2019-08-01

Identifying the Genetic Architecture and Cellular Pathways of Inherited Axonopathies

Dana Marie Bis-Brewer

University of Miami, littlemissbis@gmail.com

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation


https://scholarlyrepository.miami.edu/oa_dissertations/2364
UNIVERSITY OF MIAMI

IDENTIFYING THE GENETIC ARCHITECTURE AND CELLULAR PATHWAYS OF INHERITED AXONOPATHIES

By

Dana Marie Bis-Brewer

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Coral Gables, Florida

August 2019
UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

IDENTIFYING THE GENETIC ARCHITECTURE AND CELLULAR PATHWAYS OF INHERITED AXONOPATHIES

Dana Marie Bis-Brewer

Approved:

Stephan Züchner, M.D., Ph.D. Mustafa Tekin, M.D.
Professor and Chair of Professor of Human Genetics
Human Genetics

Anthony Griswold, Ph.D. Guillermo Prado, Ph.D.
Research Assistant Professor of Dean of the Graduate School
Human Genetics

Stefan Wuchty, Ph.D.
Associate Professor of
Computer Science
Inherited axonopathies are a group of disorders unified by a common pathological mechanism: length-dependent axonal degeneration. Progressive axonal degeneration can lead to both Charcot-Marie-Tooth type 2 (CMT2) and hereditary spastic paraplegia (HSP) depending on whether the peripheral or central nerves are affected, respectively. Historically, CMT2 and HSP are treated as distinct disorders, but their increasingly apparent clinical and genetic overlap challenges this classification. Either CMT2 or HSP can be caused by mutations within a single gene, yet what determines whether the peripheral or central nerves are affected in each patient remains unclear. Currently, ~30% of CMT2 cases and ~60% of HSP cases have been genetically diagnosed, and only ~25% of clinical exome sequencing results in molecular diagnosis. The percentage of genetically undiagnosed inherited axonopathy cases suggests that novel disease genes remain unidentified. Furthermore, the general rate of exome sequencing diagnosis indicates that rare mutational mechanisms or modes of inheritances may be overlooked in standard whole-exome analysis.

The primary aims of my dissertation project were, firstly, to investigate possible rare genetic mechanisms and expand the genetic architecture of inherited axonopathies to address the diagnostic gap and, secondly, to explore cellular
pathways that differentiate peripheral versus central nervous system involvement. To catalog the current genetic variant landscape of CMT, I co-developed a web-based platform to catalog, rate, and discuss any variation observed in a CMT case. I analyzed 872 whole exome datasets from patients with CMT and HSP to explore a cumulative mutational burden and di/oligogenic inheritance across known disease genes and to identify risk alleles in both known and novel disease genes. This analysis lead to the identification of a candidate risk allele in EXOC4 for CMT, further expanding the genetic architecture of IA beyond highly penetrant monogenic alleles. To identify rare genetic mechanisms, I screened 96 hereditary motoneuron probands for uniparental isodisomy by scanning whole exome datasets for very long regions of homozygosity. From this pilot screen, I identified one HSP case with complete isodisomy of chromosome 16 harboring a homozygous mutation in FA2H, inherited from one heterozygous parent. To continue looking for rare genetic mechanisms, I screened each possible open reading frame of genome-wide 3' UTR sequences to classify candidate genes at risk for harboring cryptic amyloidogenic elements. To investigate cellular pathways involved in IA, I analyzed a protein-protein interaction dataset and an RNA-sequencing dataset. The PPI analysis revealed a topological relationship between CMT and HSP as well as cellular pathways putatively involved across the IA spectrum. RNA-sequencing of human motor neuron axonal compartment revealed a significant enrichment of nuclear-encoded mitochondrial genes and IA disease genes.
DEDICATION

This work is dedicated to my mother, Eugenia Bis, whose strength in living with Limb-Girdle Muscular Dystrophy has inspired me, to my father, Stephen Bis, whose everlasting commitment to her taught me dedication, to my son, Shackleton Brewer, whose imminent arrival refreshed my aspirations, and to the rest of my family and friends who encouraged my success. In particular, I would like to thank my husband, Tommy Brewer, for always believing in me and supporting me on this journey.
ACKNOWLEDGEMENTS

I would like to thank Dr. Stephan Züchner for his outstanding mentorship during my graduate studies. His continued guidance, availability, and encouragement were vital to my development as a young research scientist. Dr. Adriana Rebelo for being the pillar of strength who we all rely upon and my ‘lab mom.’ Cima Saghira for our countless hours of troubleshooting and running software together. Dr. Michael Gonzalez for setting me up for success and continuing to mentor me long after his graduation. Dr. Feifei Tao for helping me learn to program and collaborating on projects. Dr. Karolina Janczura for her understanding, encouragement, and friendship. Dr. Alex Abrams for his deep and insightful scientific questions. Elena Buglo and Steve Courel for their wonderful company and collaborations. Dr. Mario Saporta, Dr. Renata Maciel, Dr. Rebecca Schüle, Lisa Abreu, Dr. Matt Danzi, and Dr. Andrea Cortese – I truly enjoy working with all of you! To my dissertation committee members, Dr. Mustafa Tekin, Dr. Juan Young, and Dr. Stefan Wuchty, for their time, commitment, and guidance during my graduate studies. To the Human Genetics and Genomics program faculty and staff for providing me with a wonderful education, especially Dr. Bill Scott, Dr. Derek Dykxhoorn, and Dr. Susan Blanton. To Dori McLean for her immense help and support in all matters.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF FIGURES</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
</tbody>
</table>

## CHAPTER 1 Introduction to Inherited Axonopathies: Background, Clinical Features, and Genetics

1.1 Overview of the Nervous System ........................................ 1
1.2 Clinical Overview of Inherited Axonopathies ........................ 5
1.3 Molecular Mechanisms of Inherited Axonopathies ................. 7
1.4 Genetics of Inherited Axonopathies .................................. 10
1.5 The “Dark Matter” of Clinical Genomics .............................. 12
1.6 Beyond Mendelian Inheritance ........................................... 14
1.7 Network Biology Approaches to Locus Heterogeneity ............ 19
1.8 Summary ................................................................. 20

## CHAPTER 2 Variant Pathogenicity Evaluation in the Community-Driven Inherited Neuropathy Variant Browser

2.1 Background ............................................................. 22
2.2 Materials and Methods ................................................ 23
2.3 Results ................................................................. 25
2.4 Discussion .............................................................. 31

## CHAPTER 3 Homozygosity Mapping of Whole-Exome Sequencing Detects Uniparental Isodisomy in Hereditary Spastic Paraplegia

3.1 Background ............................................................. 36
3.2 Materials and Methods ................................................ 37
3.3 Results ................................................................. 39
3.4 Discussion .............................................................. 46

## CHAPTER 4 Gene-based Association Analysis and Cumulative Mutational Burden in Inherited Axonopathies

4.1 Background ............................................................. 49
4.2 Materials and Methods ................................................ 49
4.3 Results ................................................................. 55
4.4 Discussion .............................................................. 62
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Starting Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Identification of Cryptic Amyloidogenic Elements in Genome-Wide Screen</td>
<td>65</td>
</tr>
<tr>
<td>5.1</td>
<td>Background</td>
<td>65</td>
</tr>
<tr>
<td>5.2</td>
<td>Materials and Methods</td>
<td>66</td>
</tr>
<tr>
<td>5.3</td>
<td>Results</td>
<td>68</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
<td>75</td>
</tr>
<tr>
<td>6</td>
<td>Enrichment of Transcripts Related to Mitochondrial Function and Microtubule-Based Axonal Transport within the Human Motor Neuron Axonal Transcriptome</td>
<td>76</td>
</tr>
<tr>
<td>6.1</td>
<td>Background</td>
<td>76</td>
</tr>
<tr>
<td>6.2</td>
<td>Materials and Methods</td>
<td>77</td>
</tr>
<tr>
<td>6.3</td>
<td>Results</td>
<td>78</td>
</tr>
<tr>
<td>6.4</td>
<td>Discussion</td>
<td>83</td>
</tr>
<tr>
<td>7</td>
<td>Network-based Analysis of Inherited Axonopathies Reveals Cellular Pathways and Molecular Overlap</td>
<td>84</td>
</tr>
<tr>
<td>7.1</td>
<td>Background</td>
<td>84</td>
</tr>
<tr>
<td>7.2</td>
<td>Materials and Methods</td>
<td>85</td>
</tr>
<tr>
<td>7.3</td>
<td>Results</td>
<td>92</td>
</tr>
<tr>
<td>7.4</td>
<td>Discussion</td>
<td>104</td>
</tr>
<tr>
<td>8</td>
<td>Conclusions</td>
<td>114</td>
</tr>
<tr>
<td>8.1</td>
<td>Summary of Findings</td>
<td>114</td>
</tr>
<tr>
<td>8.2</td>
<td>Future Perspectives</td>
<td>120</td>
</tr>
</tbody>
</table>

References                                                                 | 122           |
# LIST OF FIGURES

## CHAPTER 1 Introduction to Inherited Axonopathies: Background, Clinical Features, and Genetics
- Figure 1.1 Basic neuronal structure (motor neuron) ........................................ 2
- Figure 1.2 Overview of Inherited Axonopathy (IA) ...................................... 5
- Figure 1.3 Molecular mechanisms of Inherited Axonopathy (IA) ............... 10
- Figure 1.4 The diagnostic heritability gap of Inherited Axonopathy (IA) .... 13

## CHAPTER 2 Variant Pathogenicity Evaluation in the Community-Driven Inherited Neuropathy Variant Browser
- Figure 2.1 Overview of Inherited Variant Browser homepage and search options ................................................................. 27
- Figure 2.2 Overview of query result table................................................. 28
- Figure 2.3 Inherited Neuropathy Variant Browser descriptive statistics ...... 30
- Figure 2.4 Inherited Neuropathy Variant Browser usage examples .......... 32

## CHAPTER 3 Homozygosity Mapping of Whole-Exome Sequencing Detects Uniparental Isodisomy in HSP
- Figure 3.1 Distribution of regions of homozygosity within screening cohort 41
- Figure 3.2 Uniparental isodisomy screening flowchart ......................... 42
- Figure 3.3 Isodomy mapping from whole-exome sequencing ................... 44
- Figure 3.4 Regions of homozygosity in gene-poor region of chr13 .......... 45

## CHAPTER 4 Gene-based Association Analysis and Cumulative Mutational Burden in Inherited Axonopathies
- Figure 4.1 Quality control of joint-called variant set ............................... 55
- Figure 4.2 Gene-based association analysis of CMT cases ...................... 57
- Figure 4.3 EXOC4 carrier risk .............................................................. 57
- Figure 4.4 Cumulative mutational burden across known disease genes.. 59
- Figure 4.5 Digenic/oligogenic inheritance across known disease genes.. 62

## CHAPTER 5 Identification of Cryptic Amyloidogenic Elements in Genome-Wide Screen
- Figure 5.1 NEFH Frameshift Variants in CMT2-Affected Families .......... 69
- Figure 5.2 Identification of CAEs Encoded by NEFH and NEFL 3′ UTRs 70
- Figure 5.3 Genome-wide scan for cryptic amyloidogenic elements of the human 3′-UTRef ................................................................. 73
CHAPTER 6 Enrichment of Transcripts Related to Mitochondrial Function and Microtubule-Based Axonal Transport within the Human Motor Neuron Axonal Transcriptome
   Figure 6.1 Transcriptome profiling of iPSC-derived motor neuron axonal and somatodendritic compartments (AXC and SDC) ............................... 81

CHAPTER 7 Network-based Analysis of Inherited Axonopathies Reveals Cellular Pathways and Molecular Overlap
   Figure 7.1 Inherited axonopathy proto-module connectivity and expanded module networks ................................................................. 97
   Figure 7.2 Biological validation of inherited axonopathy DIAMOnDs ...... 98
   Figure 7.3 Disease proto-modules separation and connections .............. 99
   Figure 7.4 Interactions within the expanded inherited axonopathy modules ... ................................................................................................... 101
   Figure 7.5 DIAMOnD ribosomal protein interactions with disease seed proteins .......................................................................................... 102
   Figure 7.6 Pathway analysis of inherited axonopathy modules ............ 105
# LIST OF TABLES

<table>
<thead>
<tr>
<th>CHAPTER 3</th>
<th>Homozygosity Mapping of Whole-Exome Sequencing Detects Uniparental Isodisomy in HSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1</td>
<td>Segregation analysis of potential UPD events ................................................. 44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 4</th>
<th>Gene-based Association Analysis and Cumulative Mutational Burden in Inherited Axonopathies (in progress)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4.1</td>
<td>Summary of cumulative mutational burden across known IA disease genes ........................................ 60</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Summary of digenic/oligogenic variation across known IA disease genes ................................................ 61</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 5</th>
<th>Identification of Cryptic Amyloidogenic Elements in Genome-Wide Screen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 5.1</td>
<td>List of aggregation disease genes predicted to contain a 3'-UTR CAE .................. 74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 6</th>
<th>Enrichment of Transcripts Related to Mitochondrial Function and Microtubule-Based Axonal Transport within the Human Motor Neuron Axonal Transcriptome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 6.1</td>
<td>Mitochondrial genes overexpressed in axons................................................. 82</td>
</tr>
<tr>
<td>Table 6.2</td>
<td>Neurological disease genes overexpressed in axons...................................... 82</td>
</tr>
</tbody>
</table>
CHAPTER 1
Introduction to Inherited Axonopathies: Background, Clinical Features, and Genetics

1.1 Overview of the Nervous System

The human nervous system is morphologically divided into two major regions: the central nervous system (CNS), including the brain and spinal cord, and the peripheral nervous system (PNS), including the cranial and spinal nerve fibers and ganglion. The brain integrates sensory information and coordinates a response, while the spinal cord is the channel for conducting signals between the brain and body. The PNS can be functionally decomposed into the sensory and motor components. The sensory system brings information to the CNS (afferent) for processing while the motor system delivers processed information away from the CNS (efferent) to the effector cell for stimulus response.¹

The tissue of the nervous systems contains two basic cell types: neurons and glial cells. Glia are non-neuronal cells that provide support, protection, and nutrients to neurons.¹ The neuron is the functional unit of the nervous system and communicates information by receiving, conducting, and transmitting impulses to each other and to effector cells.¹² The structure of a neuron allows the cell to transmit electrical impulses within the nervous system (Figure 1.1). Neurons are comprised of a soma (cell body), which contains the nucleus, and multiple processes (extensions), called the axon and dendrites. Dendrites branch off from the soma and are responsible for receiving input signals while the axon is the fiber
that connects a neuron with its target and transmits the signal (Figure 1.1). Some neurons contain an insulating layer, called the myelin sheath, which surrounds the axon to increase the speed of electrical signal conduction. Glial cells are responsible for forming the myelin sheath: oligodendrocytes in the CNS and Schwann cells in the PNS.

**Figure 1.1 Basic neuronal structure (motor neuron)**


Within the CNS, neurons are organized into regions of gray and white matter. Gray matter is large aggregates of neuronal cell bodies and dendrites that is arranged as sheaths or nuclei, whereas white matter is a primarily composed of myelinated axons that can form large fiber bundles named tracts. The PNS does not organize into anatomical regions of gray and white matter; however, the neurons are still organized in a similar fashion. Within the PNS, a cluster of
neuronal cell bodies is called a ganglion while a bundle of axons is referred to as a nerve.²

The PNS can also be classified into the somatic and autonomic nervous systems. The somatic nervous system controls voluntary skeletal movement and consists of the sensory (afferent) and motor (efferent) nerves, which are coordinated by specific regions of the CNS. The sensory neurons relay information about sensation from receptors in the skin, skeletal muscles, tendons, and joints to the spinal cord through the dorsal root. Ascending pathways carry the peripheral sensations up the spinal cord to the brain. The dorsal column system is one ascending pathway that begins with the axon of the dorsal root ganglion and ultimately terminates on the postcentral gyrus of the cerebral cortex. After processing, the motor response output from the cortex travels down two descending pathways. The corticospinal tract is the major descending pathway that controls skeletal muscle movement, composed of the upper motor neurons (UMNs) and terminating on the lower motor neurons (LMNs). The UMN’s cell body is in the frontal lobe and synapses with the LMN’s cell body in the ventral horn of the spinal cord, which then project to the peripheral skeletal muscle.²

Together, the UMNs and LMNs are known as the pyramidal motor system. Damage to the pyramidal system results in varying clinical symptoms, depending on the location of the lesion. Signal disruption between the LMNs and muscle cells cause gradually muscle weakening and wasting and can result in various diseases including progressive muscular atrophy, spinal muscular atrophy, and motor neuropathies.³ When the signal disruption occurs between the UMN and LMN,
muscles develop spasticity and tendon reflexes become overactive resulting in diseases such as primary lateral sclerosis and hereditary spastic paraplegia.\textsuperscript{3} Other diseases, such as amyotrophic lateral sclerosis and progressive bulbar palsy, include both UMN and LMN involvement.\textsuperscript{3} Though motor neuron diseases (MND) can be classified by the motor neuron most heavily affected, most MND can have both UMN/LMN involvement.\textsuperscript{3}

UMN and LMN both contain extremely long axons; UMN extend from the brain down the spinal cord (1-3 feet or more) and LMNs extend from the spinal cord to the skeletal muscles (3-5 feet). Length-dependent degeneration of these long axons can be due to a group of hereditary disorders known as inherited axonopathies. Inherited axonopathies are traditionally classified into two broad genetic disorders: Hereditary Spastic Paraplegia (HSP) and Charcot-Marie-Tooth disease (CMT) depending on upper or lower motor neuron involvement (Figure 1.2). However, due to their increasingly apparent genetic and clinical overlap, HSP and CMT are frequently classified as a spectrum of upper and lower motor neuron axonopathies.
Figure 1.2 Overview of Inherited Axonopathy (IA)

(A) Overview of affected nervous systems. CMT affects the peripheral nervous system (blue) while HSP affects the central nervous system (green). (B) Detailed view of the sites of long axon degeneration in IA (HSP: green, CMT: blue). Modified from OpenStax Anatomy and Physiology.

1.2 Clinical Overview of Inherited Axonopathies

Charcot-Marie-Tooth disease

Charcot-Marie-Tooth disease (CMT) is representative of inherited neuropathies affecting the peripheral nervous system.\(^4,5\) Initially reported in the late 19\(^{th}\) century by physicians Jean Martin Charcot, Pierre Marie, and Howard Henry Tooth, CMT is now the most common inherited neurological disorder and affects an estimated 1 in 2500 people.\(^6\) CMT is highly genetically heterogeneous with over 80 genes involved\(^7\) and all modes of inheritance observed.\(^4\) Additionally, CMT also exhibits clinical heterogeneity with variability in disease progression rate, age of onset, and nerve conduction velocity results.\(^4\)
The most typical CMT phenotypes are distal muscle weakness and atrophy, high arch of the foot, decreased tendon reflexes, and sensory loss. Symptoms are a result of the progressive peripheral neuropathy of motor and/or sensory nerves. Inherited peripheral neuropathy is caused by genetic defects in either of the principal cells of the peripheral nervous system (PNS): Schwann cells or axons of peripheral neurons. One CMT classification system, axonal or demyelinating, is based on the affected cell. Peripheral neuronal axons transmit information via electrical signals to and from the central nervous system. Schwann cells form and maintain the myelin sheath around the axons. The myelin sheath provides the necessary insulation to enable quick propagation of electrical signals along the axon. Schwann cells are also involved in maintenance and regeneration of the axon itself. Due to the interaction between Schwann cells and axons, secondary axonal degeneration also occurs in demyelinating CMT.

Nerve conduction velocities determine the primary CMT type. Demyelinating CMT (CMT1) exhibits diminished velocities, axonal CMT (CMT2) displays normal velocities with reduced amplitudes, while intermediate CMT (CMTi) shows velocities between CMT1 and CMT2. CMT is also classified by inheritance mode: CMT1 and CMT2 are autosomal dominant, CMT4 is autosomal recessive with either demyelination or axonopathy, and CMTX is X-linked with axonopathy. Each subtype is additionally subdivided based on genetic etiology.
Hereditary spastic paraplegia disease

Hereditary spastic paraplegia (HSP) is a genetic neurological disorder that affects the central nervous system. Physicians Adolph Strumpell and Maurice Lorrain were the first to describe the disease in the late 19th century. Today, HSP is estimated to affect ~3-9/100,000 in most populations, demonstrates all modes of inheritance, and affects more than 80 genetic loci. HSP results from genetic defects primarily in the axons of the corticospinal tracts within the central nervous system. The corticospinal tracts are comprised of upper motor neurons and are a component of the pyramidal motor system, which is responsible for voluntary movement. Similar to CMT, HSP also displays high variability in age of onset, disease progression rate, and degree of symptom severity.

In general, HSP is clinically characterized by bilateral lower extremity weakness and spasticity. Based on the absence or presence of additional neurological features, HSP is classified as pure or complicated respectively. Pure HSP may include other clinical features such as mild sensory loss in the lower limbs, high foot arches, urinary symptoms, and mild decline in cognition. Complicated HSP is the appearance of spastic paraplegia with features such as vision loss, hearing loss, mental retardation, ataxia, peripheral neuropathy, or epilepsy. Both types of HSP are subdivided based on genetic diagnosis.

1.3 Molecular Mechanisms of Inherited Axonopathies

CMT2 and HSP both are genetic diseases that cause progressive degeneration of long axons. While they affect distinct cells within the nervous system, these
diseases share many clinical, genetic, and pathological features. The number of genes discovered to cause axonopathies along with the functional characterization of these genes has improved our understanding of the biological mechanisms (Figure 1.3). From this improved understanding, it has become evident that the pathological mechanisms of CMT2 and HSP overlap, including disruption of axonal transport, mitochondrial dynamics, mitochondrial regulation, membrane trafficking, and organelle shaping. The axons of upper motor neurons and peripheral nerves can reach a meter in length. They require adaptable cellular machinery to distribute molecules and maintain neuronal homeostasis, and a specialized neuronal cytoskeleton to maintain this transport.

Within CMT2, a decrease in cytoskeleton stability occurs through mutations in the neurofilaments (NEFL and NEFH), a filament within the inner nuclear envelope (LMNA), and neuro-specific tubulin (TUBB3). The most common cause of HSP is the microtubule-severing protein spastin which affects the stability of the microtubule cytoskeleton and disrupts anterograde and retrograde axonal transport in mice. Microtubule motor-associated proteins, DCTN1 and KIF5A, are essential to axonal transport and cause CMT and HSP, respectively. Mitochondrial dynamics are affected in CMT2 through mutations in GDAP1 and MFN2, and in HSP through mutations in SPG7, SPG13, and HSPD1. Altogether, these genes affect mitochondrial integrity, fusion, fission, and fragmentation, which are pathogenic likely due to impaired in axonal transport and energy production. Mutations within the small GTPase, RAB7, cause CMT2 through disruption of membrane trafficking via intracellular vesicles. Within HSP, the largest subgroup
of genes is involved in membrane trafficking processes. Membrane trafficking is effected through perturbation of endoplasmic reticulum shape (*RTN2, SPAST, ATL1, REEP1*), endosomal membrane shape (*SPG8*), lysosomal trafficking (*SPG20*), as well as less characterized proteins localizing to endosomes.\(^{15}\) Interestingly, mutations within the gene encoding myelin proteins can also cause CMT2 and HSP. In CMT2, *MPZ* mutations can cause axonal loss possibly through disruption of signaling from Schwann cells required for axonal intracellular trafficking.\(^{27}\) Three myelin genes (*PLP1, FA2H*, and *GJA12/GJC2*) can cause HSP through impairment of oligodendrocyte support to axonal transport and proper myelin formation.\(^{15}\)

The characterized molecular mechanisms do not entirely overlap for inherited axonopathies. Unique to CMT2 are dysfunctions of aminoacyl-tRNA synthetases (*GARS, YARS, AARS, HARS, MARS, and KARS*) and small heat shock proteins (*HSP27* and *HSP22*).\(^{15}\) The exact mechanism through which the tRNA synthetases cause disease and why peripheral nerves are susceptible to mutations within them remains unclear.\(^{15}\) Since several studies have demonstrated that aminoacylation activity is not affected by mutations, a recent hypothesis is that the tRNA synthetases reduce protein translation by effecting translation elongation.\(^{28}\) Mutations in heat shock proteins, which are involved in protein refolding to protect cells from stress, may lead to protein aggregation and disrupted axonal transport. Additionally, mutated heat shock proteins have been shown to stabilize the microtubule cytoskeleton and destabilize neurofilament-light protein.\(^{29-31}\) Likewise, lipid metabolism is implicated in HSP alone through
mutations with \textit{CYP7B1}, \textit{DDHD1}, \textit{DDHD2}, \textit{GBA2}, \textit{PNPLA6}, \textit{CYP2U1}, and \textit{B4GALNT1}. Both signaling and cellular membrane composition are likely affected by altered lipid metabolism, however its role in axonal degeneration remains unclear.\footnote{18}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{MolecularMechanisms.png}
\caption{Molecular Mechanisms of Inherited Axonopathy (IA)}
\textbf{Legend:}
- CMT only
- HSP only
- CMT and HSP
- CMT2
- CMTi

Molecular mechanisms involved and underlying affected genes involved in IA. Unmarked blue text is CMT only genes (all subtypes); unmarked red text is HSP only genes (all subtypes); black dots mark CMT and HSP genes; solid green dots mark CMT2 genes; and open green dots mark CMTi genes. \textit{Modified from Timmerman et al. 2012. Experimental Neurology 246:14-25.}

\subsection*{1.4 Genetics of Inherited Axonopathies}

As with many other Mendelian diseases, the introduction of next-generation sequencing (NGS) revolutionized the genetic diagnosis of IA with over 76 genomic loci and 58 corresponding genes for HSP and over 80 genes for CMT.\footnote{7,32} Though
IA can be caused by an overwhelming amount of genetic defects, it is noteworthy
that a handful of genes are responsible for the bulk of HSP and CMT cases. More
than half of all CMT cases are caused by five genetic mutations: \textit{PMP22}
duplication (39.5\%), \textit{PMP22} point mutation (1.4\%), \textit{GJB1} (10.8\%), \textit{MFN2} (2.8\%),
and \textit{MPZ} (3.1\%).\textsuperscript{33} For autosomal dominant (AD) demyelinating CMT (CMT1), the
most commonly mutated genes are: \textit{GJB1, PMP22, MPZ, EGR2, LITAF, NEFL,} or
\textit{PMP2}.\textsuperscript{33} Unlike CMT1, AD axonal CMT (CMT2) and autosomal recessive (AR)
axonal and/or demyelinating forms (CMT4) are caused by many individually rare
genes that typically affect only a handful of families.\textsuperscript{33,34} The most common cause
of CMT2 (~20\%) is mutations within the outer mitochondrial membrane protein,
\textit{MFN2}\textsuperscript{35}, while the most common CMT4 genes are \textit{GDAP1} and \textit{SH3TC2}.\textsuperscript{34} The
AD HSPs lead mostly to the pure form of disease and are linked to 19 spastic gait
(SPG) genes.\textsuperscript{32} The most common AD genes are SPG4/\textit{SPAST}, SPG3A/\textit{ATL1},
SPG31/\textit{REEP1}, and SPG10/\textit{KIF5A}. Complicated HSPs more frequently occur in
autosomal recessive (AR) families and are linked to 57 loci and 52 genes. The
most common AR genes are SPG11/\textit{KIAA1840}, SPG5A/\textit{CYP7B1}, SPG7, and
SPG15/\textit{ZFYVE26}. Though rare, X-linked and mitochondrial inheritance are also
observed.

In the past 5 years alone >15 novel HSP/HSP-related genes and >20 CMT
genes have been reported. However, the number of families identified in the initial
and follow-on papers is typically low; for example, the original \textit{REEP2} publication
had 2 families with a follow-on of a single family.\textsuperscript{36,37} This has led to the concern
that we are in an asymptotic situation where, even with many new genes, the
diagnostic yield may not get close to 100%. This potential gap in heritability is observed in other rare disorders as well and might be referred to as “dark matter” of clinical genomics. This dissertation will largely focus on potential causes and efforts to overcome these challenges.

1.5 The “Dark Matter” of Clinical Genomics

It is usually assumed that IA is caused by Mendelian mechanisms and that eventually nearly all patients will receive a single-gene diagnosis. However, this is not necessarily true. Related motor neuron disorders illustrate a diverse situation: inherited ataxias are also highly heterogeneous Mendelian disorders whereas amyotrophic lateral sclerosis (ALS) is largely not explained by Mendelian genes. The proportion of Mendelian genes is even lower in late onset neurodegenerative diseases, such as Parkinson and Alzheimer disease.

The reported diagnostic yield for exome sequencing in the general clinical setting ranges from 25 to 50%. In HSPs, Schüle et al. identified a molecular diagnosis in 46% of families (240/519 families), despite extensive whole-exome sequencing efforts, with low success in simplex families (28%). Similarly, a 20% diagnostic yield was reported in a cohort of 98 previously unsolved HSP families analyzed by a custom sequencing panel of 70 HSP and HSP-related genes. In CMT, Fridman et al. reported a 60.4% (997/1652 patients) genetic diagnosis rate; while 91.2% of CMT1 patients received a genetic diagnosis, only 42% of CMT2 cases received genetic confirmation.
While the genetics of certain neurodegenerative diseases are deemed ‘complex’, the field, thus far, has not applied similar models of inheritance to IA. We believe that it is too early to tell whether non-Mendelian effects have a major contribution to IA. There are several valid pro-Mendelian hypotheses as to how to close the diagnostic gap, and the coming years will allow us to test these ideas (Figure 1.4). These include non-coding regions of the genome, unorthodox types of mutations (such as repeat expansions) and digenic inheritance models. However, the search for risk genes and alleles will likely contribute to the understanding of IA in multiple ways, from exploring oligogenic causation, gene/environment interactions, and phenotype modifying genes; thus, expanding the inheritance models in IA.

Figure 1.4 The diagnostic heritability gap of Inherited Axonopathy (IA)
A) Despite unprecedented success in the identification of additional Mendelian genes, the diagnostic yield may not get close to 100% in IA, but rather reach an asymptotic ceiling. B) Areas that are potentially understudied in IA thus far for cost and technical challenges. These include uncommon Mendelian causation, but also modifier and oligogenic risk alleles. The colored bars represent genes, the black lines connecting genes represent noncoding regions, and each ‘X’ represents a mutation. Modified from Bis-Brewer et al. 2018. Frontiers in Neurology 9(958).
1.6 Beyond Mendelian Inheritance

Based on the assumption of fully penetrant alleles, traditional Mendelian disease analysis focuses on the rare DNA variation that segregates within a family. However, these locus-specific family studies treat Mendelian traits as distinct entities and disregard a more comprehensive genetic model for human disease in which variants of varying effect size as well as environmental influences contribute to disease.\(^{42}\) The challenge is the unexpected large amount of variation in the human genome on a population level, where >99% of all variants show a minor allele frequency of <1%.\(^{43}\) While many of these variants are without phenotypic consequence, some certainly are very harmful, and a considerable number must have effect sizes that are below the threshold of a Mendelian gene but contribute significantly to phenotypic expression. Identification of strong effect sizes in the background of mostly minor effects is the next big challenge in human genetics. Recent method developments in statistical genetics allow for unbiased genome-wide screens for non-Mendelian alleles, and surprisingly, are able to re-identify \textit{bona fide} Mendelian genes as well.\(^{44}\) The application to IA genomics will eventually generate a more complete genetic architecture of the disease.

\textit{Reduced penetrant and risk alleles}

Contrary to general expectations for IA families, asymptomatic carriers are not infrequent, in which case, the genotype is said to be incompletely penetrant.\(^{45,46}\) Reduced and age-dependent penetrance is a diagnostically challenging situation observed in autosomal dominant IA, especially SPG3A and \textit{MFN2}, which can lead
to misinterpretation of inheritance patterns due to asymptomatic carriers and exclusion of the disease-causing allele.\textsuperscript{47-50} Additionally, sex-dependent penetrance is suspected in HSP genes \textit{SPAST} and \textit{ATL1} based on the excess of affected males.\textsuperscript{51} Incomplete penetrance can also manifest in autosomal recessive disorders when the primary mutation leads to varying phenotypic effects depending on the secondary mutation.\textsuperscript{45} For example, novel compound heterozygous mutations in \textit{SPG11} caused an atypical late onset and mild form of \textit{SPG11} in HSP and compound heterozygous mutations in \textit{SCO2} led to a phenotypic expansion including early-onset CMT2.\textsuperscript{52,53}

Risk alleles are another form of variation that does not conform to standard Mendelian inheritance and have been defined as variants with smaller effect sizes that are part of a multifactorial model of disease causation.\textsuperscript{45} However, since the possibility of risk alleles is only recently recognized in rare Mendelian disease, the line between penetrance and risk is often blurred. In this context, risk alleles more broadly refer to rare variants that may lead to a less severe, later-onset form of disease or contribute to an individual’s susceptibility to disease, likely through an oligogenic model. For example, heterozygous mutations in \textit{SPG7} were identified as a potential susceptibility factor for late-onset neurodegenerative disorder.\textsuperscript{54} Similarly, heterozygous mutations in \textit{MME} were recently shown to predispose carriers to late-onset axonal neuropathy.\textsuperscript{55} In MME, the comparison of the ‘rare variant load’ of missense and loss of function changes in late-onset CMT to the general population showed a significant enrichment of such variation.\textsuperscript{55}
Systematic identification of rare variant associations are usually limited by low statistical power unless sample sizes or variant effect sizes are very large.\textsuperscript{56} To illustrate, >60,000 cases (and an equal number of controls) would be necessary to detect a disease association for a rare variant (0.1\% frequency) with an odds ratio of 2.0 for a disease with a 5\% population prevalence.\textsuperscript{56} Fortunately, powerful study designs can alleviate the sample size requirement to more reasonable numbers.\textsuperscript{57} One approach that can be explored in IA is the gene-based variant burden test which collapses the number of minor alleles into one genetic score (gene), thus reducing multiple testing and increasing power.\textsuperscript{56,58} One successful example of this approach was the identification of a new ALS gene, \textit{TBK1}, in 2869 sporadic ALS patients.\textsuperscript{44} Remarkably, other known ALS genes showed strong associations, indicating that additional variation in known familial ALS genes also contribute to sporadic ALS forms.\textsuperscript{44} The rare variant association studies are particularly useful for identifying risk genes and novel gene associations.

\textit{Modifier alleles}

An increasing number of exceptions to the fundamental “one gene, one phenotype” paradigm are being published across Mendelian phenotypes.\textsuperscript{59} The oversimplified view that phenotypic expression, even for classically monogenic disorders, is driven exclusively by mutations at a single locus is being replaced by the concept of genetic modification.\textsuperscript{60} Though several types of genetic modification are possible, the simple definition is the effect of one allele on the phenotypic outcome of a second allele.\textsuperscript{60} If the primary allele is sufficient to cause disease, then the
secondary allele is a “modifier” that modulates phenotypic expression, such as disease severity or progression.

Given the high clinical variability observed across IA patients, genetic modification of the primary allele was anticipated. Over a decade ago in HSP, intragenic polymorphisms were suggested to modify the age at onset of SPAST mutations. More recently, SPAST deletions spanning the adjacent DPY30 gene were shown to have significantly reduced age at onset. Furthermore, in a study of a large Cuban spinocerebellar ataxia type 2 (SCA2) cohort, 33% of the residual age at onset variance was attributed to genetic modifiers. In CMT1A, a polymorphism in miR-149 was recently associated with onset age and severity.

Another study design that increases the statistical power for association testing of rare variants is the extreme phenotype sampling (EPS) approach. Based on the assumption that rare causal variants are more likely found in the extremes of a quantitative trait such as age of onset or severity of a symptom, EPS can increase the power to detect rare variants over random sampling. For example, Emond et al. utilized an extreme phenotype sampling approach to identify an association between rare coding variants in DCTN4 and time to first Pseudomonas infection (measure of cystic fibrosis severity). In CMT1A, Tao et al. identified SIPA1L2 as a genetic modifier of muscle strength impairment. In vitro knock down of SIPA1L2 in Schwannoma cells lead to a significant reduction in PMP22 expression, offering a potential pathway for therapeutic strategies. Application of EPS to an HSP cohort may also reveal modifier alleles that contribute to disease.
**Oligogenic inheritance**

Digenic or oligogenic inheritance refers to instances when one primary allele is insufficient to cause disease, instead requiring the combined consequence of multiple alleles. Evidence of oligogenic inheritance has emerged in other neurological disorders. In both sporadic and familial amyotrophic lateral sclerosis (ALS) cases, patients harboring two or more rare variants had lower survival or earlier age at onset, suggesting that the combined effect of rare variants affects ALS development and progression. Similarly, over 30% of Parkinson’s disease (PD) patients carried additional rare variants in Mendelian PD genes and had younger ages at onset. In CMT, an increased rare variant burden was observed in two cohorts of inherited neuropathy cases, which was followed up in vivo zebrafish experiments. In zebrafish, more severe phenotypic outcomes were observed as a consequence of increased mutational burden in neuropathy genes, consistent with a positive genetic interaction mechanism of oligogenic inheritance.

Demonstrating oligogenic inheritance from family studies is challenging without experimental models. However, one trending approach to assessing oligogenic inheritance – which has been explored in Parkinson, ALS, Frontotemporal Dementia, Congenital Hypothyroidism, Inherited Neuropathy, and more – is to evaluate the mutational burden across known disease genes through Fisher’s exact test or logistic regression. However, caution should be used with this approach as Koegh et al. warns that systematic bias can lead to the apparent
enrichment of ‘oligogenic’ variants in familial cases and controlling such bias is essential for investigating an oligogenic role in neurodegenerative diseases.\textsuperscript{76}

1.7 Network Biology Approaches to Locus Heterogeneity

High-throughput (HT) technologies and computational methods have revolutionized the study of biology, genomics, and genetic disease. Biological network analysis is a systems-level approach to analyzing and interpreting HT data.\textsuperscript{77} Network biology arises from the idea that cellular components function through interactions with other cellular components.\textsuperscript{78} Notably, protein-protein interaction networks (PINs) have become a well-studied and valuable tool for biology. The first systematic PINs were published in the early 2000s based on HT assays such as yeast two-hybrid screens and affinity purification coupled with mass spectrometry.\textsuperscript{79} Based on the hypothesis that a disease phenotype is caused by the interaction of various pathophysiological processes, PINs have been successful in the study of disease mechanisms and disease gene discovery.\textsuperscript{78,79}

For example, a recent PIN analysis of HSP genes identified 3 candidate genes found to be mutated within a cohort and expanded the mechanistic understanding of HSP by revealing links to other neurodegenerative disorders.\textsuperscript{80} Increasing evidence suggests that disease genes interact with each other and cluster within a network neighborhood to form mechanistically linked disease module.\textsuperscript{81} Though the high amount of locus heterogeneity present in IA complicates clinical diagnosis, it does provide an opportunity to study the overarching biological pathways through analysis of molecular networks. The identification and characterization of disease modules can help resolve differences between normal and disease pathways,
reveal common biological functions among related diseases, and identify novel genes and pathways related to disease.\textsuperscript{78,81-83} With the rapid pace of gene discovery in HSPs, network analysis will continue to be a powerful approach for deciphering the complex interactions underlying the phenotype.

1.8 Summary

Inherited axonopathies (IA) are a group of hereditary disorders that have been historically classified into separate diseases, Hereditary Spastic Paraplegia and axonal Charcot-Marie-Tooth disease, yet are currently unified into a disease spectrum based on overlapping genetic defects and common pathological mechanisms. Although considerable progress has been made in understanding the genetic basis of IA, a large fraction of families remains without genetic diagnosis. In my dissertation, I focus on exploring rare genetic mechanisms that may contribute to the observed diagnostic gap as well as approaches to understanding the molecular etiology of IA. In Chapter 2, I develop a web-based community platform for discussing and interpreting Variants of Uncertain Significance (VUS) in CMT. In Chapter 3, I screen a cohort of HSP cases for uniparental isodisomy by performing homozygosity mapping on whole-exome sequencing data. In Chapter 4, I explore risk alleles and oligogenic inheritance in a large cohort of IA WES. In Chapter 5, I perform a genome-wide screen to identify additional potential instances of a rare genetic mechanism: translation of cryptic amyloidogenic elements. In Chapter 6, I investigate the molecular etiology of IA by analyzing RNA-sequencing data of the axons of induced pluripotent stem cell
(iPSC)-derived human motor neurons. In Chapter 7, I use IA’s locus heterogeneity to build and characterize disease gene modules. Together, I hope these findings will provide novel insights into inherited axonopathies that can eventually inform novel diagnostics and therapeutics.
CHAPTER 2

Variant Pathogenicity Evaluation in the Community-Driven Inherited Neuropathy Variant Browser¹

2.1 Background

Charcot-Marie-Tooth disease (CMT) is an umbrella term for inherited neuropathies affecting an estimated 1 in 2500 people. Over 120 CMT and related genes have been identified and clinical gene panels often contain more than 100 genes. Such a large genomic space will invariably yield variants of uncertain clinical significance (VUS) in nearly any person tested. This rise in number of VUS creates major challenges for genetic counseling. Additionally, fewer individual variants in known genes are being published as the academic merit is decreasing, and most testing now happens in clinical laboratories, which typically do not correlate their variants with clinical phenotypes. For CMT, we aim to encourage and facilitate the global capture of variant data to gain a large collection of alleles in CMT genes, ideally in conjunction with phenotypic information. The Inherited Neuropathy Variant Browser provides user-friendly open access to currently reported variation in CMT genes. Geneticists, physicians, and genetic counselors can enter variants detected by clinical tests or in research studies in addition to genetic variation

¹ The data presented in this chapter were previously published in Human Mutation (Volume 39, Issue 35). I wrote the manuscript, assisted in user interface design, and collected, curated, and analyzed the data.
gathered from published literature, which are then submitted to ClinVar bi-annually. Active participation of the broader CMT community will provide an advance over existing resources for interpretation of CMT genetic variation.

2.2 Materials and Methods

Data Collection, Standardization, and Curation

The CMT genetic variation data was collected from multiple sources including the Human Gene Mutation Database, the Inherited Peripheral Neuropathies Mutation Database, results published by the Inherited Neuropathies Consortium, the supplemental data published by Athena and Quest Diagnostics, and manual data entry from collaborators. All collected and submitted data conform with institutional review boards and ethical guidelines. The initial set of genes was selected based on the published literature; however, users have the ability to add new genes as they are described as causes of CMT and related disorders. For phenotypic data, the CMT neuropathy score was imported from the Inherited Neuropathies Consortium’s natural history study when available. Additionally, each variant was annotated with the minor allele frequency from the Exome Aggregation Consortium.

The coding and protein sequence variant annotation follows recommendations of the Human Genome Variation Society. HUGO Gene Nomenclature Committee’s Multi-symbol checker tool was used to validate all gene symbols. Each variant’s syntax is standardized automatically with Mutalyzer Syntax Check, and the protein sequence variation was populated with Mutation
The Variant Effect Predictor was used to fill in missing variant effect data. All publications require a valid PubMed ID, and the title, author, and reference metadata is automatically retrieved and saved in the database.

Continuing the community-driven nature of this database, CMT researchers and clinicians have begun to adopt individual genes to curate. In the initial release, each adoptee verified the genetic variation data. For future curation, bi-annual reports of new submissions will be generated for manual review and curation.

**Data Submission**

The Inherited Neuropathy Variant Browser is designed for simple and fast variation upload. User registration and log-in is required for any variant submissions. The submitter must complete all required submission fields including: variant in HGVS notation, gene (drop-down selection of current CMT genes), protein notation, variant type, genotype, and NCBI mRNA reference sequence ID. For compound heterozygous variants, the submitter should select the “Add Compounded Variant” button to submit the variants together.

For data curation purposes, the data source is also required. The possible data sources are: clinical report (including clinical lab name), published paper (including Pubmed ID), or research finding (indicated by Sanger sequencing or Next-Generation Sequencing). A comment box allows for any additional information pertinent to the variant that the submitter wishes to include.

Each variant submission is assigned a randomly generated family ID to allow for genotypic data. This family ID is displayed in the public browser. The
submitter has the option to input a private family ID during submission. The private ID, along with the matched public ID, will be available only to the submitter in his/her account page. The purpose of the optional private family ID is to assist submitters with data entry tracking to minimize duplicate entries.

2.3 Results

Graphical User Interface

The variant browser is accessible to users through an online graphical user interface (http://hihg.med.miami.edu/neuropathybrowser). The website consists of four different main tabs: Home, Sign Up, Contact Us, and Log In (Figure 2.1). The Home tab is the homepage of the website and contains most of the functionality. It contains a quick search option along with three other multi-select boxes in order to filter for a gene, variant type, or data source (Figure 2.1 A-B). For the “gene” filter, users can select a single gene, multiple genes, or all genes. The “variant type” filter is grouped into three categories: loss-of-function (LoF), missense, and others. The “LoF” category includes frameshift insertions and deletions (INDELs), stop-gained, and splice altering variants; the “missense” category includes in-frame INDEL, stop-lost, and missense variant; the “others” category includes 3' untranslated region (UTR), 5' UTR, intergenic, noncoding exons, intronic, upstream of gene, and downstream of gene variants. Finally, the “data source” filter includes three options: clinical report, published paper, or research finding.

Query results are displayed in graph and table format. For each gene, the graphic result contains three viewing options: Allele, CMTNS (CMT neuropathy...
score)\textsuperscript{80}, or Exome Aggregation Consortium minor allele frequency (ExAC MAF)\textsuperscript{86} (Figure 2.1 C). Each plot uses the ‘Mutations Needle Plot’ and displays the known protein sites and domains, retrieved from NCBI. The default “allele” plot displays the protein location of each allele, count of observations, and variant type. The “CMTNS” and “ExAC” plots require available data for at least one variant within the gene in order to be displayed. The “CMTNS” plot shows CMTNS for each allele and the associated age at onset. The “ExAC” plot displays the ExAC MAF (release 0.3.1). Currently, not all genes have the full set of annotated information available.

The table result consists of seven sortable columns: gene, variant, protein notation, variant type, rating, links and data source (Figure 2.2). The data is grouped by primary variant into a gray expandable row (Figure 2.2 A). The expanded view shows detailed information about each variant (Figure 2.2 B). In the expanded view, each variant is grouped by the public family ID (Figure 2.2 C) and the observed genotype and zygosity (Figure 2.2 D). For example, the variant in Figure 2.2 E is shown twice to indicate its homozygous zygosity (HOM). Alternatively, Figure 2.2 F displays the variant in the compound heterozygous state (HET). For easy submission, additional families with the same genotype can be added by clicking the button next to the zygosity status (Figure 2.2 G). Each family’s data source can be viewed, edited, or deleted in the expanded view (Figure 2.2 H). Lastly, each individual variant can be rated through the variant rating system, described below (Figure 2.2 I).
Figure 2.1 Overview of Inherited Neuropathy Variant Browser homepage and search options

A) The quick search field displays an autocomplete dropdown option and can be queried by: gene name, variant coding position, variant protein position, publication title, publication author, variant type, and data source. B) The multi-select searches enable multiple selections across gene, variant type, and data source. These fields allow users to conduct complex queries, such as all missense and LoF variants reported in published papers in *MFN2*, *GJB1*, and *GDAP1*. C) For single gene queries, variants can be displayed in three plots: allele, CMTNS, and ExAC. These plots can be toggled or hidden by selecting the buttons above each plot.

**Variant Rating System**

The variant pathogenicity rating system is based upon the standardized terminology of the American College of Medical Genetics and Genomics (ACMG). The average rating of a variant is represented as a 5-star system in the table of query result. Each star corresponds to the ACMG terminology: benign, likely benign, uncertain significance, likely pathogenic, pathogenic (starting from 1
star rating). Registered users can rate a variant within the result table and can provide additional comments about their rating (Figure 2.2 I). A history of all ratings and comments will be maintained for each variant. Importantly, any individual rating allows for free-text commenting. This creates a track record of evidence in support of the specific rating.

Figure 2.2 Overview of query result table
Overview of query result table. A) The collapsed view of each variant is returned by default. This view displays a high-level overview of the variant including gene symbol, cDNA sequence, protein notation, variant type, rating, link outs to ExAC (E), OMIM (O), ClinVar (C), and NCBI (N), and the data sources. B) The expanded row view shows additional information including: C) the family-level information for each variant and D) the observed variant zygosity and relevant genotype. E) Variants observed in the homozygous state is displayed twice and is marked as ‘HOM’ while (F) variants observed in the compound heterozygous state are paired with the compounded variant and marked as ‘HET.’ G) The simple-add button allows a user to quickly add another family with the same observed variant(s). H) Buttons to view additional user comments, edit variant information, or delete a variant. I) In the collapsed view, the user can view the top-level variant rating. The star button allows the user to rate the variant and display the variant rating history.
**Database Summary Statistics**

The Inherited Neuropathy Variant Browser currently contains 3,809 unique variants within 82 genes. The genes currently contain a median of 16 variants, an interquartile range from 4.0 to 69.5, and a maximum of 720 variants in *GJB1* (Figure 2.3 A). A total of 4,558 unrelated families exist with the following genotypes: 2,244 heterozygous, 528 homozygous, and 301 compound heterozygous (Figure 2.3 B). Currently, 1,475 families were reported from clinical laboratories without genotypic information. All new submissions will require genotypic data. The currently available phenotypic data highlights the clinical variability of CMT with an age of onset range of 9 to 75 years and a CMT neuropathy score of 2 to 24 (Figure 2.3 C). The majority of variants are coding (Figure 2.3 D). However, we encourage the submission of non-coding variation to cover the full spectrum of the genetic architecture of CMT. As whole-genome sequencing usage increases we expect the amount of non-coding variation to expand. The variant rating system is becoming a popular feature with 179 ratings to date (Figure 2.3 E).
Figure 2.3  Inherited Neuropathy Variant Browser descriptive statistics.
A) Boxplot depicting the variant counts per CMT gene. B) Pie chart of variant zygosity counts. The unknown zygosity variants predominantly come from clinical lab reports that do not contain this information. C) Violin plots displaying available phenotypic data for a variant. D) Bar chart of variant functional consequences. Coding consequences are highlighted in blue. E) Bar chart of user ratings per gene.

Usage Examples

In order to illustrate use cases for the variant browser, we provide two examples that highlight current and future advantages.

1. Locus and allelic frequency are determinants of expected maximum tolerated population allele frequencies and directly affect pathogenicity evaluation. In addition, genic sub-regions, such as protein domains and exons, have been shown valuable in determining genetic tolerance and thus pathogenicity. For example, non-synonymous variants are observed across nearly all of GJB1,
while such variants are clustered at specific protein domains in DNM2 (Figure 2.4 A). By cataloging the genetic variation (and phenotypic details) observed in CMT genes, CMT-specific genetic tolerance metrics can be developed. Genetic tolerance metrics support clinical VUS interpretation as well as reveal insights into the biological mechanisms of disease genes. \(^{96}\)

2. The Variant Rating System is a low threshold tool that allows clinicians and researchers to share and record supporting evidence for variant interpretation. Figure 2.4 B shows a real example of a commentary from a user regarding conflicting data: the frequency of a variant in a healthy population (indicates benign) and abnormal electrophysiological data (indicates pathogenic). For quick interpretation from the table view, the variant is rated as two-star “likely benign” while the detailed view shows users’ comments. The intuitive user interface allows for simple and fast user interactions, which will greatly improve the community’s ability to interpret VUS.

2.4 Discussion
At a time when clinical multi-gene testing in highly heterogeneous Mendelian diseases is becoming a standard, the burden of VUS has risen to a point where it obstructs high quality diagnosis in individual patients. Neurologists and genetic counselors working with CMT patients are confronted by this issue daily. While there are potential solutions conceivable to address this problem, including future functional genomics platforms, a sensible and cost effective way forward is the comprehensive and public mapping of disease associated variation to disease-
Figure 2.4 Inherited Neuropathy Variant Browser usage examples.  
A) Allele plots displaying the missense and loss-of-function variants within GJB1 and DMN2. The number of families with an observed variant is shown on the y-axis and is emphasized by the size of the needle head. The protein position and domain information is displayed on the x-axis. Blue regions in DMN2 highlight the clustering of variation. B) Detailed variant rating history view displaying user comment and user ratings in collapsed table view.

causing genes. We have collected CMT specific genetic variation, along with genotypic and phenotypic data when available, from published literature, clinical lab reports, and our own in-house data. We then created the interactive, web-based Inherited Neuropathy Variant Browser (INVB) accessible to, and relying on participation from, CMT researchers and clinicians to view the collected CMT variation. We have implemented an interactive rating system of genetic variation based on standardized terminology from the ACMG.91 The rating history for each variant, along with comments from raters, will be maintained to provide a rich perspective to users. The Inherited Neuropathy Consortium is encouraging its global membership to submit observed pathogenic variation, VUS and
polymorphisms to the INVB. This platform enables a joint effort by the global CMT expert community to store, share, and discuss genetic variation to resolve VUS.

While the frequency of a DNA variant within an unaffected population helps to categorize benign variation, only repeated observation of a variant in affected families will strengthen pathogenic classification. The continued aggregation of both disease-associated and benign variation will remain essential to further resolve VUS.97 Currently, 25%-30% of variants in databases can be clearly classified as benign or pathogenic.98 We see an opportunity to create enthusiasm and attention in the CMT field to participate in the important task of collecting individual genetic test results from many clinics and different parts of the world. We were able to assemble an international group of experts committed to working together to classify variation and we hope the CMT community will serve as a positive example of joint enterprise within genomic medicine. The INC, the largest CMT-related consortium, has designated the interactive INVB as the CMT genetic tool of choice. The international character of the INC is attracting users from different geographic locations with diverse ancestries, such as the new Asia Oceanic Inherited Neuropathies Consortium. As minor allele frequencies can vary widely between ancestral populations, uniquely rare variants in one population might be less rare in another, thus excluding such an allele as pathogenic.

Similarly, large-scale population data, such as ExAC, a collection of over ~60,000 exomes from patients unaffected by severe pediatric diseases, has revealed previously reported pathogenic alleles as too common relative to the prevalence of CMT.99 We therefore incorporated ExAC data into the INVB.86
Benign variants falsely assigned as pathogenic have been revealed and reported disease variant penetrance has been re-evaluated, such as p.Arg468His in *MFN2*. This variant has been implicated in variant screening studies at least four times; yet, the relatively high count of 265 heterozygous alleles out of 115,542 ExAC chromosomes excludes the variant as a causative, high penetrance allele in a dominant rare disease.¹⁰⁰-¹⁰³

Though variant databases excel in providing overall allele frequencies, most capture genetic variants without listing individual patients and their genotypes; however, this information may contribute to variant interpretation.¹⁰⁴ Without knowing the heterozygous, homozygous or compound heterozygous state, it may be difficult to determine causality in a recessive disorder. Furthermore, the pathogenicity of a compound heterozygous pair of alleles is directly influenced by each involved allele. For CMT specifically, genotypic information is particularly important, because several genes, for example *GDAP1* and *HSPB1*, can cause disease in dominant and recessive inheritance modes depending on the specific variant.¹⁰⁵-¹⁰⁷ Finally, genotypic information may well guide future oligogenic models of inheritance in individual patients and influence predictable phenotypic expression. For these reasons, we require both the observed zygosity and phenotypic information with each variant submission. Beyond variant interpretation, storage of these datasets will enable future genotype-phenotype correlation analyses.

In order to benefit the wider scientific community, all collected data has been submitted to ClinVar – a public archive of the interpreted clinical significance of
variants for reported conditions at the National Center for Biotechnology Information.\textsuperscript{108} ClinVar was developed to provide access to the interpretation of human variation and is an effort critical to genomic medicine.\textsuperscript{109} As avid supporters of the ClinVar initiative, we recommend the entire community to submit variation data directly to the database. However, we have implemented direct submission to the INVB to collect data inappropriate for ClinVar (such as variants based solely on computational predictions) or from users who have chosen not to submit to ClinVar (such as CMT researchers outside of the United States). As we do not want to add to the issue of multiple fragmented data collection efforts with yet another database, we will perform 2-way synchronization with ClinVar at least bi-annually. Users should look to ClinVar for the most up-to-date variant database and utilize the INVB as a tool to view and discuss this data. We hope the simple and easy-to-use interface of the Inherited Neuropathy Variant Browser will facilitate communication about variant pathogenicity status among CMT researchers and clinicians and provide improved diagnostic abilities.
CHAPTER 3

Homozygosity Mapping of Whole-Exome Sequencing Detects Uniparental Isodisomy in Hereditary Spastic Paraplegia

3.1 Background

Despite the advances whole exome sequencing (WES) has brought to clinical genetics, a large portion of cases remains unresolved. Complicated modes of inheritance, such as uniparental disomy (UPD), can obscure a diagnosis under standard WES analysis. UPD is the inheritance of both chromosomal homologs from one parent. The affected region can be confined to a portion of the chromosome (segmental) or can span the entire chromosome (complete), and is an identical copy of one parental allele (isodisomy) or distinct alleles from the same parent (heterodisomy). Isodisomy can generate homozygosity of a deleterious recessive mutation from a heterozygous carrier-parent. Both clinical and research settings are impacted by undiagnosed isodisomy. Genetic counseling for the proband’s parents would not be accurate without knowledge of the isodisomic event. Isodisomy would also cause incongruent segregation analysis. In a research laboratory, where Sanger-confirmed segregation analysis is standard, an

---

2 The data presented in this chapter were previously published in Molecular Genetics & Genomics Medicine (Volume 5, Issue 3). I wrote the manuscript and performed all the work presented, with the exception of patient sample collection and clinical examination.
incongruent result will require time and resources to resolve. Accordingly, an accessible method to investigate these events is increasingly important to reach a diagnosis. Due to the contiguous stretches of homozygosity created by isodisomy, it is now detectable from single nucleotide polymorphism-based (SNP) data. A novel computational algorithm, $H^3M^2$, has been adapted for the sparse, irregular SNP distribution of WES.

In this study, we evaluate the feasibility of applying the $H^3M^2$ algorithm to the detection of UPD based on index WES data. We systematically studied the exomes of a phenotypically heterogeneous cohort of unresolved cases (n=96 families) to reveal UPD events hindering a diagnosis and to evaluate the prevalence of UPD in recessive MND and ATX. One hereditary spastic paraplegia case harbored homozygous regions spanning 80% of chromosome 16. A homozygous disease-causing mutation in the SPG35 disease gene was then identified within this region. This study demonstrates the ability to detect UPD in exome data of index patients. Our results suggest that UPD is a rare mechanism for recessive MND and ATX.

### 3.2 Materials and Methods

*Ethical Compliance*

The patients were collected from the University of Antwerp, the University Hospital Tübingen, and the University of Miami, and all participating individuals gave informed consent prior to initiating this study in agreement with the institutional review boards.
*Patient Inclusion*

We studied 96 unrelated cases from simplex families with a spectrum of rare neurodegenerative disorders from the ataxia-motoneuron spectrum and no reported consanguinity. Four consanguineous index cases were additionally included in the study as a positive control for the detection method. The cohort was derived from unresolved cases in the GENESIS database and included the following phenotypes: early onset ataxia (50 index cases), hereditary spastic paraplegias (35 index cases), and amyotrophic lateral sclerosis/frontotemporal dementia (11 index cases).\textsuperscript{116}

*Whole exome sequencing and variant analysis*

Whole exome sequencing was performed at the John P. Hussman Institute for Human Genomics at the University of Miami. Samples were enriched with the SureSelect Human All Exon 50 Mb kit (Agilent), underwent standard sample preparation for the Illumina Hiseq2000 platform and were processed using the GENESIS/GEM.app analysis pipeline.\textsuperscript{116} BAM files were retained as input for the H\textsuperscript{3}M\textsuperscript{2} isodisomy mapping. Each putative isodisomic region was closely investigated for high quality (read depth≥10 and genotype quality≥35), rare (1000 Genomes Project\textsuperscript{117} minor allele frequency≤0.01 and Exome Aggregation Consortium\textsuperscript{43} minor allele frequency≤0.01), homozygous, pathogenic variants within known disease genes gathered from the Genetic Testing Registry and Online Mendelian Inheritance in Man.
Isodisomy mapping

Regions of homozygosity were detected using $H^3M^2$ with recommended parameters (DNorm=100000, P1=0.1, P2=0.1). This algorithm takes a BAM file as input to calculate the ratio between the alternate allele reads and the total coverage at each polymorphic position (based on the 1000 Genomes Project). This ratio is used to determine the genotype state at each SNP. It is a discrete state Hidden Markov Model that includes a parameter to account for sequencing and alignment errors and a parameter within the transition probabilities matrix to account for the inconsistent distance between SNPs within WES.\textsuperscript{115}

3.3 Results

We analyzed 96 patients from non-consanguineous, simplex families with inherited motoneuron disease and ataxia (MND and ATX) to identify potential isodisomic events. The $H^3M^2$ algorithm was applied to each patient’s BAM file to detect long regions of homozygosity (Figure 3.1). Such extended homozygous regions could indicate uniparental isodisomy of a chromosome. Based on standard clinical analysis, homozygous regions below 2.5 Mb were filtered from the analysis.\textsuperscript{114} We used a minimum ROH length threshold to improve the signal-to-noise ratio for cohort-level distribution summaries. All cases contained at least one homozygous region $> 2.5$ Mb (Figure 3.1 A). Similar to previously published strategies, we defined “putative isodisomy” as the presence of a homozygous region over 10 Mb on just one chromosome.\textsuperscript{114} From the screened cohort, 29 cases contained at least one homozygous region greater than 10 Mb. Of these, five contained homozygous
regions on multiple chromosomes, indicating hidden consanguinity. The remaining 24 cases contained one or more homozygous regions over 10 Mb on a single chromosome (Figure 3.1 E) and were thus classified as putative isodisomy cases. To validate the putative UPD regions we selected high quality (read depth≥10 and genotype quality≥75), rare (1000 Genomes Project\textsuperscript{117} minor allele frequency≤0.001 and Exome Aggregation Consortium\textsuperscript{43} minor allele frequency≤0.001), homozygous variants within the possible UPD site for Sanger-based segregation analysis in 5 cases with available parental DNA (Figure 3.2). Additionally, parental DNA of 2 cases with consanguineous family background and 1 case without reported consanguinity yet multiple runs of homozygosity (ROH) > 10 Mb were available for genotyping.

The isodisomic mapping revealed a strong signal on chromosome 16 comprised of multiple large ROHs which sum to 72 Mb and cover 80% of the chromosome in one HSP case (SPG35, MIM #612319) (Figure 3.3 A-B), indicating putative complete isodisomy. During the same time as this study, this case was undergoing independent investigation for UPD at the University Hospital Tübingen. UPD was molecularly verified by excluding genomic deletions with multiplex ligation-dependent probe amplification assay and performing fragment analysis of PCR amplicons of microsatellite marks within the index patient and both parents. A homozygous p.Trp176* mutation was identified within \textit{FA2H} (MIM *611026, NM_024306.4), a known SPG35 disease gene. Sanger sequencing confirmed that only the father was a carrier of the identified variant.\textsuperscript{118}
Figure 3.1 Distribution of regions of homozygosity within screening cohort
Data are represented as “violin plots,” depicting the distribution of ROH counts per individual [(A) 2.5 – 5 Mb, (B) 5 – 10 Mb, and (C) >10 Mb] and the total ROH length (D) within the entire cohort (excluding the UPD16 case), the consanguineous controls, and the identified UPD16 case. The width of each “violin” represents the 90°-rotated kernel density trace and its reflection and the white dot shows the median. The “pie chart,” (E) displays the distribution of cases within the cohort by the amount of ROH greater than 10Mb each contains.
Figure 3.2 Uniparental isodisomy screening flowchart
Flowchart describing UPD screening in 96 sporadic cases and 4 consanguineous controls. Gray indicates consanguinity and false positives, blue indicates possible UPD, and orange highlights confirmed UPD.
In the remaining 4 cases with putative UPD and available parental DNA, segregation analysis was not consistent with UPD (Table 3.1). After closer inspection of these 4 cases, case 01 contains a 9.1 Mb ROH on a separate chromosome. Though this ROH is below our defined UPD threshold, it may indicate hidden consanguinity. All consanguineous controls contained homozygous regions greater than 10 Mb on multiple chromosomes as expected (Figure 3.3 C-D). Segregation analysis of each ROH greater than 10 Mb was performed on 2 consanguineous controls and 1 suspected consanguineous case with available parental DNA. Tested SNVs within the ROHs of both consanguineous controls and suspected consanguineous case were compatible with recessive inheritance.

Notably, we identified a region enriched for ROH on the q arm of chromosome 13, a gene poor region (average gene density below 3 genes per Mb) from 52.9 – 90.7 Mb.\textsuperscript{119} ROHs in this region were present in a total of 18 index cases in our cohort: 13 of 24 putative UPD (2 with parental DNA), 3 of 5 suspected consanguinity, and 2 of 4 consanguineous controls (Figure 3.4). The 3 cases with suspected consanguinity still contain ROH on multiple chromosomes even if the chromosome 13q ROH are removed. Segregation analysis of ROHs within this region was possible in 3 samples (cases 03 and 04 in Table 1; 1 consanguineous control), and revealed that the ROH were not compatible with UPD. The remaining 9 cases with putative UPD, but without parental DNA, contained ROH on separate chromosomes.
Figure 3.3 Isodisomy mapping from whole-exome sequencing
Isodisomy mapping of (A) the UPD16 case and (B) a consanguineous control. All ROH greater than 2.5 Mb are shown with a color overlay and ROH greater than 10 Mb are highlighted in yellow. (A) displays the complete isodisomic event on chromosome 16, while (B) shows multiple ROH greater than 10 Mb dispersed to multiple chromosomes.

<table>
<thead>
<tr>
<th>Segregation analysis of potential UPD events</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Case</th>
<th>Homozygous</th>
<th>Region</th>
<th>Tested Variant</th>
<th>Index</th>
<th>Father</th>
<th>Mother</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>1</td>
<td>192312018-203317394</td>
<td>Chr1:202698171C/T</td>
<td>T/T</td>
<td>C/T</td>
<td>C/T</td>
<td>recessive inheritance</td>
</tr>
<tr>
<td>02</td>
<td>12</td>
<td>93964833-109547815</td>
<td>Chr12:108912996G/A</td>
<td>A/A</td>
<td>G/A</td>
<td>G/A</td>
<td>recessive inheritance</td>
</tr>
<tr>
<td>03</td>
<td>13</td>
<td>61013944-72440942</td>
<td>Chr13:71725803G/A</td>
<td>A/A</td>
<td>G/A</td>
<td>G/A</td>
<td>recessive inheritance</td>
</tr>
<tr>
<td>04</td>
<td>13</td>
<td>80910758-92101290</td>
<td>Chr13:80940229G/A</td>
<td>A/A</td>
<td>G/A</td>
<td>G/A</td>
<td>recessive inheritance</td>
</tr>
<tr>
<td>UPD16</td>
<td>16</td>
<td>70284990-83842988</td>
<td>Chr16:74760209C/T</td>
<td>T/T</td>
<td>C/T</td>
<td>C/C</td>
<td>compatible with UPD</td>
</tr>
</tbody>
</table>

Table 3.1 Segregation analysis of potential UPD events
Segregation analysis for 5 sporadic cases with possible UPD based on the presence of a homozygous region greater than 10 Mb isolated to a single chromosome.
Figure 3.4 Regions of homozygosity in gene-poor region of chromosome 13
Chromosome 13 region showing A) the previously reported gene poor region across chromosome 13 (in orange), B) homozygous regions greater than 10 Mb; blue regions are sporadic index cases, gray regions are consanguineous control index cases; C) UCSC hg19 Known Gene track.
3.4 Discussion

Even though WES is a powerful tool for diagnosing genetic diseases, approximately 75% of cases remain undiagnosed after WES analysis. In order to increase the amount of diagnoses, WES should be analyzed for rare disease-causing mechanisms whenever possible. In this study, we applied the novel $H^3M^2$ algorithm to an index patient’s WES and successfully detected complete uniparental isodisomy. Because of $H^3M^2$’s basic hardware requirements and quick computational speed, it can rapidly be used on existing WES or readily incorporated into an exome analysis pipeline. Incorporation of such a ROH detection tool into an exome analysis pipeline is preferable to manual detection as it allows for objective, scalable, and high-throughput results while accounting for inconsistent distances between SNPs. $H^3M^2$ achieves more accurate ROH identification by including a parameter which accounts for sequencing and alignment errors. Additionally, $H^3M^2$ is independent of genotype-caller errors as it calculates the $B$-allele frequency directly from the alignment file. Isodisomy mapping is a powerful technique as it provides a simple way to extract information critical to a diagnosis solely from a proband’s WES. This method is particularly equipped to detect complete isodisomy from a single exome; however, due to small segmental background homozygosity in the general population, it does not fully capture small segmental isodisomies without trios analysis. Though trio studies are the most informative analysis for all uniparental disomies, they are not always available or financially practical. Therefore, isodisomy mapping provides
efficient means of distinguishing, which cases warrant further confirmation of
uniparental isodisomy from already available patient data.

The criteria for defining a putative isodisomy event will likely need to be
adjusted to improve the rate of true positives. In addition to homozygous regions
over 10 Mb on a single chromosome, the cumulative percentage of homozygous
regions across a chromosome may be considered. Although the \( H^3M^2 \) algorithm is
designed to handle the non-uniform distribution of the whole exome, our
experience indicates that gene poor regions may produce false positive calls for
UPD. Half of the detected putative UPD cases \( (n=12/24) \) were located within a
gene poor region on chromosome 13q. Segregation analysis did not support UPD
in two of these cases (Table 3.1). We suspect these ROH are primarily generated
due to unusually long non-informative regions of the genome. In this study, we
identified the chromosome 13q region initially by the high frequency of ROH within
the region and followed by a literature search. We did not observe additional
regions with high ROH frequency suggestive of a false-positive region. As our
experience with this method grows, this and possibly other gene poor regions, will
likely have to be excluded from analysis.

Beyond detection of complete isodisomy, we were also able to identify
potential hidden consanguinity in approximately 5% of our cohort. Hidden
consanguinity is not usually addressed during WES analysis as it is not considered
to be a frequent event. Our results challenge this standard view and indicate that
distant consanguinity may occur more often than expected. As such, this approach
additionally helps to unravel hidden consanguinity within cases without a known 
family history.

We sought to evaluate the prevalence of UPD within a cohort of recessively 
inherited motoneuron diseases and ataxias (MND and ATX). Only complete 
isodisomy was considered for the prevalence analysis, as this screening method 
is not well suited for detection of small segmental isodisomies. We observed one 
complete isodisomy within 96 screened cases (1.04%) which is not significantly 
different from the four complete isodisomies observed within the Deciphering 
Developmental Disorders (DDD) cohort\textsuperscript{113} (\(P\)-value is 0.356 [Fischer’s exact test]). 
However, like the DDD cohort, our cohort is significantly different than the general 
UPD rate of 1 in 3500 births\textsuperscript{1} (\(P\)-value is 0.0267 [binomial test]). Our results 
suggest that complete isodisomy occurs more frequently in recessively inherited 
MND and ATX than in the general population. Notably, this frequency will likely 
increase when disease-causing segmental isodisomies are considered. However, 
considering the recently reported cases of UPD in spastic paraplegia type 35 
cases, it is unclear whether our observed frequency is driven by a SPG35-specific 
phenomenon.\textsuperscript{118}

In summary, isodisomy mapping can detect complete isodisomy from WES 
and can be easily implemented into a bioinformatics workflow. Clinically, isodisomy 
should be especially suspected in cases with homozygous variants in the absence 
of a consanguineous family history and prompt further analysis. Currently, 
complete isodisomy appears to be a rare mechanism of disease in recessively 
inherited MND and ATX.
CHAPTER 4

Gene-based Association Analysis and Cumulative Mutational Burden in Inherited Axonopathies

4.1 Background

Inherited axonopathies (IA) are rare, clinically and genetically heterogeneous diseases that lead to length-dependent degeneration of the long axons in central and peripheral nervous systems. Mendelian high-penetrance alleles in over one hundred different genes have been shown to cause IA; yet, more than 50% of patients do not receive a genetic diagnosis. A more comprehensive spectrum of genes and alleles is warranted, including causative and risk alleles, as well as oligogenic inheritance. Through international collaboration, IA exome studies are beginning to be sufficiently powered to perform rare variant burden analysis. After extensive quality control, our cohort contained 343 CMT cases, 515 HSP cases, and 935 non-neurological controls. We assessed the cumulative mutational burden across disease genes, explored the evidence for oligogenic inheritance, and perform an exome-wide rare variant burden analysis.

4.2 Materials and Methods

Ethical Compliance

The patients were collected from the University of Miami, Children’s Hospital of Philadelphia, the University Hospital Tübingen, and McGill University, and all
participating individuals gave informed consent prior to initiating this study in agreement with the institutional review boards.

Sample Criteria

Families included in the study are affected by inherited axonopathy: either Charcot-Marie-Tooth disease (CMT) or Hereditary Spastic Paraplegia (HSP). CMT cases were diagnosed with CMT (type 1, 1A, 2, 4, or intermediate), distal hereditary motor neuropathy, hereditary sensory autonomic neuropathy, or hereditary sensory neuropathy, while HSP cases were diagnosed with pure HSP or complicated HSP. To avoid bias due to relatedness, only the proband from each family was included. Based on this criteria, 396 CMT probands and 569 HSP probands were selected for the study. Controls were supplied from the Children's Hospital of Philadelphia and did not display a neurological phenotype. In total, we received 938 unrelated samples with reported similar ancestry and demographics as the cases.

Whole exome sequencing

Whole exome sequencing was performed at the John P. Hussman Institute for Human Genomics at the University of Miami (CMT and HSP cases), the Montreal Neurological Institute and Hospital (HSP cases), and at the Children’s Hospital of Philadelphia (controls). Samples were enriched with the SureSelect Human All Exon 50 Mb kit (Agilent), versions 3, 4 or 5 for cases and version 5 for controls. All samples were sequenced on the Illumina HiSeq2000 platform and were processed
according to the Genome Analysis Tool Kit (GATK v.3.3) germline whole exome best practices. Briefly, samples were aligned to the 1000 Genomes Project b37+decoy reference genome (hs37d5) with BWA MEM (v.0.7.12). Following alignment, duplicate reads were marked with Picard and sorted with Samtools. Base recalibration was performed using GATK's recommend known variant and indel sites. Variant discovery was completed with GATK HaplotypeCaller to produce individual GVCFs. Case and control GVCFs were merged for multi-sample joint genotype calling with GATK GenotypeGVCFs. To mitigate genotyping errors caused by the heterogeneity of sequence capture kits, the combined multi-sample GVCF was filtered for genomic regions contained within each capture (at least 70% reciprocal overlap required, n=155,686). GATK Variant Quality Score Recalibration was performed on the called variants, and the 99.5 and 99.0 tranches were selected for SNPs and indels, respectively.

Quality control

Whole exome sequencing was performed at different sites with multiple capture kits; thus, strict quality measures will be applied to reduce sample bias. Sample quality control was performed at multiple points during the pipeline. After alignment, each sample was evaluated based on Picard MarkDuplicates output metrics: percent duplication and estimated library size. PLINK (v1.07) was used to perform sex and relatedness quality control. Sex assignments were verified based on X chromosome inbreeding coefficients (minor allele frequency $\geq 0.01$, missing rate $\leq 0.05$, F estimates $\leq 0.2$ yields female calls, F estimates $\geq 0.8$ yields male
calls). Pairwise relatedness estimation was computed with a trimmed set of independent variants (n=55,246 variants, autosomal variants only, minor allele frequency ≥ 0.01, missing rate ≤ 0.05, and linkage disequilibrium r^2 ≤ 0.35). Pairwise relatedness estimates were used to filter one sample from each following pair: second-degree and higher relatives (π-hat ≥ 25%), duplicated samples, contaminated samples, and unclear relationships. Finally, each sample was evaluated based on the VCFtools metrics per sample missing rate (missing rate ≤ 0.02) and per sample average depth (average depth ≥ 15). Samples with poor average genotypic depth across all autosomes were removed (depth ≤ 15). Altogether, 54 samples (38 CMT cases, 55 HSP cases, 7 controls) were removed from the analysis based on quality control.

**Variant annotation and filtering**

The following QC thresholds were applied for the variant call set with VCFtools: (i) GATK truth sensitivity 99.5% for single nucleotide variants (SNVs) and 95% for indels, (ii) call rate ≥ 90% in the 1,849 quality controlled samples. 1,149,823 out of 17,429,532 variants called passed the QC thresholds. The 90% call rate threshold was set to minimize batch effects between the WES sequencing site and target capture regions. Ensembl’s Variant Effect Predictor Human GRCh37 was used to annotate the passing variants with population allele frequency and variant consequence. The following criteria was applied: (i) MAF ≤ 0.01 in the Exome Aggregation Consortium (ExAC), (ii) not co-located with existing variant with global AF > 0.01 from 1000 Genomes Project Phase 3 data, (iii) protein-changing variants
according to the merged hg19 Reference Sequence (RefSeq) and Ensembl cache (stop-gain/loss, splice acceptor/donor, frameshift indels, transcript ablation, and non-synonymous variants). Out of 1,149,823 QC-passing variants, 217,793 were selected for further analysis (Figure 4.1).

*Gene-based rare variant association analysis*

Protein coding regions were defined based on the human GRCh37.75 gene annotation (n=22,810). Exome-wide gene-based association tests were performed using PLINK/SEQ suite. To detect risk and protective variants, we employed the two-sided C-alpha test to each protein coding region (MAF ≤ 0.02). Briefly, the C-alpha test statistic compares the distribution, rather than the mean, of rare variants for a given region. An adaptive permutation was used to assess the empirical significance of each association test (phenotypic label swapping across all samples with genes dropped from further permutation if test will clearly not reach significance). PLINK/SEQ provides an additional statistic, the i-statistic, which gives a sense of the smallest possible empirical p-value. Following recommended protocol, we removed tests with an i-statistic greater than $10^{-3}$. This filtering step is comparable to removing low minor allele frequency SNPs in genome-wide association studies. Bonferroni correction was applied to the genes for which there was adequate power for association.
**Cumulative mutational burden across disease genes**

To compare the mutational burden across known disease genes (CMT: \(n=88\), HSP: \(n=95\)), the number of rare variants (non-synonymous or loss-of-function at ExAC MAF \(\leq 0.01\) and \(\leq 0.001\)) within disease genes was computed for each inherited axonopathy case. CMT cases were assessed for variation in CMT genes, and HSP cases were assessed for variation in HSP genes. Following a similar method\(^72\), the average number of rare variants was compared between CMT or HSP cases and control samples using a non-parametric Mann-Whitney-Wilcoxon test. We calculated the statistical likelihood of this outcome by performing an affection status permutation procedure with 10,000 iterations.

**Digenic and oligogenic inheritance across disease genes**

After observing a significant mutational burden across disease genes, we next assessed the exact number of variant-carrying genes in each sample. For cases and controls, the number of disease genes carrying at least one qualifying variant (non-synonymous or loss-of-function at ExAC MAF \(\leq 0.01\) and \(\leq 0.001\)) was calculated. Samples that carried qualifying variants in a single gene were classified as monogenic while samples that carried qualifying variants in \(\geq 2\) genes were classified as di/oligogenic and samples that carried qualifying variants in \(\geq 3\) genes were classified as oligogenic. The frequency of cases and controls classified as each inheritance type was organized into 2x2 contingency tables and assessed by Fisher’s exact test.
**Figure 4.1 Quality control of joint-called variant set**
First three principal components of cases and controls to explore potential site-specific bias.

### 4.3 Results

**Association of EXOC4 with CMT cases**

Exome-wide association analysis was performed at 17,637 protein coding loci by the C-alpha test. The PLINK/SEQ suite computes an estimate of the minimal achievable $p$-value for a locus, the $i$-statistic. To correct for multiple testing, we followed recommended protocol to filter out loci with an $i$-statistic greater than $10^{-3}$ before Bonferroni correction.\(^{121}\) Filtering based on the $i$-statistic is recommended to remove genes without power to find an association. Based on the 2,145
remaining loci, the $p$-value threshold for an experiment-wide significance (alpha=0.05) was 2.3 x $10^{-5}$. After filtering results by the PLINK/SEQ i-statistic and applying Bonferroni multiple-testing correction, three genes, $KDM5A$ ($p$-value=9.9x$10^{-7}$, OR=3.6), $EXOC4$ ($p$-value=6.9x$10^{-6}$, OR=2.6), and $CEP78$ ($p$-value=2.3x$10^{-5}$, OR=4.4), reached experiment-wide significance (Figure 4.2). $KDM5A$ and $EXOC4$ both contained a single allele in cases that drove the association: $KDM5A$ NM_001042603.1:c.11T>G and $EXOC4$ NM_021807.3:c.1648G>A. Sanger sequencing confirmed the driver allele in $EXOC4$ (Figure 4.2 A-B) and revealed a false call in $KDM5A$. We did not follow up with $CEP78$ since the gene did not contain a single driving allele. At the gene-level, cases are 2.6 (95% CI: 1.28-5.37) times more likely to carry a heterozygous mutation than controls, while at the driver allele-level, cases are 9.07 times more likely to be a heterozygous carrier (95% CI: 2.94-28.01; Figure 4.2 C). Control carrier status was not significantly different from GnomAD population frequency, indicating that this allele was adequately covered in our control cohort (Figure 4.2 D).

**Increased mutational burden across known disease genes in IA cases**

After stringent quality control, 343 CMT cases, 515 HSP cases, and 931 controls were included in analysis. Inherited axonopathy cohorts were independently tested for a mutational burden (an excess of rare variants) across known disease genes. Two smaller CMT patient cohorts ($n=40$ and $n=32$) have been previously shown to harbor a rare non-synonymous mutational burden, which may potentially contribute to or modify phenotypic variability. In our CMT and HSP cohorts, we
Figure 4.2 Gene-based association analysis of CMT cases
Q-q plot of the observed p-values from gene-based association analysis. Blue line indicates multiple testing correction threshold. Known CMT genes with nominal significance are annotated.

Figure 4.3 EXOC4 carrier risk
identified a significant mutational burden (Mann-Whitney, nominal $p$-value $\leq 0.05$) in each tested variant set (non-synonymous and loss-of-function variation at ExAC MAF $\leq 0.001$ or $\leq 0.01$; Figure 4.4). As a further test of our observations, we repeated the mutational burden comparison with permutated case/control status over 10,000 iterations. We found that each tested variant set remained statistically significant (empirical $p$-value $\leq 0.05$), thus supporting that the mutational burden found across disease genes is specific to each inherited axonopathy cohort.

Di/oligogenic inheritance suggested in IA cases

Next, we sought to determine whether the observed mutational burden was more likely to follow a monogenic (single gene), digenic (two genes), or oligogenic (more than two genes) inheritance. Unlike the mutational burden, the significance of each inheritance type was influenced by the minor allele frequency (Figure 4.5). HSP cases showed consistent evidence for oligogenic inheritance ($\geq 3$ genes) of non-synonymous (NS) variation and monogenic inheritance (1 gene) of loss-of-function (LoF) variation at both ExAC MAF $\leq 0.01$ and $\leq 0.001$ (Fisher’s exact, $p$-value $\leq 0.05$). HSP cases also displayed significant di/oligogenic inheritance ($\geq 2$ genes) of NS variation at the less common ExAC MAF $\leq 0.001$ ($p$-value $\leq 0.05$). Furthermore, di/oligogenic inheritance of both NS and LoF variation for HSP cases is suggested at ExAC MAF $\leq 0.01$ ($p$-value $= 0.0598$ and $0.0572$, respectively). Evidence for inheritance types in CMT was not as consistent as in HSP, possibly due to a lower CMT sample size. At ExAC MAF $\leq 0.01$, CMT cases demonstrated
monogenic inheritance for LoF variation and oligogenic inheritance for NS variation ($p$-value $\leq 0.05$) with potential di/oligogenic inheritance for NS variation ($p$-value = 0.521). Lastly, at ExAC MAF $\leq 0.001$, CMT cases only showed significant evidence for monogenic inheritance of NS variation ($p$-value $\leq 0.05$) with potential evidence for oligogenic NS inheritance and monogenic LoF inheritance ($p$-value = 0.0641 and 0.0536, respectively.). Table 4.2 summarizes the counts of mutated genes in cases and controls.

**Figure 4.4 Cumulative mutational burden across known disease genes**

(A&B) Distribution of the number of qualifying non-synonymous (A) and loss-of-function (B) variants across HSP and CMT disease genes with varying ExAC minor allele frequency thresholds. (* $\leq 0.05$, ** $\leq 0.0001$, Mann-Whitney p-value).

(C&D) Difference of case and control average qualifying variants per sample with 95% confidence intervals.
Table 4.1  Summary of cumulative mutational burden across known IA disease genes
Descriptive statistics summarizing the central tendency and variability of rare non-synonymous and loss-of-function variation in known CMT and HSP disease genes. Additionally included are the Mann-Whitney test p-value and the ratio of case/control means.

<table>
<thead>
<tr>
<th>disease</th>
<th>consequence</th>
<th>exac maf</th>
<th>pvalue</th>
<th>case/control ratio</th>
<th>mean case</th>
<th>stddev case</th>
<th>median case</th>
<th>n case</th>
<th>mean control</th>
<th>stddev control</th>
<th>median control</th>
<th>n control</th>
</tr>
</thead>
<tbody>
<tr>
<td>cmt</td>
<td>lof</td>
<td>0.001</td>
<td>6.62E-07</td>
<td>2.07</td>
<td>0.140</td>
<td>0.068</td>
<td>0.386</td>
<td>0.345</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>lof</td>
<td>0.01</td>
<td>1.49E-06</td>
<td>1.92</td>
<td>0.157</td>
<td>0.082</td>
<td>0.401</td>
<td>0.368</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>lof</td>
<td>0.05</td>
<td>1.07E-05</td>
<td>1.78</td>
<td>0.162</td>
<td>0.091</td>
<td>0.406</td>
<td>0.379</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>ns</td>
<td>0.001</td>
<td>7.81E-08</td>
<td>1.29</td>
<td>1.683</td>
<td>1.302</td>
<td>1.280</td>
<td>1.257</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>ns</td>
<td>0.01</td>
<td>3.79E-05</td>
<td>1.16</td>
<td>3.025</td>
<td>2.605</td>
<td>1.771</td>
<td>1.695</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>ns</td>
<td>0.05</td>
<td>9.11E-04</td>
<td>1.11</td>
<td>3.457</td>
<td>3.106</td>
<td>1.894</td>
<td>1.830</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>357</td>
</tr>
<tr>
<td>hsp</td>
<td>lof</td>
<td>0.005</td>
<td>5.00E+00</td>
<td>3.24</td>
<td>0.449</td>
<td>0.131</td>
<td>0.758</td>
<td>0.399</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>lof</td>
<td>0.01</td>
<td>5.00E+00</td>
<td>3.24</td>
<td>0.474</td>
<td>0.146</td>
<td>0.770</td>
<td>0.418</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>lof</td>
<td>0.05</td>
<td>5.00E+00</td>
<td>3.24</td>
<td>0.474</td>
<td>0.146</td>
<td>0.770</td>
<td>0.418</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>ns</td>
<td>0.001</td>
<td>5.00E-10</td>
<td>1.26</td>
<td>3.524</td>
<td>2.807</td>
<td>2.093</td>
<td>1.750</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>ns</td>
<td>0.05</td>
<td>8.50E-09</td>
<td>1.20</td>
<td>4.206</td>
<td>3.511</td>
<td>2.198</td>
<td>1.945</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>515</td>
</tr>
</tbody>
</table>
### Table 4.2  Summary of digenic/oligogenic variation across known IA disease genes

Descriptive statistics summarizing the count of samples carrying a qualifying variant in 1 (monogenic), 2 (digenic), 2+ (di/oligogenic), and 3+ (oligogenic) known disease genes. Additionally included are the p-values from Fisher’s exact test of case and control carrier counts, odds ratios, and the carrier proportions.

<table>
<thead>
<tr>
<th>disease</th>
<th>consequence</th>
<th>inheritance</th>
<th>exact maf</th>
<th>pvalue</th>
<th>odd ratio</th>
<th>carrier count</th>
<th>n</th>
<th>carrier proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>case</td>
<td>control</td>
<td>case</td>
</tr>
<tr>
<td>cmt</td>
<td>lof</td>
<td>1</td>
<td>0.001</td>
<td>&lt;0.0001</td>
<td>3.61</td>
<td>43</td>
<td>34</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>lof</td>
<td>2</td>
<td>0.001</td>
<td>0.8246</td>
<td>1.04</td>
<td>2</td>
<td>5</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>lof</td>
<td>1</td>
<td>0.01</td>
<td>&lt;0.0001</td>
<td>3.06</td>
<td>49</td>
<td>46</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>lof</td>
<td>2</td>
<td>0.01</td>
<td>0.6334</td>
<td>1.31</td>
<td>2</td>
<td>4</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>ns</td>
<td>2+</td>
<td>0.001</td>
<td>&lt;0.0001</td>
<td>1.66</td>
<td>170</td>
<td>329</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>ns</td>
<td>3+</td>
<td>0.001</td>
<td>0.0027</td>
<td>1.57</td>
<td>79</td>
<td>143</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>ns</td>
<td>2+</td>
<td>0.01</td>
<td>0.0033</td>
<td>1.5</td>
<td>284</td>
<td>672</td>
<td>357</td>
</tr>
<tr>
<td>cmt</td>
<td>ns</td>
<td>3+</td>
<td>0.01</td>
<td>0.0013</td>
<td>1.52</td>
<td>200</td>
<td>425</td>
<td>357</td>
</tr>
<tr>
<td>hsp</td>
<td>lof</td>
<td>1</td>
<td>0.001</td>
<td>&lt;0.0001</td>
<td>3.21</td>
<td>140</td>
<td>97</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>lof</td>
<td>2</td>
<td>0.001</td>
<td>&lt;0.0001</td>
<td>6.88</td>
<td>22</td>
<td>6</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>lof</td>
<td>1</td>
<td>0.01</td>
<td>&lt;0.0001</td>
<td>3.02</td>
<td>145</td>
<td>107</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>lof</td>
<td>2</td>
<td>0.01</td>
<td>&lt;0.0001</td>
<td>6.13</td>
<td>26</td>
<td>8</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>ns</td>
<td>2+</td>
<td>0.001</td>
<td>&lt;0.0001</td>
<td>1.79</td>
<td>267</td>
<td>350</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>ns</td>
<td>3+</td>
<td>0.001</td>
<td>&lt;0.0001</td>
<td>2.33</td>
<td>143</td>
<td>132</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>ns</td>
<td>2+</td>
<td>0.01</td>
<td>&lt;0.0001</td>
<td>1.81</td>
<td>429</td>
<td>683</td>
<td>515</td>
</tr>
<tr>
<td>hsp</td>
<td>ns</td>
<td>3+</td>
<td>0.01</td>
<td>&lt;0.0001</td>
<td>1.49</td>
<td>320</td>
<td>448</td>
<td>515</td>
</tr>
</tbody>
</table>
Figure 4.5 Digenic/oligogenic inheritance across known disease genes
(A&B) Difference in proportion of samples carrying non-synonymous (A) and loss-
of-function (B) variants in multiple HSP and CMT disease genes with varying ExAC minor allele frequency thresholds. (C&D) Difference in proportion of case and control carriers with 95% confidence intervals. (* <= 0.05, ** <= 0.0001, two proportion z-test p-value).

4.4 Discussion

As the cost and availability of next-generation sequencing continues to drop, we are now reaching large enough sample sizes to apply statistical approaches to rare diseases. In this study, we sought to assess the mutational burden and di/oligogenic involvement of rare variation in a cohort of inherited axonopathies as well as identify potential risk loci. We performed an unbiased exome-wide rare variant burden analysis with the C-alpha test. After filtering results and performing
Sanger sequencing, *EXOC4* stands out as a candidate CMT gene. *EXOC4* is involved in vesicle transport and membrane tethering in polarized cells and is expressed in Schwann cells.\textsuperscript{123} In a human neuropathy mouse model (CMT4B1, *Mtmr2*-null mouse), Exoc4 (Sec8) formed a complex with Mtmr2 and Dlg1 to coordinate homeostatic control of myelination.\textsuperscript{123} Exoc4 is abundantly expressed at the *Drosophila* neuromuscular junction (NMJ) and required for in vivo regulation of synaptic microtubule formation.\textsuperscript{124} Furthermore, results suggest that Exoc4 plays a central role in oligodendrocyte membrane formation through the regulation of vesicular transport of myelin proteins, indicating potential central and peripheral nervous system involvement.\textsuperscript{125}

In addition to identifying *EXOC4* as a candidate CMT gene, we were also able to re-identify several established monogenic CMT genes, including *MME*\textsuperscript{55}, *MORC2*\textsuperscript{126} and *MFN2*\textsuperscript{35}, the most common CMT2 gene. This is despite a general effort to exclude patients with *MFN2* and other common CMT genes from exome analysis. Similarly, known familial ALS genes showed strong associations in a gene-based rare variant burden analysis of sporadic ALS cases.\textsuperscript{52,53} We hypothesize that this indicates the presence of additional risk alleles in these known CMT genes that contribute to the phenotype.

We also observed a significant mutational burden across CMT and HSP disease genes in cases compared to non-neurological controls. This enrichment of rare, damaging alleles may contribute to risk, severity, and clinical heterogeneity. Previous *in vivo* zebrafish studies observed increased phenotypic severity when pairs of neuropathy genes were inactivated.\textsuperscript{72} We observed non-
synonymous variation significantly distributed across 2 or 3 disease genes, indicating digenic/oligogenic inheritance. Oligogenic inheritance remains underexplored in rare diseases because of functional validation challenges. However, additional variants in multiple disease genes may either have a combinatorial effect on the same biological pathways or a destabilizing effect on the entire disease module. In summary, statistical methods traditionally reserved for more ‘common’ phenotypes are becoming increasingly available for rare disease genetics and will help to comprehensively define the genetic architecture of heterogeneous rare neurodegenerative disorders.
CHAPTER 5

Identification of Cryptic Amyloidogenic Elements in Genome-Wide Screen

The story of the toxic protein aggregation of the neurofilament heavy gene (*NEFH*) is primarily the work of Dr. Adriana Rebelo, who alongside Dr. Stephan Züchner, formulated the original hypothesis and experimental design. Dr. Rebelo identified the families harboring pathogenic *NEFH* variants, hypothesized the 3’UTR aggregation pathomechanism, and performed the cellular experiments to test the hypothesis. I contributed to the project by performing a genome-wide screen of 3’UTR sequences in each reading frame to identify additional cryptic amyloidogenic elements. The following data presented in this chapter were previously published in the *American Journal of Human Genetics* (Volume 98, Issue 4), with myself as a co-author.\(^\text{20}\)

5.1 Background

Abnormal protein aggregation is associated with a wide range of neurodegenerative diseases including Huntington disease, Parkinson disease, and Alzheimer disease.\(^\text{127-130}\) In motor neuron diseases, including amyotrophic lateral sclerosis (ALS), giant axonal neuropathy, and Charcot-Marie-Tooth disease (CMT), abnormal aggregation and misassembly of neurofilament (NF) have been reported; however, the causes of NF accumulation remains unclear.\(^\text{131-134}\) NFs are a class of intermediate filaments comprised of light (NEFL)-, medium (NEFM)-, and
heavy (NEFH)-molecular weight subunits. They are expressed exclusively in neurons and provide structural support for axonal diameter, growth, and trafficking. In addition to NF accumulation, mutations in NFs have also been associated with multiple neurodegenerative diseases. Here, we report the first mutations in NEFH to cause axonal neuropathy (CMT2) as well as a novel protein aggregation mechanism caused by the translation of the 3' UTR of NEFH. The identified NEFH mutations are distinct frameshift variants that lead to the loss of the terminating codon and translation of an extra 40 amino acids within the 3'UTR. In silico aggregation prediction suggested that the terminal 20 residues were amyloidogenic, which was confirmed experimentally by serial deletion analysis. A genome-wide scan of additional cryptic amyloidogenic elements within the 3' UTR identified NEFL and FUS, in which mutations are known to cause aggregation in CMT and ALS, respectively.

5.2 Materials and Methods

Families
The families were identified as part of our ongoing genetic studies in CMT. The families were ascertained in Austria and Great Britain. Participants were recruited, enrolled, and sampled according to the protocols of the institutional review board at the Universities of London and Vienna. A complete description of the study was provided to the subjects, and written informed consent was obtained. Whole blood was collected from all participants by venipuncture. Affection status was
determined by consensus of physicians and clinical staff experienced in clinical CMT research and was based on medical records and in-person evaluation.

**Sanger and Exome Sequencing**

We performed exome sequencing in 269 index individuals with autosomal-dominant CMT. 16 CMT-affected families were from the UK, and 48 families were from Austria. The remaining families were from several other countries from Europe and North and South America. The SureSelect Human All Exon 50 MB Kit (Agilent) was used for in-solution enrichment, and the HiSeq 2500 instrument (Illumina) was used to produce 100 bp paired-end sequence reads. The Burrows-Wheeler aligner, Picard, and the Genome Analysis Toolkit were used to align sequence reads and call variants. These data were imported into GENESIS (formerly GEM.app) for further analysis.\(^{116,142}\) Variants were filtered for those that segregated in an autosomal-dominant fashion and met the “strict” criteria, which required that variants be rare (minor allele frequency < 0.05\% in the National Heart, Lung, and Blood Institute [NHLBI] Exome Sequencing Project [ESP] Exome Variant Server [ESP6500]), be present in fewer than three families within GEM.app (~4,300 exomes), be conserved (GERP score > 2 or PhastCons score > 0.6), and have sufficient quality scores (genotype quality > 75). Mutations in known CMT-associated genes were absent in both Austrian and UK families. NEFH-variant calls identified were validated by conventional Sanger sequencing. In addition, we had access to 5,200 control samples from the GENESIS database. These samples broadly included 1,824 neuromuscular disorders (of which 470 were peripheral
neuropathies), 286 cardiomyopathies, 188 dementia disorders, and 509 deafness disorders.

5.3 Results

Identification of NEFH Frameshift Variants in CMT2-Affected Families

We performed whole-exome sequencing (WES) on three family members (UK1) affected with autosomal dominant axonal peripheral neuropathy (CMT2, Figure 5.1 A). WES processing and analysis was performed with the GENESIS software. After a strict variant filtering approach, we identified a heterozygous frameshift variant in NEFH that co-segregated within the family across three generations. The variant (c.3010_3011delGA [p.Asp1004Glnfs*58] at chr22:29,886,637 from hg19) creates a frameshift mutation that removes the original stop codon, thus causing continued translation of an alternative open reading frame (ORF) with an extension of 40 amino acids to the C-terminal tail of NEFH (Figure 5.2 A). As this mutation occurs near the end of the protein, the major functional domains remain intact (Figure 5.1 C). We identified an additional CMT2 family (AT1) with a heterozygous frameshift mutation in NEFH that co-segregated within four affected individuals after screening whole exomes from an additional 322 CMT families (Figure 5.1 B). Interestingly, this variant also creates a stop-loss mutation which results in translation of the identical ORF (Figure 5.2 A).
Figure 5.1 *NEFH* Frameshift Variants in CMT2-Affected Families

Asterisks indicate probands. (A) Pedigree and Sanger sequence traces of the CMT-affected family carrying the *NEFH* variant c.3010_3011delGA (p.Asp1004Glnfs*58). Abbreviations are as follows: M, mutant c.3010_3011delGA allele; and +, wild-type allele. (B) Pedigree and Sanger sequence traces of the CMT-affected family carrying the *NEFH* variant c.3017_3020dup (p.Pro1008Alafs*56). Abbreviations are as follows: M, mutant c.3017_3020dup allele; and +, wild-type allele. (C) Diagram shows *NEFH* domains and variants associated with diseases. Coding KSP deletions and insertions from reported ALS individuals are represented by triangles. CMT frameshift variants from families UK1 and F2 are indicated by arrows.
Figure 5.2 Identification of CAEs Encoded by NEFH and NEFL 3’ UTRs

(A) Clustal Omega multiple-sequence alignment of wild-type NEFH and frameshift variants harbored by the CMT-affected families. Translation of the 3’ UTR open reading frames (ORFs) is illustrated. (B) TANGO score of NEHF 3’ UTR ORFs. (C) Consensus sequence of positive residues (asterisks) for all aggregation predictors tested for NEFH 3’ UTR ORF3.

Identification of Cryptic Amyloidogenic Elements Encoded by NEFH 3’ UTR

Based on NFs known tendency to aggregate in neurodegenerative diseases, we assessed the intrinsic aggregation propensity of extended NEFH amino acid sequence caused by the stop-loss mutation in the two identified families. We analyzed all 3 potential NEFH ORFs using a web-based aggregation prediction tool TANGO (Figure 5.2 B). The wildtype NEFH protein was not predicted to aggregate while ORF2 and ORF3 returned high aggregation scores, suggesting
that any frameshift variant which results in a stop-loss mutation could potentially cause protein aggregation. Additional aggregation propensity prediction tools were used to analyze the extended 3'UTR sequence including AGGREGSCAN\textsuperscript{144}, FoldAmyloid\textsuperscript{145}, and PASTA 2.0\textsuperscript{146}; all tested tools detected an overlapping amyloidogenic region within the mutant protein extension (Figure 5.2 C). We refer to this aggregate-prone amino acid sequence as a cryptic amyloidogenic element (CAE).

\begin{center}
\textit{Genome-wide Analysis of the Human 3' UTR-Encoded CAE}
\end{center}

In order to investigate additional potential candidate genes whose 3' UTRs encode CAEs, we performed a bioinformatics aggregation-prediction analysis of all human 3' UTR sequences. Human 3' UTR sequences were acquired from the UTRef section of UTRdb, a curated collection of eukaryotic 5' and 3' UTRs. The UTRef section contains 34,619 3' UTR sequences from genes retrieved from RefSeq transcripts\textsuperscript{147}. We translated these sequences into the three forward open reading frames to simulate stop-loss mutations caused by either missense (frame 1) or frameshift (frames 2 and 3) mutations. After we filtered out amino acid sequences with over 90\% similarity and genes of uncertain function (LOC symbols), approximately 12,400 genes per reading frame were annotated with the aggregation-prediction programs TANGO and PASTA. Next, sequences were filtered for highly stringent threshold aggregation scores: above 200 for TANGO and below −4 for PASTA. These score cutoffs were based on the aggregation-prediction scores obtained for the NEFH 3' UTR. Sequences that lack an
alternative stop codon were filtered out because they would most likely be degraded by the nonstop-decay mechanism. It has been reasoned that the stability of stop-loss mRNAs and/or proteins decreases as the distance between the mutated stop codon and the next alternative stop codon increases.\textsuperscript{148} Therefore, only sequences containing an alternative stop codon within a stretch coding for 50 amino acids were considered. After applying these filter criteria, we obtained 4,861 genes (approximately 1,600 genes per reading frame) containing a 3’ UTR sequence with a high potential for aggregation if translated (Figure 5.3).

Although our results suggest that a large number of genes have the potential to cause aggregation as a result of a stop-loss mutation, several physiological factors that vary in different intracellular micro environments (such as temperature, pH, pressure, and protein concentration) influence aggregation.\textsuperscript{149} Moreover, the frequency of stop loss caused by a missense mutation within the stop codon (frame 1) is very low: 0.027\% (609 of 2,207,918 variants) as observed in the NHLBI ESP Exome Variant Server (ESP6500). Finally, in order to cause significant protein-aggregation disease, the protein must be present in cells that can be negatively affected by aggregations (such as postmitotic neurons), and the presence of the protein must overwhelm the cell’s ability to clear aggregations.

Next, we filtered for disease-associated genes that had previously been reported to cause aggregation and that encode a predicted CAE in any 3’ UTR reading frame. We obtained a list with the top 21 high-risk aggregation genes (Table 5.1). Although variants in these genes have already been shown to cause aggregation, stop-loss mutations resulting in translation of a CAE from a 3’ UTR
have not been reported yet. Because the genes listed encode proteins prone to aggregation, they might be more susceptible to aggregation by our proposed mechanism.

Figure 5.3 Genome-wide scan for cryptic amyloidogenic elements of the human 3'‐UTRef.
Schematic illustrating the bioinformatics approach and filtering steps for a genome-wide scan of all possible 3'-UTR translations. 3'-UTR sequences were obtained from the UTRef database and putative 'LOC' genes were removed. Each 3'-UTR sequence was translated into all 3 possible open reading frames. Translated sequences with >90% redundancy were removed followed by filtering for sequence 5-51 amino acids in length. The remaining protein sequences were annotated with TANGO and PASTA aggregation scores and filtered for positive predictive scores.
Table 5.1 List of aggregation disease genes predicted to contain a 3’-UTR CAE
Known aggregation disease genes identified in the genome-wide screen that are predicted to contain a cryptic amyloidogenic element in at least one open reading frame of the 3’-UTR.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene MIM Number</th>
<th>Disease</th>
<th>Phenotype MIM</th>
<th>TANGO</th>
<th>PASTA</th>
<th>TANGO</th>
<th>PASTA</th>
<th>TANGO</th>
<th>PASTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOA2</td>
<td>107670</td>
<td>Apolipoprotein A-I deficiency</td>
<td>143890</td>
<td>182.8</td>
<td>no</td>
<td>0.0</td>
<td>no</td>
<td>0.0</td>
<td>no</td>
</tr>
<tr>
<td>ATXN2</td>
<td>601517</td>
<td>Spino cerebellar ataxia 2</td>
<td>183090</td>
<td>989.1</td>
<td>-7.9</td>
<td>76.5</td>
<td>no</td>
<td>853.0</td>
<td>-7.7</td>
</tr>
<tr>
<td>B2M</td>
<td>109700</td>
<td>Amyloidosis</td>
<td>105200</td>
<td>no</td>
<td>-7.9</td>
<td>2265.7</td>
<td>-17.6</td>
<td>181.1</td>
<td>no</td>
</tr>
<tr>
<td>FGB</td>
<td>134830</td>
<td>Hypofibrinogenemia, congenital</td>
<td>202400</td>
<td>358.3</td>
<td>-3.5</td>
<td>998.2</td>
<td>-13.9</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>FUS</td>
<td>137070</td>
<td>Amyotrophic lateral sclerosis (ALS)</td>
<td>608030</td>
<td>323.0</td>
<td>no</td>
<td>no</td>
<td>48.4</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>IAPP</td>
<td>147940</td>
<td>Type 1 diabetes</td>
<td>125853</td>
<td>no</td>
<td>no</td>
<td>541.1</td>
<td>-10.9</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>LYZ</td>
<td>153450</td>
<td>Amyloidosis</td>
<td>105200</td>
<td>183.1</td>
<td>no</td>
<td>no</td>
<td>962.9</td>
<td>-6.5</td>
<td>160.7</td>
</tr>
<tr>
<td>NOTCH3</td>
<td>600276</td>
<td>Cerebral arteriopathy</td>
<td>125310</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>566.0</td>
<td>-5.1</td>
<td>no</td>
</tr>
<tr>
<td>PRNP</td>
<td>176640</td>
<td>Prion disease</td>
<td>606688</td>
<td>0.0</td>
<td>no</td>
<td>896.1</td>
<td>-10.3</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>RH0</td>
<td>180380</td>
<td>Retinitis pigmentosa</td>
<td>613731</td>
<td>697.1</td>
<td>-7.0</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>SAA1</td>
<td>104750</td>
<td>Amyloidosis</td>
<td>105210</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>326.2</td>
<td>-4.1</td>
<td>no</td>
</tr>
<tr>
<td>SNCA</td>
<td>163890</td>
<td>Parkinson’s disease</td>
<td>168801</td>
<td>4.4</td>
<td>no</td>
<td>1442.5</td>
<td>-9.0</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>SOD1</td>
<td>147450</td>
<td>Amyotrophic lateral sclerosis (ALS)</td>
<td>105400</td>
<td>no</td>
<td>no</td>
<td>568.1</td>
<td>-7.0</td>
<td>6.7</td>
<td>no</td>
</tr>
<tr>
<td>TARDBP</td>
<td>605078</td>
<td>Amyotrophic lateral sclerosis (ALS)</td>
<td>612069</td>
<td>714.6</td>
<td>-5.8</td>
<td>no</td>
<td>no</td>
<td>464.4</td>
<td>-10.0</td>
</tr>
<tr>
<td>NEFL</td>
<td>162280</td>
<td>Charcot-Marie Tooth</td>
<td>607734</td>
<td>445.6</td>
<td>-8.7</td>
<td>no</td>
<td>no</td>
<td>5.5</td>
<td>-2.9</td>
</tr>
<tr>
<td>BSCL2</td>
<td>606158</td>
<td>Neuropathy, distal hereditary motor</td>
<td>600794</td>
<td>126.7</td>
<td>no</td>
<td>0.6</td>
<td>no</td>
<td>1.2</td>
<td>no</td>
</tr>
<tr>
<td>MAPT</td>
<td>157140</td>
<td>Parkinson disease</td>
<td>168600</td>
<td>no</td>
<td>no</td>
<td>156.6</td>
<td>6.7</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>TTR</td>
<td>176300</td>
<td>Amyloidosis</td>
<td>105210</td>
<td>0.0</td>
<td>no</td>
<td>0.0</td>
<td>112.3</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>
5.4 Discussion

We describe two dominant axonal peripheral neuropathy families with heterozygous frameshift variants in *NEFH*, resulting in stop-loss mutations that extend the NEFH protein by 40 amino acids via translation of the 3’ UTR. The extended protein was determined by *in silico* aggregation prediction tools to carry a stretch of amino acids with high propensity to aggregate, or cryptic amyloidogenic elements (CAEs). Through extensive functional studies, we demonstrated that the *NEFH* aggregates are caused specifically by translation of the CAE. Though NF aggregation is a well characterized feature in several neurodegenerative diseases, our results indicate a novel pathological mechanism by which protein aggregates can form. Our genome-wide aggregation-prediction analysis of the entire human 3’ UTR collection showed that a large number of genes could potentially be affected if they are translated into a specific reading frame encoding a CAE. We experimentally validated aggregation induced by translation of a CAE from the FUS 3’ UTR, confirming that this is not a NF-exclusive phenomenon. It is commonly implied that stop-loss mutations are associated with a clinical phenotype through a loss-of-function mechanism; however, our studies have revealed the importance of investigating protein aggregation propensity due to CAEs encoded by the 3’ UTR as a toxicity-inducing mechanism.
CHAPTER 6

Enrichment of Transcripts Related to Mitochondrial Function and Microtubule-Based Axonal Transport within the Human Motor Neuron

Axonal Transcriptome

The story of the human motor neuron axonal transcriptome is primarily the work of Dr. Renata Maciel who, alongside Dr. Mario Saporta, formulated the original hypothesis and experimental design, differentiated human induced pluripotent stem cells into spinal motor neurons, and developed the platform to isolate the axonal compartment of the motor neurons. I contributed to the methods and results sections of the manuscript and to bioinformatics analysis of the data. I performed the RNA-seq processing, differential gene expression analysis between axonal and soma compartments, and statistical analysis and pathway analysis of the differentially expressed genes. The following data presented in this chapter were previously published in *Experimental Neurology* (Volume 307), with myself as second author.

6.1 Background

Local modulation of protein synthesis within the neuronal axon and synaptic terminal is required for many forms of long-term synaptic plasticity and may have a role in axonal outgrowth and regeneration regulation. Local protein translation is achieved via selective transport of messenger RNAs from the neuronal cell body to the axon. Recently, disrupted axonal mRNA transport and
translation has emerged as a pathomechanism in neurodegenerative diseases, including Spinal Muscular Atrophy (SMA) and Amyotrophic Lateral Sclerosis (ALS). It is possible that mRNA axonal transport is disrupted as well in inherited axonopathies, especially given the significant disruption of axonal transport in several types of axonal Charcot-Marie-Tooth disease (CMT). To identify mRNA species that are actively transported to axons and may be dysregulated in human axonopathies, we mapped the axonal transcriptome of several healthy human induced pluripotent stem cell (iPSC)-derived motor neurons using permeable inserts to obtain large amounts of pure axonal material for RNA isolation and sequencing. Interestingly, transcripts of several genes known to cause inherited axonopathies were significantly localized to the axons of human motor neurons, suggesting the importance of their local expression in axonal maintenance. The human motor neuron axonal transcriptome map assembled in this study will be a useful resource to further investigate the dysregulation of mRNA localization in human neurodegenerative diseases.

6.2 Materials and Methods

RNA sequencing and data processing

Whole transcriptome sequencing was performed at the Sequencing Core of the John P. Hussman Institute of Human Genomics at the University of Miami. Following standard Illumina protocols, 80 to 100 ng of total RNA were ribo-depleted via Agilent Bioanalyzer and subsequent library construction was performed using the TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero. Libraries were
single-end sequenced on Illumina’s HiSeq2000 sequencer to produce 75 base pair reads. Quality control metrics for the raw reads were assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Total reads per sample showed low variance, ranging from 39.8 – 42.9 million reads. Reads were aligned to Ensembl’s human GRCh37.75 genome using STAR. HTSeq was used for gene quantification of the aligned reads against the GENCODE reference gtf file for the human GRCh37.75 genome.\textsuperscript{162}

RNA-seq analysis

Gene-level differential expression analysis was performed using DESeq2.\textsuperscript{163} Differentially expressed genes were determined by an adjusted p-value threshold of 0.05 and log fold change greater than 1.5 or less than 0.5. The DAVID Functional Annotation Tool was used for enrichment analysis of the differentially expressed gene sets.\textsuperscript{164,165} The intersection of known CMT disease genes and mitochondrial genes with the differentially expressed genes was analyzed by Chi-square with Yate’s correction test.

6.3 Results

\textit{Differential gene expression between axons and neuronal cell bodies from human motor neurons}

Whole transcriptome sequencing was performed on ribosomal-depleted RNA from the axonal and somatodendritic compartments of the motor neuron cultures (n = 3 for each compartment). Sequencing was performed in two separate batches.
Batch 1 included axonal and somatodendritic RNA from sample 1, while batch 2 included axonal and somatodendritic RNA from samples 2 and 3. Regularized log-transformation was applied to the count data using the DESeq2 R module. Correlation analysis was performed on genes with an average transformed count greater than 1 (n = 26902, Figure 6.1 A); however, of the 26902 transcripts identified, only 14,129 transcripts had enough data for statistical analysis of differential expression by DESeq2. The correlation between technical replicates demonstrates minimal variation across different cell lines and for each compartment. The degree of variation is highest between replicate 1 with replicates 2 and 3, reflecting the batch structure of sequencing runs (Figure 6.1 A).

Sequence read alignment produced 24,989 detectable genes at which each technical replicate contained at least one non-transformed aligned read (Figure 6.1 B). Although more than nineteen thousand genes were expressed in both compartments, 3,458 and 1,702 genes are exclusively expressed in axonal and somatodendritic compartments, respectively (Figure 6.1 B, left). A comparison of genes with the highest averaged expression in each compartment (n = 1000) revealed a significant overlap of 859 genes (Figure 6.1 B, right), consistent with what has been shown in a similar study using mouse neuronal cultures (Briese et al. 2015). Nonetheless, almost 150 transcripts were enriched in either the axonal or the somatodendritic compartment.

To explore the transcriptomic similarities between AXC and SDC, the most highly expressed genes (q-value <= 0.1, n = 1000) underwent unsupervised hierarchical clustering (Figure 6.1 C) revealing that transcripts from three different
lines are grouped by location, despite batch differences. Using thresholds fold change \( \geq 1.5 \) or fold change \( \leq 0.5 \) and \( q \)-value \( \leq 0.05 \), we identified 2,297 differentially expressed genes (Figure 6.1 D). Approximately 60 genes were identified as having a greater than 1.5-fold increase in their expression in the axonal compartment (57 axonal genes, FC \( \geq 1.5 \)).

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of differentially expressed gene sets (log fold-change \( \geq 1.5 \) or \( \leq 0.5 \) and \( q \)-value \( \leq 0.05 \)) on SDC revealed enrichment for pathways associated with nucleus and cytoplasm metabolism (Figure 6.1 E). Conversely, in the axonal compartment pathways related to microtubule-based axonal transport and mitochondrial metabolism, including oxidative phosphorylation, respiratory chain complex, ATP synthesis and NADH dehydrogenase activity (Figure 6.1 F), were enriched.

*Enrichment analysis reveals several transcripts associated with neurological disorders enriched in axons of human motor neurons.*

The Broad Institute’s Human MitoCarta2.0 inventory of mitochondrial localized proteins was used to test for an enrichment of mitochondrial genes within the differentially expressed genes in AXC (\( n = 958 \) genes after removing “possible” and “non-localized” proteins). Furthermore, we used Online Mendelian Inheritance in Man, OMIM database to filter for neurological disease- and/or axonal transport-related genes. We identified 6 genes related to genetic axonopathies, including *BICD2, DST, KIF1C, KIF5C, MPZ* and *mt-ATP6*. Both the mitochondrial
and neurological disease-associated genes were significantly enriched within the differentially expressed axonal gene set (Chi-square with Yate’s correction test, \( p \)-value \( \leq 0.0001 \); Table 6.1 and 6.2, respectively).

Figure 6.1 Transcriptome profiling of iPSC-derived motor neuron axonal and somatodendritic compartments (AXC and SDC).
(A) Correlation analysis of regularized log transformed read counts for biological replicates (rows 1 and 2) and compartments derived (row 3) from single patient. (B) Venn diagrams displaying the total detectable genes (read count >= 1) in all 3
biological replicates of each compartment (left), and the 1000 genes with the highest average read count of each compartment (right). (C) Unsupervised clustering of the 1000 genes with the highest averaged regularized-log-transformed read counts, annotated with patient cell number and subcellular compartment. (D) MA-plot showing the log2 fold change between compartments (y-axis) versus the gene abundance (x-axis), highlighting significant genes in red (q < 0.1). Red dotted lines represent thresholds fold change used for differentially expressed genes (>= 1.5 or fold change <= 0.5). Neurological disorders- and axonal transport-related genes are labeled (KIF1C, KIF5C, DST, BICD2 and MT-ATP6). (F & G) Gene Ontology enrichment analysis of differentially expressed gene sets colored by significance: (F) upregulated in soma (log fold-change < 0.5 and q-value < 0.05) and (G) upregulated in axon (log fold-change > 2 and q-value < 0.05).

**Table 6.1 Mitochondrial genes overexpressed in axons**

<table>
<thead>
<tr>
<th>Mitochondrial gene</th>
<th>Overexpressed</th>
<th>Not overexpressed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>947</td>
</tr>
<tr>
<td>Non mitochondrial gene</td>
<td>45</td>
<td>28786</td>
</tr>
</tbody>
</table>

Chi-square (Yates correction) 47.951
P-value < 0.0001
Odds ratio 7.54

* Fold change ≥ 1.5

**Table 6.2 Neurological disease genes overexpressed in axons**

<table>
<thead>
<tr>
<th>Neurological disease genes</th>
<th>Overexpressed</th>
<th>Not overexpressed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>162</td>
</tr>
<tr>
<td>Non neurological disease genes</td>
<td>52</td>
<td>28779</td>
</tr>
</tbody>
</table>

Chi-square (Yates correction) 49.194
P-value < 0.0001
Odds ratio 15.89

* Fold change ≥ 1.5

Table 6.1 Mitochondrial genes overexpressed in axons
Two-by-two contingency table and chi-square results of mitochondrial genes differentially localized to the axonal compartment.

Table 6.2 Neurological disease genes overexpressed in axons
Two-by-two contingency table and chi-square results of neurological disease genes differentially localized to the axonal compartment.
6.4 Discussion

Limited access to affected cell types has hindered research into the understanding of neurodegenerative disease pathophysiology and treatment development. iPSCs-derived motor neurons are highly relevant for studying the disease mechanisms of such diseases, including Amyotrophic Lateral Sclerosis (ALS), Spinal Muscular Atrophy (SMA), Charcot-Marie-Tooth disease and Hereditary Spastic Paraplegia (HSP). As disruption of axonal mRNA transport to and local protein translation at the distal axons has been identified as a pathomechanism in motor neuron disorders\textsuperscript{156,167-169}, understanding the role of axonal RNA trafficking and expression may lead to insights into biological pathways and therapeutic targets. To explore the use of the human motor neuron axonal transcriptome, we searched for transcripts associated with neurological diseases which feature axonal degeneration as a primary pathology (based on the Online Mendelian Inheritance in Man database) and for mitochondrial-localized genes (MitoCarta2.0 inventory). Interestingly, we found significant enrichment of both mitochondrial genes and inherited axonopathy disease genes (\textit{BICD2, DST, KIF1C, KIF5C} and \textit{MT-ATP6}), most of which have a direct function in microtubule-based axonal transport. Our findings support the hypothesis that the high metabolic and axonal transport demands of long motor axons require local axonal protein translation. In future studies, this same approach can be used to identify improperly localized mRNA within human motor neurons containing a genetic defect known to cause inherited axonopathy.
CHAPTER 7
Network-based Analysis of Inherited Axonopathies Reveals Cellular
Pathways and Molecular Overlap

The data presented in this chapter were previously published in Scientific Reports (Volume 5, Issue 3). I wrote the manuscript and performed all the work presented.

7.1 Background
Inherited axonopathies represent a spectrum of disorders unified by the common pathological mechanism of length-dependent axonal degeneration\textsuperscript{15,16,170}. Progressive axonal degeneration can lead to both Charcot-Marie-Tooth type 2 (CMT2) and Hereditary Spastic Paraplegia (HSP) depending on the affected neurons: peripheral motor and sensory nerves or central nervous system axons of the corticospinal tract and dorsal columns, respectively. Inherited axonopathies display an extreme degree of genetic heterogeneity of Mendelian high-penetrance genes.\textsuperscript{7,10,171} High locus heterogeneity is potentially advantageous to deciphering disease etiology by providing avenues to explore biological pathways in an unbiased fashion.\textsuperscript{78} Here, we investigate ‘gene modules’ in inherited axonopathies through a network-based analysis of the Human Integrated Protein-Protein Interaction rEference (HIPPIE) database.\textsuperscript{172} We demonstrate that CMT2 and HSP disease proteins are significantly more connected than randomly expected. We define these connected disease proteins as ‘proto-modules’ and show the
topological relationship of these proto-modules by evaluating their overlap through a shortest-path based measurement. In particular, we observe that the CMT2 and HSP proto-modules significantly overlapped, demonstrating a shared genetic etiology. Comparison of both modules with other diseases revealed an overlapping relationship between HSP and hereditary ataxia and between CMT2+HSP and hereditary ataxia. We then use the DIseAse Module Detection (DIAMOnD) algorithm\textsuperscript{81} to expand the proto-modules into comprehensive disease modules. Analysis of obtained disease modules reveals an enrichment of ribosomal proteins and pathways likely central to inherited axonopathy pathogenesis, including protein processing in the endoplasmic reticulum, spliceosome, and mRNA processing. Furthermore, we determine pathways specific to each axonopathy by analyzing the difference of the axonopathy modules. CMT2-specific pathways include glycolysis and gluconeogenesis-related processes, while HSP-specific pathways include processes involved in viral infection response. Unbiased characterization of inherited axonopathy disease modules will provide novel candidate disease genes, improve interpretation of candidate genes identified through patient data, and guide therapy development.

7.2 Materials and Methods

Disease Genes

Charcot-Marie-Tooth type 2 (CMT2) and Hereditary Spastic Paraplegia (HSP) input-proteins, also known as “seed protein”, were selected through an extensive
literature review by domain experts from the Rare Diseases Clinical Research Network (RDCRN). Specifically, the CMT2 gene list was obtained from the Inherited Neuropathy Consortium and the HSP gene list was obtained from the Alliance for Treatment in HSP and PLS, and the Clinical Research in ALS and Related Disorders for Therapeutic Development consortium. Non-syndromic CMT2 genes (n=65) and both complex and pure HSP genes (n=96) were included as seed proteins for disease module construction. To evaluate the topological distance between CMT2 and HSP in a broader context, we selected seven additional diseases for topological separation analysis. We chose diseases that vary in degree of expected shared etiology with inherited axonopathies. Such considerations are based on clinical symptoms and shared disease genes, with hereditary ataxia expected to be the closest and cancer expected to be the most distance diseases. Seed proteins for comparison of diseases (cancer (CA), deafness (DFN), cardiomyopathy (CM), muscular dystrophy (MD), Parkinson’s disease (PD), amyotrophic lateral sclerosis (ALS), and hereditary ataxia (ATX)) were collected from the Online Mendelian Inheritance in Man (OMIM, updated 10-25-2017).¹⁷³

Molecular Interactions
The Human Integrated Protein-Protein Interaction rEference (HIPPIE) is a publically available, scored protein-protein interactions (PPI) dataset.¹⁷⁴ HIPPIE integrates data from BioGRID (genetic interactions removed), DIP, HPRD, IntAct,
MINT, BIND, and MIPS databases. HIPPIE provides a curated scoring scheme that is based on the experimental technique to identify the PPI, the number of studies that have reported the PPI, and the reproducibility of the PPI across model organisms. The HIPPIE confidence score ranges from 0 to 1 and reflects the quality of experimental evidence supporting each PPI. The HIPPIE authors have predefined confidence levels based on the quartiles of all confidence scores: medium confidence (0.63 – second quartile of the HIPPIE score distribution) or high confidence (0.73 – third quartile) [http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/information.php#sources](http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/information.php#sources). To balance network coverage and reasonable interaction confidence, we accounted for the interactions with at least a medium confidence score of 0.63. The most recent update includes approximately 273,900 experimentally determined PPIs between 17,000 human proteins.\(^{172}\) The HIPPIE (v2.0) dataset was pre-filtered to remove interactions from non-human sources and below the predefined medium-confidence score threshold in order to create a more stringent global interactome of 204,215 interactions between 15,690 proteins. The filtered HIPPIE protein-protein interaction network thus obtained is scale-free, as demonstrated by a power-law decay in the corresponding degree distribution.

**Disease Proto-Module Construction**

A goal of network analysis is to define and study a disease module. We first define a disease proto-module since our knowledge of both the disease-causing proteins
and all human protein interactions is incomplete.\textsuperscript{175} Disease proto-modules are considered clusters of highly connected seed proteins that were constructed by retrieving the subgraph of directly connected seed proteins from the filtered HIPPIE network. The connectivity of a proto-module is quantified by the size of the largest connected component (LCC) and shortest distance (SD) between proteins.\textsuperscript{175} While LCC indicates the number of directly connected disease proteins, SD measures the localization of disease proteins that do not directly interact. Utilizing publically available Python scripts, we calculate the LCC, SD, and significance of each disease proto-module.\textsuperscript{175} To assess the significance of LCCs, expected null distributions were calculated by randomizing sets of proteins of equal seed list size. We determined the expected distribution of LCCs after 10,000 randomizations and calculated a proto-module specific z-score.\textsuperscript{175} As previously described, a proto-module is obtained if the corresponding z-score was $\geq 1.6$ with a significance $p$-value $\leq 0.05$.\textsuperscript{175} To assess the SDs, the distribution of SDs after 10,000 randomizations was tested using a Mann-Whitney-U test.\textsuperscript{175}

**Disease Proto-Module Expansion**

The DIseAse MOdule Detection (DIAMOnD) algorithm is one of the latest and most utilized disease-gene prediction methods.\textsuperscript{81,176} We selected the highly cited DIAMOnD algorithm for this study based on the previous successful application to a diverse range of diseases.\textsuperscript{83,177-180} Specifically, DIAMOnD expands each proto-module to identify the putatively complete disease module.\textsuperscript{81} Based on the
assumption that disease proteins interact more significantly with each other, DIAMOnD explores the topological neighborhood of the seed proteins based on the significance of connections to disease proteins. Briefly, DIAMOnD calculates the connectivity significance for all proteins that interact with the seed proteins; ranks the proteins according to \( p \)-values; incorporates the highest ranked protein into the set of seed proteins; and repeats the process to expand the disease seed set. Each disease proto-module was expanded using DIAMOnD recommended settings, including weighting the seed proteins to 10 and running 200 iterations. The core inherited axonopathy module was constructed from the intersection of the CMT2 and HSP expanded modules.

**Disease Proto-Module Separation**

The proximity of diseases within a network is useful for understanding their biological and clinical overlap. The network-based separation between two disease proto-modules \( A \) and \( B \) was assessed by comparing the mean shortest distances of seed proteins within each proto-module \( (d_{AA} \text{ and } d_{BB}) \) to the mean shortest distance between the proteins of each proto-module \( (d_{AB}) \). In particular, Menche et al. defined the network-based separation as \( d_{AB} - \frac{d_{AA} + d_{BB}}{2} \) and calculated a z-score to quantify the difference between the observed separation and random expectation and analytically calculate a corresponding \( p \)-value for each z-score. A positive z-score indicates separation of disease pairs while a negative z-score indicates overlap. The \(|z\text{-score}| \) must be \( \geq 1.6 \) for a disease pair.
to be more/less overlapping than randomly expected with a significance \( p \)-value \( \leq 0.05 \).\textsuperscript{175}

**Semantic Similarity Within Expanded Module**

The biological functions of the expanded modules were assessed with Gene Ontology (GO) semantic similarity measurements using the GOSemSim R package.\textsuperscript{181-183} The semantic similarity was computed using the graph-based Wang method, which computes the semantic similarities between terms based on both their topology within the GO graph structure and their relations with ancestor terms.\textsuperscript{183} Semantic similarities of multiple GO terms were combined using the Best-Match Average (BMA) strategy which calculates the average of all maximum similarities on each pair of GO terms.\textsuperscript{183} A systematic evaluation of the metrics for GO based semantic similarity found that the BMA is the best combination approach since it yields the highest resolutions and does not show the undesired behaviors of the maximum approach or the average approach.\textsuperscript{184} Before computing the semantic similarities of GO terms, we filtered the GO terms based on annotation evidence codes to produce a high quality dataset. Specifically, we removed Inferred from Electronic Annotation (IEA) evidence codes. Although IEA are the most common evidence codes, they also represent weak associations as they are based on electronic curation without human curation. Since the input for this analysis (DIAMOnD proteins) is based on a dataset of physical interactions, we
also removed the Inferred from Physical Interaction (IPI) evidence codes to remove bias and circular reasoning from the analysis.

**Functional Enrichment Analysis of Disease Module**

Functional enrichment analysis was performed for GO Biological Process (BP) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, and Wikipathways.\(^{181,185,186}\) Over-Representation Analysis (ORA) of protein clusters, based on GO BP semantic similarity were analyzed using the DAVID Functional Annotation Tool.\(^ {164}\) ORA of disease modules for KEGG pathways and Wikipathways was performed using WebGestalt and were filtered for FDR \(\leq 0.2\).\(^ {187}\)

Pathway ORA was performed on 6 sets: the expanded CMT2 module, the expanded HSP module, the intersection of the expanded CMT2 and HSP modules (AXONO core module), the union of the expanded CMT2 and HSP modules (AXONO spectrum module), the difference of the expanded CMT2 and HSP modules (CMT2-only module), and the difference of the HSP and CMT2 modules (HSP-only module). Because of gene redundancy in pathway sets, the Cytoscape plugin Enrichment Map was used to visualize the similarity between enriched pathway sets with a Jaccard coefficient greater than 0.1.\(^ {188}\)
7.3 Results

Constructing the inherited axonopathy modules

Disease module construction requires an input seed list of disease proteins and a molecular interaction network. We curated 65 non-syndromic axonal Charcot-Marie-Tooth 2 (CMT2) and 96 Hereditary Spastic Paraplegia (HSP) input seed genes through an extensive literature review. As a protein-protein interaction data source, we used the Human Integrated Protein-Protein Interaction rEference (HIPPIE, v2.0) that integrates interaction data from high quality databases and provides a high-performing scoring system.\(^{172}\) The HIPPIE confidence score is based on the amount and the reliability of supporting experimental evidence for each interaction.\(^{174}\) Applying the predefined medium confidence score (0.63 – second quartile of the HIPPIE score distribution), we created a medium-confidence global interactome consisting of 204,215 interactions between 15,690 human proteins.

In an interactome, disease associated proteins are expected to cluster into specific network neighborhoods through physical interactions.\(^{78,189}\) The inherited axonopathy proteins follow this expectation by forming significantly connected disease modules (Figure 7.1). The largest connected component (LCC) of a subgraph measures the number of seed proteins in the largest subgraph composed of direct interactions between seed proteins. In our analysis, the CMT2 and HSP LCCs are significantly larger than expected by random chance (CMT2: \(n=35\), \(z\)-score=17.1; HSP: \(n=16\), \(z\)-score=4.0) (Figure 7.1 A&C), indicating that the
CMT2 and HSP disease proteins form disease proto-modules composed of proteins with high biological and functional similarity. Another measure of connectivity is the shortest distance (SD) between seed proteins. SD also includes fragmented seed proteins that do not connect to proteins in the LCC, providing a more global measure of connectivity than LCC alone. As the knowledge of human protein interactome and disease etiology is incomplete, seed proteins are expected to fragment within the interactome, pointing to larger distances. However, we observed that SD distributions of CMT2 and HSP seed genes are shifted towards 1.0, indicating direct connections between seed proteins. In comparison to the expected distribution based on randomized protein sets (Figure 7.1 B&D) the observed means are significantly smaller (Mann Whitney U Test; CMT2: \( p \)-value=2.7E-16; HSP: \( p \)-value=5.0E-05). Taken together, the significant connectivity measurements demonstrate that the CMT2 and HSP disease proteins form agglomerations that can be expanded and further studied (Figure 7.1).

The significantly connected CMT2 and HSP proto-modules were next used as seed modules to expand into putatively “complete” disease modules using the DIseAse Module Detection (DIAMOnD) method.\textsuperscript{81} The DIAMOnD algorithm identifies the topological neighborhood of the seed proteins based on the significance of connections to seed proteins and has been successfully used to study asthma and cardiovascular endophenotypes.\textsuperscript{83,177} After performing 200 iterations of the DIAMOnD method, an additional 53 CMT2 proteins and 61 HSP proteins were incorporated into each LCC, which were originally fragmented and
not directly connected (Figure 7.1 E&F). To limit the incorporation of false positives into each inherited axonopathy module, the DIAMOnD proteins were evaluated for biological evidence by comparing the Gene Ontology Biological Process (GO BP) terms of the DIAMOnD proteins to the GO BP terms of the known disease proteins. The GO BP terms were compared by hierarchical clustering based on semantic similarity (Figure 7.2)\textsuperscript{181}, allowing us to infer the biological relationship between DIAMOnD proteins and disease proteins. As indicated by the semantic similarity clustering, DIAMOnD proteins are functionally related to the disease seed proteins, suggesting that we limited the incorporation of false positives beyond the true biological limit. To understand the underlying biological function, we performed GO BP Over Representation Analysis (ORA) of the proteins within each cluster. In particular, we confirmed known cellular processes including axonal transport, microtubule dysregulation, and mitochondrial involvement (Figure 7.2). Taken together, these results indicate that the expansion of proto-modules leads to the identification of proteins that are functionally similar to disease proteins and are involved in biological processes that are relevant to the disease. After topologically validating the disease modules with connectivity measures and biologically validating the incorporated proteins through GO BP analysis, we present the final CMT2 and HSP disease modules in Figure 7.1 E&F. The networks display the connections between disease proteins (in color) and DIAMOnD proteins (in gray). The DIAMOnD proteins will improve our understanding of the disease by adding
power to pathway analysis and containing new candidate disease genes and targets.

**Topological relationship of disease modules**

Disease modules that are located in adjacent network areas will likely share proteins, interactions, and pathways involved in disease pathogenesis. The disruption of overlapping functional modules will likely result in shared clinical characteristics in each disease. The topological relationship between inherited axonopathy proto-modules (i.e. CMT2, HSP) was quantified through the shortest distances between modules (Figure 7.3). The close relationship within the spectrum of inherited axonopathies (between CMT2 and HSP) is confirmed in our analysis by significantly overlapping disease proto-modules (network distance=-0.15; z-score=-2.6) (Figure 7.3 A). We next explored the relationships of the inherited axonopathies with cancer (CA), deafness (DFN), cardiomyopathy (CM), muscular dystrophy (MD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), and hereditary ataxia (ATX) (Figure 7.3 A). Utilizing disease genes curated from the OMIM database (updated 10-25-2017), we compared the corresponding disease-specific proto-modules with CMT2, HSP, and their union (HSP&CMT2). All compared diseases were significantly separated from each axonopathy module except for ALS and ATX. Notably, the ALS proto-module was not significantly separated nor overlapped with the CMT2, HSP, and HSP&CMT2 proto-modules. The ATX proto-module overlapped significantly with both the HSP and HSP&CMT2 proto-modules, yet significantly separated from the CMT2 proto-
module. Since the network separation parameter depends on the number of proteins in each disease and the number of direct interactions between disease proteins, we provided the count of seed proteins (ranging from 9 to 72) and the amount of direct connections between disease proteins (ranging from 6 to 52 connections) (Figure 7.3 B).

Ribosomal proteins are enriched within the inherited axonopathy modules

The CMT2 and HSP expanded modules both contained an unexpected number of ribosomal proteins incorporated as DIAMOnD proteins (CMT2 $n=47$; HSP $n=48$; overlap $n=40$) (Figure 7.4). To qualitatively determine if the enriched ribosomal proteins are an artifact of the global network structure, we examined the underlying protein interactions (Figure 7.4 A&B). The interactions between ribosomal DIAMOnD proteins and disease seed proteins (green nodes connecting to orange nodes) are similar to the interactions between disease seed proteins (orange nodes connecting to orange nodes), suggesting true biological interactions instead of a topology-biased artifact. We then quantified the direct interactions between disease seed proteins and ribosomal DIAMOnD proteins -
Figure 7.1  Inherited axonopathy proto-module connectivity (A-D) and expanded module networks (E-F).

(A-B) Increased observed largest connected component (LCC) size of CMT2 (A) and HSP (B) disease seed genes in comparison to the expected, random distribution. (C-D) Increased shortest distance between inherited axonopathy disease genes in comparison to the expected, random expectation. (E-F) First degree interactions between known Mendelian disease genes (colored) and DIAMOnD genes (gray) for CMT2 (E) and HSP (F). Disease specific genes are labeled, and direct edges between disease genes are more heavily weighted.
**Figure 7.2 Biological validation of inherited axonopathy DIAMOnDs**

(A: CMT2, B: HSP). To limit the incorporation of false positives into each inherited axonopathy module, the DIAMOnD proteins were evaluated for biological evidence by comparing Gene Ontology Biological Process (GO BP) terms of the DIAMOnD proteins to the GO BP terms of the known disease proteins. Heatmaps (A & B) display the Wang semantic similarity of Gene Ontology Biological Process terms between disease gene (teal bars) and DIAMOnD gene (gray bars). To understand the underlying
Figure 7.3 Disease proto-modules separation and connections.

(A) Significance of topological relationships between axonopathy proto-modules (Charcot-Marie-Tooth type 2 only: CMT2; Hereditary Spastic Paraplegia only: HSP; combined HSP+CMT2) and control diseases (Cancer: CA, Deafness: DFN, Cardiomyopathy: CM, Muscular Dystrophy: MD, Parkinson’s Disease: PD, Amyotrophic Lateral Sclerosis: ALS, and Hereditary Ataxia: ATX). Gray dotted lines indicate the threshold for significance (|Z-scores| >= 1.6). Negative (positive) Z-scores indicate overlapped (separated) modules. (B) Direct connections between HSP+CMT2 disease proteins and comparison disease proteins. Disease names and input seed protein count (highlighted by red bar) for each disease compared are indicated outside the circle. Direct interactions between disease proteins of HSP+CMT2 are displayed on the outer gray bar (highlight by orange bar).
CMT2: minimum: 1, maximum: 22, mean: 4.6; ribosomal DIAMOnD proteins - CMT2: minimum: 1, maximum: 22, mean: 4.6; HSP: minimum: 1, maximum: 15, mean: 4.2 (Figure 7.4 C&D). To further verify that ribosomal protein interactions are not an artifact of the HIPPIE network, we expanded each comparison disease proto-module and quantified the first-degree interactions between disease seed proteins and ribosomal DIAMOnD proteins (Figure 7.5). We found that neurologically related diseases contained more ribosomal DIAMOnD proteins than the non-neurological control disease, demonstrating disease-specific enrichment. Furthermore, in a recent study of the human motor neuron transcriptome, 81 ribosomal proteins showed significant localization (q-value ≤ 0.1) to the axon with modest expression levels (median=0.89) (Figure 7.4 E&F).\textsuperscript{150} In the Venn diagram in Figure 7.4 E, we observed that 50 of the axonal ribosomal DIAMOnDs were significantly localized to the axon, with 37 of these present in both the CMT2 and HSP sets of DIAMOnDs. In Figure 7.4 F, we show that the distributions of expression level log fold changes (LFCs) of the ribosomal DIAMOnDs (shared CMT2-HSP: n=37, HSP only: n=7, CMT2 only: n=6) and the ribosomal non-DIAMOnDs (n=31) were highly similar, corroborating the biological relevance of the network analysis results. Eight ribosomal proteins were reported as differentially localized to the axons in comparison to the soma (q-value ≤ 0.1; LFC ≤ 0.05 or ≥ 1.5).\textsuperscript{150} RPL26, RPL28, RPL7A, and RPS6KA1 are DIAMOnD proteins while RPL31, RPL7L1, RPS6KC1, and RSL24D1 are not (Figure 7.4 F). Significant
Axonal localization of 50 ribosomal DIAMOnD proteins demonstrates a biological significance, with potential to explain the enrichment of ribosomal proteins within neurologically-related disease network.

Figure 7.4 Interactions within the expanded inherited axonopathy modules. (A-B) First degree interactions within CMT2 (A) and HSP (B) expanded modules. DIAMOnD proteins are separated into ribosomal proteins (green) and other (blue). Size of DIAMOnD nodes are proportional to DIAMOnD order of incorporation – smaller nodes were incorporated earlier. (C-D) First degree interaction counts between disease protein and ribosomal DIAMOnD proteins. (E) Relationships between HSP ribosomal DIAMOnD proteins, CMT ribosomal DIAMOnD proteins, and ribosomal proteins significantly localized to the axons of human motor neurons. (q-value ≤ 0.1) (F) Distribution of log fold changes between axonal and soma differential localization.
To verify that ribosomal protein interactions are not an artifact of the HIPPIE network, we expanded each comparison disease proto-module and quantified the first degree interactions between disease seed proteins and ribosomal DIAMOnD proteins (Charcot-Marie-Tooth type 2: CMT2; Hereditary Spastic Paraplegia: HSP; Cancer: CA, Deafness: DFN, Cardiomyopathy: CM, Muscular Dystrophy: MD, Parkinson’s Disease: PD, Amyotrophic Lateral Sclerosis: ALS, and Hereditary Ataxia: ATX).

**Pathway analysis of the inherited axonopathy spectrum modules**

Functional analysis of the inherited axonopathy spectrum modules was performed by an Over-Representation Analysis (ORA) of WikiPathways and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Our results show both unique and shared pathways across inherited axonopathies (Figure 7.6). Pathway enrichment analysis was performed on 6 sets of proteins: the expanded

<table>
<thead>
<tr>
<th>Disease Name</th>
<th>First degree interaction count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMT2</td>
<td>51</td>
</tr>
<tr>
<td>HSP</td>
<td>47</td>
</tr>
<tr>
<td>ATX</td>
<td>42</td>
</tr>
<tr>
<td>PD</td>
<td>38</td>
</tr>
<tr>
<td>DFN</td>
<td>30</td>
</tr>
<tr>
<td>MD</td>
<td>28</td>
</tr>
<tr>
<td>ALS</td>
<td>28</td>
</tr>
<tr>
<td>CM</td>
<td>12</td>
</tr>
<tr>
<td>CA</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 7.5 DIAMOnD ribosomal protein interactions with disease seed proteins.
CMT2 module (n=254 proteins), the expanded HSP module (n=295 proteins), the intersection of the expanded CMT2 and HSP modules (AXONO core module, n=152 proteins), the union of the expanded CMT2 and HSP modules (AXONO spectrum module, n=397 proteins), the difference of the expanded CMT2 and HSP modules (CMT2-only module, n=102 proteins), and the difference of the HSP and CMT2 modules (HSP-only module, n=143 proteins)(Figure 7.6 A). We simplified the 6 protein sets to 3 protein sets for pathway classification and visualization purposes and accounted for the CMT2-dominant, HSP-dominant sets and the intersection of CMT2-expanded and HSP-expanded networks (AXONO). We assigned each enriched pathway to the most influenced disease module (CMT2-dominant, HSP-dominant, or AXONO) based on the combination of module sets that was enriched for a pathway (Figure 7.6 B). For example, pathways that were enriched in the CMT2 only set, but not in the HSP only set, were categorized as a CMT2-dominant pathway. The pathway results can be ranked by the highest count of observed input proteins within the pathway set or the highest enrichment score within the pathway set, each averaged across the 6 disease module sets. The Cytoplasmic Ribosomal Proteins (wiki) and Ribosome (KEGG) pathways were the most highly enriched with the highest count of input proteins within each module set, likely due to the high amount of ribosomal DIAMOnD proteins. Excluding the ribosomal protein pathways, the top 5 pathways (based on the observed count of input proteins) are: broad axonopathy spectrum: mRNA Processing, Protein processing in the endoplasmic reticulum, Viral carcinogenesis, Spliceosome, and
Apoptosis; **CMT2-dominant**: HIF-1 signaling pathway, Aminoacyl-tRNA biosynthesis, Biosynthesis of amino acids, Carbon metabolism, and Glycolysis and Gluconeogenesis; and **HSP-dominant**: Epstein-Barr virus infection, Herpes simplex infection, and Antigen processing and presentation. The top 5 most enriched pathways, excluding the ribosomal protein pathways, are: **broad axonopathy spectrum**: Aminoacyl-tRNA biosynthesis, Hedgehog signaling pathway, Pathogenic Escherichia coli infection, Parkin-Ubiquitin Proteasomal System pathway, and Estrogen signaling pathway; **CMT2-dominant**: Cori Cycle, Pentose phosphate pathway, Biosynthesis of amino acids, Glycolysis and Gluconeogenesis, and Carbon metabolism; and **HSP-dominant**: Epstein-Barr virus infection, Antigen processing and presentation, and Herpes simplex infection. The enriched pathways were simplified for easier understanding by grouping pathways together based on overlapping proteins (Figure 7.6 C).

### 7.4 Discussion

The goal of this study was to compare the spectrum of inherited axonopathies by identifying and characterizing the disease modules of axonal Charcot-Marie-Tooth disease (CMT2) and Hereditary Spastic Paraplegia (HSP). To our knowledge, this is the first study to apply an unbiased network analysis approach to CMT2 or inherited axonopathies combined.
Figure 7.6 Pathway analysis of inherited axonopathy modules. (A) Venn diagram of disease gene sets used for pathway enrichment including the expanded CMT2 module, the expanded HSP module, the intersection of the expanded CMT2 and HSP modules (AXONO core), the union of the expanded CMT2 and HSP modules (AXONO spectrum), the difference of the expanded CMT2 and HSP modules (CMT2-only), and the difference of the HSP and CMT2 modules (HSP-only). (B) KEGG and WikiPathways significantly enriched within each disease module set. (C) Closely related pathways based on redundancy in proteins involved in each pathway.
Inherited axonopathies form significantly connected disease modules

Disease genes are not distributed randomly within biological interaction networks but form local connections with each other as they interact to perform related biological functions.\textsuperscript{83,189,190} We began our analysis by identifying these localized gene neighborhoods of inherited axonopathy proteins. Since the genetic etiology of inherited axonopathies is not completely known, our analysis is limited by incomplete input seed lists. However, we found that CMT2 and HSP genes were significantly more connected to each other than expected by chance (Figure 7.1 A-D). The identification of these significant disease-specific proto-modules highlights the robustness of network analysis (Figure 7.1 E-F).

Overlapping disease modules are expected to share clinical characteristics, pointing to physical interaction of proteins, and by extension, common biological pathways.\textsuperscript{175} Since CMT2 and HSP have been categorized as inherited axonopathies based on their clinical and genetic overlap, we anticipated and observed a significant overlap between CMT2 and HSP disease proto-modules within the network (Figure 7.2 A). We found a significant separation of CMT2, HSP, and BOTH proto-modules from cancer, deafness, cardiomyopathy, muscular dystrophy, and Parkinson’s disease; an unclear relationship with amyotrophic lateral sclerosis; and discordant distances to ataxia (Figure 7.2 A). A previous study constructed an “HSPome” from published and candidate HSP proteins, and observed a significant overlap between HSP with Alzheimer’s disease, Parkinson’s disease (PD), and Amyotrophic Lateral Sclerosis (ALS).\textsuperscript{80} However, in our study,
HSP was significantly separated from PD and had an unclear relationship with ALS (Figure 7.2 A). The difference in topological relationships between studies is likely caused by differences in the underlying data sources and input seed lists. The “HSPome” was created from iRefIndex, HumanNet, and STRING with 43 known HSP proteins as input seed list, whereas we used the HIPPIE database and increased our input seed list to 95 proteins. In particular, 40 seeds overlapped between the two inputs.\textsuperscript{191-193} Interestingly, the inherited axonopathy proto-modules appear slightly “less” separated, based on z-score, from PD and ALS than the other non-neurological controls (Cancer, Deafness, Cardiomyopathy, and Muscular Dystrophy). Each axonopathy module showed an unclear relationship with ALS (neither significantly overlapped nor separated), which should be clarified in the future with additional seed proteins. Notably, the HSP proto-module is highly overlapping with the ATX proto-module while the CMT2 proto-module is significantly separated. This overlap supports the ataxia-spasticity disease spectrum and warrants further investigation.\textsuperscript{194}

\textit{Ribosomial proteins in neurological pathogenesis}

The CMT2 and HSP disease modules were identified by expanding the seed protein sets with the recommended 200 iterations of the DIseAse Module Detection (DIAMOnD) method.\textsuperscript{81} Exploration of the DIAMOnD proteins revealed that close to 25\% are ribosomal proteins (Figure 7.4). Ribosomal proteins (RPs) surround a catalytic core of ribosomal RNA (rRNA) to form ribosomes.\textsuperscript{195,196} RPs have
universal roles in ribosome assembly and function, such as assisting in the proper folding of rRNA.\textsuperscript{196,197} Beyond these core roles, RPs have demonstrated extra-ribosomal functions both within the ribosome system, such as regulating and balancing individual RP synthesis with rRNA, and outside of the ribosome system, such as triggering apoptosis in response to disruption in ribosome synthesis.\textsuperscript{197} RPs are responsible for various phenotypes across organisms: the \textit{Minute} phenotype in \textit{Drosophila} (delayed development; short, thin bristles; and impaired fertility), gene-specific abnormalities caused by morpholino-induced knock-downs in zebrafish, developmental defects in mice, and 25\% of Diamond-Blackfan anemia cases in humans.\textsuperscript{196,198,199} Additionally, RPs were shown to be major interactors with kinase substrates of LRRK2 - a familial and sporadic PD protein.\textsuperscript{200} Specifically, phosphodeficient ribosomal small protein 15 (RPS15) rescued LRRK2 neurotoxicity by affecting translation.\textsuperscript{200} Impaired protein translation has been repeatedly implicated in CMT with mutations within five aminoacyl-tRNA synthetases identified to date.\textsuperscript{28} Furthermore, RPs have tissue-specific expression and have been experimentally shown to localize to nerves \textit{in vivo}, including mature peripheral nerve axons and cortical tract axons.\textsuperscript{201,202} Locally synthesized RPs can dynamically change specific ribosomal compositions which may allow axons to fine-tune mRNA selectivity.\textsuperscript{202} In this study, we show that RPs are appreciable interactors of inherited axonopathy disease proteins. Interestingly, an enrichment of ribosomal proteins was recently observed within the axonal transcriptome from fibroblast-derived human motor neurons (Figure 7.4 E&F).\textsuperscript{150} Forty-six of the
DIAMOnD RPs show modest localization between axon and cell body in a human motor neuron model and an additional four were significantly differential localized to the axon. Furthermore, we observe increased RP involvement in neurological disease controls compared to non-neurological disease (Figure 7.5). Taken together, our results support the emerging role of translational dysregulation in diverse neurologic diseases. We speculate that RPs may play an important function in neurologic diseases via their extra-ribosomal roles, including alerting the cell to stress.

*Pathway analysis provide insights into inherited axonopathy disease mechanisms*

Lastly, we compared the 6 inherited axonopathy spectrum module sets through pathway overrepresentation analysis, and assigned each enriched pathway to the most influenced disease module (CMT2-dominant, HSP-dominant, or AXONO) (Figure 7.6). Since each disease module set contained many ribosomal proteins, the most enriched pathways were the Cytoplasmic Ribosomal Proteins (Wikipathway) and Ribosome (KEGG) pathways. Aside from the ribosomal pathways, we identified pathways known to be involved in inherited axonopathies, such as aminoacyl-tRNA biosynthesis and protein processing in the endoplasmic reticulum (ER). Protein processing in the ER was significantly associated with each module involving the “axonopathy core” set, and is important in both CMT2 and HSP pathophysiology. The ER is the primary site for protein/lipid biosynthesis and intracellular calcium storage, and is tightly coordinated with the mitochondria to
maintain the energetically demanding mechanisms of long axons.\textsuperscript{203} Many affected axonopathy genes localize to the ER or mitochondria and disrupt the organelles’ coordination.\textsuperscript{204} For example, \textit{atlastin-1} can cause both CMT2 and HSP through abnormal ER shaping and mitochondrial transport.\textsuperscript{15,204}

Pathways that are not directly implicated in the axonopathy pathogenesis were also identified. The AXONO module was enriched for the Hedgehog signaling pathway and the Estrogen signaling pathway. Sonic hedgehog (Shh) is essential for central nervous system development and is expressed within injured sciatic nerve.\textsuperscript{205} Recently, reduced Shh signaling was shown to disturb axon regeneration via misregulated myelin degradation.\textsuperscript{205} Estrogen signaling has also been implicated in axon regeneration by accelerating peripheral nerve regeneration and increasing motoneuron participation.\textsuperscript{206,207} Another unexpected pathway enriched within the AXONO module was Pathogenic \textit{Escherichia Coli} infection. The \textit{E. Coli} pathways were significantly overlapping with Parkin-Ubiquitin Proteasomal System pathway through the tubulin proteins in each pathway (Figure 7.6 C). Alteration of the gut microbiome has been suggested to contribute to neurodegenerative diseases, such as Parkinson’s and Alzheimer’s disease.\textsuperscript{208} Furthermore, gut inflammation and microbiome alterations may play a role in the development or progression of motor neuron disease and ALS.\textsuperscript{208-210}

Distinct pathways were revealed between the CMT2-dominant and HSP-dominant modules. The CMT2-dominant module predominantly involved overlapping glucose metabolism pathways: HIF-1 signaling, glycolysis and
gluconeogenesis, Cori cycle, and pentose phosphate pathway (Figure 7.6 B). The disturbance of bioenergetics via mitochondrial dysfunction is a well-documented cause of CMT2.\textsuperscript{35,105,211} Furthermore, mutations in the gene encoding pyruvate dehydrogenase kinase isoenzyme 3 (PDK3) cause CMT, a protein which links the glycolytic cascade to the Krebs cycle.\textsuperscript{212} Based on our results, the involvement of glycolytic metabolism may be underappreciated in CMT2 pathophysiology. We speculate that glucose metabolism could contribute to CMT2 through reduced vesicular transport. Vesicular glycolysis is necessary and sufficient for the energy requirements of fast axonal transport (FAT).\textsuperscript{213} FAT depends on ATP generated by glycolysis, not by the mitochondria, and is reduced when glycolysis is perturbed.\textsuperscript{213} FAT alteration is linked to multiple neurodegenerative diseases, and FAT restoration is likely to contribute to neuronal protection.\textsuperscript{214}

The HSP-dominant modules were enriched for immune-related pathways: Epstein-Barr virus (EBV) infection, Herpes simplex (HSV) infection, and Antigen processing and presentation. EBV and HSV are members of the \textit{Herpesviridae} family and have large, complex KEGG pathways which involve multiple sub-pathways including, for example, antigen processing and presentation. For a more detailed understanding of which EBV and HSV processes were most involved, we investigated the GO BP terms of the genes involved in the EBV and HSV pathways. The EBV pathway genes were enriched for membrane organization, protein targeting, and protein insertion into the mitochondrial membrane during apoptosis induction. The HSV pathway genes were enriched for transcription from
the RNA polymerase II promoter. These results implicate antiviral mechanisms within HSP pathogenesis. Growing evidence documents the important roles of the mitochondria in antiviral immunity, including participation in signaling cascades and inflammation activation.\textsuperscript{215,216} Since mitochondrial dysfunction is known to cause HSP, mitochondrial involvement may link HSP to innate immune responses.\textsuperscript{217} Interestingly, EBV and human herpesvirus 6 (HHV-6) have been associated with multiple sclerosis (MS).\textsuperscript{218} HSP and MS both cause axonal loss of the corticospinal tract, share clinical similarities, and have concurred within families.\textsuperscript{219-221}

\textit{Concluding remarks}

In this study, we extracted insights about the spectrum of inherited axonopathies through the application of network-based tools to publically available protein interaction data. Starting with lists of known CMT2 and HSP disease genes, we determined significantly connected disease proto-modules from the global human interactome. The CMT2 and HSP proto-modules were topologically overlapping, supporting the consideration of these diseases as a related spectrum. We show the inherited axonopathies have an unclear (neither overlapping nor separated) relationship with ALS. These relationships will be elucidated as the molecular etiology of motor neuron diseases is further understood. We found a significant overlap between HSP and inherited ataxias, which is consistent with the ataxia-spasticity disease spectrum, while CMT2 was significantly separated from the
ataxias. This demonstrates the utility of network-based approaches in understanding the relationships between spectrums of closely related diseases. Each proto-module was expanded to identify additional proteins likely involved in disease pathogenesis, thus increasing the power of pathway analysis and providing candidate disease genes. We show that the inherited axonopathies are enriched for interactions with ribosomal proteins, which were recently shown to be preferentially localized to the axonal compartment of motor neurons. Pathway analysis revealed biological processes enriched for each axonopathy independently, such as glucose metabolism for CMT2 and innate immunity for HSP, as well as for general axonopathies, such as new signaling pathways. Additionally, the identified candidate disease genes may guide exploration into drug target identification. Sixty of the identified DIAMOnD proteins are present on the updated list of druggable genes: 49 DIAMOnDs may be targeted by a small molecule, 18 DIAMOnDs may be targeted by a biotherapeutic (monoclonal antibody/enzyme or other protein), and 1 DIAMOnD is involved in absorption, distribution, metabolism, and excretion of a compound. Follow-up functional studies will be required to evaluate the potential involvement of DIAMOnD proteins in disease and their therapeutic potential. Overall, our study provides useful insights into inherited axonopathies relationships and disease mechanisms, and demonstrates the opportunities available through network-based approaches.
CHAPTER 8

Conclusions

8.1 Summary of Findings

Length dependent degeneration of the long axons of motor (both upper and lower) and sensory neurons leads to a spectrum of disease known as inherited axonopathies (IA). On one end of the spectrum, predominantly central nervous system involvement causes HSP, characterized by bilateral lower extremity weakness and spasticity. Conversely, CMT results when the peripheral nervous system is primarily affected, causing distal muscle weakness and atrophy, decreased tendon reflexes, and sensory loss.

The advent of next-generation sequencing (NGS) has revolutionized the genetic diagnosis of IA with over 76 genomic loci and 58 corresponding genes for HSP and over 80 genes for CMT. The success of NGS has revealed several genes that can cause CMT, HSP, or a mixed phenotype, including NIPA1, REEP1, ATL1, SPAST, KIF1A, CCT5, BSCL2, and DNM2. Functional characterization of IA disease genes has also provided an improved understanding of the pathological mechanisms, including disruption of axonal transport, mitochondrial dynamics, mitochondrial regulation, membrane trafficking, and organelle shaping.

Despite these large advances in understanding IA etiology, the genetic diagnostic yield remains below 50% in IA families (with even less success in simplex families). My dissertation focused on illuminating the diagnostic gap that
remains in IA by exploring both the genetic causes and their molecular consequences. In this final chapter, I will summarize the important findings from my dissertation work and explain the importance of considering rare genetic mechanisms in Mendelian disease diagnostics.

*Characterizing and expanding the mutational spectrum of inherited axonopathies*

The first several chapters of my dissertation pertained to various approaches for addressing the “dark matter” of genetic diagnostics. The decrease in cost and increase in availability of NGS has created a deluge of human variation data, which is both exciting and burdensome. Interpretation of human variation includes annotations such as the variant’s functional consequence on the protein and its predicted pathogenicity, the allele frequency in large healthy populations, and the biological effects from functional assays. Standards and guidelines have been developed by the American College of Medical Genetics and Genomics (ACMG) to help clinical geneticists classify variants into 5 categories: pathogenic, likely pathogenic, uncertain significance, likely benign, and benign.\(^{175}\)

Despite immense efforts, many novel variants fall into the “uncertain significance” category, referred to as “Variants of Uncertain Significance” (VUS). Eventual interpretation of VUS will surely lead to more genetic diagnoses for IA families, helping to close the diagnostic gap through continued standard Mendelian analysis. To aid in VUS interpretation for CMT, I developed a web-based browser to store, rate, and communicate observed patient variation. We hope this provides a rich resource of detailed variant information that can be used in the future to
perform genotype-phenotype correlations and to define mutation-intolerant regions of disease proteins.

Next, my work yielded advancements in establishing the occurrence and prevalence of rare genetic mechanisms. First, I screened a cohort of probands from simplex HSP/ataxia families for uniparental isodisomy – the inheritance of identical copies of one parental chromosome due to meiosis II error or post-zygotic duplication of a chromosome. This genetic mechanism can create a homozygous variant from a single heterozygous carrier parent. If trios data is being analyzed, the homozygous variant may be considered a false call and discarded due to inconsistent inheritance. Even if the homozygous variant is not rejected, the risk of reoccurrence in another offspring is altered by uniparental isodisomy. I detected chromosomal isodisomy by screening WES data for homozygous regions greater than 10 Mb. I identified one case with complete isodisomy of chromosome 16 (1/96). On chromosome 16, the patient harbored a homozygous disease-causing mutation in known disease gene, FA2H. Though this variant was identified during standard analysis, we rejected it because Sanger analysis showed only one parent is a carrier. This study showed that, although uniparental isodisomy may be a rare cause of hereditary motoneuron disease, this mechanism can be reasonably detected from WES data and can potentially increase the diagnostic yield.

Next, I contributed to defining the prevalence of a previously unrecognized rare genetic mechanism: translation of cryptic amyloidogenic element (CAE) in the 3’ UTR. Frameshift mutations can cause loss of a protein’s natural stop codon, leading to continued translation into the 3’ UTR until the next stop codon. We
identified and functionally validated a CAE in the novel disease gene, \textit{NEFH}, in two dominant CMT2 families. To assess the potential prevalence of this disease mechanism, I screened the 3’ UTRs of all human genes for CAE in all open reading frames (ORFs). Ultimately, I found \( \sim 4,800 \) genes that could contain a CAE in at least one ORF. Although my results suggest that this mechanism should be much more common, several additional variables would affect a CAE’s disease-causing propensity including intracellular microenvironments, frequency of stop-loss mutations, and cellular expression. Nevertheless, this study identified a novel CMT gene, an unrecognized disease mechanism, and an avenue by which a "loss-of-function" mutation can behave in a dominant manner.

Lastly, I expanded the expected mutational spectrum of IA by exploring risk alleles and di/oligogenic inheritance in a large WES cohort. IA is typically approached through the lens of strict Mendelian inheritance; however, recent findings have begun to push these boundaries, such as an increased mutational burden and modifier alleles in inherited neuropathies.\textsuperscript{67,72} I performed an exome-wide gene-based rare variant burden analysis in 343 CMT cases and 935 non-neurological controls and identified \textit{EXOC4} as a candidate risk gene for inherited neuropathy. Further functional analysis is required to confirm the biological effect of the p.Gly550Arg allele, but it is a promising candidate based on previous studies. Additionally, I replicated the previously described mutational burden across CMT disease genes in a much larger cohort and observed a similar mutational burden across HSP disease genes. Combined with the nominally significant results of known CMT disease genes in the exome-wide analysis, these
results suggest that weaker alleles, in addition to the highly penetrant Mendelian alleles, exist in known disease genes. Furthermore, these additional alleles may indicate di/oligogenic inheritance. Indeed, I observed a significant proportion of cases harbored rare, non-synonymous variation in two or more disease genes. Taken together, results from this study demonstrate a wide mutational spectrum in known disease genes that remains to be functionally characterized.

Understanding the molecular relationships and consequences of inherited axonopathies

In addition to exploring the extent of both rare genetic mechanisms and an expanded mutational spectrum in IA, my dissertation also addressed understanding the underlying pathological mechanisms that unite and separate HSP and CMT. To begin studying these mechanisms in IA, we mapped the transcriptome of healthy iPSC-derived human motor neurons. Research into neurodegenerative disease pathophysiology and treatment development is hindered by limited access to human neurons. iPSC technology holds great potential for filling this research gap. After mapping the human motor neuron axonal transcriptome, we explored its uses by searching for transcripts of known neurological disease genes and nuclear-encoded mitochondrial genes. Interestingly, we found significant enrichment of mitochondrial genes and inherited axonopathy disease genes (BICD2, DST, KIF1C, KIF5C and MT-ATP6), most of which have a direct function in microtubule-based axonal transport. Our findings support the hypothesis that the high metabolic and axonal transport demands of
long motor axons require local axonal protein translation and will be a useful resource for investigating the dysregulation of mRNA localization in human neurodegenerative diseases.

Next, I sought to understand the molecular connections between IA disease genes. One advantage of high locus heterogeneity is the ability to interrogate the biological pathways underlying the disorder through a network biology approach. Genes and their products form complex networks of functionally related proteins that interact with each other to accomplish similar biological mechanisms, thus forming cellular pathways. Network medicine capitalizes on these interactions and hypothesizes that perturbation of a single gene product will propagate along the entire network. The interactions between a set of disease-causing genes can be summarized into a disease module for further study. I used publically available protein-protein interaction data to define the observed disease proto-modules for both CMT and HSP and demonstrated their topological overlap, indicating that CMT and HSP are likely to share disease pathomechanisms and clinical characteristics. Each proto-module was expanded to include 200 candidate disease genes. The expanded modules were enriched for interactions with ribosomal proteins, which were differentially localized to the human motor neuron axonal compartment as well. Lastly, pathway analysis of the unique and overlapping sets of disease module genes identified glucose metabolism pathways for CMT2 and innate immunity pathways for HSP. A deeper understanding of biological pathways in turn will help continue novel gene discovery, investigation
of relationships between human diseases, and identification of potential therapeutic targets.

### 8.2 Future Perspectives

Historically, many movement disorders have been clinicogenetically classified based on the predominant phenotype of the first gene locus.\(^\text{194}\) These classification systems have similar shortcomings, including erroneously assigned loci, duplicated loci, missing loci, and unconfirmed loci.\(^\text{194,226}\) Furthermore, these classification systems suggest that movement disorders are distinct and isolated when they in fact exist on a phenotypic spectrum. For example, HSP exists on a spectrum with both CMT2 and with inherited ataxias: not only do these disorders share clinical symptoms, such as prominent lower extremity spasticity, but they can also be caused by mutations within the same genes.\(^\text{170,194,227}\) Next-generation sequencing greatly facilitated the appreciation of these genetic overlaps by providing an unbiased approach that broke through the prior clinical and diagnostic preconceptions.\(^\text{194}\)

Awareness and consideration of phenotypic expansions will be essential for both individual genetic diagnoses as well as revealing common pathways underlying neurodegeneration.\(^\text{228}\) As the phenotypic spectrum broadens across the neurologic community, these historical classifications are being reconsidered.\(^\text{170}\) To address this issue, Synofzik and Schüle have proposed a mechanism based classification system for the ataxia-spasticity spectrum, based on unbiased modular phenotyping, that captures nuanced phenotypic expression,
opens ataxia and spasticity to a multisystem neuronal dysfunction, and help to prioritize research on shared pathways.$^1$94

A mechanism based classification system can also help form larger cohorts for statistical-based approaches by removing the historical barriers that currently isolate cases. Additionally, data aggregation and collaboration will be necessary to gather the required amount of rare disease cases to overcome power limitations caused by small sample sizes and genetic heterogeneity. Considering the challenges in genetic diagnostics across Mendelian diseases, it is likely that genetic mechanisms beyond traditional Mendelian inheritance will have a sizeable impact on increasing genetic diagnoses.

In this dissertation project, I considered CMT and HSP jointly as a phenotypic spectrum, called Inherited Axonopathies, and provide further evidence of the molecular overlap between these disorders. The identification of unifying biological pathways across phenotypic spectrums will aid in developing broadly-acting therapeutic interventions. Additionally, my dissertation results show the necessity of considering rare genetic mechanisms in order to increase genetic diagnostic yield, which is essential for novel therapeutic approaches such as gene replacement, antisense oligonucleotides, and forthcoming gene editing. Lastly, I demonstrated that IA contains a mutational burden across known disease genes which may contribute to phenotypic variability as well as risk alleles for more common forms of neuropathy and neurodegeneration.
REFERENCES


