identifying genetic disease-causing variants, was performed.

When individuals' genomes or exomes are sequenced, FASTQ files containing reads with sequences and corresponding quality scores are produced. At least 30 samples should be used in the subsequent steps of GATK analysis, in case of exomes as a machine learning process is applied. By using bwa-mem algorithm the sequences of the FASTQ files were mapped to the human genome build (reference assembly) producing SAM (Sequence Alignment Map) files. These files then were transformed to BAM (Binary Alignment Map) files containing the same data in compressed representation.

The following steps of data preprocessing were performed to clean the data and minimize bias and systematic errors of the technique.

First, the duplicate reads were located and tagged with Picard tools. Duplicates may arise during library preparation (due to PCR amplification) or by the incorrect detection of a single amplification cluster as multiple clusters. Such duplicates may cause bias in subsequent analysis as some reads may be amplified better than others. Therefore, the duplicates are marked and only one is kept with the higher quality. Around indels, local realignment was performed. This is crucial for reducing mismatches across reads and transforming regions with misalignments into clean reads with indels, which can be easily identified during variant discovery. Base quality score recalibration (BQSR) is the final step in data preprocessing. Systematic errors in base calling are empirically modeled, considering factors like the preceding base, read position, sequencing cycle number, and more. The initial base quality scores (expressed in Phred scale), which indicate the confidence in the

correctness of the called base at each position, are then adjusted, leading to more accurate base quality scores and, consequently, more precise variant calls.

Variant calling was performed by using HaplotypeCaller followed by GenotypeGVCFs tool of GATK.[50] At the first step, gVCF files were produced containing records for all sites regardless of variant call or not. Then, during Joint Genotyping, spanning records of gVCF files were merged and new records were created.

Finally, variant filtering was performed using variant quality score recalibration (VQSR). A machine learning algorithm was applied, utilizing resources with known SNVs and indels, to identify annotation profiles for true variants. The variant distribution is modeled, and the variants were grouped into clusters. Based on this clustering, a score was assigned to each site, indicating the likelihood of it being a true variant. This score was expressed as the variant quality score log-odds (VQSLOD). When a tranche sensitivity threshold, expressed as a percentage, was specified, the program determined the VQSLOD value that corresponded to the given percentage of variants in the training set. It then used this value as a threshold to filter the variants in the target file, assigning 'PASS' to the FILTER field for variants with a VQSLOD above this threshold. 99.9% tranche sensitivity threshold was selected for SNVs and 99% for indels.

Quality of Variant calling format files

Initially, VCF files were evaluated by using VariantEval tool of GATK or bcftools v.1.9.

One metric that can be monitored is the number of detected indels and SNVs. This number is influenced by various factors such as sequencing quality and coverage depth, making it challenging to establish a strict range. Nonetheless, it serves as a valuable metric for identifying potential anomalies or quality issues within the analysis pipeline.

Other important metrics to consider are those that can reveal biases in the generated VCF files due to artifacts during variant discovery. One such metric is the ratio of transitions (Ti, nucleotide changes between two purines [A ↔ G] or two pyrimidines [C ↔ T]) to transversions (Tv, nucleotide changes between a purine and a pyrimidine [A or G ↔ C or T]) SNVs. Since methylcytosines in the genome are commonly converted to thymines through deamination (transition), the Ti/Tv ratio is expected to be around 2 for WGS and between 3 and 3.3 for WES. This is because exomes can be enriched for CpG islands, often found in promoter regions (depending on the length of the flanking regions that are sequenced) with higher levels of methylcytosines, resulting in a stronger bias toward transitions. Moreover, coding regions are under stronger evolutionary constraint, and transitions (especially synonymous variants) are more tolerated than transversions.

Another relevant ratio is the indel ratio, where the proportion of insertions to deletions is typically around 1 for common variants, and between 0.2 and 0.5 for rare variants. Deviation from these expected values may indicate biases during the analysis.

Information about these metrics can be found here: https://gatk.broadinstitute.org/hc/en-us/articles/360035531572-Evaluating-the-quality-of-a-germline-short-variant-callset

Variant Effect Predictor

The obtained VCF files were annotated using the Variant Effect Predictor (VEP) tool to enrich raw variant calls with additional knowledge, improving the assessment of variants likely to have a damaging effect on function. VEP is a freely available open-source software by Ensembl with significant advantages over other variant annotation tools, particularly for nonsense, frameshift, or splicing variants.[83] Annotation and prioritization are enhanced through access to a comprehensive collection of genomic annotations, including gene symbols (HUGO Gene Nomenclature Committee [HGNC]), exon and intron numbering, existing variant identifiers (e.g., Database for single nucleotide polymorphisms [dbSNP]), variant classes, Consensus coding DNA sequence (CCDS) transcript identifiers, Human Genome Variation Society (HGVS) nomenclature, Ensembl protein identifiers, variant consequences (predicted using a rule-based approach), allele frequencies from the 1000 Genomes Project[84] and gnomAD[85] databases, and deleterious predicted scores from tools like Polymorphism Phenotyping v2 (PolyPhen-2)[86] and Sorting Intolerant From Tolerant (SIFT).[52,87] The output is further enriched by dbNSFP plugin,[53] which provides annotations for missense variants, including pathogenicity predictions from multiple algorithms (e.g. MutationTaster,[88] MutationAssessor,[89] Protein Variation Effect Analyzer [PROVEAN],[90] Rare Exome Variant Ensemble Learner [REVEL][91] etc.) and conservation scores (e.g., Genomic Evolutionary Rate Profiling [GERP++_RS][92]). Additionally, CADD-phred scores were also retrieved for both SNVs and indels by installing the CADD plugin.[56]

Filtering out strategy for Single nucleotide variants and small insertions and deletions

A filtering strategy was applied to the annotated VCF files to identify potential disease-causing variants. At each step, variants that did not meet specific criteria were excluded, narrowing the analysis to those variants that could be disease-causative (Figure 10).[83]

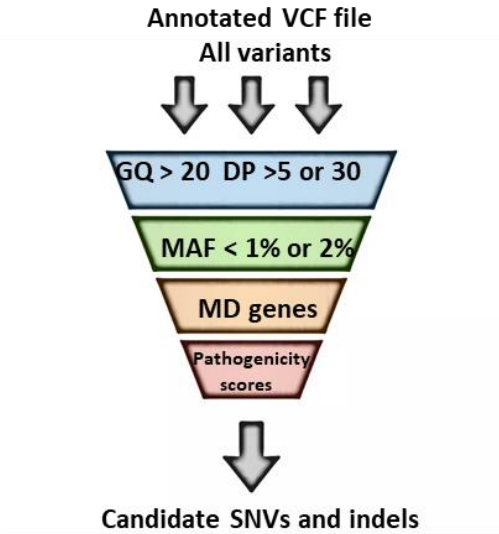

Figure 10: The filtering strategy that was followed in several analyses of this study using WES or WGS data. At each step, specific criteria are applied. *GQ*: Genotype Quality, *DP:* Read Depth, *MAF*: Minor Allele Frequency, *MD*: Movement Disorders

Filter based on quality criteria

The initial filters applied to minimize false positive calls were based on quality criteria. For exome sequencing, the depth of coverage (DP) annotation (the number of reads supporting the alleles) was not used as a model parameter due to the extreme variation in the depth of exome targets. GC-content and mappability issues are often regarded as factors that negatively influences coverage uniformity in WES.[93] GC-content can affect both hybridization capture of exons and sequencing. The default value which is applied by variant callers is DP ≥ 5. In case of WGS data, only sites with a DP exceeding 30 were considered. The genotype quality (GQ) can also be used as a quality-based filter. Genotype quality indicates the confidence that the given genotype is accurate, with higher GQ values reflecting greater confidence in the genotype assignment. Variants with GQ < 20 were filtered out for increasing the genotype likelihood. Variants with a DP of 30 or higher and/or a GQ of 20 or higher were selected using bcftools v1.9.

Filter based on the rarity of the variants

The next filter applied focused on one of the most crucial criteria for identifying variants that are likely to impact the gene product. This criterion is based on the rarity of the variant within the population. Rare variants are more likely to be

disease-causing since they tend to have a more significant functional effect on the gene product compared to common polymorphisms.[83] Hereditary rare diseases such as dystonia and ataxia, are typically caused by very rare variants. The approach commonly used is based on filtering variants using a subset of variants from frequency databases.[94] The recommended starting point is to exclude variants with a minor allele frequency (MAF) greater than 1%, where MAF is defined as the frequency of the second most common allele for a variant in the population.[83] This threshold ensures that only rare variants are selected. Additionally, variants that are absent from allele frequency databases were also considered as very rare and assessed further.

Specifically, the mutant allele frequencies of the hereditary movement disorders in the Swedish population can be estimated by using the Hardy-Weinberg Equilibrium (HWE) equations. HWE states that allele and genotype frequencies in a large, randomly mating population remain constant from generation to generation in the absence of evolutionary influences such as mutation, migration, natural selection, genetic drift, or non-random mating. Although HWE may not hold perfectly due to factors like population structure and migration in Sweden, it can be used for determining the MAF cutoff for dystonia, ataxia and PD. The Hardy-Weinberg equation is $p^2 + 2pq + q^2 = 1$, where p^2 represents the frequency of individuals homozygous for the major allele, $2pq$ represents the frequency of heterozygous individuals, and q^2 represents the frequency of individuals homozygous for the minor allele; together, these frequencies sum to 1, meaning they account for the entire population's genotypes. p and q sum also to 1; this means that in a population, the total frequency of the two alleles for a particular gene adds up to 100%. Dystonia has a prevalence of P = 35.1/100,000. For autosomal dominant mode of inheritance, which is the mode of inheritance of the family under investigation in Paper I, the frequency of affected individuals can be approximated as 2q since the disease allele frequency q is small and the normal allele frequency p is close to 1; solving $2q = 0.000351 \Rightarrow q = 0.000351 / 2$ gives an allele frequency q of approximately 0.000176 (or 0.018%). Because of the expected dominant mode of inheritance, carriers are affected individuals, so there is no separate carrier frequency. For progressive ataxia, P = 6.0/100,000 in Sweden. For autosomal dominant ataxias, based on HWE equation, the MAF cutoff should be approximately 0.003%. In contrast, for autosomal recessive ataxias, where two copies of the variant are required, the MAF can be higher and is estimated as the square root of the prevalence. In this case, MAF cutoff would be approximately 0.775%. In PD, the upper bound of the prevalence in Sweden is 20,000/10,000,000. At most 10% of this prevalence has a genetic cause. Therefore, the proportion of PD patients in Sweden carrying a genetic cause is estimated to be 20/100,000. For autosomal dominant inheritance, a MAF threshold of approximately 0.0001 (or 0.01%) would be reasonable while for autosomal recessive inheritance, the MAF threshold could be approximately 0.0141 (1.41%).

In the analyses that was conducted in Papers I,II,III, variants with MAF ≤ 1% or variants that are not included in the non-Finnish European population of genome aggregation database (gnomAD) v.4.0 exomes and genomes,[85] in 1000 Genomes project phase 3,[84] and in SweGen variant frequency dataset of SweFreq,[79] were selected by using VEP filtering tools or Python scripts. In the analysis of Parkinson's patients' data (Paper IV), a 2% cutoff was applied to capture biallelic and compound heterozygous variants as well as variants in *GBA1* which are not rare. Additionally, a 5% threshold was used in the analysis conducted in Paper III to identify common variants present in affected family members but absent in unaffected members.

Choosing a 1% MAF threshold or higher, can improve sensitivity by allowing the inclusion of rare and potentially disease-causing variants that might be slightly more common than strict prevalence-based estimates suggest. This prevents missing true pathogenic variants (false negatives), especially in disorders with incomplete penetrance and late onset, where pathogenic alleles can be present at higher frequencies in the population without always causing disease. In other words, a slightly higher MAF cutoff increases the chance of catching all relevant variants, even if that means reviewing more variants that might ultimately be benign.

However, when interpreting genetic variant frequencies from frequency databases, several limitations must be considered. gnomAD[85] includes data from approximately 807,000 individuals, with an average age of 54 years, and some individuals with Mendelian disease may still be included which can complicate variant interpretation. For example, an increasing number of dystonia-associated alleles have been seen in the gnomAD database.[95] The presence of sequencing artifacts or mosaicism are possible explanations. There is also population bias. Some populations are overrepresented, while others are underrepresented. This can lead to incorrect interpretation of variants as rare or common. The 1000 Genomes Project,[84] with about 2,500 individuals and low sequencing coverage, offers diverse but limited population representation and reduced sensitivity for very rare variants. SweFreq[79] focuses on the Swedish population, sequencing around 1,000 individuals with an average age of 65.2 years, but its smaller sample size may limit detection of rare variants and regional genetic diversity. Since not all individuals in these databases are healthy and some may have undiagnosed, late-onset or incompletely penetrant diseases, this can lead to an overestimation of how common certain pathogenic variants are in the general population, making it harder to interpret their true disease risk. Therefore, cautious interpretation and use of complementary data sources are essential for accurate variant pathogenicity assessment.

Filter based on *in-house* gene lists

For index cases without a family history or probands whose family members did not consent to provide samples for genetic analysis, a genetic diagnosis would not be feasible without first selecting and prioritizing variants located in genes known to

be associated with the movement disorders studied in this project. Even when family samples were available for identifying new genetic causes, a sufficient number of family members were needed to provide samples, as was the case with the families analyzed in Paper I and III. To address this, *in-house* gene lists were compiled. For Paper I, a dystonia gene list was compiled from HPO[78] entry on dystonia (HP:0001332, accessed in 2020), and by retrieving information from MalaCards,[80] OMIM,[77] and GeneReviews[81] databases. Additional genes mentioned in review articles on dystonia[14,44,96-100] were also included. The list contained 535 dystonia genes, 5 RNA genes and 7 dystonia loci. For Paper II and III an ataxia gene list with 1020 genes compiled by HPO[78] entry on ataxia (HP:0001251, accessed in 2021) and by ataxia gene panels (Clinical genetics in Lund https://genpaneler.genetiklund.se/ataxi/ and Blueprint Genetics https://blueprintgenetics.com/tests/panels/neurology/ataxia-panel/), while for Paper III genes associated with neuropathy (HP:0009830, accessed in 2021) were also searched. For Paper IV, the HPO[78] entry on Parkinsonism with favorable response to dopaminergic medication (HP:0002548, accessed in 2024) and newly discovered PD genes resulted in a list of 44 PD genes. The VCF files were filtered based on these gene lists by using Python scripts.

Filter based on pathogenicity predictions

The variants were also filtered based on pathogenicity prediction and conservation scores from several predictive tools as for example PolyPhen-2,[86] SIFT[87] and GERP++_RS.[92] In most cases, they were prioritized based on their CADD-phred score[56] (Paper I,II and IV). CADD predicts a continuous phred-like score that ranges from 1 to 99, with higher values indicating more deleterious cases. For example, variants with CADD-phred score above 30 are predicted to be the 0.1% most deleterious possible substitutions in the human genome. It can also predict deleteriousness of both SNVs and indels of coding and non-coding regions. Synonymous and intronic variants evaluated with freely available tools (TraP-score,[69] PredictSNP2,[70] GWAVA[71]). SpliceAI[68] was used for the prediction of variants' effect on splicing.

Classification of the variants

The variant classification was performed based on the guidelines published by the ACMG in 2015.[60] We only considered candidate disease-causing variants those that were classified Pathogenic, Likely Pathogenic or Variants of uncertain significance (VUS) by the classifiers used (final verdict of VarSome ACMG Implementation of VarSome.com[82] and Franklin by Genoox). In Paper II, manual classification was also performed. We also searched for the candidate variants in ClinVar[76] and OMIM[77] databases.

Candidate SNVs and small indels were curated with IGV[59] in order to manually review artifacts like read-end artifacts caused by local misalignments near indels, incorrect alignments in low-complexity regions, low-quality base calls, strand bias artifacts, and paralogous read alignments poorly represented in the reference genome.[101]

Copy number variants analysis

ExomeDepth for Whole exome sequencing data

ExomeDepth was used for the analysis of CNVs from exomes. ExomeDepth uses read depth information to detect CNVs in exome sequencing experiments. The key approach involves comparing the test exome to a corresponding aggregate reference set. This reference set should include exomes from the same batch and be optimized for each individual exome.[57] ExomeDepth assumes that the CNV of interest is not present in the aggregate reference set. Therefore, related individuals should be excluded from this set. This also means that ExomeDepth may overlook common CNVs if they are present in the aggregate reference. ExomeDepth is particularly well-suited for detecting rare CNVs, making it ideal for analyzing rare Mendelian disorders.[57] When calling CNVs on the X chromosome, mismatched genders between the test sample and reference samples can cause inaccuracies. It is important to ensure proper gender matching by not comparing female test samples to male reference samples and vice versa.

First BAM files are parsed and the getBAMCounts function creates an object of the GRanges class. A fundamental concept behind ExomeDepth is that each exome should be compared not to all other exomes, but to a carefully selected set of exomes that are highly correlated with the test exome. The first step, therefore, is to choose the most suitable reference samples. For each genomic region (or bin), the read counts of the reference samples are combined as a weighted sum to create a reference profile. This reference profile represents the expected "normal" coverage for that region. Then beta-binomial model is applied to the full set of exons which provides a more accurate representation of the variability in read counts across different genomic regions of the test sample and then the CNVs are called by running the underlying hidden Markov model.[57] The calls can be ranked by using the BF column. The BF column stands for Bayes factor, which indicates the statistical support for each CNV.[57] It represents the log10 of the likelihood ratio between the data supporting the CNV call and the null hypothesis (normal copy number). A higher value indicates greater confidence in the presence of a CNV. Common CNVs were filtered out, as the goal was to detect rare CNVs. This was

achieved by identifying overlaps with a set of common CNVs reported in Conrad et al., 2010.[102]

GATK for Whole genome sequencing data

GATK run in 'cohort' mode was used for the CNV analysis from WGS data (https://gatk.broadinstitute.org/hc/en-us/articles/360035531152--How-to-Call-common-and-rare-germline-copy-number-variants). First, a list of intervals is created and then the read alignments that overlap with these intervals are counted. PreprocessIntervals divides WGS intervals into equally sized bins across the reference genome. Binning involves creating uniform intervals across the entire reference. CollectReadCounts compiles the raw integer counts of reads that overlap each interval. At the next step intervals can be annotated with covariates to aid in filtering and explicit modeling and in the removal of intervals with outlier read counts. To determine which regions to exclude, AnnotateIntervals assigns labels to the given intervals based on GC content, and, if provided with the appropriate optional resource files, also based on mappability and segmental duplication content. FilterIntervals then refines the intervals list by applying the annotations and other customizable thresholds. Then the ploidy of each contig is modeled and called. DetermineGermlineContigPloidy estimates the ploidy at the contig level for both autosomal (e.g., human chr20) and allosomal (e.g., human chrX) contigs. The tool determines the baseline ploidy for each contig using sample coverages and ploidy priors, which provide the prior probabilities for each ploidy state. During this process, the tool also generates global baseline coverage and noise data. The next step is the heart of the workflow. The copy number for each interval is modeled. Given the high computational demand of model fitting, the analysis is breaking into smaller parts. Picard IntervalListTools breaks up the intervals list into roughly balanced lists and then GermlineCNVCaller learns a denoising model for each scattered shard while consistently detecting CNVs across all shards. The tool simultaneously models systematic biases and CNVs, enabling the sensitive detection of both rare and common CNVs. Finally, PostprocessGermlineCNVCalls consolidates the results from GermlineCNVCaller, performs segmentation, and determines the copy number states. The tool generates sample calls for each interval and segment in VCF format and processes one sample at a time.

CNVs were then curated with IGV[59] in order to visually inspect the findings, cross-check for noise or artifacts and manually validate and investigate CNV breakpoints by visualizing signals like split reads (one read maps partially at two places), soft-clipping (clipped bases at breakpoint boundaries), discordant read pairs (paired-end reads with unexpected insert sizes or orientations). DECIPHER[58] was used for visualizing the CNVs, checking the frequency and linking them to phenotypes and clinical conditions.

Short tandem repeats analysis

ExpansionHunter

ExpansionHunter utilizes a predefined variant catalog that outlines the genomic locations and structures of targeted loci. The locus structure is defined with a subset of standard regular expression syntax. For each locus, the tool extracts the relevant reads from a BAM file and realigns them based on a graph-based model that represents the structure of the locus. A sequence graph is made up of nodes representing sequences, with directed edges that describe how these sequences can be linked to form different alleles. The realigned reads are then used to genotype the variants at that locus. Genotyping is carried out by examining how the reads align through the graph, using the observed alignment patterns to infer which alleles are supported (present) and which are not.[61] Repeats are genotyped as described in Dolzhenko et al. (2017)[62] (Figure 11), while SNVs and indels are genotyped using a simple Poisson-based model.

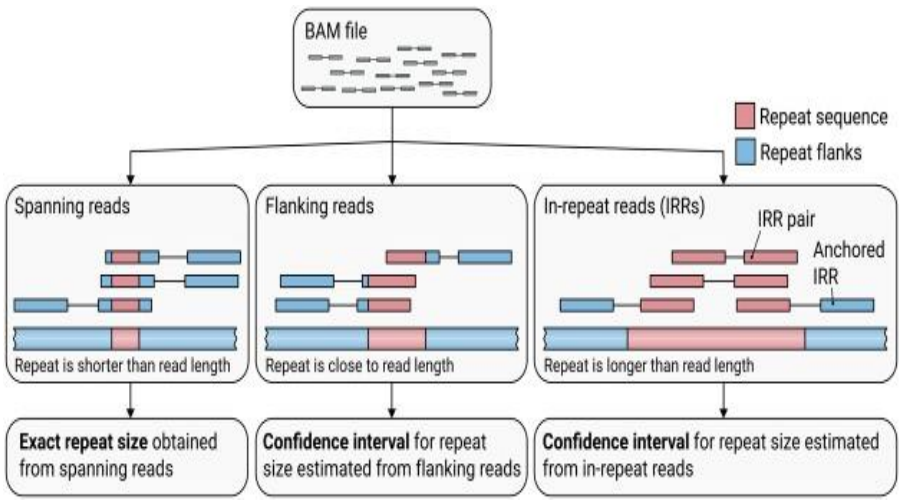

Figure 11: Genotyping of repeats. To quantify repeat lengths, reads are classified into 3 categories. (1) spanning reads, which fully encompass the repeat; (2) flanking reads, which include the repeat and the flanking sequence on one side; and (3) in-repeat reads (IRRs), which are entirely contained within the repeat. For repeats shorter than the read length, the repeat length based on spanning and flanking reads is estimated. For repeats longer than the read length, their lengths are determined by identifying and counting the IRRs. In-repeat reads whose mates are anchored to the repeat flanking regions are used to estimate repeat sizes up to the fragment length. When no additional evidence of long repeats with the same motif is found elsewhere in the genome, pairs of in-repeat reads can also be employed to estimate the size of longer repeats that exceed the fragment length. Reproduced from Dolzhenko *et al.* 2017[62] licensed under CC-BY 4.0

ExpansionHunter Denovo

STRs not previously associated with disease were assessed with ExpansionHunter Denovo in Paper III. The tool generates genome-wide STR profiles that include two categories of in-repeat reads (IRRs): anchored IRRs and paired IRRs.[64] Anchored IRRs are reads where one read maps within the repeat region, while its mate aligns confidently to the flanking genomic sequence. In contrast, paired IRRs are read pairs where both reads consist entirely of repeat sequences sharing the same motif.

Repeats that are longer than the read length tend to produce anchored IRRs, while repeats that exceed the fragment length of the DNA library generate both anchored and paired IRRs.[64] The genomic positions where the anchored mates align approximate the locations of loci containing repeat expansions, and the number of IRRs reflects the length of the expanded repeat.[64]

Information about anchored IRRs is captured in an STR profile for each repeat motif (e.g., CAG) by grouping together regions where anchored IRRs are located close to one another, along with the total count of anchored IRRs observed in those regions (Figure 12). It is important to note that the reported positions refer to the anchor reads, not the IRRs themselves, since IRRs often have unreliable alignments.[64] In contrast, the source location of paired IRRs cannot be determined when the genome contains multiple long repeats with the same motif. As a result, STR profiles report only the total number of paired IRRs per repeat motif, without assigning them to specific genomic regions.

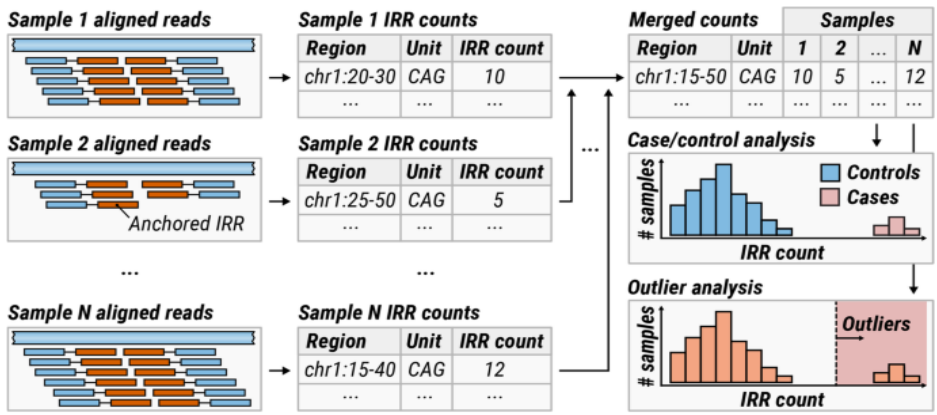

Figure 12: Principle of ExpansionHunter Denovo. (Left) Anchored IRRs are identified by scanning all aligned reads. (Middle) The counts of these IRRs are then compiled into STR profiles. (Right) STR profiles from all samples are combined. If the dataset includes defined case and control groups, IRR counts are compared between them at each locus. If no such grouping exists, an outlier analysis is conducted instead. Reproduced from Dolzhenko *et al.* 2020,[64] licensed under CC BY 4.0.

Phasing Variant calling format files with Beagle

Phasing refers to the computational process of determining the arrangement of alleles on each homologous chromosome, thereby establishing which variants are co-inherited on the same haplotype. In diploid organisms, such as humans, each individual possesses two copies of each chromosome. While sequencing data typically provides the genotype at each variant locus, it does not inherently specify the chromosomal origin of these alleles; that is, which variants reside on the same chromosome copy.[103] Phasing addresses this limitation by resolving genotypes into phased haplotypes, effectively reconstructing the linear sequence of alleles on each homologous chromosome separately.[103]

Beagle is a widely used software tool for genotype phasing and imputation (filling in missing genotypes). It uses a sophisticated statistical model based on hidden Markov models (HMMs) and reference haplotypes (1000 Genomes reference panel) to infer the most likely phased haplotypes from genotype data.[65] Two-stage phasing substantially reduces computational time, particularly when a large proportion of variants are of low frequency, which is common in sequencing data. In the first stage, phasing is performed on the relatively small subset of high-frequency variants, creating a haplotype scaffold. This scaffold is then used in the second stage to impute alleles at low-frequency variants. By excluding the low-frequency variants from the computationally intensive iterative phasing algorithm, the two-stage approach achieves greater efficiency.[65]

Identity-by-descent analysis with hap-ibd

Identity-by-descent (IBD) refers to the phenomenon where two or more individuals inherit the same segment of DNA from a common ancestor without recombination. IBD analysis involves detecting and quantifying these shared genomic segments to infer relatedness between individuals or populations.

IBD segments are identical stretches of DNA inherited from a recent common ancestor, distinguishing them from identity-by-state (IBS), where alleles may be identical but not necessarily inherited from a common source.[66] The IBD analysis can also be used for mapping disease genes. IBD analysis helps identify regions of the genome that are shared among individuals with the same disease. These shared regions likely contain the gene responsible for the disease, since they were inherited from a common ancestor.

hap-ibd is a software tool designed for efficient and accurate detection of IBD segments between individuals using phased haplotype data. hap-ibd begins by identifying short exact matching haplotype segments, called seeds, between pairs of

phased haplotypes.[66] These seeds are short stretches where two haplotypes are identical with no mismatches. Seeds serve as anchors or starting points for building longer shared segments. Once a seed is found, hap-ibd attempts to extend this seed in both directions, allowing for a small number of mismatches to accommodate, genotyping errors or mutations.[66] This process expands the shared segment beyond the seed to capture the full length of the IBD segment. After extension and merging, hap-ibd applies length and quality thresholds to filter out false positives. The output reports the final set of high-confidence IBD segments defined by these extended seed regions.[66]

Co-segregation analysis

According to the ACMG guidelines, the co-segregation of a variant with disease within a family can serve as supporting, moderate or strong evidence for pathogenicity (criterion PP1). Conversely, the lack of co-segregation between a variant and disease can support a benign classification (criterion BS4).[60] A set of straightforward, quantitative guidelines for analyzing the co-segregation of variants with disease were proposed in 2016.[104]

The model relies on the following assumptions:

- The condition follows a fully penetrant inheritance pattern, with no instances of phenocopies.
- The variant allele is rare enough that all occurrences within the pedigree originate from a single common ancestor, rather than multiple independent introductions.

Segregation analysis can be conducted either within a single family with several affected members or across multiple families sharing the same variant and condition. A probability is calculated to assess how likely it is that the observed pattern of variant segregation and affected status in the pedigree(s) arose by chance.[104] The probability is $N=(1/2)^m$, where m represents the number of informative meioses for the variant being evaluated in the context of co-segregation. For the proband who carries the variant the probability is considered as 1. Unaffected family members can also be informative. In that case the probability of an unaffected non-carrier, is 1 minus the probability that the individual is a carrier.[104] The smaller the N probability, the stronger the evidence that the variant co-segregates with the condition in affected individuals rather than being distributed randomly.[104] Specific probability thresholds define whether the evidence for pathogenicity is considered supporting, moderate, or strong (Table 2).

Table 2: Proposed co-segregation evidence for each ACMG pathogenicity evidence level. From Jarvik and Browning 2016,[104] modified.

Evidence level	*N* from data from single family	*N* from data from multiple families
PP1_Supporting	≤1/8 (≤0.125)	≤1/4 (≤0.25)
PP1_Moderate	≤1/16 (≤0.06)	≤1/8 (≤0.125)
PP1_Strong	≤1/32 (≤0.03)	≤1/16 (≤0.06)

Methods used in Paper I

WES raw data were aligned to the human genome reference build hg19/GRCh37. We analyzed the annotated variant lists by identifying all rare variants shared by the two affected individuals and absent in the three unaffected family members. After excluding common population variants, 740 rare variants remained that were shared by the affected pair and selected for further assessment. From these, 306 were not present in the unaffected members. Two filtering strategies were applied: first, we prioritized variants in genes listed in our expanded dystonia gene panel; second, we considered variants in all genes, including those rare not previously associated with dystonia, with CADD_phred score[56] above 20.

We then screened for rare *TOR1AIP2* variants in over 280 exomes from patients with other movement disorders, 15 exomes from dystonia patients, and in the Munich Exome Server, which contains over 1,000 dystonia exomes and more than 25,000 exomes from individuals with various conditions or relatives of dystonia patients. This led to the identification of a second family (Family 2) from Slovakia, with a similar movement disorder and a different candidate variant in *TOR1AIP2*.

The functional impact of the LULL1 (luminal domain-like LAP1) p.(Arg412Gly) variant (LULL1^{R412G}) on TorsinA binding was assessed through both in vitro co-purification assays and *in-silico* modelling. Results were compared to those of wild-type protein interactions and to the DYT1-associated TorsinA$^{\Delta E303}$ variant, revealing a similarly impaired binding profile.

Methods used in Paper II

Clinical examination

All patients with a clinical diagnosis of hereditary ataxia were interviewed using a standardized checklist, either during a research visit to our clinic or during home visits conducted by a study physician and a nurse. Clinical records were reviewed

to obtain results from brain imaging, nerve conduction studies, cerebrospinal fluid analyses, and genetic testing. A standardized neurological examination protocol specific to ataxia was used, and symptom severity was assessed using the Scale for the Assessment and Rating of Ataxia (SARA).[105] Family pedigrees were also constructed.

Genetic analysis

Before the start of this study, some patients had already received a confirmed genetic diagnosis through analyses conducted as part of their routine clinical evaluation. These individuals had been tested for repeat expansions associated with SCA1, 2, 3, 6, 7, while others were examined for specific genes based on their clinical phenotype. More recently, some patients also underwent gene panel testing, WES or WGS.

As part of this study, patients without a prior genetic diagnosis were investigated using WES or WGS, with the choice of method guided by their clinical presentation, family history and availability.

The **bioinformatic analysis** was carried out as described above for SNVs, indels, CNVs and known STRs.

Post NGS-phenotyping

Variants with a high likelihood of pathogenicity were reviewed in collaboration with clinicians, bioinformaticians, and a medical geneticist. These candidate findings were subsequently confirmed or rejected through post-NGS phenotyping. If the findings did match orthogonal validation testing was performed if necessary and feasible.

A variant was considered disease-causing based on the following criteria:

- The family history was consistent with the expected mode of inheritance for the suspected disorder.
- The proband and affected family members exhibited a clearly defined syndrome, characterized by a specific pattern of neurological and/or non-neurological features. In familial cases, the phenotypes of affected individuals were compared both within the family and against descriptions in existing literature.
- The genetic findings and supporting database evidence were carefully re-evaluated.

When needed, patients were re-examined by clinicians after these discussions to document additional clinical signs or symptoms not recorded initially.

Neuropathology reports were available for some of the families or for individuals who had been followed for a longer time at our clinic. In some instances, family members with similar symptoms or unaffected family members were also tested for co-segregation analysis.

Methods used in Paper III

Clinical examination

Affected and unaffected family members from five families were studied clinically. Some of them had been followed continuously by the clinician authors since the first publication of two families in 2014,[45] others were re-approached. Data was collected with a standardized interview and neurological examination, and the International Cooperative Ataxia Rating Scale (ICARS) to examine and compare specific ataxia symptoms,[106] and/or the SARA[105] scale were used.

Genetic analysis

WGS data from Families 1 and 2 were analyzed to identify variants in known ataxia and neuropathy genes, as well as through a gene-agnostic approach using co-segregation analysis, based on who we assumed to be affected.

IBD analysis was used to identify genomic segments shared by affected family members but absent in unaffected individuals by first phasing the VCF files with Beagle[65] and then running hap-ibd,[66] as was described above. Identified SNVs and CNVs, after running GATK,[50,51] VEP[52] and GATK for calling CNVs (https://gatk.broadinstitute.org/hc/en-us/articles/360035531152--How-to-Call-common-and-rare-germline-copy-number-variants.), were cross-correlated with genomic haplotypes shared among affected individuals. Known STR expansion loci were also examined and in the variant catalog were included recently described pathogenic repeats in *FGF14*[107,108] and *NOTCH2NLC*.[109] STR loci not previously linked to disease were detected as was described above with ExpansionHunter Denovo. All findings were manually curated and correlated with shared haplotypes.

From this analysis, a trinucleotide repeat expansion in *ZFHX3* emerged as the only notable finding. This expansion was subsequently confirmed through its inclusion in the ExpansionHunter[61,62] variant catalog.

Screening of larger datasets

We analyzed *ZFHX3* repeats in 89 individuals from our ataxia series datasets (25 WES and 64 WGS), and in 90 WES and 60 WGS datasets from individuals with other neurological disease from our research database. We queried the SweGen dataset with short-read WGS data (Illumina) from 1,000 unrelated Swedish individuals.[79] The architecture of normal repeats in 300 alleles (from our 90 WES and 60 WGS in-house datasets) and in all 2,000 alleles from SweGen was analyzed with ExpansionHunter[61,62] and REViewer.[63]

Analysis of shared haplotype of the five families

IBD predictions by adding WGS data from members of Families 3, 4, and 5, to investigate a possible relation between all five families and to confirm a shared haplotype were performed. Within the IBD genomic region that included *ZFHX3,* we looked for low frequency SNVs in population databases (gnomAD NFE[85] and SweGen[79]) and their occurrence among affected and unaffected members of all five families was compared.

Long-read sequencing

Long-read genome sequencing was conducted on DNA samples from individuals III:1 and IV:1 of Family 1, who exhibited significant differences in disease severity and age at onset (37 and 15 years, respectively). For comparison, PacBio long-read data from 27 individuals provided by the Human PanGenome Consortium[110] were also utilized. We manually analyzed the data to accurately determine the number and structure of repeats in *ZFHX3* and to rule out structural variations within the 16q22 SCA4 locus.

Neuropathology

A clinical postmortem examination was performed on individual IV:1 from family 1 who died unexpectedly at the age of 28 years in order to elucidate the cause of his death. Neuropathological examination of the peripheral nervous system as well as of the CNS was performed according to standard practices at the Department of Pathology, Region Skåne.

Methods used in Paper IV

The **bioinformatic analysis** was carried out as described above for SNVs, indels and CNVs.

SNVs and small indels in *GBA1* were manually reviewed using IGV[59] to ensure that the variant calls represented true positives, rather than artifacts caused by misaligned reads from the highly homologous pseudogene *GBAP1*. This verification was performed using the BLAT tool[111] integrated within IGV.

For the two patients carrying the *CHCHD2* frameshift variant, IBD analysis was conducted to assess possible relatedness. VCF files covering all autosomal chromosomes were phased and imputed using the Michigan Imputation Server 2,[112] with the HRC r1.1 2016 reference panel.[113] Following phasing, hap-ibd v1.0 was employed to perform the IBD analysis.[66] The resulting data were then curated using custom *in-house* scripts.

To confirm and pinpoint the breakpoints of the *SNCA* duplications, SNV array analysis was carried out at Centogene in Rostock, Germany, utilizing the Infinium™ Global Diversity Array with Cytogenetics kit (Illumina). This kit features around 1.8 million SNV markers distributed across the entire genome and covers 99.9% of exons.

To assess findings related to monogenic variants, we utilized clinical and genetic data from the Malmö Diet and Cancer Study (MDCS; https://www.malmo-kohorter.lu.se/malmo-diet-cancer-mdc), a population-based cross-sectional cohort established to explore potential associations between dietary and lifestyle factors and cancer risk.[114] A total of 29,387 individuals with complete genetic and clinical data were included in this cohort, of whom 695 were diagnosed with Parkinson's disease.

This thesis has been partially produced using the generative AI model ChatGPT, which was used primarily for spelling and grammar review in line with the guidelines of Lund University. I have processed the text and take full responsibility for the content.

Statistical analysis

The diagnostic yield **(**proportion of analyses that provide a definitive diagnosis**)** was determined in papers II and IV.

Variant evaluation in Paper IV involved calculating chi-squared and Fisher's exact test statistics and odds ratios based on allele frequencies from the study and the MDCS (https://www.malmo-kohorter.lu.se/malmo-diet-cancer-mdc) cohorts combined. Due to the low number of occurrences, the Yates correction was applied to the chi-squared p-values.

Ethical aspects

The studies of this thesis were approved by the Regional Ethics Review Board, Lund, Sweden. Biological samples were collected and stored according to ethics committee approval and applicable laws and regulations. Informed written consent was obtained by all patients prior to inclusion in the studies. Genetic data analyzed in our studies, was retrieved with written consent from the patients as approved by the Swedish Ethical Review Authority. In the consent, it was mentioned that the patient would be contacted again only if there were findings explaining the cause of their neurological disease. Secondary findings were not investigated.

All individual genetic data were treated as highly sensitive and were analyzed exclusively on secure servers designated for sensitive data. Individual-level datasets are stored within the healthcare provider's secure server infrastructure, which is specifically designed for the protected storage of medical and genetic information.

Results

Paper I

Family 1

The index patient from Family 1 was a 63-year-old woman who had begun experiencing episodes of rapid horizontal head movements at the age of 37. These episodes were accompanied by dystonic twisting of the upper body, characterized by protrusion of the right shoulder and sudden, ballistic elevation of the right arm. Over time, similar movements began affecting the left shoulder and arm. Clinical examination revealed dystonia in the right arm and/or mild chorea in the left. Her brain MRI showed no abnormalities, and surface electromyography was conducted. Her son reported involuntary muscle twitching around the mouth, noticeable only when smiling. At age 27, he began experiencing episodic dystonic movements.

Family 2

The index patient of Family 2 was a Slovakian woman who, at the age of 38, developed involuntary choreatic movements, predominantly affecting her face and body. While walking, she exhibited reduced right arm swing accompanied by posturing and dystonic stiffness in the right leg, along with mild hemichorea on the left side. Brain MRI was normal. Her 10-year-old son showed finger hyperkinesias consistent with dystonia. During gait assessment, he demonstrated reduced right arm swing, dystonic stiffness of the right leg, and mild hyperkinesias in the left arm and hand. His brain MRI was also normal. Video recordings of her 42-year-old brother revealed signs suggestive of dystonic posturing combined with distal weakness in his right arm.

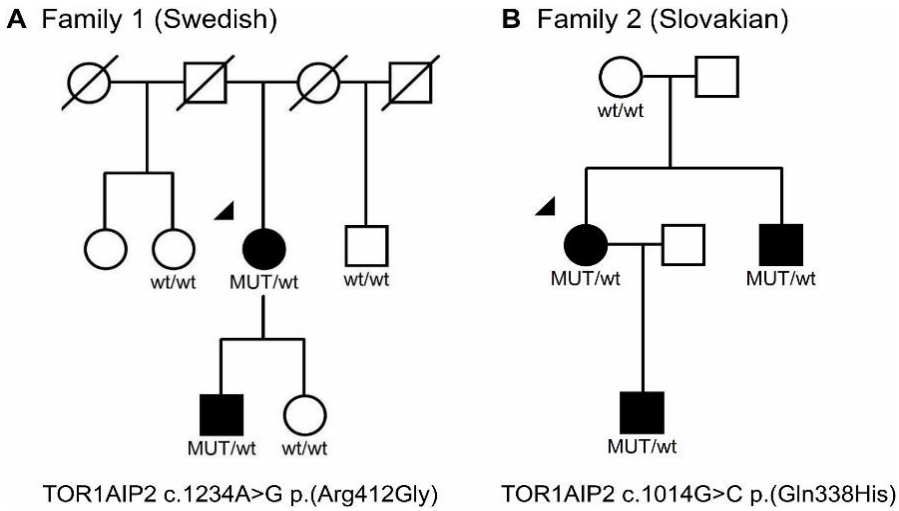

Figure 13: Pedigrees of the two families with dystonia and rare variants in *TOR1AIP2*. Black symbols denote symptomatic individuals. MUT: variant; wt: wild type. Reproduced from Kafantari *et al.*2025,[115] licensed under CC BY 4.0.

Genetic analysis

In Family 1, WES analysis of five members (as shown in Figure 13), targeting on genes from our expanded dystonia gene panel identified nine rare variants shared between the affected mother and son. Five of these variants were also found in unaffected family members. Among the remaining four, one was consistently predicted to be benign across multiple *in-silico* tools, and two were synonymous changes.

An exome-wide analysis revealed 41 rare variants with a CADD-phred score[56] above 20 that were present in the affected pair but absent in the three unaffected relatives. Of these, seven missense variants were located in genes previously associated with neurological phenotypes, but the patients' clinical presentation did not match the corresponding disorders.

Both targeted and exome-wide approaches identified the same heterozygous variant, c.1234A>G (p.Arg412Gly) in *TOR1AIP2* (NM_001199260.2, ENST00000609928.6; rs753337635). This variant had a CADD-phred score of 23.8 and was extremely rare, present in only 7 out of 1,180,028 Non-Finnish European alleles in the gnomAD exome database (v4.0.1; allele frequency: 0.000005932). It was absent from the gnomAD genome dataset, the 1000 Genomes Project,[84] all other ethnic groups in gnomAD[85], and from over 25,000 in-house exomes analyzed in Munich, Germany.

No *TOR1AIP2* variants were detected in 15 exomes from patients with dystonia or rare movement disorders at Lund University. However, a search of more than 1,000 dystonia exomes from the Munich Exome Server identified a second rare *TOR1AIP2* variant, c.1014G>C (p.Gln338His; Q338H), present in only 20 of 1,614,092 gnomAD alleles (allele frequency: 0.00001239; CADD-phred score: 17.50). This variant co-segregated with the disease in Family 2. Structural modelling indicated that the Q338 residue does not participate in TorsinA binding and is located on the surface of LULL1's luminal domain.

Functional assays

The R412G variant is a substitution of a basic amino acid with a long side chain by a neutral amino acid without side chain. The R412 residue of LULL1 forms hydrogen bonds to the neighboring E302 residue of TorsinA. This or the adjacent residue E303 can be deleted in the recurring p.(delGlu303) 3bp in-frame deletion (ΔE302/303) which is the most common cause of DYT1[116] (Figure 15). So, the R412G variant disrupts the same binding interface with the ΔE302/303. This suggested that by weakening LULL1 binding, the variant likely affects TorsinA activation. The variant's effect on its binding to TorsinA was investigated experimentally. The co-purification assay revealed that TorsinA^{wt}-LULL1^{R412G} (reflecting the situation in Family 1) molecularly phenocopies TorsinA$^{\Delta E303}$-LULL1^{wt} (Figure 14).

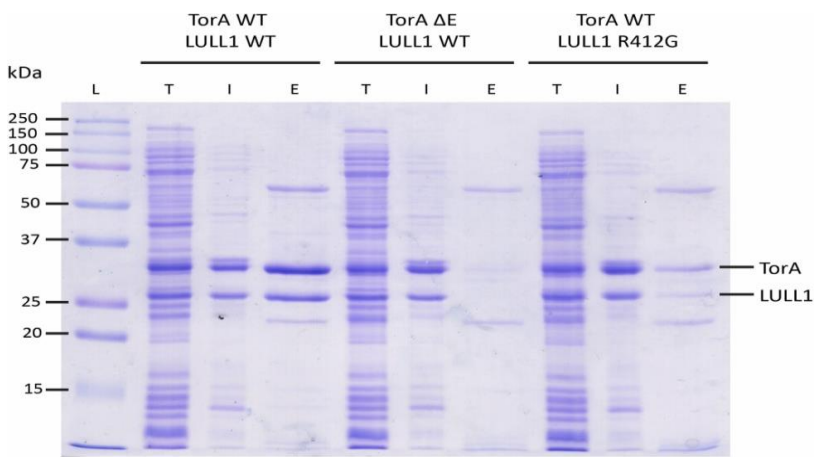

Figure 14: Co-purification assay. Binding between LULL1 and TorsinA was examined using a Ni-affinity co-purification with recombinant, bacterially expressed protein. TorsinA is polyhistidine-tagged whereas LULL1 is not. Complex formation is observed by an approximate 1:1 ratio of TorsinA and LULL1 in Ni eluates as visualized on SDS-PAGE. Whereas wild type TorsinA (TorsinA^{WT}) complexes wild type LULL1 (LULL1^{WT}) in the eluate, E303-deleted TorsinA (TorsinA$^{\Delta E303}$) abolishes binding with LULL1^{WT}. LULL1^{R412G} results in markedly reduced complex formation, phenocopying the TorsinA$^{\Delta E303}$/LULL1 ^{WT} results. L, protein marker; T, total lysate; I, insoluble fraction; E, Ni eluate. Reproduced from Kafantari *et al.*2025,[115] licensed under CC BY 4.0.

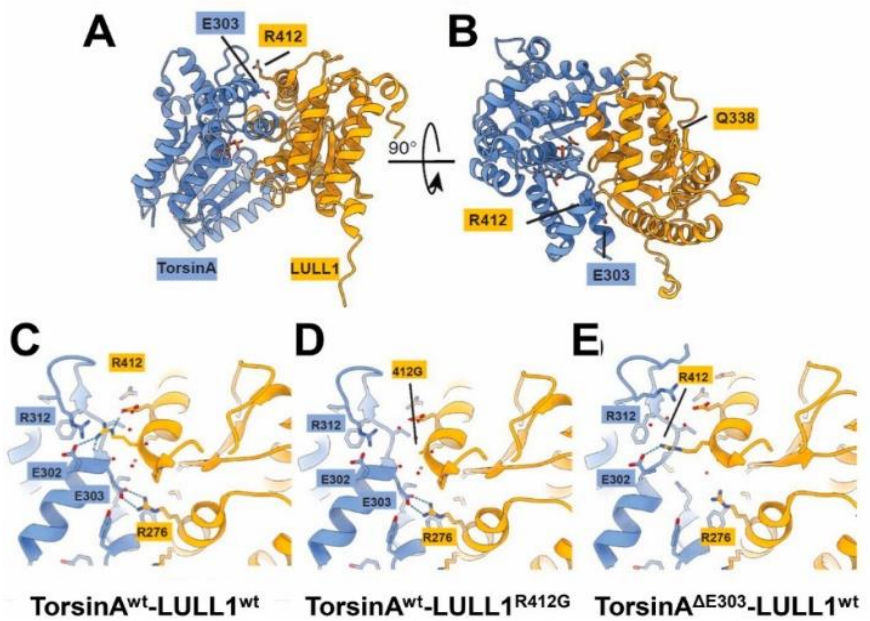

Figure 15: The binding site of TorsinA to LULL1. A and B: Overview of the TorsinA:LULL1 complex (Protein Data Bank code: 5J1S). LULL1 in orange, TorsinA in blue. The LULL1 variants' positions described here (Family 1 LULL1^{R412G}, Family 2: LULL1^{Q338H}) and of the recurrent TorsinA$^{\Delta E303}$ variant causing DYT1 are shown. C: Close-up of the wild type (wt) TorsinA:LULL1 interface, indicating that the LULL1 R412 residue forms several hydrogen bonds with TorsinA. D: The R412G variant in LULL1 (Family 1) disrupts hydrogen bonding at R412. E: In DYT1, lack of the TorsinA E303 residue (ΔE303 deletion) disturbs hydrogen bonds with LULL1 residues in the vicinity of LULL1 R412. Therefore, LULL1^{R412G} likely phenocopies TorsinA$^{\Delta E303}$, affecting the binding of TorsinA to LULL1. The LULL1^{Q338H} variant is not expected to interfere with TorsinA binding and awaits further molecular analysis. Reproduced from Kafantari *et al.*2025,[115] licensed under CC BY 4.0.

Paper II

A total of 158 individuals with ataxia were identified and contacted.[117] Of these, 91 patients were examined as part of this study, with four subsequently excluded due to alternative diagnoses. This resulted in 87 patients with hereditary ataxia being included for both genetic and clinical analysis. Prior to inclusion, 27 patients from 19 families already had a confirmed genetic diagnosis, and two additional relatives were identified as presymptomatic carriers. Within this study, 15 patients from 11 families received a new genetic diagnosis. By the end of the study, a confirmed genetic diagnosis had been established for 44 out of 87 individuals (50.6%), representing 30 out of 76 families (39.5%) (Figure 16).

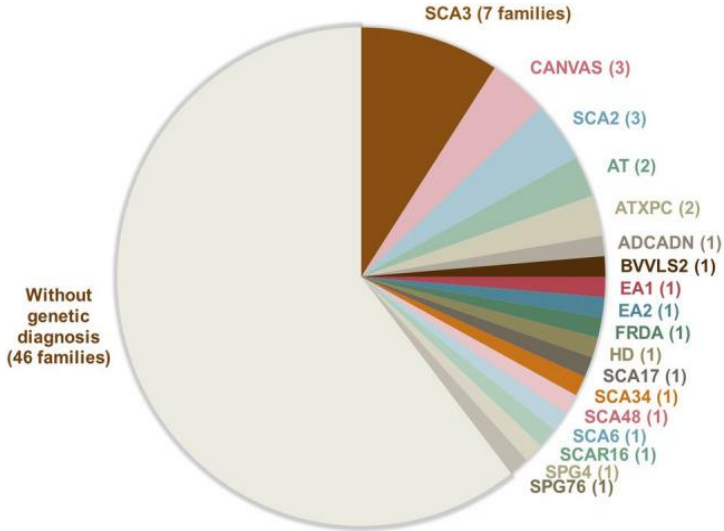

Figure 16: Subtypes of ataxia encountered in this study. Molecular diagnoses of ataxia were established for the study participants. Patients with SCA3 were recruited from a broader geographical region, as they participated in multicenter studies focused on this disorder. Additionally, nine patients with Friedreich ataxia from seven families within our hospital's catchment area had been enrolled in a previous study and were therefore not recontacted. *ADCADN:* Autosomal dominant cerebellar ataxia, deafness, and narcolepsy, *AT*: ataxia telangiectasia, *ATXPC:* ataxia-pancytopenia syndrome, *BVVLS2:* Brown–Vialetto–Van Laere syndrome-2, *CANVAS*: cerebellar ataxia, neuropathy and vestibular areflexia, *EA*: episodic ataxia, *FRDA*: Friedreich's ataxia, *HD*: Huntington's disease, *SCA*: spinocerebellar ataxia, *SCAR16*: spinocerebellar ataxia autosomal recessive 16, *SPG:* spastic paraplegia. Reproduced from Gorcenco *et al.*2024,[118] licensed under CC BY 4.0.

Nine of the eleven genetic diagnoses and their associated disease-causing variants identified in this study are presented in Figure 17.

One proband was diagnosed with Brown-Vialetto-Van Laere syndrome type 2, carrying a homozygous *SLC52A2* variant c.968T>C p.(Leu323Pro). In two families, heterozygous variants in *SAMD9L* c.2956C>T p.(Arg986Cys) and c.2640C>A p.(His880Gln) respectively, were detected, both causing ataxia pancytopenia syndrome.[119]

A family diagnosed with Spinocerebellar ataxia 34 carried a variant of uncertain significance (VUS), *ELOVL4* c.511A>C p.(Ile171Leu). Heterozygous variants in this gene have been associated with SCA34.[120,121] Following co-segregation analysis, the ACMG criterion PP1_Strong was met based on five informative meioses,[104] leading to reclassification of the variant as 'likely pathogenic.'

Two unrelated index cases exhibited different variants in *STUB1*. The first carried a heterozygous c.107T>C p.(Leu36Pro) variant associated with Spinocerebellar ataxia 48, while the second, of German descent, was homozygous for c.761G>A p.(Arg254His), which causes the autosomal recessive Spinocerebellar ataxia 16.[122]

The heterozygous *STUB1* c.107T>C p.(Leu36Pro) variant in the familial proband was classified as VUS, but the ACMG PP4 criterion was applied because 'the patient exhibited a well-defined syndrome with minimal overlap with other clinical presentations'.[123,124] However, PP4 provides only supporting evidence for pathogenicity. This patient was examined neuropathologically and showed intraneuronal inclusions that were considered relatively specific for SCA48. The combination of the genetic finding, clinical observations, neuropathology and family data made us set the diagnosis of SCA48. Genetic analyses of additional family members have since the publication of Paper II shown findings in line with this diagnosis.

An autosomal dominant spastic paraplegia 4 diagnosis was made in a proband carrying a heterozygous frameshift mutation in *SPAST*, c.722del p.(His241Profs*13). Additionally, in an index case of Turkish origin with a family history suggestive of an autosomal recessive disorder, a homozygous splice donor variant *CAPN1* c.759+1G>A was identified, confirming a diagnosis of spastic paraplegia 76.[125]

In two index cases from separate families, biallelic expansions of the *RFC1* gene involving extended AAGGG repeats (replacing the normal AAAAG pentanucleotide repeats) were found; these expansions have been associated with cerebellar ataxia with neuropathy and vestibular areflexia syndrome (CANVAS).[126] Although the exact number of pentanucleotide repeats on each allele could not be determined, the presence of the variation was confirmed through orthogonal methods.

Finally, one index case was diagnosed with Huntington's disease, carrying 36 uninterrupted CAG trinucleotide repeats in the *HTT* gene. This diagnosis was confirmed based on the clinical presentation.

In eight index cases, although candidate disease-causing variants were identified, a definitive diagnosis could not be established. This was due to several factors, such as a mismatch between the clinical phenotype and the identified genotype, or because the detected variants had not previously been reported in the literature as causes of ataxia (Figure 17). In some families, further co-segregation analyses involving relatives of the probands are still ongoing.

For example, one patient was found to carry a heterozygous variant in the *POLR3B* gene, c.1568T>A p.(Val523Glu), but their clinical presentation did not align well with diseases typically associated with *POLR3B* mutations (Figure 17). In another case, expanded repeat lengths were detected in the *RFC1* gene; however, uncertainty remains about the significance of this finding because the patient had a family history of autosomal dominant motor polyneuropathy, whereas CANVAS, a condition linked to *RFC1* expansions, is inherited in an autosomal recessive manner (Figure 17).

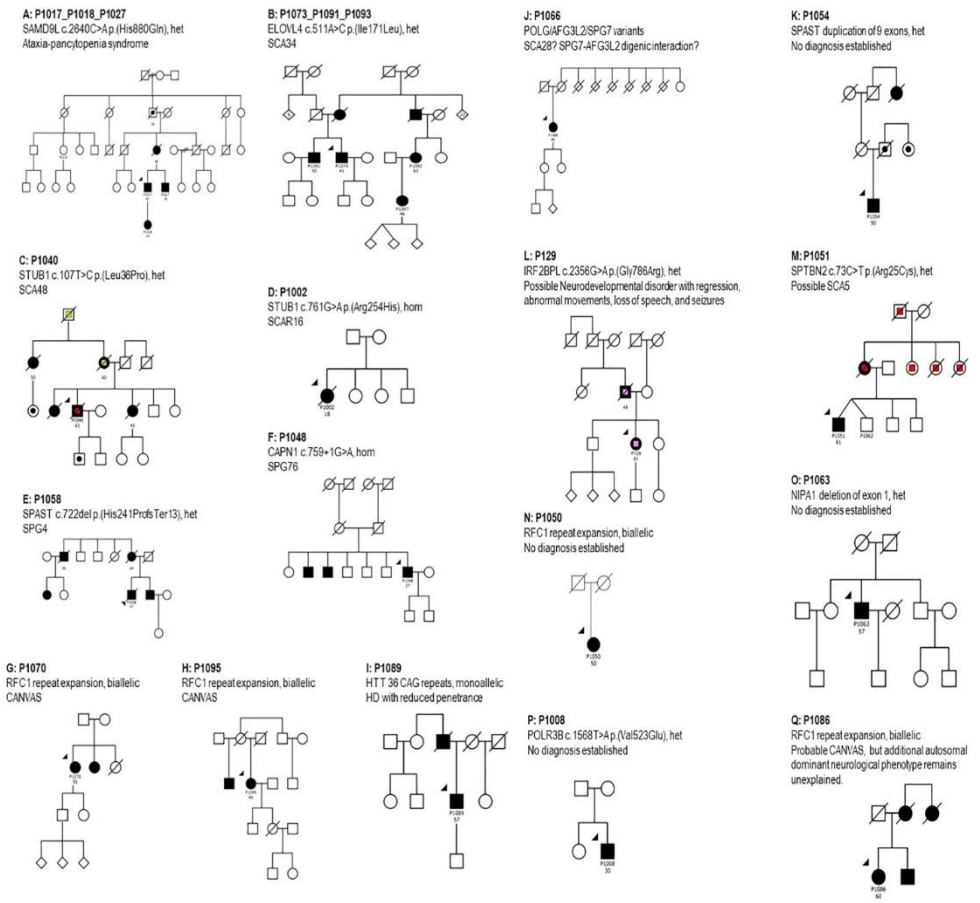

Figure 17: Pedigrees of patients and families examined genetically within this study. Round symbols indicate females, square symbols males; diagonal line indicates that the individual is deceased; patient identifiers and age at onset in years are provided below symbols; solid black symbols indicate ataxia, black dots indicate possible ataxia (acc. to family history); yellow color indicates possible dementia, red color indicates dementia. *CANVAS*: cerebellar ataxia, neuropathy and vestibular areflexia; *HD*: Huntington's disease, *het*: heterozygosity, *hom*: homozygosity, *SCA28*: spinocerebellar ataxia 28, *SCA34:* spinocerebellar ataxia 34, *SCA48*: spinocerebellar ataxia 48, *SCA5*: spinocerebellar ataxia 5, *SCAR16*: autosomal recessive spinocerebellar ataxia 16, *SPG4*: spastic paraplegia 4, *SPG7:* spastic paraplegia 7, *SPG76*: spastic paraplegia 76. Reproduced from Gorcenco *et al.*2024,[118] licensed under CC BY 4.0.

Paper III

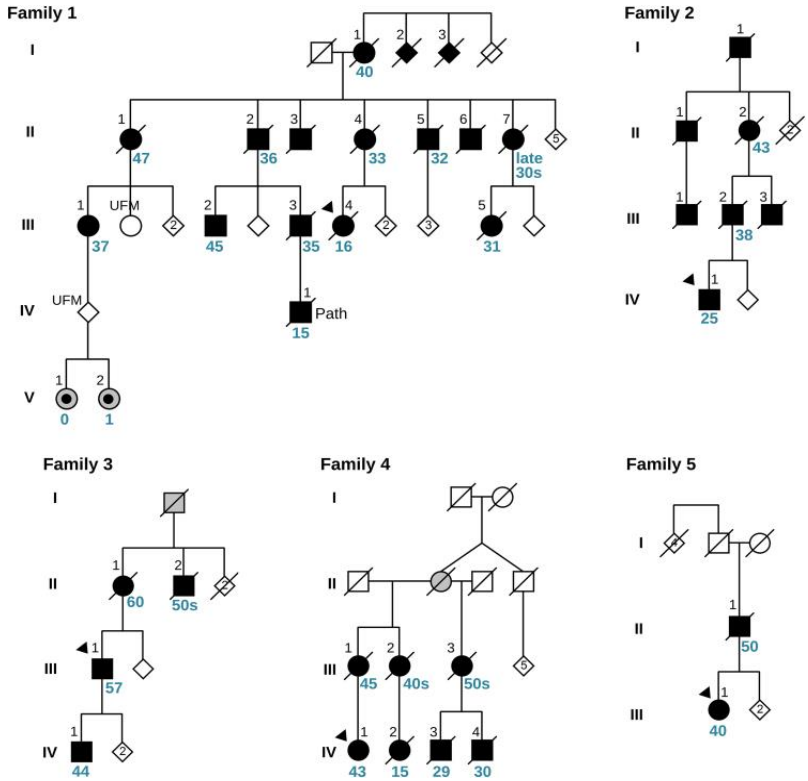

Figure 18: Family pedigrees. They were illustrated using standard symbols. Individuals affected by cerebellar ataxia with sensory and autonomic neuropathy are represented by black symbols. Probands are marked with black triangles. Gray symbols denote family members reported to have gait and balance issues based on family history. Gray symbols with a central black dot indicate two siblings who were evaluated in this study but diagnosed with a neurodevelopmental disorder distinct from cerebellar ataxia with sensory and autonomic neuropathy, and who do not carry *ZFHX3* repeat expansions. Individual identifiers are shown in black numbers above the symbols. For family members with available information, age at symptom onset is indicated below the symbols in blue. To maintain confidentiality, the sex and gender of unaffected individuals were masked, sibling order was partially altered, and some unaffected family members were omitted. UFM stands for unaffected family member who underwent genetic analysis in this study. Reproduced from Wallenius *et al.*2024,[127] licensed under CC BY 4.0.

No pathogenic SNVs, indels, CNVs, or previously reported STRs that co-segregate with the disease were found in Families 1 or 2. When including the two individuals with a severe neurodevelopmental phenotype from Family 1 (V:1 and V:2), no common genetic cause was identified, leading to their exclusion from further analyses. IBD analysis revealed 199 genomic segments shared among all affected individuals analyzed in Family 1. The shared region on chromosome 16q22 was refined to a 394 kbp segment when the proband from Family 2 (IV:1) was included.

Within this region, a GGC (poly-glycine) repeat expansion in exon 10 of the *ZFHX3* gene perfectly co-segregated with the disease in both families. Long-read sequencing determined the exact repeat expansion lengths as 57 uninterrupted GCC repeats in individual III:1 (age of onset 37 years) and 74 uninterrupted GCC repeats in IV:1 (age of onset 15 years).

Four additional ataxia-affected individuals from three more families, the probands of Families 3 through 5 and individual IV:1 from Family 3, were found to carry expanded *ZFHX3* repeats. Including these families reduced the shared genomic segment among all affected individuals to 111 kbp. Within this region, three rare intronic SNVs were detected exclusively in carriers of the repeat expansion, further refining the shared haplotype.

Predominantly synonymous GGT repeats and a non-synonymous AGT repeat encoding serine interruptions were observed within the GGC repeat region in all non-expanded alleles. Most alleles had the exact structure shown in Figure 19, while the remaining alleles exhibited various deviations. In contrast, expanded alleles showed no detectable interruptions. Since the repeat region often exceeds the 150 bp length of short-read sequencing, short reads alone cannot fully rule out interruptions. However, long-read sequencing of two affected individuals confirmed a complete absence of interruptions, with GGC being the sole repeat unit.

Thirty-eight affected members from Families 1 to 5 exhibited similar clinical features, all presenting with autosomal-dominant cerebellar ataxia accompanied by sensory and autonomic neuropathy. They experienced gait or balance disturbances with ages at onset ranging from 15 to 60 years, and the average age at symptom onset was 37.6 years. This pattern suggested anticipation, as in all but two cases, the affected offspring developed symptoms earlier than their affected parent.

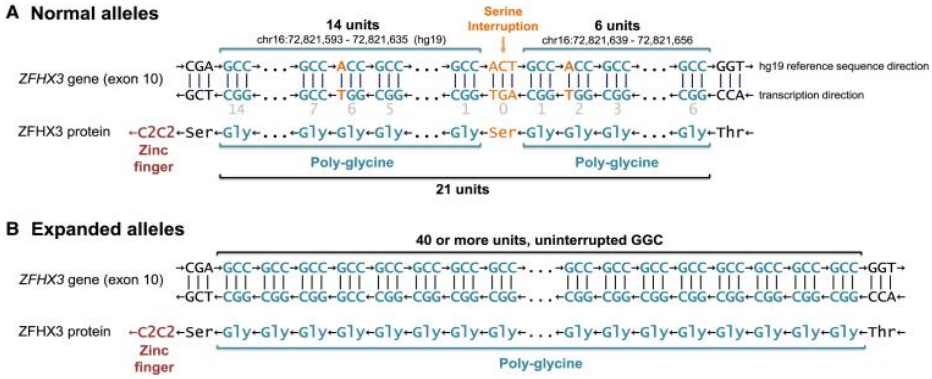

Figure 19: The *ZFHX3* locus contains the repeat expansion encoding a poly-glycine tract. In the hg19 reference genome, *ZFHX3* is encoded on the negative strand. For clarity, the figure shows the sequence from the negative strand perspective, which is the viewpoint used throughout unless otherwise stated. (A) In most non-expanded alleles, the repeat region comprises exactly 18 GGC units, two GGT units, and a single AGT interruption. These alleles are reported to have a total repeat length of 21. The positions of the GGT units—which also code for glycine but disrupt the repetitive GGC sequence at the DNA level—are indicated relative to the AGT interruption (shown as light gray numbers). Such a normal allele produces a protein containing 20 glycine residues interrupted by one serine at position 7. This poly-glycine segment is followed by a serine residue and a C2C2-type zinc finger motif, which typically functions as a sequence-specific DNA-binding element. (B) In contrast, none of the expanded alleles from affected individuals in the five families showed any interruptions; the repeat region consisted solely of uninterrupted GGC units. Reproduced from Wallenius *et al.*2024,[127] licensed under CC BY 4.0.

Nerve cells of the myenteric plexus in the esophagus of Individual IV:1 from family 1contained p62 positive inclusions. Furthermore, there were abundant p62 positive intranuclear and less abundant intracytoplasmic inclusions in brainstem neurons and singular inclusions of the same types in the cerebral cortex. Alpha-synuclein positivity was also found.

Paper IV

The study involved 285 Parkinson's disease (PD) patients, with an average age at onset of 56.1 years (±16.1; range 20–82), from the MultiPark database infrastructure. Among them, 175 (61%) were female, and 60.8% reported a positive family history of PD.

The quality of the VCF files was evaluated using specific metrics to detect potential biases or artifacts. The transition-to-transversion (Ti/Tv) ratio was about 3.0, aligning with typical values for whole-exome sequencing (WES) data. On average, each sample contained approximately 3,000 indels and 35,000 SNVs.

Known pathogenic variants in established PD genes

Previously reported pathogenic variants were identified in 4 out of the 285 probands. The *LRRK2* p.(Gly2019Ser) variant was found in one proband, and family analysis confirmed that the affected mother also carried the same variant. In another family, the *VPS35* p.(Asp620Asn) variant, known to cause autosomal dominant PD, was detected (Figure 20). One patient found to carry the *SNCA* p.(Ala53Thr) variant and an *SNCA* duplication, arr[GRCh37]4q22.1(90320154_91139340)x3, was identified in another patient. This individual belonged to the previously described family from the Lister peninsula in southern Sweden.[128]

CHCHD2 p.(Phe84LeufsTer6)

A frameshift variant, p.(Phe84LeufsTer6), was identified in the *CHCHD2* gene in two PD probands. This variant had a CADD phred-score[56] of 31 and an allele frequency of 0.0001239% in the gnomAD database.[85] This variant had not been previously linked to PD. Statistical analysis revealed a strong association between the variant and PD, with carriers having 26.2 times higher odds of developing the disease; this finding was statistically significant. IBD analysis identified a shared genomic segment of 4.7 Mbp (3.452 centiMorgans) encompassing the entire *CHCHD2* gene, representing the longest shared segment between the two patients studied. The p.(Phe84LeufsTer6) carrier, patient P064, was a man who began experiencing a resting tremor in his right hand at age 67 (Figure 20). He reported no family history of Parkinson's disease. The other p.(Phe84LeufsTer6) carrier, patient P088, was a woman who developed a resting tremor around the age of 55 (Figure 20). She reported a family history of PD.

Diagnostic yield for monogenic PD

In the entire PD series, 6 (2.1%) of 285 patients had a monogenic form. A positive finding was obtained for 2 (8%) of 25 PD patients with positive family history and young onset, 2 (2.2%) of 91 early-onset PD patients, 5 (2.9%) of 173 with positive family history and 1 (0.9%) of 112 patients with negative family history.

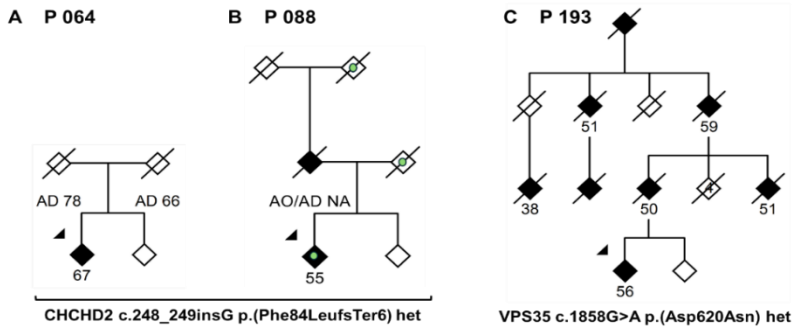

Figure 20: Pedigrees of the families with the *CHCHD2* p.(Phe84LeufsTer6) (A,B) and with the *VPS35* p.(Asp620Asn) (C) variants. Black diamond: affected with PD; white diamond: unaffected; green circle: dementia; arrowhead: proband. The age at disease onset, if known, is shown for the affected individuals. AO = age at onset, AD = age at death, NA= not available. Reproduced from Kafantari *et al.*2025,[129] licensed under CC BY 4.0.

GBA1 risk variants

We identified variants in *GBA1* known to be associated with increased PD risk in 43 out of 285 probands (15.1%). These variants were classified according to the GBA1-browser (Figure 21). Based on studies that have reported a lack of association of the p.(Thr408Met) [T369M] with PD in the Swedish population,[130,131] 10.9% of patients carried any of the other *GBA1* variants. Published odds ratios (ORs) for p.(Leu483Pro) [L444P] range from 2.62 to 30.42 indicating a strong association with increased PD risk.[132] The heterozygous p.(Ser146Leu) [S107L] variant was first linked to early-onset PD in 2019 when it was identified in two Swedish half-siblings.[133] All identified variants were confirmed to originate from *GBA1* and were not mis-mapped from the homologous pseudogene *GBAP1*.

Heterozygous *ATP7B* variants

We identified heterozygous variants in *ATP7B.*The p.(His1069Gln) and p.(Gln289Ter) in one proband each and p.(Thr991Met) in two probands.

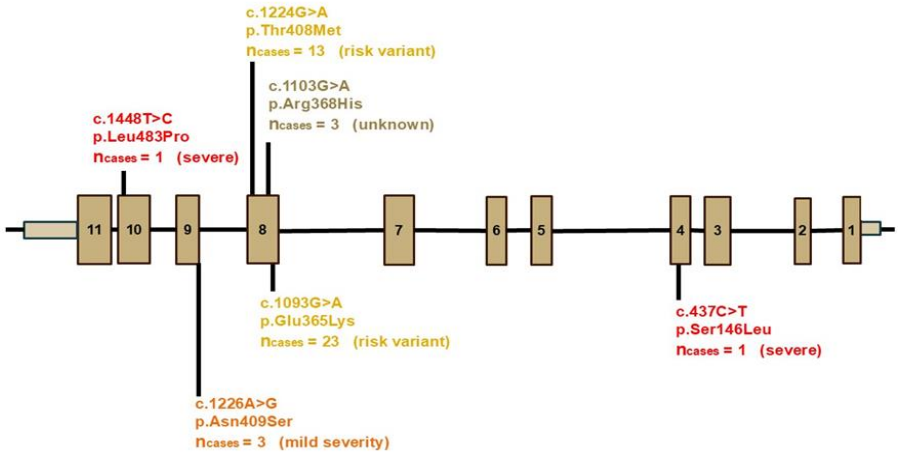

Figure 21: *GBA1* variants identified in 43 probands with PD from southern Sweden. In parenthesis, the severity of the variants based on the classification proposed by Parlar *et al.* 2023[132] (pdgenetics.shinyapps.io/gba1browser/) is shown. Image adapted with modifications from Gabbert *et al.* 2023,[134] licensed under CC BY 4.0. Reproduced from Kafantari *et al.*2025,[129] licensed under CC BY 4.0.

Variants in *ARSA*

Four heterozygous *ARSA* variants were identified in the PD series, with classifications spanning from VUS to pathogenic based on ACMG criteria. The p.(Pro351Ala), p.(Arg301Gln) and p.(Gly129Ala) variants were found in one proband each and the splicing variant *ARSA* c.465+1G>A in two probands. The p.(Gly129Ala) variant showed a strong association with PD, whereas the other three variants did not demonstrate a significant association and were relatively common in the control group.

Additional variants

We identified several variants of uncertain significance and variants with unknown PD disease associations. These included monoallelic changes in genes typically linked to autosomal recessive PD (e.g., *VPS13C, PRKN*), as well as variants in genes where only a single other variant has previously been associated with PD (e.g., *RAB32* p.Ser71Arg).

Four different heterozygous variants were identified in *PRKN*, in one proband each. The p.(Arg275Trp), which was classified as pathogenic in ClinVar, the p.(Arg275Gln) and p.(Arg256Cys), of uncertain significance, and the p.(Gly429Glu) which has not been reported in ClinVar database.

Discussion

This thesis investigates the genetic causes underlying several neurological movement disorders. A significant proportion of adult patients with severe movement disorders suffer from rare conditions that cannot be diagnosed through clinical evaluation alone. Many of these disorders have a monogenic origin and/or display clear hereditary patterns. Recent advances in genetic analysis, particularly NGS, have made genetic testing more accessible and feasible within clinical settings.

Through genetic analysis of patients and families with ataxia, dystonia, and PD, we identified numerous pathogenic variants that led to definitive genetic diagnoses. The implementation of NGS-based testing into clinical practice has brought substantial benefits to patients and their families by ending prolonged "diagnostic odysseys" and sparing them from more invasive procedures.[13]

Although causal treatment options have become available for a rapidly increasing number of rare monogenic movement disorders,[13,135] they are still lacking for many others. However, in other cases, an accurate genetic diagnosis enables personalized clinical monitoring and management of associated complications such as diabetes or cardiomyopathy and a more specific prognosis.[13] Furthermore, an increasing number of preclinical and clinical trials are investigating targeted therapies for specific genetic movement disorders.[136] Genetic counseling following diagnosis can offer risk assessment and predictive testing for family members. For those planning families, options such as prenatal and preimplantation genetic testing during in vitro fertilization (IVF), can be discussed with genetic counselors.

In addition, this work established repeat expansions in *ZFHX3* as the genetic cause of SCA4 and provided evidence supporting *TOR1AIP2* as a novel candidate gene for dystonia. Our findings also further confirmed *CHCHD2* as a PD-gene. Collectively, these discoveries expand the known genetic landscape of movement disorders by adding novel genes and variants to the list of disease-causing alterations. My work has also generated new knowledge relevant for clinicians who meet patients and who order genetic tests. These insights not only enhance future genetic diagnostic capabilities but also provide potential targets for the development of new therapies.

Paper I

Our findings support *TOR1AIP2* as a new candidate gene involved in a combined movement disorder characterized by dystonia, asymmetric arm movements with dystonic and/or hemichoreatic components, and fluttering movements of the perioral and lower facial muscles. *TOR1AIP2* encodes LULL1, a luminal domain-like LAP1 protein that interacts with TorsinA, the protein implicated in early-onset primary dystonia (DYT1). LAP1 and LULL1, encoded by *TOR1AIP1* and *TOR1AIP2* respectively, activate TorsinA by providing an arginine finger necessary for ATP hydrolysis, which TorsinA itself lacks.[116,137] They form complexes which are critical for maintaining nuclear envelope structure.

DYT1 results from defective TorsinA activation due to mutations causing protein misfolding or reduced binding to LAP1/LULL1. The identified R412G variant affects hydrogen bonding between LULL1 and TorsinA, weakening their interaction similarly to the recurrent TorsinA$^{\Delta E303}$ deletion in DYT1. Functional assays confirmed that the LULL1 variant in Family 1 significantly disrupts this binding. A second Slovakian family with a different rare *TOR1AIP2* variant (p.Q338H) showed milder symptoms but shared clinical features, including dystonia, involuntary facial movements, and proximal arm involvement.

While biallelic variants in *TOR1AIP1* cause severe early-onset dystonia,[138-141] *TOR1AIP2* has not been previously linked to human disease. However, variants in other nuclear envelope or pore proteins, such as *NUP62* and *NUP54*, have been associated with dystonia.[142,143] Elements of the clinical phenotype of our families resembles DYT1, suggesting overlapping pathogenic mechanisms involving *TOR1A* and *TOR1AIP2*-related disease.[144]

Applying the 2015 ACMG guidelines,[60] the *TOR1AIP2* c.1234A>G p.(Arg412Gly) variant meets criteria supporting pathogenicity, including rarity (PM2 supporting), co-segregation which was perfect and reached an N = 9/64 (supporting pathogenic segregation data) and strong functional evidence fulfilling the PS3 criterion. We formally classify it as variant of uncertain significance and probably disease-causing in Family 1. The c.1014G>C p.(Gln338His) variant in Family 2 also meets supporting evidence criteria based on rarity (PM2 supporting evidence), co-segregation which was perfect with N = 1/8 (PP1 supporting) and clinical similarity with Family 1, suggesting it is a probable disease cause.

Our data provides 'limited' evidence for associating *TOR1AIP2* with this neurological disorder.[145] Identification of additional patients with similar symptoms and deleterious *TOR1AIP2* variants will be important to confirm this link, therefore we have placed our findings in GeneMatcher. Limitations of the work in Paper I, include the use of exome rather than whole-genome sequencing, which may have resulted in missing causative variants outside captured regions, and an incomplete

understanding of the disease mechanism for the p.Q338H variant, which requires further study.

Paper II

Our results show the complexity of the interpretation of genetic findings and the high variability between the phenotypes of different forms of ataxia even in a small geographic area as Southern Sweden. The most common form of autosomal dominant ataxia in our patient series was SCA3 followed by CANVAS and SCA2. We also detected conditions that are not typically categorized as genetic ataxias, including two cases of hereditary spastic paraplegia, one case of Huntington's disease and one case of Brown–Vialetto–Van Laere syndrome-2. While it is possible that clinicians recorded the "wrong" ICD-10 code for these patients, all of them exhibited ataxia upon examination. Additionally, hereditary ataxia and hereditary spastic paraplegias not only have overlapping phenotypes and shared genetic causes but also involve similar disease mechanisms and cellular pathways.[146] This highlights the challenge, and at times the impossibility, of grouping neurological disorders based on combined symptoms, as the most prominent clinical feature can differ between individuals with the same genetic cause or shift over time.

WES initially was chosen for reasons of cost and availability at the time. Motivated by the finding that intronic *RFC1* variants were a relatively frequent cause of recessive ataxia, we transitioned to WGS. Based on the short-read WGS data that was used, our approach could not reliably determine the length of the *RFC1* pentanucleotide repeats. A key limitation of our study is the use of short-read WES. It is possible that additional patients would have received a genetic diagnosis with short-read WGS and/or long-read sequencing, but we used the methods that are currently available in most clinical diagnostic settings.

Most patients from our study continue to be monitored at our clinic, where we regularly reassess diagnostic approaches. This involves testing for newly identified genetic causes of ataxia that so far may not be detected by WES or WGS, reanalyzing data using updated gene panels, and potentially incorporating new techniques like long-read sequencing in the near future. We also aim to expand family pedigrees and test additional relevant relatives whenever possible. For patients without an identified genetic cause, non-genetic factors that have not yet been evaluated or discovered cannot be completely ruled out.

One of the limitations of our study is the relatively small sample size, which is a common issue in rare disease research. Additionally, since our findings are based on a selected patient cohort, they may not accurately reflect the true prevalence of genetic ataxia subtypes in Sweden. Neurologists were not blinded to the genetic

results and were also informed about variants of uncertain significance, which could potentially increase the risk of false-positive diagnoses. However, we believe this risk is low within the context of a research study and view the benefits of multidisciplinary discussions as crucial for achieving accurate diagnoses.

Paper III

In five Swedish families, we found that expansions of exonic GGC trinucleotide repeats in *ZFHX3*, which produce poly-glycine tracts, are associated with autosomal-dominant cerebellar ataxia, accompanied by sensory and autonomic neuropathy. The affected individuals presented with variable onset of gait and balance difficulties. The presence of anticipation in these families pointed to the repeat expansions as the underlying genetic cause.

We observed a clear separation between normal alleles, which ranged from 14 to 26 repeats with 21 being the most frequent, and pathogenic expansions, which extended from 42 up to 74 repeats. Although short-read sequencing becomes less reliable when expansions approach the maximum read length of about 150 base pairs (roughly 50 repeats), we were still able to confidently identify non-expanded alleles. Notably, all 2,300 non-expanded alleles examined contained interruptions within the repeat sequence, known to contribute to repeat length stability.[8] We propose that the initial loss of these interruptions may have triggered the progressive expansion of the repeats over generations.[147]

The family trees of all five families originated from the limited geographic area of Skåne could not be linked on the basis of existing genealogical information. A common founder event many generations ago is indicated by the fact that the affected members of the five families, shared a genomic segment of at least 111 kbp in length within which the *ZFHX3* repeat expansion was located. The actual founder event might be spontaneous mutations that eliminated or reduced the number of repeat interruptions to an allele. This allele without interruptions then was prone to expand in length over successive generations. The oldest members of families 4 and 5 did not have any neurological symptoms, which is compatible with this hypothesis. A relatively high prevalence of this disorder, at least among individuals with ataxia from southern Sweden, is suggested as we identified five families with *ZFHX3* repeat expansions. At the same time, the causative expansion was identified by other research groups in three separate Swedish families,[148] in the original Utah family that was described 25 years ago and in an unrelated Iowa family of Swedish descent,[149] as well as in additional related individuals from Utah and in families from northern Germany.[150] A common ancestral founder haplotype was found. [150,151]

Limitations of the study constitute that we did not confirm that additional unaffected members have normal repeat lengths, for example by an orthogonal method such as

repeat-primed PCR and that we have not been able to determine the exact lengths of the repeat expansions in all affected family members. The presumed inverse association of repeat length with age at symptom onset or severity of clinical symptoms, as we have observed in long-read sequencing data from only two individuals, was validated from other groups. [150,152]

The functional effect of the expanded *ZFHX3* trinucleotide repeat on transcription or translation, or the effect of an expanded poly-glycine repeat on protein function was not investigated experimentally. *ZFHX3* is involved in neuronal differentiation[153-156] and is thought to play a role in how cerebellar neurons respond to oxidative stress,[157] potentially linking variation in this gene to a cerebellar disorder.

Mutant ZFHX3 with longer-than-normal poly-glycine stretches and lacking a serine interruption, may be present in the eosinophilic inclusions observed in the affected individual we describe here and that such mutant ZFHX3 proteins could have toxic effects. RNA toxicity, where transcripts with expanded uninterrupted repeats sequester RNA-binding proteins, leading to the formation of RNA foci is another possible pathomechanism.[158]

Paper IV

We conducted a comprehensive WES analysis of both SNVs and CNVs across 44 genes associated with PD in a cohort of 285 probands from southern Sweden. This cohort also included groups in which the genetic contribution to disease etiology is presumed to be higher. Despite this enrichment, we identified disease-causing variants associated with monogenic PD in only 6 out of 285 probands (2.1%). Previously, our group reported a low frequency (0.59%) of known monogenic variants in a large Swedish PD cohort (n = 2,206), screened specifically for *SNCA* and *LRRK2* variants and/or duplications; a subset of 154 probands in the current study was included in that earlier analysis.[37] A recent large-scale screening study of 8,301 North American patients identified 272 individuals (3.3%) with monogenic PD genotypes, considering only disease-causing variants,[159] and a baseline genetic analysis conducted in United Kingdom revealed known pathogenic variants in 69 of 718 screened PD families (9.6%).[160] These frequencies are higher than those reported for monogenic causes of PD in Sweden.

We identified variants in *SNCA*, *LRRK2*, and *VPS35*, genes that are recognized as well-established contributors to monogenic forms of PD.

The p.(Ala53Thr) substitution in *SNCA* was the first genetic alteration linked to PD.[161]

Copy number gains, including duplications and triplications of *SNCA*, have been documented in PD patients from diverse populations.[162,163] In 2007, a large Swedish family (Lister family) was identified to carry a duplication in the *SNCA*.[128] The identified duplication of our study encompassed four genes: *SNCA*, *SNCA-AS1*, the adjacent *MMRN1*, and approximately 91 kb of *CCSER1*. Comparison of high-resolution SNV array data from this patient with that of a confirmed carrier of the Lister showed that they are members of the same family.

We found one proband (0.35%) with *LRRK2* p.(Gly2019Ser). We have previously observed that the *LRRK2* p.(Gly2019Ser) variant is significantly less frequent in Sweden[37,164] and other Northern European populations compared to regions studied elsewhere.

To our knowledge, this is the first study to identify the *VPS35* p.(Asp620Asn) variant in a PD patient of Swedish descent.

We detected the *CHCHD2* frameshift mutation p.(Phe84LeufsTer6) in two unrelated probands from a confined area within southern Sweden covered by our center. Genetic analysis revealed that these individuals did not share extended genomic segments beyond 3.4 cM, suggesting that although they have somewhat increased genetic similarity compared to the broader population, they are unlikely to be closely related since close relatives typically share much larger genomic regions.[165] Previous studies have shown that truncating mutations like frameshifts and nonsense variants in *CHCHD2* lead to loss of function, probably through nonsense-mediated mRNA decay, and are linked to mitochondrial impairment.[166] It is therefore probable that the p.(Phe84LeufsTer6) variant found in both probands causes a similar loss-of-function effect.

Both *GBA1* and *ARSA* encode important enzymes in the lysosomal ceramide pathway and are involved in lysosomal storage disorders which frequently present with neurological complications, such as parkinsonism. Biallelic pathogenic variants in *ARSA* cause metachromatic leukodystrophy.[167] Heterozygous variants in *ARSA* were first linked to Parkinson's disease in 2019;[168] however, the gene's significance in PD remains uncertain due to a lack of consistent replication in independent PD cohorts.[169]

In this study, *GBA1* variants were identified in 15.1% of patients. The frequencies of the p.(Thr408Met) and p.(Glu365Lys) variants in our cohort were 4.5% and 8%, respectively. When combining data from previous studies on White/Caucasian populations, the frequencies of p.(Thr408Met) and p.(Glu365Lys) were 2.03% and 4.07%, respectively.[134] Notably, two previous studies have found that p.(Thr408Met) is not associated with PD in the Swedish population.[130,131] Therefore, after excluding this variant, only 10.9% of patients carried one of the other *GBA1* variants, mostly driven by the p.(Glu365Lys) which was found in 23 probands and which is known to only confer a very small risk increase. The *GBA1* p.(Asn409Ser) variant was found in 1% of our probands, a frequency somewhat higher than that

reported in a previous Swedish study (0.62%).[170] A combined frequency of 1.59% in studies of White/Caucasian populations has been reported.[134] The *GBA1* p.(Leu483Pro) variant was detected in 0.35% of our case series, which is notably lower than the 2.15% reported in another Swedish cohort.[170] A combined frequency of 1.53% reported in studies of White/Caucasian populations.[134] The p.(Arg368His) variant, of uncertain significance, was identified in three unrelated probands (1%) and has only been rarely reported in previous studies. The *GBA1* p.(Ser146Leu) variant was identified in one proband (0.35%) who developed symptoms at age 66 and reported having one affected parent and another affected relative. This variant has been previously reported in a few Gaucher disease patients as part of a compound heterozygous combination with other severe *GBA1* mutations. Additionally, it was found in two Swedish half-siblings with Parkinson's disease who carried the variant in a heterozygous state and exhibited a severe PD phenotype.[133]

It has been proposed that heterozygous variants in *ATP7B* may manifest phenotypically as PD; [171] however, this association requires further investigation. Heterozygous variants in the *PRKN* gene are also considered potential genetic risk factors[172] and have been found to occur more frequently in PD cases than in controls.[173] However, uncertainties persist regarding the exact relationship between heterozygous *PRKN* variants and PD.

In our study, the largest WES based investigation of PD conducted in Sweden, although the case series was enriched for patients with a highly likely monogenic disease cause, clearly disease-causing variants were identified in only 2.1% of cases (6 individuals) within the *SNCA*, *LRRK2*, *VPS35*, and *CHCHD2* genes. Our results support the pathogenic role of truncating variants in *CHCHD2*, thereby strengthening the evidence linking this gene to autosomal dominant PD. The low detection rate may be due in part to the limitations of whole-exome sequencing, which targets only coding regions and does not capture non-coding, deep intronic, or regulatory variants. Additionally, it underscores the complex etiology of PD, which is often not driven solely by monogenic variants, even among patients with early onset or a positive family history.

Conclusions

This thesis highlights the power and necessity of NGS and advanced bioinformatic analysis to determine the genetic causes and to identify new findings of dystonia, hereditary ataxia, and PD. Our approach is based on the integration of genetic analysis, careful clinical phenotyping and collaborations with other clinical, genetic and experimental experts. Across all patients and families, a critical enabler of these findings was the use of advanced bioinformatic pipelines. These included tools for variant calling, annotation and pathogenicity prediction, as well as specialized algorithms for detecting STRs and CNVs. These computational strategies were essential in filtering, prioritizing and classifying candidate variants.

First, we identified *TOR1AIP2* as a novel candidate gene associated with dystonia with hemichorea and hemiballism. These findings expand the spectrum of nuclear envelope-associated movement disorders and illustrate the importance of considering gene-gene interaction networks in genetic diagnostics. Second, in a cohort of 76 families with progressive hereditary ataxia, a genetic diagnosis was reached in approximately 40% of cases through whole exome and genome sequencing. This underscores the clinical utility of genetic testing in ataxia, especially when combined with careful phenotypic re-evaluation and re-appreciation of VUS. These results also highlight the diagnostic benefit of sustained interdisciplinary collaboration. A major genetic breakthrough was the discovery of exonic GGC repeat expansions in *ZFHX3* in five affected Swedish families. The findings establish *ZFHX3* as the causative gene for SCA4 and demonstrate the importance of tools capable of detecting STRs. In PD, sequencing of 285 patients including also probands with early-onset and/or familial disease revealed pathogenic variants in known PD-associated genes. However, clearly disease-causing monogenic forms were found in only 2.1% of cases. This underscores the genetic complexity of PD, which often involves factors beyond monogenic variants, forms that are especially uncommon in the Swedish population.

In conclusion, this work emphasizes the indispensable role of integrative genomic and clinical analysis in the diagnosis of movement disorders. It demonstrates that comprehensive genetic testing, combined with expert interpretation and detailed phenotyping, significantly enhances diagnostic yield and contributes to the discovery of novel disease causes.

Future perspectives

Moving forward, several important avenues should be prioritized to enhance diagnosis, understanding, and treatment of neurological movement disorders:

1. Expansion of genetic testing panels

The identification of novel candidate genes, such as *TOR1AIP2* for dystonia-hemichorea/hemiballism, and the discovery of pathogenic repeat expansions in *ZFHX3* causing SCA4, emphasize the need to continuously update and expand genetic testing panels. Integrating such newly implicated genes and genetic variants into routine diagnostics will improve the detection rate of hereditary neurological movement disorders.

2. Functional validation of genetic variants

Accurate interpretation of genetic variants remains a major challenge, especially for variants of uncertain significance. Future research must focus on robust functional validation using molecular and cellular assays and development of animal models to elucidate mechanisms by which the *TOR1AIP2* identified variants, the *CHCHD2* p.(Phe84LeufsTer6) variant and the *ZFHX3* repeat expansions contribute to disease pathogenesis.

3. Investigation of disease mechanisms to enable targeted therapies

Understanding how genetic variants and repeat expansions alter neuronal function and contribute to neurodegeneration will pave the way for developing targeted therapies. For example, elucidating how disrupted nuclear envelope interactions lead to dystonia, or how *ZFHX3* expansions cause plausibly intranuclear inclusions and autonomic dysfunction in SCA4, will guide the development of innovative therapies.

4. Exploring Digenic Inheritance in Parkinson's Disease

A promising direction for future research would be to investigate samples from PD patients to explore genetic causes beyond monogenic factors. Specifically, examining the potential role of digenic inheritance could offer new insights into the complex genetic architecture of PD. By considering the interactions between variants in multiple genes, rather than focusing solely on single mutations, could uncover novel genetic contributions to disease risk and improve our understanding of PD's etiology.

Acknowledgments

The journey of pursuing a PhD can be long and sometimes difficult. Along the way, I have learned not only about my research topic but also about the importance of collaboration. The support and encouragement from supervisors, colleagues, friends, and family have been essential pillars throughout this demanding yet rewarding experience.

First of all, I would like to express my sincere gratitude to my main supervisor, Prof. Andreas Puschmann, for his invaluable guidance, encouragement, and support throughout the course of this research. I am truly grateful for his patience, constant motivation and dedication, which greatly contributed to the completion of this work.

I would also like to express my appreciation to my co-supervisors, Dr. Kristina Karrman and Assoc. Prof. Carola Hedberg Oldfors, for their valuable advice, constructive feedback, and support throughout my research.

I would like to sincerely thank my colleagues from the Clinical Neurogenetics Group of Lund University, Joel Wallenius, MSc, for his valuable help regarding my bioinformatic questions and revising my papers and my thesis, Christin Karremo, BSc, for taking good care of the patients and the samples, Assoc. Prof. Andreea Ilinca for giving me the opportunity to analyze data from patients with stroke, Dr. Sorina Gorcenco for our perfect collaboration on preparing our paper, Dr. Ashraf Yahia for providing help with my papers and thesis and Dr. Sigurd Dobloug for his support.

I am also truly grateful to my previous teachers in bioinformatics Assoc. Prof. Björn Canbäck, and Assoc. Prof. Dag Ahren, for their support during my bioinformatics masters program in Lund. Without their guidance, this journey would never have started.

Special thanks to all my good colleagues and friends from the Wigerthuset, for being kind, supportive and always willing to help Trinette van Vliet, PhD, Monica Scharfenort, PhD, Inês Santos, Elise Dirckx Lundberg, Camilla Månsson, Linnea Edman, Elisabeth Jirström Hiersemann, PhD, Arja Andersson, Anna Wedding, Gunilla Thorneman, Dr. Sara Hall, Dr. Sotirios Grigoriou, Ulrika Mundt-Petersen, PhD, Prof. Per Odin (Head of our Department).

I would also like to thank all co-authors and editors of the papers that are included in this thesis and beyond for the excellent collaboration and the constructive discussions that we had.

Many thanks to my beloved friends Alex, Elena, Dimitris, George, Aris, Johan for the support and encouragement that they provided during this work.

Finally, I would like to thank my mother, my sister, my nephew, my uncles, aunts and cousins for always believing in me.

This thesis is dedicated to my father…

Funding

This study was funded by research grants from Region Skåne, Skåne University Hospital, the Swedish government (via ALF, avtal för läkarutbildning och forskning), Lund University, The Swedish Parkinson Foundation (Parkinsonfonden), Hans-Gabriel och Alice Trolle-Wachtmeisters Stiftelse för Medicinsk Forskning, Hardebos Foundation, Bundy Academy, SCA network, and infrastructure support from MultiPark—a strategic research environment at Lund University and The Swedish Parkinson Academy, all in Sweden.

References

1 Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019). https://doi.org/10.1093/nar/gky955

2 Forero, D. A. & Patrinos, G. P. *Genome plasticity in health and disease*. (Academic Press is an imprint of Elsevier, 2020).

3 Alberts, B. *Molecular biology of the cell*. Sixth edition. edn, (Garland Science, Taylor and Francis Group, 2015).

4 Alberts, B. *Molecular biology of the cell*. 4th edn, (Garland Science, 2002).

5 Naidoo, N., Pawitan, Y., Soong, R., Cooper, D. N. & Ku, C. S. Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Hum Genomics* **5**, 577-622 (2011). https://doi.org/10.1186/1479-7364-5-6-577

6 Marian, A. J. Clinical Interpretation and Management of Genetic Variants. *JACC Basic Transl Sci* **5**, 1029-1042 (2020). https://doi.org/10.1016/j.jacbts.2020.05.013

7 Kulkarni, S. *Clinical genomics*. (Elsevier, 2014).

8 Depienne, C. & Mandel, J. L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am J Hum Genet* **108**, 764-785 (2021). https://doi.org/10.1016/j.ajhg.2021.03.011

9 Habbane, M. *et al.* Human Mitochondrial DNA: Particularities and Diseases. *Biomedicines* **9** (2021). https://doi.org/10.3390/biomedicines9101364

10 Fahn, S. Classification of movement disorders. *Mov Disord* **26**, 947-957 (2011). https://doi.org/10.1002/mds.23759

11 Shipton, E. A. Movement disorders and neuromodulation. *Neurol Res Int* **2012**, 309431 (2012). https://doi.org/10.1155/2012/309431

12 Ashizawa, T. & Xia, G. Ataxia. *Continuum (Minneap Minn)* **22**, 1208-1226 (2016). https://doi.org/10.1212/CON.0000000000000362

13 Yeow, D., Rudaks, L. I., Siow, S. F., Davis, R. L. & Kumar, K. R. Genetic Testing of Movements Disorders: A Review of Clinical Utility. *Tremor Other Hyperkinet Mov (N Y)* **14**, 2 (2024). https://doi.org/10.5334/tohm.835

14 Balint, B. *et al.* Dystonia. *Nat Rev Dis Primers* **4**, 25 (2018). https://doi.org/10.1038/s41572-018-0023-6

15 Albanese, A. *et al.* Phenomenology and classification of dystonia: a consensus update. *Mov Disord* **28**, 863-873 (2013). https://doi.org/10.1002/mds.25475

16 Grutz, K. & Klein, C. Dystonia updates: definition, nomenclature, clinical classification, and etiology. *J Neural Transm (Vienna)* **128**, 395-404 (2021). https://doi.org/10.1007/s00702-021-02314-2

17 Albanese, A. *et al.* Definition and Classification of Dystonia. *Mov Disord* **40**, 1248-1259 (2025). https://doi.org/10.1002/mds.30220

18 Hellberg, C., Alinder, E., Jaraj, D. & Puschmann, A. Nationwide prevalence of primary dystonia, progressive ataxia and hereditary spastic paraplegia. *Parkinsonism Relat Disord* **69**, 79-84 (2019). https://doi.org/10.1016/j.parkreldis.2019.10.028

19 Pana, A. & Saggu, B. M. in *StatPearls* (2025).

20 Koptielow, J. *et al.* Genetic Update and Treatment for Dystonia. *Int J Mol Sci* **25** (2024). https://doi.org/10.3390/ijms25073571

21 Klein, C., Lohmann, K., Marras, C. & Munchau, A. in *GeneReviews((R))* (eds M. P. Adam *et al.*) (1993).

22 Hafiz, S. & De Jesus, O. in *StatPearls* (2025).

23 Jayadev, S. & Bird, T. D. Hereditary ataxias: overview. *Genet Med* **15**, 673-683 (2013). https://doi.org/10.1038/gim.2013.28

24 Pilotto, F., Del Bondio, A. & Puccio, H. Hereditary Ataxias: From Bench to Clinic, Where Do We Stand? *Cells* **13** (2024). https://doi.org/10.3390/cells13040319

25 Salman, M. S. Epidemiology of Cerebellar Diseases and Therapeutic Approaches. *Cerebellum* **17**, 4-11 (2018). https://doi.org/10.1007/s12311-017-0885-2

26 Williams, C. T. & De Jesus, O. in *StatPearls* (2025).

27 Sandford, E. & Burmeister, M. Genes and genetic testing in hereditary ataxias. *Genes (Basel)* **5**, 586-603 (2014). https://doi.org/10.3390/genes5030586

28 Perlman, S. in *GeneReviews((R))* (eds M. P. Adam *et al.*) (1993).

29 Fan, H. & Chu, J. Y. A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics* **5**, 7-14 (2007). https://doi.org/10.1016/S1672-0229(07)60009-6

30 Hayes, M. T. Parkinson's Disease and Parkinsonism. *Am J Med* **132**, 802-807 (2019). https://doi.org/10.1016/j.amjmed.2019.03.001

31 Kalia, L. V. & Lang, A. E. Parkinson's disease. *Lancet* **386**, 896-912 (2015). https://doi.org/10.1016/S0140-6736(14)61393-3

32 Kouli, A., Torsney, K. M. & Kuan, W. L. in *Parkinson's Disease: Pathogenesis and Clinical Aspects* (eds T. B. Stoker & J. C. Greenland) (2018).

33 Dorsey, E. R., Sherer, T., Okun, M. S. & Bloem, B. R. The Emerging Evidence of the Parkinson Pandemic. *J Parkinsons Dis* **8**, S3-S8 (2018). https://doi.org/10.3233/JPD-181474

34 von Campenhausen, S. *et al.* Prevalence and incidence of Parkinson's disease in Europe. *Eur Neuropsychopharmacol* **15**, 473-490 (2005). https://doi.org/10.1016/j.euroneuro.2005.04.007

35 Jia, F., Fellner, A. & Kumar, K. R. Monogenic Parkinson's Disease: Genotype, Phenotype, Pathophysiology, and Genetic Testing. *Genes (Basel)* **13** (2022). https://doi.org/10.3390/genes13030471

36 Smith, L. & Schapira, A. H. V. GBA Variants and Parkinson Disease: Mechanisms and Treatments. *Cells* **11** (2022). https://doi.org/10.3390/cells11081261

37 Puschmann, A. *et al.* Low prevalence of known pathogenic mutations in dominant PD genes: A Swedish multicenter study. *Parkinsonism Relat Disord* **66**, 158-165 (2019). https://doi.org/10.1016/j.parkreldis.2019.07.032

38 Hutchinson, M., Spanaki, C., Lebedev, S. & Plaitakis, A. Genetic basis of common diseases: the general theory of Mendelian recessive genetics. *Med Hypotheses* **65**, 282-286 (2005). https://doi.org/10.1016/j.mehy.2005.02.034

39 Rizzo, G. *et al.* Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. *Neurology* **86**, 566-576 (2016). https://doi.org/10.1212/WNL.0000000000002350

40 Armstrong, M. J. & Okun, M. S. Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA* **323**, 548-560 (2020). https://doi.org/10.1001/jama.2019.22360

41 Pal, G. *et al.* Genetic Testing in Parkinson's Disease. *Mov Disord* **38**, 1384-1396 (2023). https://doi.org/10.1002/mds.29500

42 Satam, H. *et al.* Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology (Basel)* **12** (2023). https://doi.org/10.3390/biology12070997

43 Gorcenco, S. *et al.* New generation genetic testing entering the clinic. *Parkinsonism Relat Disord* **73**, 72-84 (2020). https://doi.org/10.1016/j.parkreldis.2020.02.015

44 Zech, M. *et al.* Monogenic variants in dystonia: an exome-wide sequencing study. *Lancet Neurol* **19**, 908-918 (2020). https://doi.org/10.1016/S1474-4422(20)30312-4

45 Wictorin, K. *et al.* Autosomal dominant cerebellar ataxia with slow ocular saccades, neuropathy and orthostatism: a novel entity? *Parkinsonism Relat Disord* **20**, 748-754 (2014). https://doi.org/10.1016/j.parkreldis.2014.03.029

46 Ygland Rodstrom, E. & Puschmann, A. Clinical classification systems and long-term outcome in mid- and late-stage Parkinson's disease. *NPJ Parkinsons Dis* **7**, 66 (2021). https://doi.org/10.1038/s41531-021-00208-4

47 Brolin, K. *et al.* Insights on Genetic and Environmental Factors in Parkinson's Disease from a Regional Swedish Case-Control Cohort. *J Parkinsons Dis* **12**, 153-171 (2022). https://doi.org/10.3233/JPD-212818

48 Astrom, D. O. *et al.* High risk of developing dementia in Parkinson's disease: a Swedish registry-based study. *Sci Rep* **12**, 16759 (2022). https://doi.org/10.1038/s41598-022-21093-8

49 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009). https://doi.org/10.1093/bioinformatics/btp324

50 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11-11 10 33 (2013). https://doi.org/10.1002/0471250953.bi1110s43

51 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011). https://doi.org/10.1038/ng.806

52 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016). https://doi.org/10.1186/s13059-016-0974-4

53 Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894-899 (2011). https://doi.org/10.1002/humu.21517

54 Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-241 (2016). https://doi.org/10.1002/humu.22932

55 Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med* **12**, 103 (2020). https://doi.org/10.1186/s13073-020-00803-9

56 Schubach, M., Maass, T., Nazaretyan, L., Roner, S. & Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res* **52**, D1143-D1154 (2024). https://doi.org/10.1093/nar/gkad989

57 Plagnol, V. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**, 2747-2754 (2012). https://doi.org/10.1093/bioinformatics/bts526

58 Foreman, J. *et al.* DECIPHER: Improving Genetic Diagnosis Through Dynamic Integration of Genomic and Clinical Data. *Annu Rev Genomics Hum Genet* **24**, 151-176 (2023). https://doi.org/10.1146/annurev-genom-102822-100509

59 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011). https://doi.org/10.1038/nbt.1754

60 Richards, S. *et al.* 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. *Genet Med* **17**, 405-424 (2015). https://doi.org/10.1038/gim.2015.30

61 Dolzhenko, E. *et al.* ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754-4756 (2019). https://doi.org/10.1093/bioinformatics/btz431

62 Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res* **27**, 1895-1903 (2017). https://doi.org/10.1101/gr.225672.117

63 Dolzhenko, E. *et al.* REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Med* **14**, 84 (2022). https://doi.org/10.1186/s13073-022-01085-z

64 Dolzhenko, E. *et al.* ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol* **21**, 102 (2020). https://doi.org/10.1186/s13059-020-02017-z

65 Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* **108**, 1880-1890 (2021). https://doi.org/10.1016/j.ajhg.2021.08.005

66 Zhou, Y., Browning, S. R. & Browning, B. L. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *Am J Hum Genet* **106**, 426-437 (2020). https://doi.org/10.1016/j.ajhg.2020.02.010

67 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021). https://doi.org/10.1093/gigascience/giab008

68 Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e524 (2019). https://doi.org/10.1016/j.cell.2018.12.015

69 Gelfman, S. *et al.* Annotating pathogenic non-coding variants in genic regions. *Nat Commun* **8**, 236 (2017). https://doi.org/10.1038/s41467-017-00141-2

70 Bendl, J. *et al.* PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. *PLoS Comput Biol* **12**, e1004962 (2016). https://doi.org/10.1371/journal.pcbi.1004962

71 Ritchie, G. R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-296 (2014). https://doi.org/10.1038/nmeth.2832

72 Brandon, M. C. *et al.* MITOMAP: a human mitochondrial genome database--2004 update. *Nucleic Acids Res* **33**, D611-613 (2005). https://doi.org/10.1093/nar/gki079

73 Castellana, S. *et al.* MitImpact 3: modeling the residue interaction network of the Respiratory Chain subunits. *Nucleic Acids Res* **49**, D1282-D1288 (2021). https://doi.org/10.1093/nar/gkaa1032

74 Bianco, S. D. *et al.* APOGEE 2: multi-layer machine-learning model for the interpretable prediction of mitochondrial missense variants. *Nat Commun* **14**, 5058 (2023). https://doi.org/10.1038/s41467-023-40797-7

75 Molder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res* **10**, 33 (2021). https://doi.org/10.12688/f1000research.29032.2

76 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018). https://doi.org/10.1093/nar/gkx1153

77 Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* **15**, 57-61 (2000). https://doi.org/10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G

78 Kohler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Res* **45**, D865-D876 (2017). https://doi.org/10.1093/nar/gkw1039

79 Ameur, A. *et al.* SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur J Hum Genet* **25**, 1253-1260 (2017). https://doi.org/10.1038/ejhg.2017.130

80 Rappaport, N. *et al.* MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res* **45**, D877-D887 (2017). https://doi.org/10.1093/nar/gkw1012

81 in *GeneReviews((R))* (eds M. P. Adam *et al.*) (1993).

82 Kopanos, C. *et al.* VarSome: the human genomic variant search engine. *Bioinformatics* **35**, 1978-1980 (2019). https://doi.org/10.1093/bioinformatics/bty897

83 Jalali Sefid Dashti, M. & Gamieldien, J. A practical guide to filtering and prioritizing genetic variants. *Biotechniques* **62**, 18-30 (2017). https://doi.org/10.2144/000114492

84 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015). https://doi.org/10.1038/nature15393

85 Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92-100 (2024). https://doi.org/10.1038/s41586-023-06045-0

86 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010). https://doi.org/10.1038/nmeth0410-248

87 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081 (2009). https://doi.org/10.1038/nprot.2009.86

88 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-362 (2014). https://doi.org/10.1038/nmeth.2890

89 Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118 (2011). https://doi.org/10.1093/nar/gkr407

90 Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012). https://doi.org/10.1371/journal.pone.0046688

91 Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016). https://doi.org/10.1016/j.ajhg.2016.08.016

92 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010). https://doi.org/10.1371/journal.pcbi.1001025

93 Barbitoff, Y. A. *et al.* Bioinformatics of germline variant discovery for rare disease diagnostics: current approaches and remaining challenges. *Brief Bioinform* **25** (2024). https://doi.org/10.1093/bib/bbad508

94 Broeckx, B. J. G., Peelman, L., Saunders, J. H., Deforce, D. & Clement, L. Using variant databases for variant prioritization and to detect erroneous genotype-phenotype associations. *BMC Bioinformatics* **18**, 535 (2017). https://doi.org/10.1186/s12859-017-1951-y

95 Indelicato, E. *et al.* Genome Aggregation Database Version 4-Allele Frequency Changes and Impact on Variant Interpretation in Dystonia. *Mov Disord* **40**, 357-362 (2025). https://doi.org/10.1002/mds.30066

96 Verbeek, D. S. & Gasser, T. Unmet Needs in Dystonia: Genetics and Molecular Biology-How Many Dystonias? *Front Neurol* **7**, 241 (2016). https://doi.org/10.3389/fneur.2016.00241

97 Siokas, V. *et al.* Risk Factor Genes in Patients with Dystonia: A Comprehensive Review. *Tremor Other Hyperkinet Mov (N Y)* **8**, 559 (2018). https://doi.org/10.7916/D8H438GS

98 van Egmond, M. E. *et al.* A post hoc study on gene panel analysis for the diagnosis of dystonia. *Mov Disord* **32**, 569-575 (2017). https://doi.org/10.1002/mds.26937

99 Jinnah, H. A. & Sun, Y. V. Dystonia genes and their biological pathways. *Neurobiol Dis* **129**, 159-168 (2019). https://doi.org/10.1016/j.nbd.2019.05.014

100 Lohmann, K. & Klein, C. Update on the Genetics of Dystonia. *Curr Neurol Neurosci Rep* **17**, 26 (2017). https://doi.org/10.1007/s11910-017-0735-0

101 Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med* **12**, 91 (2020). https://doi.org/10.1186/s13073-020-00791-w

102 Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712 (2010). https://doi.org/10.1038/nature08516

103 Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* **12**, 703-714 (2011). https://doi.org/10.1038/nrg3054

104 Jarvik, G. P. & Browning, B. L. Consideration of Cosegregation in the Pathogenicity Classification of Genomic Variants. *Am J Hum Genet* **98**, 1077-1081 (2016). https://doi.org/10.1016/j.ajhg.2016.04.003

105 Schmitz-Hubsch, T. *et al.* Scale for the assessment and rating of ataxia: development of a new clinical scale. *Neurology* **66**, 1717-1720 (2006). https://doi.org/10.1212/01.wnl.0000219042.60538.92

106 Trouillas, P. *et al.* International Cooperative Ataxia Rating Scale for pharmacological assessment of the cerebellar syndrome. The Ataxia Neuropharmacology Committee of the World Federation of Neurology. *J Neurol Sci* **145**, 205-211 (1997). https://doi.org/10.1016/s0022-510x(96)00231-6

107 Pellerin, D. *et al.* Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia. *N Engl J Med* **388**, 128-141 (2023). https://doi.org/10.1056/NEJMoa2207406

108 Rafehi, H. *et al.* An intronic GAA repeat expansion in FGF14 causes the autosomal-dominant adult-onset ataxia SCA50/ATX-FGF14. *Am J Hum Genet* **110**, 105-119 (2023). https://doi.org/10.1016/j.ajhg.2022.11.015

109 Sone, J. *et al.* Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet* **51**, 1215-1221 (2019). https://doi.org/10.1038/s41588-019-0459-y

110 Liao, W. W. *et al.* A draft human pangenome reference. *Nature* **617**, 312-324 (2023). https://doi.org/10.1038/s41586-023-05896-x

111 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002). https://doi.org/10.1101/gr.229202

112 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016). https://doi.org/10.1038/ng.3656

113 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283 (2016). https://doi.org/10.1038/ng.3643

114 Manderstedt, E. *et al.* Classic Thrombophilias and Thrombotic Risk Among Middle-Aged and Older Adults: A Population-Based Cohort Study. *J Am Heart Assoc* **11**, e023018 (2022). https://doi.org/10.1161/JAHA.121.023018

115 Kafantari, E. *et al.* TOR1AIP2 as a candidate gene for dystonia-hemichorea/hemiballism. *Parkinsonism Relat Disord* **134**, 107781 (2025). https://doi.org/10.1016/j.parkreldis.2025.107781

116 Demircioglu, F. E., Sosa, B. A., Ingram, J., Ploegh, H. L. & Schwartz, T. U. Structures of TorsinA and its disease-mutant complexed with an activator reveal the molecular basis for primary dystonia. *Elife* **5** (2016). https://doi.org/10.7554/eLife.17983

117 Gorcenco, S., Karremo, C. & Puschmann, A. Patients' Perspective in Hereditary Ataxia. *Cerebellum* **23**, 82-91 (2024). https://doi.org/10.1007/s12311-022-01505-1

118 Gorcenco, S. *et al.* Clinical and genetic analyses of a Swedish patient series diagnosed with ataxia. *J Neurol* **271**, 526-542 (2024). https://doi.org/10.1007/s00415-023-11990-x

119 Chen, D. H. *et al.* Ataxia-Pancytopenia Syndrome Is Caused by Missense Mutations in SAMD9L. *Am J Hum Genet* **98**, 1146-1158 (2016). https://doi.org/10.1016/j.ajhg.2016.04.009

120 Cadieux-Dion, M. *et al.* Expanding the clinical phenotype associated with ELOVL4 mutation: study of a large French-Canadian family with autosomal dominant spinocerebellar ataxia and erythrokeratodermia. *JAMA Neurol* **71**, 470-475 (2014). https://doi.org/10.1001/jamaneurol.2013.6337

121 Ozaki, K. *et al.* A Novel Mutation in ELOVL4 Leading to Spinocerebellar Ataxia (SCA) With the Hot Cross Bun Sign but Lacking Erythrokeratodermia: A Broadened Spectrum of SCA34. *JAMA Neurol* **72**, 797-805 (2015). https://doi.org/10.1001/jamaneurol.2015.0610

122 Ravel, J. M. *et al.* Expanding the clinical spectrum of STIP1 homology and U-box containing protein 1-associated ataxia. *J Neurol* **268**, 1927-1937 (2021). https://doi.org/10.1007/s00415-020-10348-x

123 Mol, M. O. *et al.* Clinical and pathologic phenotype of a large family with heterozygous STUB1 mutation. *Neurol Genet* **6**, e417 (2020). https://doi.org/10.1212/NXG.0000000000000417

124 Chen, D. H. *et al.* Heterozygous STUB1 missense variants cause ataxia, cognitive decline, and STUB1 mislocalization. *Neurol Genet* **6**, 1-13 (2020). https://doi.org/10.1212/NXG.0000000000000397

125 Gan-Or, Z. *et al.* Mutations in CAPN1 Cause Autosomal-Recessive Hereditary Spastic Paraplegia. *Am J Hum Genet* **98**, 1271 (2016). https://doi.org/10.1016/j.ajhg.2016.05.009

126 Cortese, A. *et al.* Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet* **51**, 649-658 (2019). https://doi.org/10.1038/s41588-019-0372-4

127 Wallenius, J. *et al.* Exonic trinucleotide repeat expansions in ZFHX3 cause spinocerebellar ataxia type 4: A poly-glycine disease. *Am J Hum Genet* **111**, 82-95 (2024). https://doi.org/10.1016/j.ajhg.2023.11.008

128 Fuchs, J. *et al.* Phenotypic variation in a large Swedish pedigree due to SNCA duplication and triplication. *Neurology* **68**, 916-922 (2007). https://doi.org/10.1212/01.wnl.0000254458.17630.c5

129 Kafantari, E., Atterling Brolin, K., Wallenius, J., Swanberg, M. & Puschmann, A. WES-Based Screening of a Swedish Patient Series with Parkinson's Disease. *Genes (Basel)* **16** (2025). https://doi.org/10.3390/genes16121482

130 Ran, C. *et al.* Glucocerebrosidase variant T369M is not a risk factor for Parkinson's disease in Sweden. *Neurosci Lett* **784**, 136767 (2022). https://doi.org/10.1016/j.neulet.2022.136767

131 Brolin, K. A. *et al.* GBA1 T369M and Parkinson's disease - Further evidence of a lack of association in the Swedish population. *Parkinsonism Relat Disord* **130**, 107191 (2025). https://doi.org/10.1016/j.parkreldis.2024.107191

132 Parlar, S. C., Grenn, F. P., Kim, J. J., Baluwendraat, C. & Gan-Or, Z. Classification of GBA1 Variants in Parkinson's Disease: The GBA1-PD Browser. *Mov Disord* **38**, 489-495 (2023). https://doi.org/10.1002/mds.29314

133 Hertz, E. *et al.* First Clinicogenetic Description of Parkinson's Disease Related to GBA Mutation S107L. *Mov Disord Clin Pract* **6**, 254-258 (2019). https://doi.org/10.1002/mdc3.12743

134 Gabbert, C. *et al.* GBA1 in Parkinson's disease: variant detection and pathogenicity scoring matters. *BMC Genomics* **24**, 322 (2023). https://doi.org/10.1186/s12864-023-09417-y

135 Jinnah, H. A. *et al.* Treatable inherited rare movement disorders. *Mov Disord* **33**, 21-35 (2018). https://doi.org/10.1002/mds.27140

136 Sartorelli, J., Ng, J., Rahim, A. A., Waddington, S. N. & Kurian, M. A. Genetic therapies for movement disorders - current status. *J Neurol* **272**, 220 (2025). https://doi.org/10.1007/s00415-025-12940-5

137 Brown, R. S., Zhao, C., Chase, A. R., Wang, J. & Schlieker, C. The mechanism of Torsin ATPase activation. *Proc Natl Acad Sci U S A* **111**, E4822-4831 (2014). https://doi.org/10.1073/pnas.1415271111

138 Ben-Haim, Y. *et al.* Generation and characterization of iPSC lines from two nuclear envelopathy patients with a homozygous nonsense mutation in the TOR1AIP1 gene. *Stem Cell Res* **56**, 102539 (2021). https://doi.org/10.1016/j.scr.2021.102539

139 Dorboz, I. *et al.* Severe dystonia, cerebellar atrophy, and cardiomyopathy likely caused by a missense mutation in TOR1AIP1. *Orphanet J Rare Dis* **9**, 174 (2014). https://doi.org/10.1186/s13023-014-0174-9

140 Fichtman, B. *et al.* Combined loss of LAP1B and LAP1C results in an early onset multisystemic nuclear envelopathy. *Nat Commun* **10**, 605 (2019). https://doi.org/10.1038/s41467-019-08493-7

141 Kayman-Kurekci, G. *et al.* Mutation in TOR1AIP1 encoding LAP1B in a form of muscular dystrophy: a novel gene related to nuclear envelopathies. *Neuromuscul Disord* **24**, 624-633 (2014). https://doi.org/10.1016/j.nmd.2014.04.007

142 Basel-Vanagaite, L. *et al.* Mutated nup62 causes autosomal recessive infantile bilateral striatal necrosis. *Ann Neurol* **60**, 214-222 (2006). https://doi.org/10.1002/ana.20902

143 Harrer, P. *et al.* Recessive NUP54 Variants Underlie Early-Onset Dystonia with Striatal Lesions. *Ann Neurol* **93**, 330-335 (2023). https://doi.org/10.1002/ana.26544

144 di Biase, L., Di Santo, A., Caminiti, M. L., Pecoraro, P. M. & Di Lazzaro, V. Classification of Dystonia. *Life (Basel)* **12** (2022). https://doi.org/10.3390/life12020206

145 Strande, N. T. *et al.* Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am J Hum Genet* **100**, 895-906 (2017). https://doi.org/10.1016/j.ajhg.2017.04.015

146 Synofzik, M. & Schule, R. Overcoming the divide between ataxias and spastic paraplegias: Shared phenotypes, genes, and pathways. *Mov Disord* **32**, 332-345 (2017). https://doi.org/10.1002/mds.26944

147 Chintalaphani, S. R., Pineda, S. S., Deveson, I. W. & Kumar, K. R. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol Commun* **9**, 98 (2021). https://doi.org/10.1186/s40478-021-01201-x

148 Paucar, M. *et al.* Spinocerebellar ataxia type 4 is caused by a GGC expansion in the ZFHX3 gene and is associated with prominent dysautonomia and motor neuron signs. *J Intern Med* **296**, 234-248 (2024). https://doi.org/10.1111/joim.13815

149 Chen, Z. *et al.* Adaptive Long-Read Sequencing Reveals GGC Repeat Expansion in ZFHX3 Associated with Spinocerebellar Ataxia Type 4. *Mov Disord* **39**, 486-497 (2024). https://doi.org/10.1002/mds.29704

150 Figueroa, K. P. *et al.* A GGC-repeat expansion in ZFHX3 encoding polyglycine causes spinocerebellar ataxia type 4 and impairs autophagy. *Nat Genet* **56**, 1080-1089 (2024). https://doi.org/10.1038/s41588-024-01719-5

151 Chen, Z. *et al.* The ZFHX3 GGC Repeat Expansion Underlying Spinocerebellar Ataxia Type 4 has a Common Ancestral Founder. *Mov Disord* **40**, 363-369 (2025). https://doi.org/10.1002/mds.30077

152 Dalski, A. *et al.* Repeat length in spinocerebellar ataxia type 4 (SCA4) predicts age at onset and disease severity. *J Neurol* **271**, 6289-6300 (2024). https://doi.org/10.1007/s00415-024-12600-0

153 Ido, A., Miura, Y. & Tamaoki, T. Activation of ATBF1, a multiple-homeodomain zinc-finger gene, during neuronal differentiation of murine embryonal carcinoma cells. *Dev Biol* **163**, 184-187 (1994). https://doi.org/10.1006/dbio.1994.1134

154 Ishii, Y. *et al.* ATBF1-A protein, but not ATBF1-B, is preferentially expressed in developing rat brain. *J Comp Neurol* **465**, 57-71 (2003). https://doi.org/10.1002/cne.10807

155 Jung, C. G. *et al.* Homeotic factor ATBF1 induces the cell cycle arrest associated with neuronal differentiation. *Development* **132**, 5137-5145 (2005). https://doi.org/10.1242/dev.02098

156 Fuller, T. D., Westfall, T. A., Das, T., Dawson, D. V. & Slusarski, D. C. High-throughput behavioral assay to investigate seizure sensitivity in zebrafish implicates ZFHX3 in epilepsy. *J Neurogenet* **32**, 92-105 (2018). https://doi.org/10.1080/01677063.2018.1445247

157 Kim, T. S. *et al.* The ZFHX3 (ATBF1) transcription factor induces PDGFRB, which activates ATM in the cytoplasm to protect cerebellar neurons from oxidative stress. *Dis Model Mech* **3**, 752-762 (2010). https://doi.org/10.1242/dmm.004689

158 Liufu, T. *et al.* The polyG diseases: a new disease entity. *Acta Neuropathol Commun* **10**, 79 (2022). https://doi.org/10.1186/s40478-022-01383-y

159 Cook, L. *et al.* Parkinson's disease variant detection and disclosure: PD GENEration, a North American study. *Brain* **147**, 2668-2679 (2024). https://doi.org/10.1093/brain/awae142

160 Towns, C. *et al.* Parkinson's families project: a UK-wide study of early onset and familial Parkinson's disease. *NPJ Parkinsons Dis* **10**, 188 (2024). https://doi.org/10.1038/s41531-024-00778-z

161 Polymeropoulos, M. H. *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045-2047 (1997). https://doi.org/10.1126/science.276.5321.2045

162 Guo, Y. *et al.* Genetic Analysis and Literature Review of SNCA Variants in Parkinson's Disease. *Front Aging Neurosci* **13**, 648151 (2021). https://doi.org/10.3389/fnagi.2021.648151

163 Konno, T., Ross, O. A., Puschmann, A., Dickson, D. W. & Wszolek, Z. K. Autosomal dominant Parkinson's disease caused by SNCA duplications. *Parkinsonism Relat Disord* **22 Suppl 1**, S1-6 (2016). https://doi.org/10.1016/j.parkreldis.2015.09.007

164 Carmine Belin, A. *et al.* Leucine-rich repeat kinase 2 (LRRK2) mutations in a Swedish Parkinson cohort and a healthy nonagenarian. *Mov Disord* **21**, 1731-1734 (2006). https://doi.org/10.1002/mds.21016

165 Huff, C. D. *et al.* Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res* **21**, 768-774 (2011). https://doi.org/10.1101/gr.115972.110

166 Chen, Y. P. *et al.* The mutation spectrum of Parkinson-disease-related genes in early-onset Parkinson's disease in ethnic Chinese. *Eur J Neurol* **29**, 3218-3228 (2022). https://doi.org/10.1111/ene.15509

167 van Rappard, D. F., Boelens, J. J. & Wolf, N. I. Metachromatic leukodystrophy: Disease spectrum and approaches for treatment. *Best Pract Res Clin Endocrinol Metab* **29**, 261-273 (2015). https://doi.org/10.1016/j.beem.2014.10.001

168 Lee, J. S. *et al.* Arylsulfatase A, a genetic modifier of Parkinson's disease, is an alpha-synuclein chaperone. *Brain* **142**, 2845-2859 (2019). https://doi.org/10.1093/brain/awz205

169 Pan, H. X. *et al.* Evaluating the role of ARSA in Chinese patients with Parkinson's disease. *Neurobiol Aging* **109**, 269-272 (2022). https://doi.org/10.1016/j.neurobiolaging.2021.08.008

170 Ran, C. *et al.* Strong association between glucocerebrosidase mutations and Parkinson's disease in Sweden. *Neurobiol Aging* **45**, 212 e215-212 e211 (2016). https://doi.org/10.1016/j.neurobiolaging.2016.04.022

171 Ilyechova, E. Y. *et al.* Case of Early-Onset Parkinson's Disease in a Heterozygous Mutation Carrier of the ATP7B Gene. *J Pers Med* **9** (2019). https://doi.org/10.3390/jpm9030041

172 Castelo Rueda, M. P. *et al.* Frequency of Heterozygous Parkin (PRKN) Variants and Penetrance of Parkinson's Disease Risk Markers in the Population-Based CHRIS Cohort. *Front Neurol* **12**, 706145 (2021). https://doi.org/10.3389/fneur.2021.706145

173 Lubbe, S. J. *et al.* Assessing the relationship between monoallelic PRKN mutations and Parkinson's risk. *Hum Mol Genet* **30**, 78-86 (2021). https://doi.org/10.1093/hmg/ddaa273

Printed by Media-Tryck, Lund 2026 NORDIC SWAN ECOLABEL 3041 0903

About the author

Efthymia Kafantari earned her Bachelor's degree in Chemistry–Biochemistry from the University of Patras in Greece. She then pursued a Master's degree in Medical Genetics at the University of Glasgow in Scotland, followed by a second Master's degree in Bioinformatics at Lund University in Sweden.

Building on this interdisciplinary training, she undertook a PhD in Clinical Neurogenetics at Lund University under the supervision of Andreas Puschmann. Her doctoral research focuses on analyzing the genomes of patients with neurological movement disorders to identify both established and novel genetic causes. Through this work, she aims to deepen our understanding of the genetic architecture of these conditions and expand the spectrum of associated genes and variants.

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Clinical Sciences, Lund

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2026:91
ISBN 978-91-8021-889-4
ISSN 1652-8220