Western University

Scholarship@Western

Electronic Thesis and Dissertation Repository

7-30-2021 2:15 PM

The Genetic Landscape of Neurodegenerative and Cerebrovascular Disease Phenotypes.

Allison A. Dilliott, *The University of Western Ontario*

Recommended Citation

Abstract

Neurodegenerative diseases are progressive, incurable conditions characterized by neuronal degeneration and protein aggregation, resulting in cognitive decline and/or motor dysfunction. Over half a million Canadians are affected with these diseases, and the number of cases is expected to rise as the aging population grows and average lifespans continue to increase. There are currently no curative treatments, and only few therapeutics are available to target disease symptoms or slow disease progression. Further, diagnosis can be challenging, relying on clinical features that are often highly heterogeneous between patients. Gaining a greater understanding of the full spectrum of genetic factors contributing to these diseases may offer the opportunity to more accurately assess risk of disease development, improve diagnosis, and identify potential therapeutic targets. By leveraging the unique study design of the Ontario Neurodegenerative Disease Research Initiative (ONDRI) — a multi-platform study characterizing neurodegenerative diseases and cerebrovascular disease (CVD) — I have made significant progress in the elucidation of overlapping genetic determinants across neurodegenerative diagnoses. Using a targeted next-generation sequencing (NGS) approach, I comprehensively genetically characterized the ONDRI cohort (n = 519), including participants diagnosed with: 1) Alzheimer's disease (AD); 2) amyotrophic lateral sclerosis (ALS); 3) frontotemporal dementia (FTD); 4) mild cognitive impairment (MCI); 5) Parkinson's disease (PD); and 6) CVD. I identified associations between common genotypes or haplotypes of high phenotypic effect and neurodegenerative disease presentation and features. I also assessed novel gene-disease relationships and the potential genetic overlap between neurodegenerative and cerebrovascular diagnoses through the assessment of rare genetic variation captured by the targeted NGS panel and gold-standard Sanger sequencing methods. Finally, I identified a novel association between Notch receptor 3 (*NOTCH3*) rare variants and CVD burden in participants with PD. The work presented throughout this Dissertation highlights the complexity of neurodegenerative disease genetic risk factors by demonstrating a large amount of overlap between specific diagnoses. The findings contribute to the longstanding effort to fully understand the genetic architecture of neurodegenerative diseases and improve therapeutic development, diagnostic tools, and progression prediction.

Keywords

Summary for Lay Audience

Neurodegenerative diseases are conditions characterized by progressive deterioration of brain cells resulting in memory loss, behavioural changes, and lack of muscle control. It was projected that more than 500,000 Canadians were affected with these diseases in 2016 — a number that will continue to rise in tandem with the aging population. Currently, neurodegenerative diseases lack treatment options with no ability to stop progression or even slow it to a manageable capacity, and accurate diagnosis can be difficult. To mitigate these issues, the Ontario Neurodegenerative Disease Research Initiative (ONDRI) study is aiming to gain a greater understanding of the risk factors and disease course of multiple neurodegenerative diseases including: 1) Alzheimer's disease; 2) amyotrophic lateral sclerosis; 3) frontotemporal dementia; 4) mild cognitive impairment; and 5) Parkinson's disease, as well as determining how cerebrovascular incidents, such as strokes, may influence neurodegeneration. Within this Dissertation, I present my work studying the DNA changes, or "genetic variants," that may be increasing the participants' risk of disease. To do this, I applied a DNA sequencing method to look for genetic changes within 80 genes that are known to contribute to the risk of neurodegenerative disease or stroke. In doing so, I determined that genetic risk for the various neurodegenerative diseases is complex, with some individuals carrying common variants that increase their disease risk and others carrying DNA variants that are much rarer. I also identified new associations between genetic variants in specific genes and individual neurodegenerative diseases. Finally, my results suggested that in addition to genetic variants that increase risk of disease, there may be genetic variants that cause individuals with the same diagnosis to present differently from one another. The application of my findings may provide further insight into what is causing neurodegenerative diseases on a molecular level and will allow for the development of new treatments and gene-sequencing-based early-diagnosis and risk-assessment tools.

Co-Authorship Statement

For all chapters included in this Dissertation, I am the first-listed author, or co-first author, on any resulting manuscripts. For all data chapters, I contributed to conceptualization, methodology, data curation and cleaning, formal analyses, writing of the original draft of the manuscript, and performing manuscript revisions. Aside from Chapter 7, which was a collaborative effort discussed further below, I wrote more than 90% of each manuscript.

My graduate supervisor, Dr. Robert A. Hegele, contributed supervision and guidance to all studies presented herein, as well as funding support. He also participated in study conceptualization, methodology, and manuscript preparation and revisions for all data chapters.

Hegele Lab core members contributed considerably to the execution of the research presented herein, including: 1) Adam D. McIntyre, who performed blood sample intake, as well as DNA extraction and isolation for ~60% of ONDRI participants; 2) Dr. Henian Cao, who prepared the ONDRI DNA samples for ONDRISeq next-generation sequencing; 3) Dr. Jian Wang, who performed the CNV breakpoint analysis and Sanger sequencing of *GBA*, presented in Chapter 5 and Chapter 6, respectively; and 4) John F. Robinson, who provided project management support for the ONDRI genomics platform. Additional Hegele Lab members also contributed to the research presented herein, including: 1) Emily C. Evans, who performed ~50% of the *APOE* genotype validation and all *MAPT* haplotyping using TaqMan allelic discrimination assays, included in Chapter 2 and Chapter 3, ~20% of DNA extractions and isolations, and ~80% of *GBA* PCR amplifications necessary for Chapter 6; 2) Abdallah Abdelhady, who assisted with the gene-based rare variant association analyses using SKAT-O in Chapter 4; 3) Kristina Zhang, who assisted with the CNV validation using whole-exome sequencing presented in Chapter 5; and 4) Alexis Vaillancourt, who assisted with the interpretation of *GBA* variants presented in Chapter 6.

Collaborators from ONDRI were also involved in all research projects presented herein. Dr. Michael J. Strong acted as the original ONDRI lead, followed by co-leadership by Dr. Douglas P. Munoz, Dr. Mario Masellis, and Dr. Richard H. Swartz. All ONDRI Investigators (**Appendix C**) contributed resources such as completing funding applications, performing

patient recruitment, and executing data acquisition and cleaning for the other platforms of ONDRI, of which the clinical, neuropsychology, and imaging data were used in studies presented herein. Specifically, Kelly Sunderland provided technical support for statistical analyses presented in Chapter 3, Chapter 4, and Chapter 6 and performed the neuropsychology data imputation in Chapter 3. Dr. Paula McLaughlin, Dr. Donna Kwan, and Dr. Angela Roberts conceptualized the framework for neuropsychology test binning in Chapter 3. Dr. Joel Ramirez analyzed the imaging data, and performed 50% of the tasks related to methodology, writing of the original manuscript draft, and manuscript revisions for the analyses of Chapter 7. Dr. Ramirez kindly agreed to the manuscript's use within this Dissertation. Dr. Malcolm Binns performed the Bayesian linear modelling included in Chapter 7.

Dedication

To those who inspired it

and will not read it.

Acknowledgments

Looking back on my years in graduate school, there are many people that I need to thank for helping and guiding me towards this final achievement. I am grateful for all of the opportunities I have been afforded and for all of the love and support I have received. I know my words here will fall short of the true appreciation that I feel, but I will do my best.

Dr. Hegele, in the summer between the second and third year of my undergraduate degree, my academic trajectory changed when I was attending an appointment with my Dad in your clinic. You were gracious enough to invite me to tour the Hegele Lab when I demonstrated interest in his genetic diagnosis and treatment, and the rest is history. The following seven years have been an absolute privilege, and I was so fortunate to be mentored by you. I cannot thank you enough for all of the opportunities you have given me. Thank you for guiding me when I was unsure, offering me independence to grow, trusting me to take the lead, celebrating my every success, and encouraging me after any speedbump. You have helped shape me into the researcher, and person, that I am today, and I will always be grateful for you.

I have also been very fortunate to spend my graduate school career in the most supportive environment. Adam, Brooke, David, Ericka, Henian, Jenn, Jian, John, and Matt, you have all played a major role in helping me achieve my goals and aspirations. I could always count on the 'adults' to let me know when there were snacks in the lunchroom and when product shows were happening. But beyond that, your patience, guidance, and encouragement allowed me to thrive. Thank you for all you have done to help me get to where I am today.

Additionally, thank you to Dr. Greg Gloor and Dr. Manuel Montero-Odasso for your mentorship as my committee members. I valued your kind and constructive feedback, and I always left our meetings with confidence to pursue my upcoming tasks. And a special thank you to Dr. Murray Huff, who often felt like an honorary committee member. I could always count on you to stop by my posters or attend my presentations — often stumping me with your excellent questions. I looked forward to the days you would pop by our office and put your nose against our window. You were a guaranteed smiling face on the fourth floor of Robarts.

Jacqueline… where do I begin? Sharing an office with you has been an absolute honour, and I look up to you in so many ways. You are a true superstar, and your work ethic and talent speak for themselves. Thank you for always being willing to give me a genetics crash course, for listening to every vent session, and for allowing me to bounce my ideas off of you, often followed by you giving the best feedback. I was so fortunate to walk through graduate school in your footsteps, and through it all, you have also become one of my closest friends. Your kindness and selflessness are inspiring, and I am so thankful that our worlds collided.

Sali, in the summer before my fourth year of undergrad, you sought out my help on your work with ONDRI. Although I was shy and nervous, you took me under your wing, taught me everything I needed to know, and instantly became one of my greatest role models. Thank you for trusting me with a project that you worked so hard at launching. I am thrilled to join your lab at McGill in the Fall, and am grateful to have you as both my mentor and friend.

I am also thankful for the other friendships I have formed throughout the wild ride of grad school. Rosettia and Michael, some of my best memories are from navigating the first two years of grad school with you (and Jacqueline too, of course). Thank you for being such wonderful friends. Between the inside jokes, wild antics, love of doggos, and life chats, you made the office feel like a second home. And the latter half of my graduate studies would not have been the same without you, Julie. I am grateful that we were able to travel together and become good friends before COVID-19 kept us apart. As well, thank you to all of the other amazing friends I have made at Robarts and Schulich: Amy, Caroline, Corey, Gabe, Nadya, Niveen, Zach, and Zainab. You all contributed to a fantastic graduate school experience!

Emily E., Abdalla, Alexis, and Kristina, I am grateful to have had the opportunity to work with and get to know you all. Extraordinary undergraduate students are invaluable, and I am so appreciative of the effort and care you put into your projects. You are now all off achieving amazing things, and I can't wait to see where your journeys take you next, with hope that our paths will cross again in the future.

It has been an honour to work with ONDRI over the last six years. I would like to thank everyone involved, including the scientists, clinicians, and scholars. Being a part of such a collaborative team has allowed me to learn and grow in many ways, and I am confident that

the opportunities I have had working on this study will stay with me throughout my career. In particular, Dr. Ekaterina Rogaeva, thank you for your continuous mentorship, and for always showering me with kindness whenever we had the opportunity to spend time together or chat. Your encouragement in some of the most intimidating situations meant the world to me. Dr. Joel Ramirez, thank you for being such a motivated collaborator, as well as a great friend during all ONDRI meetings. And Kelly Sunderland, thank you for every time you patiently walked me through a statistics explanation — teaching me more than I ever learned in a classroom — and for staying on our Zoom calls just to chat. I look forward to continued collaborations with you all! Most importantly, thank you to every ONDRI participant and their loved ones. The privilege of studying the data you provided will never be lost on me. This Dissertation would not have been possible without each one of you.

To all of my friends outside of graduate school, thank you for always listening to me talk about my research, and for giving me a great excuse not to think about it every so often. Emily J. you have been by my side more than 12 years and no matter the different directions our paths go, our friendship will always feel like home. Jayme and Marina, in you both I truly found myself [two] cheerleader[s], always right there when I need you. And Sam and Chelsey, thank you for indulging in The Bachelor with me and taking my mind off of the lab on a weekly basis.

Mom and Dad, thank you for instilling in me from a young age that I could achieve anything. Your constant confidence in my abilities has been the greatest motivation. Thank you for all you gave up and for every effort you made to allow graduate school to be a reality for me. Greg and Jamie, thank you for always reminding me not to "waste my brain" and for teaching me to take… let's call it "harsh criticism." It's a skill highly sought after in academia! But in all seriousness, thank you for always rooting for me and making me have some fun every once in a while. Rachel and Stephanie, thank you for being the sisters (in-law) I always wanted! I love you all and appreciate everything you have done for me.

Eric, I often wonder if I would have made it through the last few years of graduate school without you — let alone a worldwide pandemic. And maybe I would have, but it wouldn't have been nearly as fun. Thank you for keeping me sane and being the voice of reason I so often needed. Thank you for being my personal chef and keeping me fed. Thank you for

waking up first thing in the morning and taking Henley out without ever being asked. Thank you for going for ice cream on my hardest days, as well as my best. And thank you for reading and editing every word of this Dissertation. I am so lucky to have you as my partner, and I can't wait for our next adventure together.

And finally, tiny Hen. There were days you were a challenge far greater than this Dissertation, but I wouldn't have had it any other way.

Funding Acknowledgements

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

1000G	1000 Genomes Project
A	Adenine
Aβ	Amyloid-β
ABCA7	ATP binding cassette subfamily A member 7
ABCC6	ATP binding cassette subfamily C member 6
ACMG	American College of Medical Genetics and Genomics
AD	Alzheimer's disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
ALS	Amyotrophic lateral sclerosis
ANNOVAR	Annotate variation
APOE	Apolipoprotein E
APP	Amyloid precursor protein
ATP13A2	ATPase cation transporting 13A2
ATXN2	Ataxin 2
BAM	Binary alignment map
BED	Browser extensible data
BP	Blood pressure
bvFTD	Behavioural variant frontotemporal dementia
C	Cytosine
C9orf72	Chromosome 9 open reading frame 72
CAA	Cerebral amyloid angiopathy
CADASIL	Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy
CADD	Combined Annotation Dependent Depletion
CARASIL	Cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy
CBS	Corticobasal syndrome
CGH	Comparative genomic hybridization
CHMP2B	Charged multivesicular body protein 2B
CI	Confidence interval
ClinGen	Clinical Genome Resource
CNV	Copy number variant
COL4A1	Collagen type IV alpha 1
COL4A2	Collagen type IV alpha 2

CVD	Cerebrovascular disease
CVD ± CI	Cerebrovascular disease with or without cognitive impairment
ddNTP	Dideoxyribonucleotide
DJ1	Parkinsonism associated deglycase
DNA	Deoxyribonucleic acid
DOC	Depth of coverage
EGF	Epidermal growth factor
EOAD	Early-onset Alzheimer's disease
EOPD	Early-onset Parkinson's disease
ExAC	Exome Aggregation Consortium
fALS	Familial amyotrophic lateral sclerosis
FISH	Fluorescent *in situ* hybridization
FLAIR	Fluid attenuated inversion recovery
FTD	Frontotemporal dementia
FUS	FUS RNA binding protein
G	Guanine
GBA	Glucocerebrosidase
GBAP	Glucosylceramidase beta pseudogene 1
GCase	Glucocerebrosidase
GCEPs	Gene Curation Expert Panels
GCH1	GTP cyclohydrolase 1
GnomAD	Genome Aggregation Database
GRN	Granulin precursor
GWAS	Genome wide association study
HbA1C	Hemoglobin A1C
HDL	High density lipoprotein
HGMD	Human Gene Mutation Database
HTRA1	High-temperature requirement A serine peptidase 1
InDel	Insertion or deletion variant
LD	Linkage disequilibrium
LDL	Low-density lipoprotein
LOAD	Late-onset Alzheimer's disease
LOF	Loss-of-function
LRRK2	Leucine-rich repeat kinase 2
MAF	Minor allele frequency

MAPT	Microtubule associated protein tau
MC1R	Melanocortin 1 receptor
MCI	Mild cognitive impairment
MDS-UPDRS	Movement Disorder Society – Unified Parkinson's Disease Rating Scale
MoCA	Montreal Cognitive Assessment
MRI	Magnetic resonance imaging
mRNA	Messenger ribonucleic acid
NEFH	Neurofilament heavy chain
NGS	Next-generation sequencing
NOTCH3	Notch receptor 3
OBI	Ontario Brain Institute
OMIM	Online Mendelian in Man
ONDRI	Ontario Neurodegenerative Disease Research Initiative
OPTN	Optineurin
OR	Odds ratio
PARK2	Parkin RBR E3 ubiquitinated protein ligase
PARK7	Parkinsonism associated deglycase
PCA	Principal component analysis
PCR	Polymerase chain reaction
PD	Parkinson's disease
PINK1	PTEN induced kinase 1
PNFA	Progressive non-fluent aphasia
PolyPhen-2	Polymorphism Phenotyping v2
PPA	Primary progressive aphasia
PPMI	Parkinson Progressive Marker Initiative
PRS	Polygenic risk score
PSEN1	Presenilin 1
PSEN2	Presenilin 2
PSP	Progressive supranuclear palsy
PXE	Pseudoxanthoma elasticum
qPCR	Quantitative polymerase chain reaction
RVAA	Rare variant association analysis
sALS	Sporadic amyotrophic lateral sclerosis
SAMHD7	SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1

SD	Semantic dementia
SIFT	Sorting Intolerant from Tolerant
SKAT-O	Optimal sequence kernel association test
SNCA	α-synuclein
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOD1	Superoxide dismutase
ST-TIV	Supra-tentorial intracranial volume
T	Thymine
TARDBP	TAR DNA binding protein
UNC13A	Unc-13 Homolog A
UTR	Untranslated region
VaD	Vascular dementia
VCF	Variant calling format
VCI	Vascular cognitive impairment
VPS35	VPS35 retromer complex component
WES	Whole exome sequencing
WGS	Whole genome sequencing
WMH	White matter hyperintensity

Chapter 1 – Introduction

1.1 Overview

Neurodegenerative diseases are a collection of progressive and incurable conditions characterized by neuron degeneration and resulting in cognitive decline — also referred to as dementia — and/or motor dysfunction. It was projected that, in 2016, more than 500,000 Canadians were affected by the diseases, and worldwide prevalence is predicted to double within the next 20 years ("Chapter 3: Mapping Connections: An understanding of neurological conditions in Canada – Scope (prevalence and incidence)," 2014; Dudgeon, 2010). Similarly, when considering dementia patients alone, care costs totaled over $10 billion in 2016 and are estimated to double by 2031, including direct costs to the healthcare system and out of pocket costs of patients and their loved ones. Aside from these financial burdens, caregivers of individuals with dementia performed a collective 19.2 million hours of unpaid care across Canada in 2011 (Chambers, Bancej, & McDowell, 2016). Further, the World Health Organization has estimated that, by 2040, as the aging population is increasing, neurodegenerative diseases will overtake cancer as the second leading cause of death worldwide, only behind cardiovascular disease (Gammon, 2014).

Although encompassing a wide variety of specific diagnoses, all neurodegenerative diseases share common features, most notably, the greatest risk factor for the diseases' developments — increasing age (Hou et al., 2019). The specific diagnoses can be categorized into two main subtypes: 1) motor/movement disorders and 2) cognitive and/or behavioural disorders. The former encompasses diseases characterized by motor impairment as a result of progressive neuronal loss, including amyotrophic lateral sclerosis (ALS) and Parkinson's disease (PD). In contrast, cognitive and/or behavioural disorders are largely caused by protein aggregates within the brain and/or cerebral small vessel pathology and result in progressive cognitive decline, changes to personality, and language impairment. Alzheimer's disease (AD), frontotemporal dementia (FTD), and vascular cognitive impairment (VCI), are all examples of common cognitive and/or behavioural disorders.

A second notable similarity between neurodegenerative diagnoses is the presence of misfolded proteins and their resultant aggregates within the brain and central nervous system. Although the specific proteins involved vary between diagnoses, all neurodegenerative diseases are characterized by these aberrant accumulations (Dugger & Dickson, 2017). Common examples of the protein inclusions include amyloid plaques, neurofibrillary tau tangles, and α-synuclein-based Lewy bodies. Importantly, the aggregates present within the brains of neurodegenerative disease patients are considered the best indication of an individual's true diagnosis, as clinical presentations of the different diseases are highly variable with many overlapping features (Dugger & Dickson, 2017). Upon postmortem pathologic analysis, it is not uncommon to observe aggregates of multiple neurodegenerative diseases, indicating co-pathologies. Further, it is estimated that protein aggregation may begin 10–20 years prior to clinical onset of disease features (Katsuno, Sahashi, Iguchi, & Hashizume, 2018).

While a considerable amount of effort has been put into understanding the pathogenesis of neurodegenerative diseases, there is still an absence of appropriate treatment options. Currently, no neurodegenerative diseases are curable, and the majority of available therapeutics are only able to target specific symptoms and slow disease progression to a minimal degree (Duraes, Pinto, & Sousa, 2018). It is postulated that the lack of appropriate treatment options is a direct result of a lack of full understanding regarding the neuropathologic mechanisms of the diseases (Duraes et al., 2018).

Interestingly, neurodegenerative disease diagnosis also remains a challenge, particularly in the early stages, without the ability to accurately detect all protein aggregates while patients are still alive. Current diagnostic approaches rely on clinical features of disease; however, these are often highly variable with large amounts of heterogeneity within and between disease cohorts. As a result, misdiagnoses are, unfortunately, not uncommon (Beber & Chaves, 2013; Bicchi, Emiliani, Vescovi, & Martino, 2015; Selvackadunco et al., 2019). Yet, early and accurate diagnosis is critical, as the therapeutics that are available to patients must begin as early as possible to be most effective at mitigating symptoms and slowing disease progression (Agrawal & Biswas, 2015).

By elucidating the entire genetic landscape of the neurodegenerative disease continuum, we may not only identify more appropriate therapeutic targets by gaining a better understanding of the pathologic mechanisms involved, but we may also improve our ability to diagnose individuals, both early and accurately. Genetic analysis offers a promising opportunity for diagnosis, as heritability estimates — or the amount of risk estimated to be from genetic variation — across the diseases are relatively high (Bocchetta et al., 2016; Pang et al., 2017; Postuma et al., 2016; Strong et al., 2017). We may also be able to leverage an individual's genetic profile to gain a better understanding of how they may progress through their disease. However, to utilize genetic testing for diagnosis or progression prediction, greater understanding regarding the types of variation that contribute to neurodegeneration is necessary. Additionally, the large amount of overlap between neurodegenerative disease intermediate phenotypes and neuropathological hallmarks suggests that we must thoroughly define the potential overlap of genetic risk factors contributing to multiple diagnoses.

The Ontario Neurodegenerative Disease Research Initiative (ONDRI) is a longitudinal, observational cohort study that has aimed to fully elucidate the spectrum of neurodegenerative diseases, as well as the potential involvement of cerebrovascular disease (CVD) within the diagnoses (S. M. K. Farhan et al., 2017; Sunderland et al., 2020). With the recognition that much remains to be known regarding the large amount of genetic and phenotypic heterogeneity within and between the different neurodegenerative diseases, ONDRI proposed a novel study design providing both a longitudinal nature of assessment and follow-up, as well as simultaneous analysis of patients across five neurodegenerative phenotypes, including AD, ALS, FTD, mild cognitive impairment (MCI), and PD. The model is in contrast to the majority of prior work within the field that has taken a reductionist approach, staying within individual diagnostic silos and failing to account for the entire spectrum of neurodegeneration. ONDRI also takes a multimodal approach with multiple assessment platforms, which include neuroimaging, neuropsychology, clinical assessment, gait and balance, ocular control and morphology, and genomics, allowing for large-scale collaborative efforts to characterize the full presentation of the diseases under the ONDRI mandate.

The genomics platform of ONDRI specifically aimed to begin mitigating the gaps in the current understanding of the genetic underpinnings of neurodegenerative diseases. By leveraging the broad range of neurodegenerative diseases under study, we may be able to identify overlapping genetic risk factors not previously considered by researchers studying only individual neurodegenerative disease cohorts, thereby accounting for a portion of the missing heritability. Further, the rich phenotypic assessment of participants may allow for the identification of genetic determinants contributing to the heterogenous phenotypic intermediates observed across neurodegenerative cohorts.

1.2 Human genetics

Human genetics encompasses the study of variation in human deoxyribonucleic acids (DNA), how that variation is inherited, and the contribution of DNA variation to traits and disease. By studying our DNA, not only are we able to gain a greater understanding of how the human body functions, but also a critical knowledge regarding why and how traditional human function can go wrong, leading to important breakthroughs in disease prediction, diagnosis, and treatment.

The DNA molecule, often referred to as the building block of human life, is structured as a double-stranded helix composed of two strands of alternating phosphate and deoxyribose groups with nitrogen bases attached to each deoxyribose (Youssef, Budd, & Bielawski, 2019). There are four nitrogen groups, including adenine (A), cytosine (C), guanine (G), and thymine (T), and each base is able to form a specific covalent bond with another base, such that the two strands of DNA are held together. A always bonds with T, and C bonds with G. Together, the deoxyribose, phosphate, and base make up a single "nucleotide." Importantly, the order in which the nucleotides appear within a strand of DNA is referred to as the "DNA sequence" (Youssef et al., 2019).

Each cell in the human body contains two copies of DNA that, if stretched out, would be approximately 2 m in length; therefore, it is important that the DNA is effectively packaged within the cell nucleus. The molecules are tightly wound around histone proteins, forming chromatin that is further wound into chromosomes (Youssef et al., 2019). Humans cells are "diploid," meaning that the two copies of DNA are organized

into two sets of chromosomes. There are 22 autosomal chromosomes in each set, numbered one through 22, as well as a pair of sex chromosomes — two X in females and an X and a Y in males.

A single copy of human DNA encompasses approximately 3,200,000,000 nucleotides, or "base pairs" (bp) (Piovesan et al., 2019). The nucleotides are organized such that certain regions are protein coding, otherwise referred to as "genes." The genes are often separated by stretches of non-coding DNA of varying length, called "intergenic regions." Within a gene, each triplet of nucleotides, or "codon," encodes a specific amino acid, that when read together results in the production of a protein specific to the gene in question (Nirenberg & Matthaei, 1961). Notably, multiple different codons can encode the same amino acid, often referred to as "redundancy" in the genetic code. In this way, genes are transcribed into strands of messenger ribonucleic acid (mRNA) that are further translated into protein products. The resulting proteins will perform highly specific functions on a cellular level within the human body (Saier, 2019).

There are approximately 24,000 protein-coding genes in the human genome that are further organized into multiple "exons," referring to the regions of the gene that encode the protein product, and "introns," which are non-coding regions and are spliced out of the mRNA before it is translated into a protein (International Human Genome Sequencing, 2004). During the process of mRNA splicing, certain exons may or may not also be included or excluded resulting in different mRNA isoforms that encode variations of the protein product, referred to as "alternative splicing" and resulting in "differential transcripts" (**Figure 1.1**) (Berget, Moore, & Sharp, 1977; Chow, Gelinas, Broker, & Roberts, 1977; Moraes & Goes, 2016). The human "exome" refers to all protein-coding exons of all genes in the human genome.

In April 2003, the Human Genome Project was completed, which produced the first readout of nearly the entire human genome sequence, including the identification and mapping of almost all genes (International Human Genome Sequencing, 2004; Lander et al., 2001). The completion of this endeavour was monumental and resulted in a turning point in the field of human genetics. With the full sequence, researchers were finally able

to have a complete picture of the complexity of the human genome and could use it as a reference to aid in the identification of changes within the DNA sequence that could contribute to various traits and disease states. Although further studies were needed to fill in the small number of remaining gaps of the human genome since the Human Genome Project was released (Chaisson et al., 2015; Dolgin, 2009; Miga et al., 2020), and a recent preprint suggests the final components have now been completed (Nurk et al., 2021), multiple nearly complete versions of the human reference genome are widely available and considered highly accurate for use in human genetic studies.
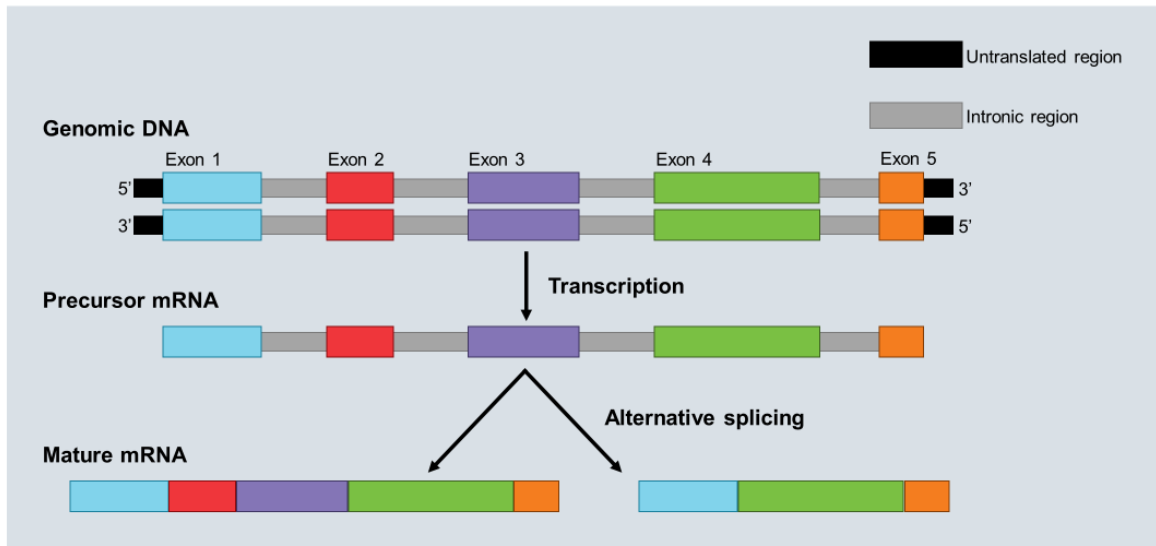
Figure 1.1 Alternative mRNA splicing of transcribed precursor mRNA to produce differential mRNA transcripts.
Genes are organized into multiple exons, which encode the resulting protein product, and introns, which are spliced out of the mRNA before it is translated into a protein. During the process of mRNA splicing, certain exons may be included or excluded resulting in different mRNA isoforms that encode variations of the protein product. The process is referred to as alternative splicing and results in differential transcripts.

1.2.1 Genetic variation

Quite remarkably, a person's DNA sequence shares > 99% similarity with any other person's DNA sequence; however, the ~1% of differences within the genome distinguish individuals from one another (Auton et al., 2015). Deviations in a person's DNA sequence from the generally accepted human reference genome are referred to as "genetic variants." These variants exist on a spectrum of consequence ranging from no effect to lethality. Variants may contribute to a certain trait, cause or increase risk of disease, or even be protective against disease. Variants also exist in a variety of sizes — ranging from single nucleotide variations affecting only one base in the genome, to structural variants affecting multiple bases or entire exons, genes, or chromosomes — and frequencies in the general population (Auton et al., 2015).

Although variants occurring in the intergenic region of the genome or intronic region of a gene may be of some consequence (F. Zhang & Lupski, 2015; Zou, Wu, Tan, Shang, & Zhou, 2020), they remain relatively difficult to interpret. In contrast, variants occurring in the human exome, or within ten nucleotides of an intron-exon junction (the "splicing region") are considered those most likely to be of consequence due to their potential direct effect on the encoded protein.

1.2.1.1 Single nucleotide variation

A "single nucleotide variant" (SNV) consists of a single base pair that deviates from the expected nucleotide based on the human reference genome. SNVs are first classified into two categories: 1) "transitions" and 2) "transversions" (Shastry, 2009b). A transition is a substitution between the nucleotides A and G — the "purine" nucleotides — or between C and T — the "pyrimidine" nucleotides. Conversely, a transversion is a substitution from one purine nucleotide to a pyrimidine, or vice-versa.

SNVs can be further classified based on their resulting "sequency ontology", referring to the impact the SNV has on a protein product of a gene (**Figure 1.2**). Although SNVs located within the intronic region of a gene are often not of consequence, some can dysregulate mRNA splicing, particularly those occurring within the splicing region (Nik & Bowman, 2019; Pagani & Baralle, 2004). mRNA splicing variants located at the

beginning of an intron are referred to as "splice-donor" variants, and mRNA splicing variants located at the end of an intron are referred to as "splice-acceptor" variants (Anna & Monika, 2018). SNVs located within an exon can have a variety of consequences on the protein. Due to redundancy in the genetic code, an SNV may result in a "synonymous" variant that does not change the amino acid at the corresponding protein position (Saier, 2019). However, if the SNV results in a new codon that does not encode the same amino acid at the corresponding position in the protein, the variant is referred to as "nonsynonymous". More specifically, a nonsynonymous SNV may be of three types: 1) a "missense" variant resulting from the substitution of one amino acid for another; 2) a "nonsense" variant resulting from the substitution of the original amino acid for a premature stop codon, and therefore, a truncated protein product; or 3) a "nonstop" variant resulting from the substitution of the original stop codon for an amino acid, and therefore, a longer-than-normal protein product.

SNVs that occur outside of the coding region or splicing region of a gene are also widely prevalent, including those within introns, untranslated regions (5' and 3' UTRs), promoters, enhancers, silencers, non-coding genes, or intergenic regions, but are much more difficult to interpret.
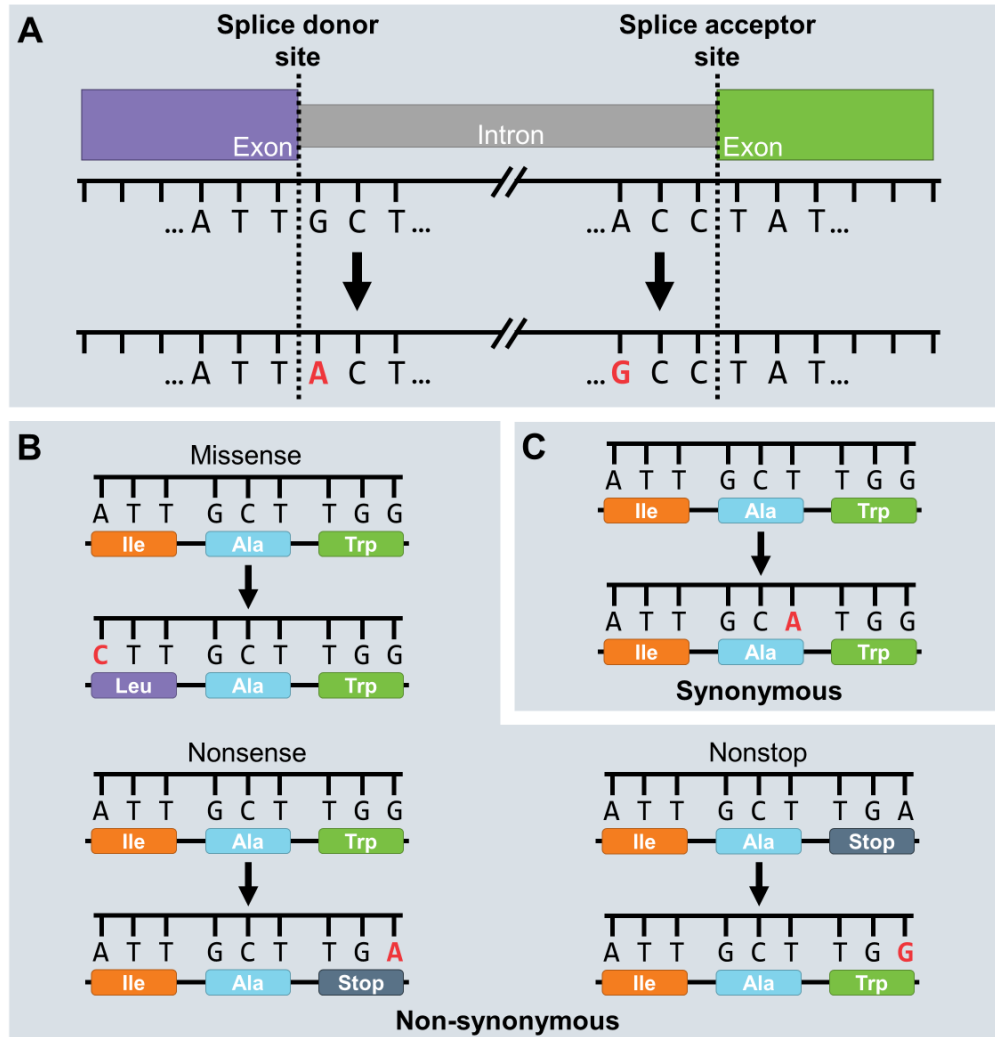
Figure 1.2 Various types of possible single nucleotide variants (SNVs), classified based on sequence ontology.
Sequence ontology refers to the impact an SNV may have on the resulting protein. **A)** A splicing variant a variant is located in an intron of a gene, but is within 10 nucleotides of the nearest exon-intron junction. Splicing variants located at the beginning of an intron are splice-donor variants, and those located at the end of an intron are splice-acceptor variants. Importantly, these variants can result in changes to how mRNA is spliced, thereby resulting in mRNA that may be missing or including entire exons that the transcript would or would not normally have, respectively. **B)** Nonsynonymous variants refer to SNVs resulting in changes to the amino acid encoded at that location. Missense variants are SNVs that result in the change of a single amino acid to another amino acid. Nonsense variants are SNVs that result in the change of a single amino acid into a stop codon. Nonstop variants are SNVs that result in the change of a stop codon into an amino acid. **C)** Synonymous variants are SNVs that do not change the encoded amino acid due to the redundancy of the genetic code.

1.2.1.2   Structural variation

Generally, all genetic variation can be included in the category of "structural variation", aside from SNVs (Ku, Loy, Salim, Pawitan, & Chia, 2010). The umbrella term encompasses variants ranging from insertions or deletions of only a few nucleotides to large-scale insertions, deletions, or transversions of entire pieces of chromosomes (**Figure 1.3**). Structural variants may be "balanced," such that the number of nucleotides is unchanged, but the location of the nucleotides changes (for example, in inversions and translocations) or the variants may be "unbalanced," resulting in the loss or gain of nucleotides (for example, in insertions, deletions, tandem repeat expansions, or copy number variants (CNVs)). Typically, the structural variants occur as a result of inaccuracies during DNA replication or recombination (Hurles, Dermitzakis, & Tyler-Smith, 2008).

It is largely accepted that structural variants tend to be more deleterious than SNVs due to the greater number of nucleotides affected (Sudmant et al., 2015). In fact, the larger the variant, the more likely it is predicted to be deleterious, with the largest structural variants (>250,000 bp) almost exclusively occurring in the general population at frequencies of <1% (Conrad, Andrews, Carter, Hurles, & Pritchard, 2006). Structural variants are also of greater likelihood to be deleterious when encompassing protein-coding genes or regulatory regions. Yet, there remains many reports of structural variants that have no effect or protective effects against certain conditions (Feuk, Marshall, Wintle, & Scherer, 2006).
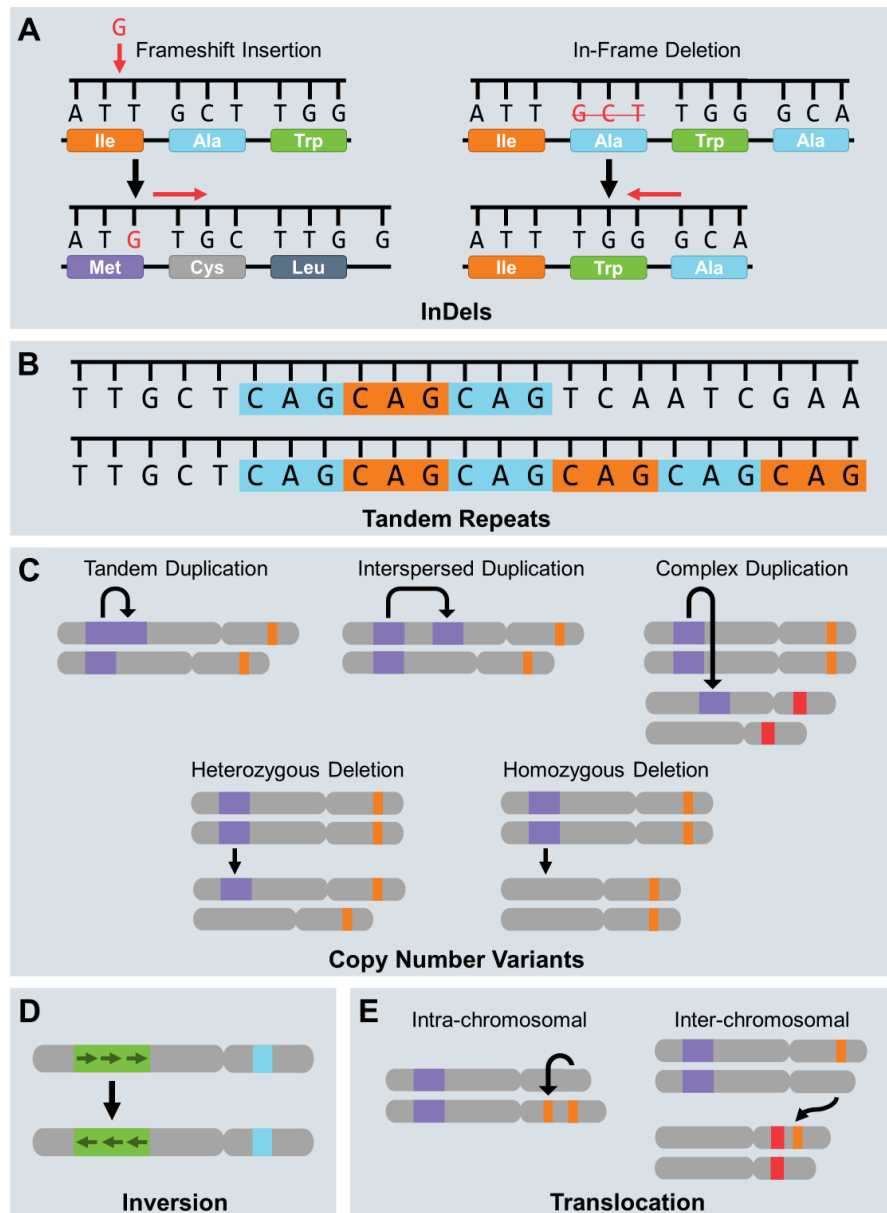
Figure 1.3 Different types of potential structural variation.
Structural variants generally encompass all variants aside from single nucleotide variants (SNVs). **A)** Insertion/deletions (InDels) range in size from 1–50 nucleotides and result in a gain or loss of nucleotides, respectively. **B)** Tandem repeats are a subset of repetitive DNA elements that consist of a repeating motif of bases in a localized region of the genome. Motifs of 1–6 nucleotides are called microsatellites, motifs > 6 nucleotides are called minisatellites. **C)** Copy number variants (CNVs) are large-scale deletions or duplications of DNA greater than 50 nucleotides in length. Duplications may occur in tandem, be interspersed to a different region of the same chromosome, or be considered complex and located on a different chromosome than the original sequence. Finally, **D)** inversions and **E)** translocations are balanced structural variants, such that copy number of the locus does not change, but the position does.

1.2.1.2.1 Insertions and deletions

"Insertions" or "deletions" of nucleotides (InDels) can range in size from 1–50 bases; although, most common are those one to three nucleotides in length, which make up ~70% of all InDels (**Figure 1.3A**) (Lin et al., 2017). Although there are multiple mechanisms that can cause InDels to occur within the genome, typically they are a result of polymerase slippage during DNA replication processes (Taylor, Ponting, & Copley, 2004).

As with all genetic variation, InDels can occur at any location, or "locus," throughout the genome, but are of particular interest when located within an exonic region of a gene. In a coding region, if an InDel's length is a multiple of three nucleotides, it is referred to as an "in-frame" variant, and it results in the insertion or deletion of one or more amino acids within a protein. In contrast, if the length of the InDel is not a multiple of three, it is called a "frameshift" variant, as all downstream nucleotides will be shifted in position resulting in disruption of the remaining codon sequence (Kuntzer, Eggle, Klostermann, & Burtscher, 2010). Ultimately, this results in variation with all downstream amino acids in the corresponding protein and often also introduces an unexpected stop codon. The premature termination of the protein may trigger a nonsense-mediated decay pathway during mRNA translation, which results in degradation of the mRNA and no production of the protein (Kuntzer et al., 2010). For these reasons, frameshift variants typically result in a large amount of biological consequence and are considered "loss-of-function" (LOF) variants.

1.2.1.2.2 Tandem repeats

Over 50% of the human genome is comprised of repetitive sequences of DNA, including many located within genes or regulatory regions (Liang, Tseng, Tsai, & Sun, 2015). "Tandem repeats" are a subset of repetitive DNA elements that consist of a repeating motif of bases in a localized region of the genome (**Figure 1.3B**). The repeats can be further subclassified into "microsatellites" with motifs of 1–6 bp and "minisatellites" with motifs of >6 bp (Hannan, 2018).

Tandem repeats are highly unstable, meaning they are prone to additional variation, typically encompassing increases or decreases to the number of repeated motifs (Fan & Chu, 2007). Although there are multiple events that may result in changes to the number of repeats, similarly to InDel variants, the main pathway is polymerase slippage during DNA replication. Briefly, during this process the two DNA strands are separated, but when they re-pair, one strand "slips" and mispairing occurs. This creates a loop of DNA at the mismatch site (**Figure 1.4**). The loops are often recognized by DNA repair systems; however, when those mechanisms fail, the number of repeats is altered (Fan & Chu, 2007).

Tandem repeat disorders are caused by tandem repeats located in specific regions or genes that have exceeded lengths considered to be of generally accepted normalcy, although these thresholds vary based on a variety of factors. Often, the disorders do not present as binary phenotypes, rather are expressed across a spectrum of disease states, typically with more severe outcomes directly correlated to the length of the associated tandem repeat (Hannan, 2018). Section 1.2.1.1.1 will introduce a hexanucleotide tandem repeat expansion in the 5' regulatory region of the chromosome 9 open reading frame 72 gene (*C9orf72*) that is the most frequently inherited genetic cause of ALS and FTD (DeJesus-Hernandez et al., 2011; Renton et al., 2011).
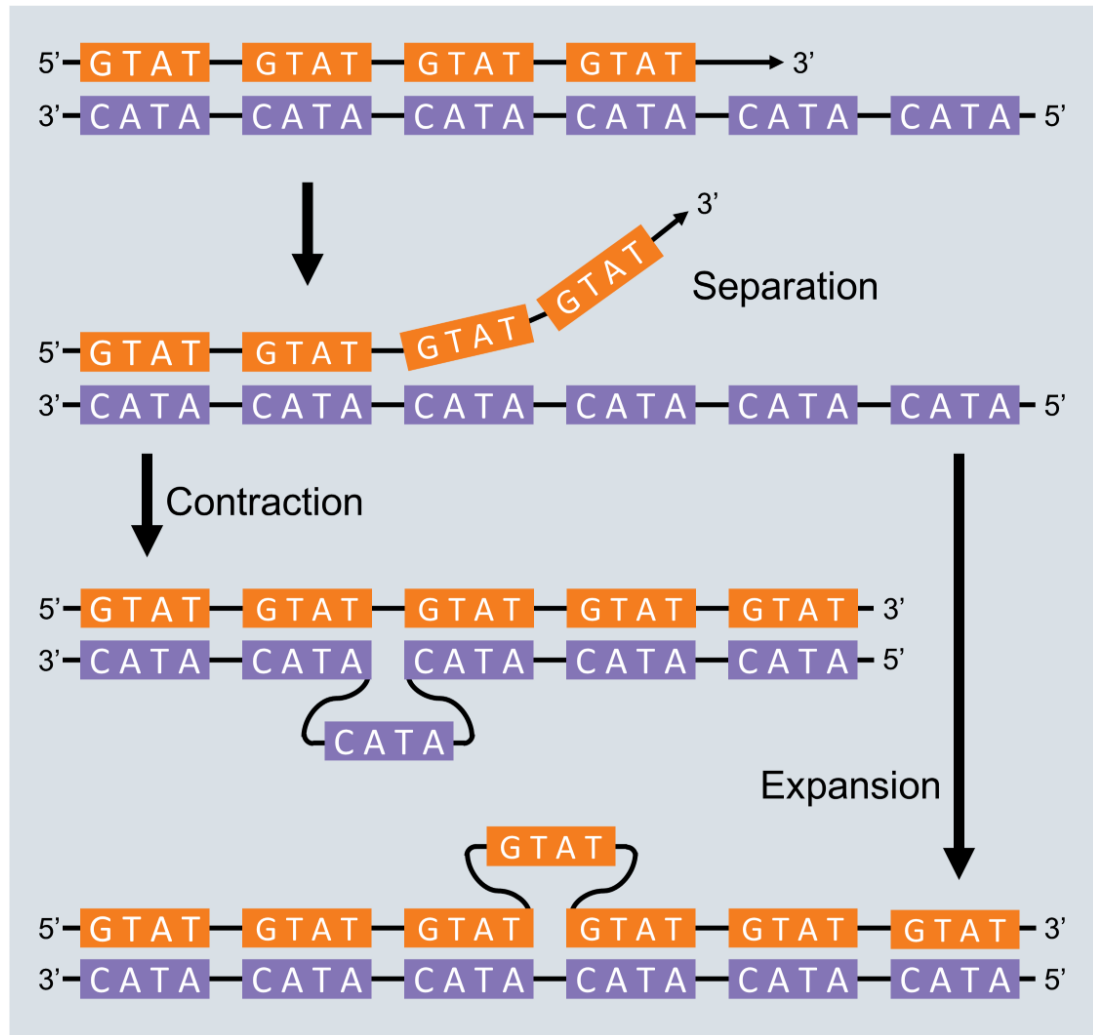
Figure 1.4 Contraction and expansion of a tandem repeat during DNA replication.
During DNA replication of an existing tandem repeat, when the DNA strands are
separated, one of the strands may slip prior to re-annealing due to the inherent instability
of these regions. Depending on which strand experiences the slippage, the tandem repeat
will contract or expand in length. Adapted from (Bush & Moore, 2012).

1.2.1.2.3   Copy number variants

CNVs are large-scale deletions or duplications of DNA of at least 50 nucleotides in length (**Figure 1.3C**) (Feuk, Carson, & Scherer, 2006). As previously described, humans are diploid, harbouring both a maternal and paternal copy of their nuclear genome. Deletions occur when there is a loss of DNA, or loss of copy number, in comparison to the reference genome, whereas duplications occur when there is repetition of DNA, or gain of copy number, in comparison to the reference genome. However, unlike tandem repeats, a CNV duplication does not always occur in tandem to the original DNA sequence — the duplication may be located in an entirely different region of the genome (Feuk, Carson, et al., 2006). Although approximately 12% of the human genome is comprised of CNVs, and they contribute to normal phenotypic variation between individuals (Redon et al., 2006), CNVs may also affect gene expression, organization, and dosage (Stranger et al., 2007). Therefore, particularly when involving genes, CNVs can increase susceptibility to disease (Shastry, 2009a).

1.2.1.2.4   Chromosome alterations

"Chromosomal alterations" were some of the first variants observed in the human genome, as identification did not require the availability of sequencing technology. Rather, the variants are identifiable using a microscope to examine all 46 chromosomes, otherwise referred to as a "karyotype" (Feuk, Carson, et al., 2006). Other cytogenetic techniques are also now routinely used to identify these variants, including fluorescent *in situ* hybridization (FISH), comparative genomic hybridization (CGH) (Thompson & Gray, 1993), and, more recently, next-generation sequencing (NGS) (Iacocca et al., 2017).

There are several types of chromosomal alterations including, but not limited to, aneuploidies and rearrangements. "Aneuploidy" refers to an abnormal number of chromosomes due to an extra or missing copy. The majority of aneuploidies are trisomic or monosomic and one third of miscarriages are due to aneuploidy, highlighting their high potential for clinical consequence (Hassold & Hunt, 2001). "Rearrangements" can be further subclassified into unbalanced rearrangements, such as deletions or insertions of

large chromosomal segments, or balanced rearrangements, such as inverted or translocated chromosomal regions (**Figure 1.3D-E**) (Morin, Eccles, Iturriaga, & Zimmerman, 2017).

1.2.1.3 Frequency of variants

The success of the Human Genome Project, combined with the quick advancement of genetic sequencing technologies over the last 20 years, has resulted in the identification of millions of variants of unknown clinical relevance. This rapidly led to the development of complete databases of genomic variation from large cohorts of ancestrally diverse individuals to fully understand the role of both common and rare variants in phenotypic presentation (Auton et al., 2015; Karczewski et al., 2020; Lek et al., 2016). Based on the frequency at which a variant is observed in each of these databases, it is assigned a "minor allele frequency" (MAF), defined as the percentage of the population that carries the variant. The use of minor in MAF refers to the second most common allele that may appear at a genetic locus, whereas the major allele refers to the first most common allele, or the allele displayed in the human reference genome. Depending on the database, or subpopulation within the database, the MAF may be in reference to the global population, a population of individuals with or without a particular trait or diagnosis, or a specific ancestral population (Auton et al., 2015; Guerreiro et al., 2018; Karczewski et al., 2020; Lek et al., 2016).

MAFs vary widely throughout the genome, often dictated by the location of the variant, the deleteriousness of the variant, and evolutionary constraints (Subramanian & Kumar, 2006). Within populations, there are also further factors at play, such as genetic selection, migration, and drift that may cause a variant to be more or less common in a specific ancestral group (Tishkoff & Verrelli, 2003; Zlotogora, 1994). For this reason, it is imperative that genetic analyses control for ancestral populations, particularly when performing genetic association analyses — which correlate differences in phenotypic frequencies to allele frequencies and are discussed further in Section 1.1.4.2 — as MAFs between ancestries can vary widely (Cardon & Palmer, 2003; Taioli, Pedotti, & Garte, 2004).

Common variants are generally defined as those with an MAF > 1% in the general population and are often referred to as "polymorphisms." Although polymorphisms exist in all variant types, a common SNV is given the specific designation of "single nucleotide polymorphism" (SNP). Typically, common variants are predicted to be of neutral or very low phenotypic effect, due to their general abundance in the genome (**Figure 1.5**) (Gibson, 2012); however, exceptions do exist, such as the apolipoprotein E (*APOE*) genotype and its high phenotypic effect in Alzheimer's disease risk, which is discussed further in Section 1.2.2.1.1.

In contrast, rare variants are classically defined as variants with an MAF ≤ 1%. Although individually not common, as a result of ancient population bottlenecks and the more-recent rapid increase in population size, the majority of variants that have been identified in the general population are rare, as new variants are introduced with each generation (Lappalainen, Scott, Brandt, & Hall, 2019). It is generally accepted that the most deleterious of genetic variants are rare, due to the effects of natural selective pressures (**Figure 1.5**). More specifically, variants of high deleterious effect will undergo negative natural selection, as they are less advantageous to survival, and the variants will become less frequent in the general population (Karczewski et al., 2020).
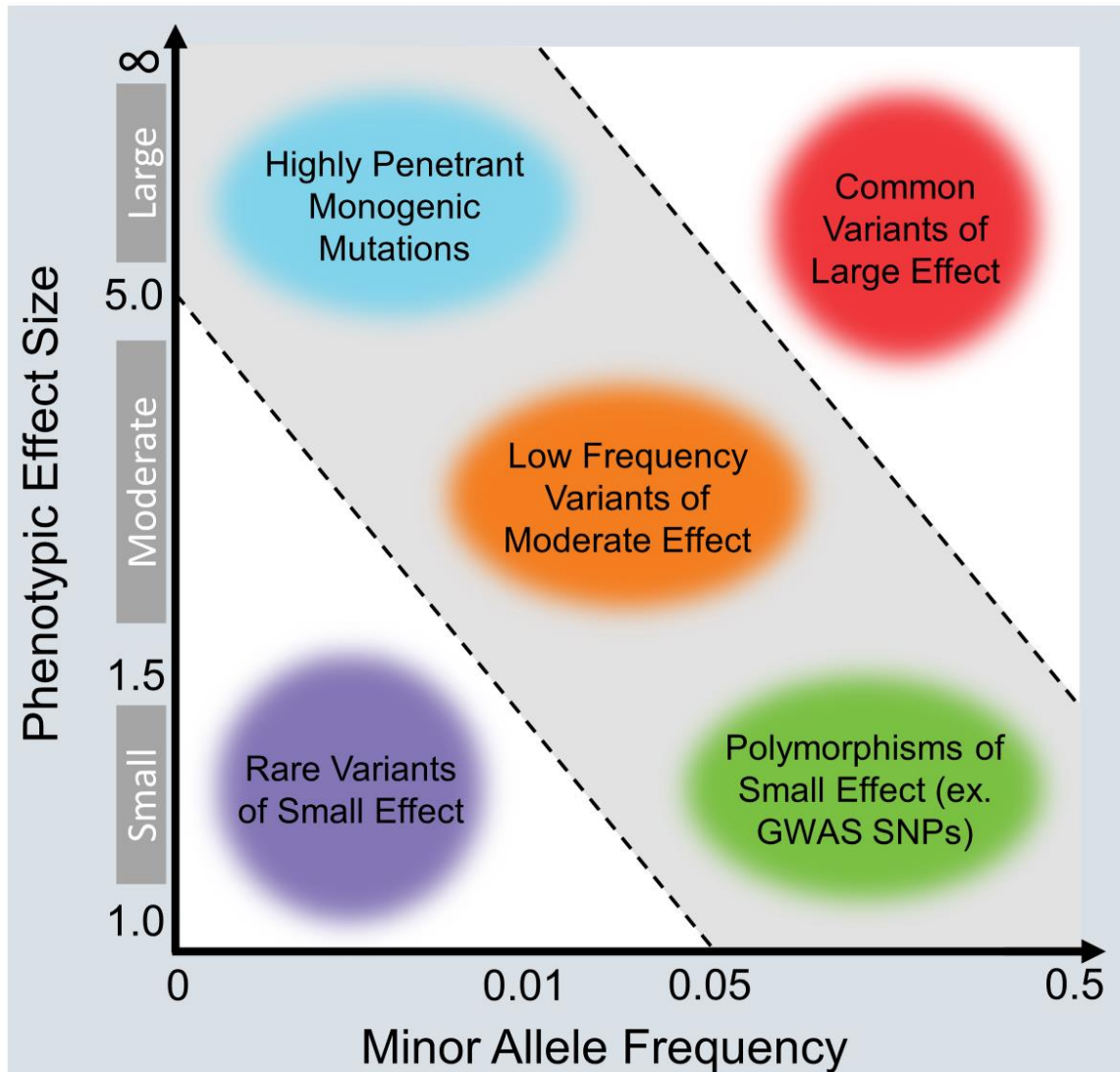
Figure 1.5 Typical range of potential phenotypic effect size, based on a variant's minor allele frequency (MAF) in the general population.

It is generally accepted that rare variants (MAF ≤ 0.01) tend to be of higher phenotypic effect, usually acting as highly penetrant monogenic disease-causing mutations. In contrast, common variants (MAF > 0.05) are typically considered to be of small phenotypic effect, as demonstrated by the low effect sizes associated with the variants in genome-wide association analyses. However, these generalizations do not always hold true and some rare variants may be of smaller phenotypic effect, yet their rarity and low impact cause them to be difficult to detect with typical genetic analytic approaches. In contrast, there are a few known cases of common variants that demonstrate high phenotypic effect, such as the apolipoprotein E (*APOE*) E4 genotype. Adapted from (Bush & Moore, 2012).

1.2.2　Transmission of genetic traits and disease

As previously described, every somatic cell in the human body contains two copies of the genome — otherwise referred to as the diploid genome — of which one copy was maternally inherited and the other paternally inherited (Georgadaki, Khoury, Spandidos, & Zoumpourlis, 2016). During fertilization, an egg cell from the biological mother containing one copy of DNA and a sperm cell from the biological father containing one copy of DNA combine to form the diploid zygote. As the zygote replicates throughout human development to form each cell in the human body, the cells all retain the same two inherited copies of the human genome (Georgadaki et al., 2016). In this way, genetic variation is inherited from an individual's biological parents. Any genetic variant present from the time the zygote was formed is called a "germline variant" and can be passed on to further offspring of an individual.

During the formation of gamete cells within the gonads, or during the early stages of embryogenesis, genetic variants may occur spontaneously as a result of errors in DNA replication, DNA damage such as double-strand breaks, or crossover interference during gamete formation (Goldmann, Veltman, & Gilissen, 2019). The resulting variants are called "*de novo* variants" and while they were not present within the genome of the biological mother or father, they can still be passed on to an individual's offspring. More simply stated, *de novo* variants are spontaneously developed within a single generation, but are able to be passed on to future generations (Goldmann et al., 2019). Prior to the introduction of large-scale sequencing techniques, the identification of disease-causing *de novo* variants was difficult, as they defy the typical inheritance patterns required when performing linkage analysis, as described further in Section 1.1.4.1. Importantly, incidence of *de novo* variants is very low, and for this reason, they are often considered to have a high likelihood of being pathogenic (Mani, 2017).

Spontaneous variants may also occur in later stages of embryogenesis or throughout an individual's development or lifespan. These "somatic variants" result in mosaicism, with only a subset of an individual's cells harbouring the variant (Poduri, Evrony, Cai, & Walsh, 2013). In a similar manner to germline *de novo* variants, somatic variants may be a result of errors in DNA replication or DNA damage from either endogenous sources,

such as oxidation, or exogenous sources, such as ultra-violet radiation or exposure to certain chemicals (Saini & Gordenin, 2018). Although they are well known for their contribution to many cancers, somatic mutations can also contribute to non-cancerous diseases as well, depending on the genes affected and the cells in which they are harboured.

1.2.2.1   Genotypes, haplotypes, and zygosity

A selection of genetic information that defines a trait or disease is referred to as a "genotype," while alternative variants of a gene, genotype, or variant, are called "alleles." The allele that is present within the human reference genome is considered the "wild-type" allele, or the "major allele." When referring to the trait or disease influenced by a genetic marker, it is called a "phenotype." Therefore, certain alleles of a genotype define a phenotype.

In some instances, groups of alleles are inherited together at a genetic locus and are called a "haplotype." Typically, haplotypes result from a lack of recombination occurring between the alleles upon gamete formation, which may be due to the close proximity of alleles. Although, some haplotypes can encompass tens of kilobases (Wall & Pritchard, 2003). In particular, the phenomenon of non-random co-occurrence of alleles is known as "linkage disequilibrium" (LD) (Lewontin & Kojima, 1960). LD can reflect population histories, evolution, and patterns of geographic subdivision and has allowed for major advancements in the study of human genetics, particularly in linkage analysis and the defining of haplotype blocks, which have aided in the mapping of SNPs in large-scale association studies (Wall & Pritchard, 2003). Approaches to genetic analysis will be discussed further in Section 1.1.4.

"Zygosity" refers to the degree to which the two alleles — each from one copy of the diploid genome — at a given locus are similar. If the two alleles are identical, the genotype is referred to as "homozygous," whereas if they are different, the genotype is "heterozygous" (Winsor, 1988). In some cases, there may be two different variants inherited at the same genetic locus, such as within the same gene, but on different alleles in which case the genotype is referred to as a "compound heterozygote." And finally, the

sex chromosomes offer interesting cases of zygosity called "hemizygosity," which describes males carrying a variant on their single X chromosomes, or females carrying a loss of one of their X chromosomes.

1.2.2.2   Monogenic inheritance of Mendelian phenotypes

"Mendelian inheritance" derives its name from the first description of the pattern by Gregor Mendel, who proposed a model for genes and inheritance as a result of his studies on pea plants. Ultimately, his work resulted in the "Law of Segregation" and "Law of Independent Assortment" (Mendel, 1865; H. Zhang, Chen, & Sun, 2017). The Law of Segregation described that upon gamete formation, only one of two gene copies are passed on to potential offspring. Closely tied to that, the Law of Independent Assortment described that alleles of different genes are independently sorted into the gametes, resulting in potential offspring diversity. Mendel was also responsible for the coining of the terms "recessive" and "dominant," which are still used today in the description of inheritance patterns (Mendel, 1865; H. Zhang et al., 2017).

Mendelian inheritance refers to a genotype at a single genetic locus driving phenotypic presentation, otherwise referred to as a "monogenic variant." The monogenic variants may be inherited in the various zygosity states described previously; however, whether they result in presentation of a phenotype depends on the "inheritance pattern" of the genotype-phenotype relationship. Monogenic phenotypes may follow a variety of inheritance patterns, including dominant, recessive, co-dominant, and sex linked (**Figure 1.6**).

"Autosomal dominant" inheritance refers to phenotypes that present when an individual carries only one copy of the genetic variant. Typically, heterozygous variant would have been inherited from a single parent, who also would have presented with the associated phenotype (Winsor, 1988). Alternatively, autosomal dominant phenotypes may result from *de novo* genetic variants, as previously described, in which case neither of the parents carried the variant or presented with the associated phenotype. Multiple mechanisms may underlie the presentation of an autosomal dominant phenotype, including dominant negativity, haploinsufficiency, and increased gene dosage (Veitia,

Caburet, & Birchler, 2018). Both dominant negative and haploinsufficiency models are a result of heterozygous LOF variants, yet differ in pathogenic mechanism. In a "dominant negative" model, the heterozygous variant results in a protein product that actively interferes with the wild-type protein product, whereas in "haploinsufficiency" models a single wild-type copy of a protein product cannot compensate for the variant protein product that cannot perform its usual function. Conversely, "increased gene dosage" models are a result of gain-of-function variants that increase the production or functioning of a gene's protein product. The excess protein production may result in dysregulation of mechanistic pathways or expression in inappropriate pathways or during inappropriate times of development (Veitia et al., 2018).

"Autosomal recessive" inheritance refers to phenotypes that present when an individual carries two copies of the genetic variant as a result of the inheritance of one copy from each parent (Winsor, 1988). In this case, each parent harbouring a single copy of the variant would be referred to as a "carrier" of the variant, but neither would present with the phenotype as they would not be homozygous for the allele.

"Autosomal co-dominant" inheritance is an interesting case in which both the variant allele and the wild-type allele are expressed. This may result in an intermediate phenotype in heterozygous carriers of the variant, otherwise known as "incomplete dominance," and more severe phenotypes in homozygous carriers (Moldovan, Banescu, & Dobreanu, 2020). Alternatively, autosomal co-dominance can result in the presentation of both phenotypes simultaneously, such as in the case of ABO blood groups.

"X-linked" inheritance refers to variants located on the X chromosome that contribute to a phenotype. X-linked inheritance may also be dominant or recessive, but it is important to recognize that these variants will have different effects in males and females (Winsor, 1988). Females have two copies of the X chromosome, so if only one of their X chromosomes harbours a variant, they may only be a carrier if the phenotype is inherited in a recessive manner. However, males only have a single X chromosome, so no matter if the phenotype is inherited in a dominant or recessive manner, if they harbour a phenotype-associated variant on the X chromosome, there is no other wild-type copy of

the allele to compensate, and they will present with the phenotype (Winsor, 1988). Similarly, "Y-linked" inheritance is the result of genetic variants present on the Y chromosome. As only males have the Y chromosome, the variants can only be paternally passed on to all male offspring.
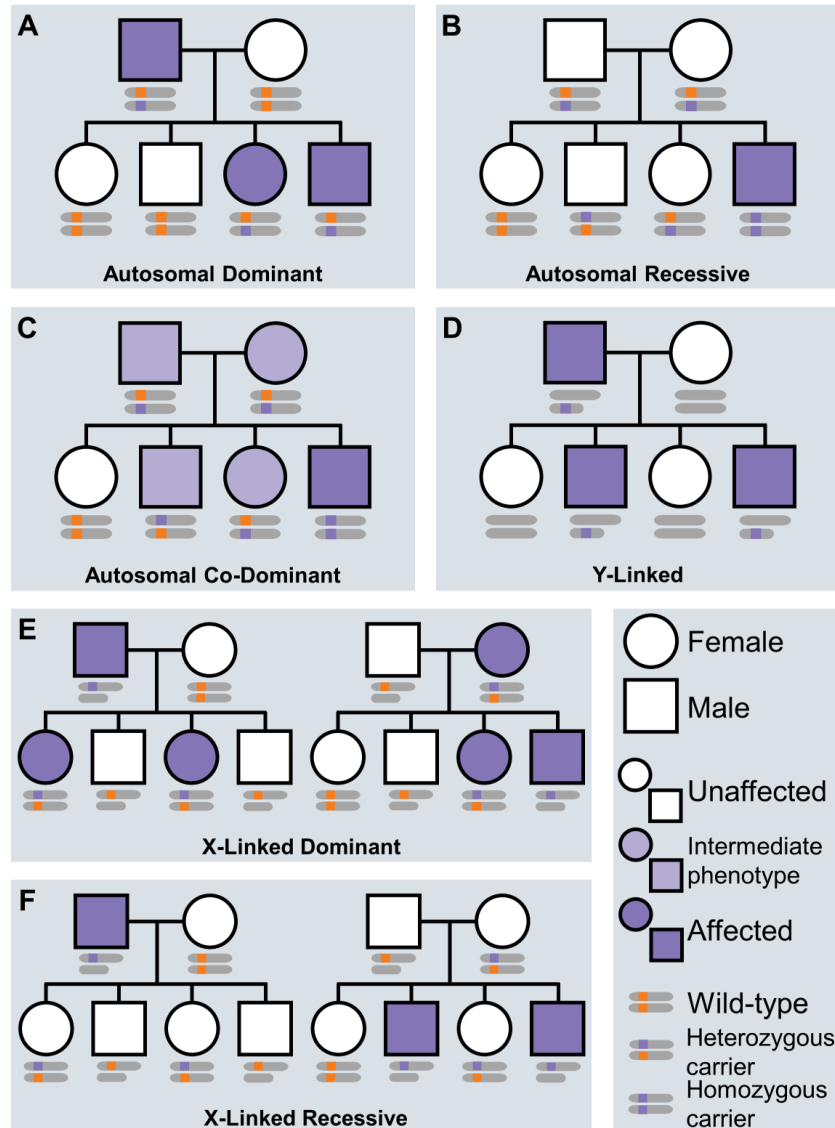
Figure 1.6 Potential inheritance patterns of Mendelian phenotypes from monogenic variants.
The inheritance pattern of Mendelian phenotypes can be determined based on pedigrees of a family, which depict those that do and do not present with the phenotype of interest. **A)** Autosomal dominant phenotypes are those that present when an individual is heterozygous for the variant of interest. **B)** Autosomal recessive phenotypes are those that only present when an individual is homozygous for the variant of interest. All heterozygotes are considered variant carriers. **C)** Autosomal co-dominance refers to phenotypes that are influenced by the presence of both the variant and wild-type allele such that it may result in an intermediate phenotype in heterozygous carriers of the variant, otherwise known as incomplete dominance, or heterozygous carriers may display both phenotypes simultaneously. **D)** Y-linked refers to phenotypes that are caused by variants on the Y chromosome, whereas X-linked phenotypes are those caused by variants on the X chromosome and may be **E)** dominant or **F)** recessive.

1.2.2.2.1   Variable penetrance

An important caveat when studying seemingly monogenic phenotypes is that there may be underlying genetic complexity complicating the analysis of phenotype-causing genetic variation, such as that introduced by incomplete penetrance. "Penetrance" is defined as the probability that individuals carrying a disease-causing genetic variant fully express the associated phenotype (Cooper, Krawczak, Polychronakos, Tyler-Smith, & Kehrer-Sawatzki, 2013). "Complete penetrance" is the term used to describe a phenotype that *always* presents itself in an individual that carries the associated genetic variant. When penetrance is "incomplete" it means that individuals may carry a variant known to cause a certain phenotype, but they may not express the expected phenotype (**Figure 1.7**). Similarly, variable expressivity of a genetic variant can complicate the seemingly straight-forward model of monogenic inheritance. "Variable expressivity" is used to describe instances where multiple individuals carry the same genetic variant, but display varying degrees of phenotypic presentation, otherwise referred to as "phenotypic heterogeneity" (**Figure 1.7**) (Schacherer, 2016). Finally, "pleiotropy" refers to genes, or even specific variants, that cause or increase risk of multiple different phenotypes that may or may not be related (Hemani, Bowden, & Davey Smith, 2018).

Variable penetrance and expressivity may be a result of a variety of mechanisms, including: influence from additional genetic variants, the age or sex of the individual, epigenetic changes, or environmental factors (Cooper et al., 2013). Due to the potential of influence from other genetic factors to variants displaying either of these phenomena, the variants begin to blur the line between true monogenic inheritance patterns and polygenic inheritance (Schacherer, 2016). For this reason, penetrance and expressivity introduce great complexity in the interpretation of genetic variants, which is an important component of genetic analyses and will be discussed further in Section 1.1.5.
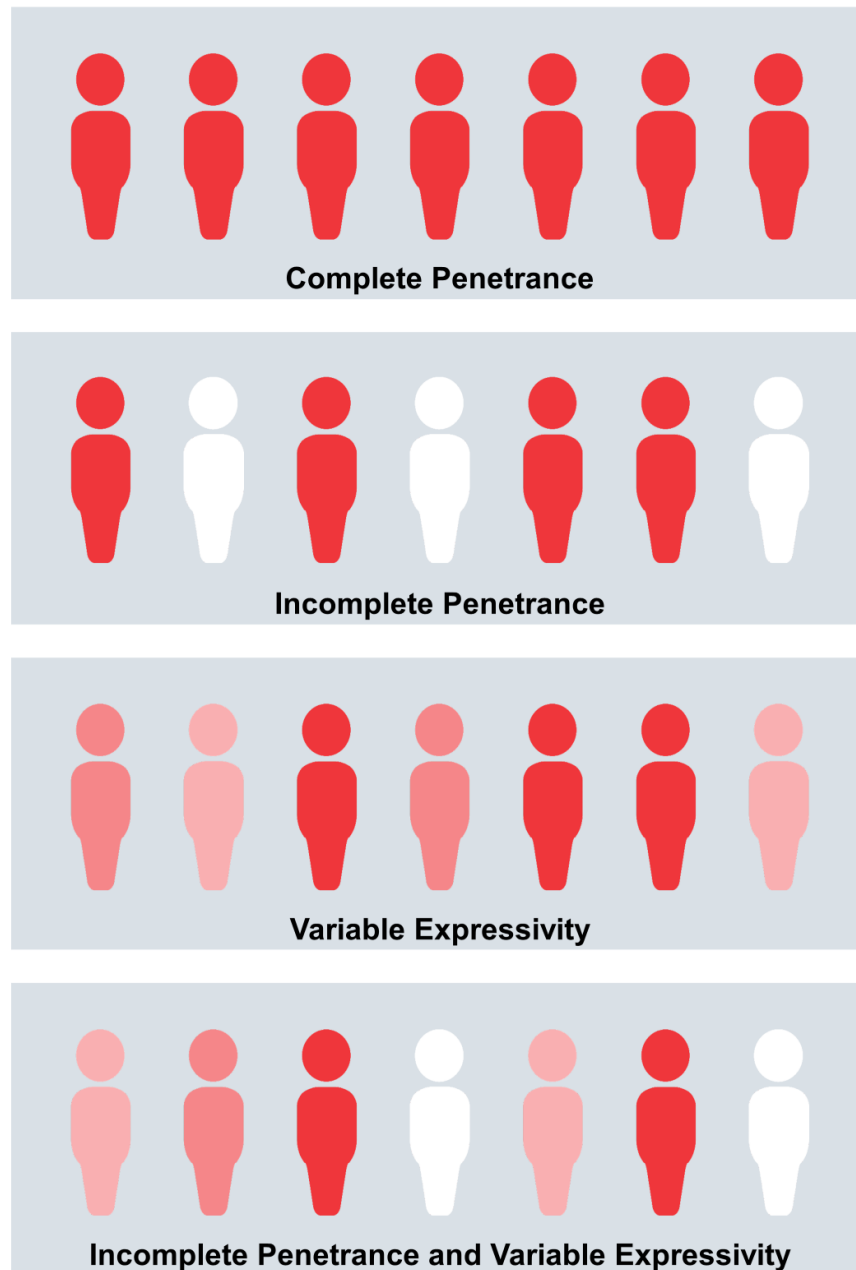
Figure 1.7 Examples of incomplete penetrance and variable expressivity of a phenotype.
Incomplete penetrance refers to phenotypes that are not always expressed when an individual carries the associated genetic variant, whereas variable expressivity refers to phenotypes that are expressed, but to varying degrees across individuals who carry the same associated genetic variant. It is also possible for a variant to exemplify both incomplete penetrance and variable expressivity.

1.2.2.3   Polygenic inheritance of complex disease

In contrast to monogenic inheritance, "polygenic inheritance" refers to traits that are influenced by more than one genetic variant, which have an additive effect on phenotypic presentation. Typically, these phenotypes are referred to as "complex" and the many genetic variants that contribute to the trait or disease are located at a variety of loci throughout the genome (Dron & Hegele, 2018). A complex phenotype may also be considered "multifactorial" if influenced by both polygenic and environmental factors.

Classically, polygenic phenotypes are considered to be influenced by SNPs of small phenotypic impact; however, variants contributing to complex traits or disease may be common or rare, and include both SNVs and structural variants. In general, rare variants tend towards larger phenotypic impact compared to common variants, yet exceptions do exist and variants of any frequency can contribute to the cumulative sum of phenotypic risk (**Figure 1.5**) (Crouch & Bodmer, 2020). The variants may also reside within coding regions of genes or in non-coding regions of the genome, and although variants located within coding regions often have larger phenotypic impact, complex phenotypes are often driven by noncoding variants that contribute to gene expression (Crouch & Bodmer, 2020).

One complication of polygenic inheritance models is the difficulty in assessing the functional impact of genetic variants and understanding their combined contribution to phenotypic presentation. Further, even when the additive impact of polygenic variants is considered, there often remains a large proportion of unexplained genetic variance, referred to as an issue of "missing heritability" (Boyle, Li, & Pritchard, 2017).

To assess overall risk from polygenic contributors to disease, a "polygenic risk score" (PRS) can be employed. Although PRSs were not used for the purposes of this Dissertation, their importance in understanding overall genetic risk of disease cannot be discounted. Overall, they are a measure of genetic risk of a specific complex phenotype based on an individual's accumulation of genetic risk variants. Typically, the weighted scores are calculated from the summation of effect size estimates of various SNPs associated with a phenotype of interest, obtained from large genome wide association

studies (GWASs) (Dron & Hegele, 2018). Although rare variants can also be included in a PRS, rare variants of small to moderate effect remain difficult to identify using GWAS, and therefore are largely unknown (Choi, Mak, & O'Reilly, 2020). The score can encompass tens to thousands of genetic variants, with larger scores often able to capture a greater amount of genetic risk. Although PRSs have allowed the ability to assess genetic risk from multiple small effect SNPs, they are not without their limitations. The design of a reliable PRS requires the results from many GWASs with large sample sizes, which still remain limited for many phenotypes. Further, it is imperative that PRSs have been replicated in multiple cohorts, as false positive claims of genetic risk using a PRS can be relatively common (Nalls et al., 2019). PRSs are also often specific to the ancestral population of the experimental cohorts with which they were designed, as common genetic architecture between ancestral populations can vary widely. Unfortunately, many of the currently designed PRSs were built using cohorts of European ancestry and fail to account for ancestral diversity (Martin et al., 2019). Finally, in many cases, PRSs have failed to fill the gap between the amount of heritability accounted for by monogenic variants' contribution to disease and overall disease heritability estimates, leaving missing heritability, which will be discussed further below. Nonetheless, PRSs are becoming a promising tool to predict individual disease risk, and can also be employed to study overlap in genetic determinants of different diseases (Bellou, Stevenson-Hoare, & Escott-Price, 2020).

1.2.2.3.1   Heritability

"Heritability" is defined as the amount of phenotypic variance in a population that is attributable to genetic variance (Manolio et al., 2009). Typically, heritability is expressed by a number ranging from zero to one, where the former describes phenotypes that are not explained by genetics at all, and the latter describes phenotypes completely explained by genetics. If a complete understanding of the genetic contributors to a certain phenotype were understood, heritability would estimate how well we could predict a phenotype based on genetic factors.

Unfortunately, many complex phenotypes still display a large gap between the estimated heritability determined by family-based studies, such as twin studies, and the phenotypic

variance explained by associated genetic variants, which is defined as the "missing heritability" (Uricchio, 2020). There are many potential explanations for missing heritability, including the proposition that there remain variants contributing to complex phenotypes that are yet to be discovered, such as common variants of smaller impact, rare variants of smaller impact that are difficult to detect even with large samples sizes, and ultra-rare variants of large impact. However, as phenotype-associated genetic variation is continuously discovered, other considerations have come to light. One hypothesis is that current heritability estimates may be inflated, potentially due to non-additive genetic effects or shared environmental influence in families. Others suggest that the non-additive genetic effects of gene-gene interactions are difficult to capture and potentially account for the missing heritability (Manolio et al., 2009). Regardless, identifying the remaining genetic factors that contribute to complex traits is imperative as we continuously move towards a precision medicine model of diagnosis, progression prediction, and treatment for many complex phenotypes.

1.2.3   Methods to identify genetic variation

The ability to identify variation within the human genome is relatively new, with methods only being developed over the last 50 years. In 1977, Frederick Sanger published his methodology for DNA sequencing, later named "Sanger sequencing," that to this day remains the gold-standard approach to obtain the exact readout of a selection of DNA (Sanger, Nicklen, & Coulson, 1977). The method also underwent important technological improvements during the completion of the Human Genome Project (Lander et al., 2001). Along with its conclusion, technological advances have allowed for a new era of genetic analysis encompassing high-throughput massively parallel DNA sequencing, referred to as NGS, as well as advancements in the ability to perform large-scale genotyping arrays.

1.2.3.1   Sanger sequencing

Sanger sequencing, otherwise referred to as "chain-termination" or the "dideoxy technique," remains one of the most reliable sequencing approaches, even almost 50 years after it revolutionized the field of human genetics (Sanger et al., 1977). Traditionally, the method utilizes analogues of the deoxyribunucleotides (dNTPs) found

in DNA that lack the 3' hydroxyl group, called dideoxyribonucleotides (ddNTPs) that result in the inability to form bonds with the next dNTP in a sequence. Using these, four parallel polymerase chain reactions (PCR) are run, each containing only one set of radiolabeled ddNTPs corresponding to a single base — A, C, G, or T — and resulting in DNA strands of each possible length being produced. The DNA strands then undergo size-separation-based capillary gel electrophoresis, and the 5' to 3' DNA sequence can be read based on the radioactive label in the corresponding position on the gel (Heather & Chain, 2016).

Over the years, methodologies have been improved, resulting in the ability to automate Sanger sequencing using machines, and the accuracy of the method remains a gold-standard in the field. Yet Sanger sequencing does not come without its limitations, as even with automation, the use of the method remains tedious, time consuming, and cost prohibitive. Sanger sequencing lacks practicality when studying large sample sizes or regions of the genome much greater than 1 kb.

1.2.3.2   Next-generation sequencing

Following the completion of the Human Genome Project, there became a striking need for the ability to sequence large cohorts of individuals quickly and at low cost. In 2005, the shift to the "next-generation" of sequencing began under the basic premise that the methods sequence DNA in a massively parallel manner (Margulies et al., 2005; Shendure et al., 2005). By shearing the DNA into small overlapping fragments, or "reads," millions of nucleotide sequences can be determined simultaneously. Not only does this speed up the process of sequencing and decrease cost, but it also results in high coverage, or "read depth", at each sequenced nucleotide, thereby increasing the confidence of the sequencing calls made by NGS methods and allowing for the determination of allele zygosity and dosage. Further, while Sanger sequencing is restricted to identifying small-scale SNVs and InDels in regions already known to be of interest, NGS offers the ability to identify a wide range of genetic variation, including large-scale CNVs (Iacocca et al., 2017). NGS can also interrogate the human genome without the bias needed for Sanger sequencing, as full exomes or genomes may be sequenced at once (Behjati & Tarpey, 2013).

NGS can be subcategorized based on specific technologies as well as the regions of the genome that are captured. Whole genome sequencing (WGS) refers to NGS of the entire human genome, providing an unbiased assessment of the genetic variation harboured by an individual. While there are obvious benefits to this approach, such as the ability to identify variants within intergenic, intronic, and regulatory regions of the genome, and the continuously decreasing costs of the sequencing itself, limitations also exist. Due to the length of the full human genome sequence, WGS data processing and analyses are computationally intensive, requiring the ability to store extremely large file sizes. Further, genetic variation located in non-coding regions of the genome remain difficult to interpret, and related genetic association analyses require very large sample sizes for the power to detect variants of interest. Alternatively, there is whole exome sequencing (WES), which targets the exonic regions of all genes in the human genome, resulting in ~1% of the content than that covered by WGS (Rabbani, Tekin, & Mahdieh, 2014). While WES obviously limits the ability to detect genetic variation to the coding regions of the genome, it is predicted that between 60–85% of disease-associated genetic variants are restricted to these regions (Rabbani et al., 2014; Ross, Dion, & Rouleau, 2020), and the decreased content coverage of WES limits both the computational resources necessary for analysis, as well as the associated costs of the sequencing. WES can also produce sequencing data of greater read depth than WGS, providing a higher degree of confidence in allele calls (LaDuca et al., 2017; Sims, Sudbery, Ilott, Heger, & Ponting, 2014). Read depth can be even greater if a targeted NGS approach is used. Targeted NGS refers to sequencing panels designed to target specific genomic content, including any genes or regions of particular interest. Targeted panels offer further decreases in cost and computational processing power, and are of particular benefit when certain genes are known or predicted to be associated with a phenotype of interest.

While NGS offers its clear advantages, there are important limitations to consider. Due to the short-read lengths of the DNA fragments being sequenced, highly repetitive regions of the genome remain difficult to map back to the human reference genome, specifically when the repeats are longer than each individual read itself (Xuan, Yu, Qing, Guo, & Shi, 2013). Further, GC-rich regions of the genome often result in decreased read depth of NGS data, and confidence in allele calls becomes uncertain (Chen, Liu, Yu, Chiang, &

Hwang, 2013). Yet, it is widely recognized that NGS approaches have revolutionized the study of human genetics and are continuously becoming more approachable. Not only are bioinformatics processes able to handle the data being refined, but costs continue to decrease, with the first genome sequence costing over $100,000,000 to complete and WGS now offered for under $1,000 per sample in some instances (Wetterstrand, 2020).

1.2.3.3   Genotyping

Unlike NGS, which sequences full genomic regions, "genotyping" is the targeting of specific genetic coordinates. Genotyping may be performed on a variety of known genetic loci at once using a microarray or may be performed on individual alleles using single SNP quantitative PCR (qPCR)-based assays, such as a TaqMan allelic discrimination assay. Although genotyping does not offer the ability to assess the full spectrum of genetic variants an individual may carry, targeting specific loci can offer the ability to scan a population for SNVs known to be pathogenic, to characterize the ancestry of a population, or to perform genome-wide association studies and identify novel phenotype-associated regions in the genome (Ragoussis, 2009).

Microarrays offer a unique approach to genetic analysis, as they can cover up to millions of SNVs. Often considered an intermediate between cytogenic analysis of chromosomal karyotypes and complete DNA sequencing, microarrays provide a scan of the entire genome to identify genomic variation (Page et al., 2007). Further, the phenomenon of LD allows for the imputation of additional SNVs that are within the same haplotype blocks of alleles that have been genotyped. The coverage of such a large amount of the genome allows for the detection of CNVs based on the signal intensity of the data; however, the locations of the SNVs on the microarray limit resolution and CNV breakpoints must be further discerned (Coughlin, Scharer, & Shaikh, 2012). Microarrays also offer the opportunity to perform GWAS, which are used to identify novel loci associated with phenotypes of interest (Ragoussis, 2009). GWAS will be discussed further in Section 1.1.4.2.1.

1.2.3.4   Repeat-primed PCR

When studying tandem repeats, obtaining an accurate detection of expansion size is imperative, as there are often well-defined thresholds for the number of repeats that cause a phenotype. But it is difficult to detect these variants with Sanger sequencing and NGS, due to the imbalances they often introduce in GC content or their inherent length. Although there are many alternatives to detect these variants — such as capillary electrophoresis, Southern blot analysis, or mass spectrometry — the most commonly used approach is repeat-primed PCR, due to its relative simplicity and feasibility (Liu, Zhang, Wang, Gu, & Wang, 2017).

Briefly, repeat-primed PCR utilizes a pair of reverse primers that can bind at various locations throughout the tandem repeat, amplifying DNA of various lengths, along with a fluorescently labelled forward primer. The resulting DNA fragments are then able to be analyzed using the fluorescence trace, creating a "ladder" and allowing for the identification of large pathogenic repeats (Renton et al., 2011; Warner et al., 1996). Unfortunately, the method is limited to only determine the length of a tandem repeat up to a certain size, for example 60 copies of a hexanucleotide repeat; however, it is still able to determine that the number of repeats are above the length threshold.

1.2.4   Approaches to identify genetic contributors of disease

The analytic approaches used to map genetic loci to associated human phenotypes can be largely categorized by two main groups, including: 1) "linkage analysis" that relies on co-segregation of genotype and phenotype throughout a pedigree and 2) "population genetics", also referred to as "association analyses", which study genetic variation in large populations of unrelated individuals. The approaches each have their own advantages and limitations, and the choice of which to utilize is dictated by the data available, the phenotype severity and transmission pattern, and the goals of the study.

1.2.4.1   Linkage analysis

As previously described, Mendel's Laws stated that alleles at different loci will segregate independently during gamete formation; however, there are genes — typically in close

proximity on the same chromosome — that are "linked", resulting in phenotypes being inherited together. Linkage analysis relies on these exceptions to Mendel's law in order to identify genetic markers that are linked to genes driving a phenotype, and to ultimately map genes to their associated phenotype (Mendel, 1865; Elston, 1998).

Prior to the availability of large-scale sequencing methods, the identification of genes contributing to a phenotype of interest relied on the use of SNPs with known positions within the genome and analyzing co-segregation of the SNPs throughout a family's pedigree (Morton, 1955). Specifically, "co-segregation" refers to genotypes that are only carried by individuals in a pedigree that also present with the phenotype being studied. A map of the SNP positions within the genome can then be generated and, based on the SNPs that co-segregated, regions of interest likely to be associated with the phenotype can be identified (Elston, 1998).

Classically, linkage analyses were used for the identification of rare, monogenic variants of high penetrance, and this application is still widely accepted, as the inherent rarity of these variants result in detection difficulty with other methods, such as GWAS (Bush & Haines, 2010). The wide-availability of NGS data coupled with linkage analysis now allows for a narrowed approach to identifying genetic markers of disease by directing which genomic loci should be prioritized for further analysis (Ott, Wang, & Leal, 2015).

Although linkage analysis has been successful in mapping disease-causing genes for decades, it is still important to recognize its limitations. Most importantly, the approach relies on the availability of genetic and phenotypic data of entire families, with larger pedigrees being most optimal (Bush & Haines, 2010). Unfortunately, large pedigrees with family members all willing and available to be involved in genetic analyses can be difficult to find, particularly in diseases with onset later in adult life. Additionally, there is always the, albeit unlikely, potential that linked genetic loci will experience recombination between them and the region will be eliminated from consideration. Finally, although not impossible, the identification of common genetic variants of small or modest phenotypic impact remains difficult with linkage analysis, due to the often lower penetrance of these variants (Bush & Haines, 2010).

1.2.4.2   Association analyses

Association analyses encompass a wide-range of approaches all aimed at detecting associations between genetic variants and phenotypes of interest at a population level. Importantly, they rely on the use of populations of unrelated individuals and the same genetic markers contributing to a phenotype across the various individuals (Rodriguez-Murillo & Greenberg, 2008). Association analyses have risen in popularity over the last two decades due to the availability and feasibility of large-scale NGS sequencing and genome-wide microarrays. The approaches are advantageous, as family data is often difficult to obtain, and the analyses offer a high level of granularity to detect variants of relatively small phenotypic impact, while simultaneously producing quantitative estimates of that impact (Rodriguez-Murillo & Greenberg, 2008).

An important caveat to association analyses is the need to accurately account for differential ancestry within the population under study. As early humans spread around the world and settled into different geographic locations, they faced differential environments, catastrophic events, and geographic isolation often resulting in gene pool isolation and, in some cases, interbreeding. All of these factors resulted in highly differential MAFs across different populations (Hellwege et al., 2017). Therefore, when performing association analyses, it is important to either have ancestrally homogenous cohorts to compare or to apply corrections for the effects of population stratification, as not to identify false positive genetic associations that are actually a result of confounding from population heterogeneity. To stratify a cohort into differential populations, statistical methods such as principal component analysis (PCA) can be used to assess the variance in SNPs harboured by individuals within a population and determine their ancestral "distance" from one another (Hellwege et al., 2017).

1.2.4.2.1   Genome wide association studies

GWAS are aimed at identifying associations between SNPs and complex traits and diseases in large population-based studies. Over the past 15 years the approach has revolutionized the study of the small phenotypic impacts of relatively common genetic variation (Visscher, Brown, McCarthy, & Yang, 2012).

The method builds upon the long acknowledged understanding that genes can be effectively mapped by relying on population-based LD; however, it has done so by leveraging the insights of the human genome brought forth by the Human Genome Project (Visscher et al., 2012). More specifically, the subsequent HapMap project (International HapMap, 2005) investigated the degree of LD between SNPs throughout the genome. The map of LD can now be exploited to ensure that the SNPs analyzed across a GWAS analysis are not redundant, as SNPs in LD would give the same signal in the analysis.

The GWAS itself requires large populations comprised of both individuals with a phenotype of interest and controls, all genotyped using a genome-wide SNP array. Analyses are then able to identify SNPs highly associated with the trait of interest, based on their frequency in the cases in comparison to the controls, and produce an effect size relative to how much impact the SNP contributes to the phenotype under study (Bush & Moore, 2012). Following the identification of significant SNPs, the results may be further analyzed to determine if the association was 'direct', meaning as a result of the SNP genotyped itself, or 'indirect', meaning as a result of another genetic variant in high LD with the SNP that was genotyped (Bush & Moore, 2012).

Any direct associations identified using GWAS contribute to the common-disease-common-variant hypothesis. It is unclear who the true founders of this hypothesis were, as multiple researchers first proposed the theory in the mid- to late-1990s (Chakravarti, 1999; Lander, 1996; Risch & Merikangas, 1996). Briefly, the hypothesis proposes that all common diseases, with a population frequency >1–5% are driven by common genetic variation. The common variation may exhibit monogenic inheritance, such is the case with *APOE* and AD, or polygenic inheritance with many SNPs of small-to-modest phenotypic effect driving phenotypic presentation. Yet, further research over the last 20 years has proven there are very few complex phenotypes that can be substantially explained solely by common variation, leading to missing heritability and giving rise to the competing common-disease-rare-variant hypothesis (Gibson, 2012).

1.2.4.2.2   Rare variant association analysis

In contrast to the common-disease-common-variant hypothesis, the rare variant counterpart suggests that rare genetic variants of high phenotypic impact are driving phenotypic presentation (Gibson, 2012). The theory has also been expanded by some to include the possibility that there are many rare variants of small-to-modest phenotypic impact that are unable to be easily detected with GWAS or linkage approaches and may account for some of the missing heritability of disease. In turn, this hypothesis has led to the development of rare variant association analysis (RVAA).

As previously explained, GWAS are largely underpowered to detect rare, disease-associated variants without extravagant sample sizes, unless a variant's phenotypic impact is very large. Further, there is difficulty performing indirect association mapping on GWAS identified disease-associated SNPs to rare variants, as two variants must have similar MAFs to be in high LD (Asimit & Zeggini, 2010). In contrast, RVAAs capitalize on the understanding that although rare variants are individually infrequent, when binned together, they may be common enough to harness the statistical power for novel associations to be made. When performing RVAA, rare variants are collapsed into variant groups, which can be dictated by a variety of factors, including, but not limited to: general genomic region, individual genes, pathways of interest, MAFs, or functional consequence (Asimit & Zeggini, 2010). Importantly, when rare variants are binned based upon specific genes or pathways, novel disease-associations may be identified.

Akin to GWAS, RVAAs also utilize large populations of individuals both with a phenotype of interest and controls; however, in contrast to GWAS, typically RVAAs employ NGS data, as microarrays cannot capture the full breadth of possible rare variation in a population (Asimit & Zeggini, 2010). Following NGS, variants are prioritized using quality control metrics, MAFs, *in silico* prediction algorithms, and/or previous disease associations — all of which are discussed further in Section 1.1.5. Rare variants are then binned into their respective groups, and various methods can be used to identify associations between the rare variant groups and the phenotype of interest. The methods vary widely and can include various types of univariate or multivariate regression models customized for the individual study, or standardized models, such as

burden tests or the optimal unified sequence kernel association test (SKAT-O) (Lee, Emond, et al., 2012).

Burden tests collapse rare variants based on genomic region under the assumption that the rare variants all contribute to the phenotype in the same direction (i.e. they all increase risk for the phenotype). The approach can be useful when the genes under study are known to contribute to the phenotype in a similar manner, but it is generally accepted that at least a portion of variants, if not the general majority identified in a region, will have little to no phenotypic impact. Unfortunately, this can introduce noise in the analysis, as well as a loss of statistical power (Wu et al., 2011). On the contrary, sequence kernel association testing (SKAT) uses a kernel function and variance component tests to identify phenotypic associations with rare variants binned, again, based on genomic region, but also accounts for variants with differential phenotypic impact magnitudes and directions (Wu et al., 2011). The approach was further refined to produce SKAT-O, which was used in the methodology of Chapter 4, and builds upon SKAT by including the ability to identify the optimal testing strategy between burden tests and SKAT to maximize statistical power. More specifically, when the effects of variants are correlated, SKAT-O will employ a burden method, rather than the SKAT approach (Lee, Wu, & Lin, 2012).

1.2.5   Genetic variant interpretation

When studying how genetic variation can contribute to disease, an important aspect to consider is interpretation of the variants. As described above, prior to RVAAs, variants are prioritized to identify those that are rare in the general population (commonly MAF < 1%), and may be further filtered based on the goals of the study. Additionally, following the identification of gene-disease associations using linkage analysis or association analysis, assessing the deleteriousness of different variants within a gene becomes the new challenge, particularly when there is interest in implementing clinical genetic testing.

One of the first steps in this process is the interpretation of identified gene-disease relationships themselves. With such rapid advances in genomic technologies, there has been an exponential increase in gene-disease correlations, yet the levels of evidence

behind these correlations may differ widely. Online Mendelian in Man (OMIM) ("Online Mendelian Inheritance in Man, OMIM®,") is a freely available resource that has amalgamated information regarding human genes and associated phenotypes in a comprehensive manner. Although the compendium only originally included genes definitively carrying Mendelian disease-causing variants, the collection has broadened to include genes with varying degrees of disease-association evidence. While this is an excellent resource to clinicians and researchers alike, standard guidelines were necessary for evaluating the strength of gene-disease correlations, especially for use in the clinical setting. In 2017, the NIH-funded Clinical Genome Resource (ClinGen) published a standard framework for the evaluation of these gene-disease correlations and efforts are ongoing to classify gene associations across many diseases by Gene Curation Expert Panels (GCEPs) (Strande et al., 2017).

When interpreting genetic variants themselves, there are multiple approaches that may be taken. A customized approach may be used, based on the goals of a study, or a validated framework may be applied, such as the American College of Medical Genetics and Genomics (ACMG) Standards and Guidelines for Interpretation of Sequence Variants (Richards et al., 2015), which is beneficial in clinical settings. Customized strategies tend towards variant filtration to prioritize those likely-to-be or known-to-be disease causing or risk associated, whereas the ACMG guidelines classify all variants using the framework of: "benign," "likely benign," "uncertain significance," "likely pathogenic," and "pathogenic."

In addition to the approach, many different factors go into interpreting genetic variants, including, but not limited to, variant zygosity, sequence ontology, MAFs, *in silico* prediction scores, and previous disease associations. To begin, it is important to consider the zygosity of the variants of interest, as for an established gene-disease association there is often an accepted inheritance model for the relationship, such as autosomal dominant or autosomal recessive, especially in cases of Mendelian inheritance. Sequence ontology refers to the type of genetic variation, typically defined by the resulting change to the protein encoded by the gene of interest. Sequence ontologies are binned into two general categories: 1) nonsynonymous, or variants that change the resulting protein code

and 2) synonymous, or variants that do not change the resulting protein code. Nonsynonymous variants can be further subdivided, based on sequence ontology, to include: 1) putative LOF variants, including nonsense, splicing, and frameshift InDels; 2) missense variants, referring to those that change a single amino acid in the protein to a new amino acid; and 3) in-frame InDels. MAFs have heretofore been comprehensively covered in this Dissertation, but it is important to recognize that these variant frequencies can be obtained from a variety of population databases. The first example of a general population database is the 1000 Genomes Project (1000G) (Auton et al., 2015), which launched in 2008 and aimed to identify most common genetic variation within populations from across the world using WGS. In total, the final 1000G dataset is comprised of 2,504 samples that have all been well defined ancestrally, allowing the data to be quite useful when performing genetic ancestral estimation. Another example of a general population database is the Exome Aggregation Consortium (ExAC) (Lek et al., 2016), which was produced by the Broad Institute of MIT and Harvard and included the WES data from over 60,000 samples. The database was later updated into the Genome Aggregation Database (gnomAD) (Karczewski et al., 2020), to reflect the inclusion of WGS data. Version 2.1.1 of gnomAD includes SNV and InDel data from WES of over 125,000 samples and structural variation data from WGS of over 15,000 samples. Version 3.1 of gnomAD includes SNV and InDel variant data from WGS of over 76,000 samples. Although larger datasets provide the obvious advantage of a greater ability to capture a wider range of human genetic variation, the choice of general population database is dependent on the specifications of the study. For example, while ExAC is a smaller dataset than either version of gnomAD, it does not include samples from individuals with neurological conditions, whereas gnomAD does include some individuals with these conditions. *In silico* predictions are algorithms that are able to predict how deleterious a nonsynonymous variant may be by considering the conservation of the affected amino acid(s) and/or the biochemical properties and location of the amino acid change. When employing *in silico* predictions, it must be understood that they are not able to definitively determine variant pathogenicity, rather they provide a prediction of how deleterious a variant may be. In fact, in many cases different *in silico* prediction algorithms may not agree and one must consider the factors contributing to

each individual algorithm in order to assess validity of prediction calls. Examples of *in silico* algorithms used within this Dissertation include: Polymorphism Phenotyping v2 (PolyPhen-2) (Adzhubei et al., 2010), Sorting Intolerant from Tolerant (SIFT) (Kumar, Henikoff, & Ng, 2009), and Combined Annotation Dependent Depletion (CADD) (Kircher et al., 2014). Finally, when interpreting genetic variation, we must consider previous disease associations of the individual variants, which may be described within the relevant literature, or captured by databases such as ClinVar (Landrum et al., 2014).

1.3   Neurodegenerative disease

Neurodegenerative diseases are a collection of progressive, debilitating conditions characterized by the degeneration of neurons, either within the brain or peripheral nervous system. Upon pathological examination, neurodegenerative disease patients present with protein aggregates within the brain that classically define their diagnosis (**Table 1.1**). Although clinical presentations differ widely between different neurodegenerative disease diagnoses, features do tend to overlap, and even within single neurodegenerative disease cohorts, such as AD, presentations between patients is highly heterogenous. Further, very few individuals have pure forms of their diagnosis; rather, many exist on a continuum of neurodegenerative phenotypes, and mixed pathologies are common (Dugger et al., 2014). Importantly, the greatest risk factor for neurodegenerative disease development is increasing age.

Neurodegenerative diseases are largely grouped into two categories: 1) motor/movement disorders and 2) cognitive and/or behavioural disorders. The former includes motor neuron diseases, such as ALS, and parkinsonisms, such as PD (Trojsi, Christidi, Migliaccio, Santamaria-Garcia, & Santangelo, 2018). As the name suggests, motor/movement disorders principally affect a patient's mobility, including decreased motor control, apraxia, and speech/swallowing difficulties, although cognitive features or mixed pathologies with cognitive and/or behavioural disorders are not infrequent. In a similar manner, cognitive and/or behavioural disorders are mainly defined by the presentation and progression of decline in cognitive functioning across a variety of domains and/or severe changes in behaviour. Yet, again, patients may also present with features of impaired movement (Trojsi et al., 2018).

A third group of neurodegenerative diseases that is often discounted, or even ignored, is vascular cognitive impairments (VCIs). Although not always defined as a neurodegenerative disease, it is important to account for the role that cerebrovascular disease (CVD) plays in neurodegeneration phenotypes. In fact, Alois Alzheimer himself, who first defined AD, suggested involvement of the circulatory system in the disease's pathology (Raz, Knoefel, & Bhaskar, 2016). Further, damage within the neuro-vasculature has been observed across neurodegenerative diagnoses, including white matter hyperintensities, cerebrovascular lesions, cerebral amyloid angiopathy (CAA), and enlarged perivascular spaces (Ramirez et al., 2015; Raz et al., 2016). However, VCIs are defined as cognitive dysfunction as a direct result of vascular dysfunction and include diagnoses such as vascular dementia and cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL).

For the purposes of this Dissertation, five neurodegenerative diseases will be described in detail (**Table 1.1**), as well as the potential involvement of CVD in neurodegeneration pathology and presentation.

Table 1.1 Breakdown of the general characteristics of various neurodegenerative disease diagnoses.

Neurodegenerative disease	Prevalence (per 1000)	Early-onset age (years)	Late-onset age (years)	Common features	Classical neuropathology hallmarks
Alzheimer's disease (AD)	13.8[1,2]	< 65	≥ 65[6]	Cognitive decline and memory impairment disrupting daily functioning	Amyloid-β plaques; neurofibrillary tau tangles; in some cases, Lewy bodies, cerebral amyloid angiopathy, and TDP-43 inclusions
Amyotrophic lateral sclerosis (ALS)	0.044[3]	< 45	≥ 45	Loss of upper and lower motor neurons resulting in progressive muscle weakness and eventual paralysis; in some cases, concurrent FTD	TDP-43 ubiquitinated inclusions; FUS ubiquitinated inclusions; p62 ubiquitinated inclusions
Frontotemporal dementia (FTD)	0.591[1,4]	< 65	≥ 65	Frontal and anterior temporal lobe atrophy resulting in severe behaviour changes, language impairment, and cognitive decline; in some cases, concurrent ALS	TDP-43 ubiquitinated inclusions; hyper-phosphorylated tau; in rare cases, FUS ubiquitinated inclusions
Mild cognitive impairment (MCI)	160[5]	< 65	≥ 65	Memory impairment and/or impaired language and executive function that does not disrupt daily functioning	Amyloid-β plaques; neurofibrillary tau tangles
Parkinson's disease (PD)	4.2[1]	< 40	≥ 60	Loss of dopaminergic neurons, resulting in bradykinesia, rigidity, resting tremor, and postural instability; in some cases, dementia	Lewy bodies (with α-synuclein aggregates); amyloid-β plaques; neurofibrillary tau tangles

Presented prevalences were per 1000 people. The AD, FTD, and PD prevalences were all based on individuals over the age of 40 years. The ALS prevalence was based on individuals of any age. The MCI prevalence was based on individuals over the age of 50 years.
[1](Ng et al., 2015); [2]("2020 Alzheimer's disease facts and figures," 2020); [3]("Chapter 3: Mapping Connections: An understanding of neurological conditions in Canada – Scope (prevalence and incidence)," 2014); [4](Hogan et al., 2016); [5](R. Roberts & Knopman, 2013).
Abbreviations: FUS, Fus RNA binding protein; p62, nucleoporin 62; TDP-43, TAR DNA-binding protein 43.

1.3.1   Motor/movement disorders

1.3.1.1   Amyotrophic lateral sclerosis (ALS)

ALS is a neuromuscular disorder characterized by the loss of upper and lower motor neurons, referring to the neurons connecting the cortex and brainstem/spinal cord and the neurons connecting the spinal cord and muscles/glands, respectively. Generally, the neuronal degeneration results in progressive, severe muscle weakness and eventual paralysis, although presentation is highly heterogeneous with mixed features of muscle atrophy, fasciculations, and spasticity. With a mean age of onset of roughly 61 years, classic cases display a median survival period of 2–4 years (Grad, Rouleau, Ravits, & Cashman, 2017). ALS is also considered relatively uncommon compared to other neurodegenerative diseases, with prevalence estimate of 0.044 per 1,000 individuals, although this considers individuals of any age ("Chapter 3: Mapping Connections: An understanding of neurological conditions in Canada – Scope (prevalence and incidence)," 2014). Upon neuropathological analysis, a classic feature of ALS is the identification of ubiquitinated inclusions in the motor neuron cytoplasm, yet the etiology and components of the inclusions can vary, with some heterogeneity attributable to specific monogenic genetic variants. In turn, potential mechanisms of disease are quite complex and may also vary between patients, suggesting a spectrum of phenotypes under the umbrella of ALS (Mejzini et al., 2019).

Generally, ALS onset begins with muscle weakness within the face, arms, and/or legs, and steadily spreads and advances throughout the body. The majority of typical ALS patients present with a spinal form of the disease with asymmetrical muscle weakness onset in the distal or proximal upper or lower limbs; however, muscle weakness spreads rather quickly to other limbs, as well as bulbar symptoms affecting speech and swallowing, and eventually respiratory symptoms, although order of onset may vary (Wijesekera & Leigh, 2009). The development of limb spasticity is also common, and bladder dysfunction and behavioural or cognitive impairments have also been reported. In fact, a report by Strong *et al.* found that based on clinically accepted diagnostic criteria, over 50% of ALS patients present with a form of FTD or other dementia, such as AD (Strong et al., 2017).

1.3.1.1.1 Genetic determinants of ALS

ALS cases are largely classified as either familial (fALS) or sporadic (sALS), referring to the ~10% of cases with a family history and ~90% of cases without a family history, respectively. Interestingly, only 40–55% of fALS cases can be explained by dominant monogenic variants in known ALS-associated genes, with variants in the genes superoxide dismutase (*SOD1*), *C9orf72*, FUS RNA binding protein (*FUS*), and TAR DNA binding protein (*TARDBP*) being most common (Mejzini et al., 2019). A small fraction of sALS patients have also been found to harbour monogenic variants in fALS-associated genes, yet over 90% of cases remain genetically unexplained in spite of sALS heritability estimates of ~60% (Mejzini et al., 2019). Importantly, patient neuropathology and clinical presentation is highly dependent on the genetic factors contributing to their ALS diagnosis, although presently, genetics fail to account for *all* disease heterogeneity.

1.3.1.1.1.1 Monogenic determinants of ALS

For the purposes of brevity, the effects of variants in the four most common genes associated with fALS will be discussed further; however, there are many other genes that have now been associated with the phenotype.

Monogenic variants in *SOD1* account for approximately 20% of fALS cases and are thought to induce excitotoxicity, oxidative stress, endoplasmic reticulum stress, mitochondrial disfunction, and prion-like propagation, although exact pathological mechanisms are still under study (Mejzini et al., 2019). Neuropathological analysis has identified aggregates of SOD1 within immunoreactive inclusions of the motor neurons of ALS patients harbouring *SOD1* pathogenic variants, yet the aggregates have also been observed in sALS patients without *SOD1* variants and in patients carrying a *C9orf72* repeat expansion or variants in other ALS-associated genes (Bosco et al., 2010; Mejzini et al., 2019). *TARDBP* encodes the protein TDP-43, which is the main component of ubiquitinated inclusions in ALS patients that are distributed in neurons throughout the central nervous system. Normally, TDP-43 is involved in mRNA stability, processing, and transport, but when mutated the protein no longer localizes to the nucleus and instead aggregates in the cytoplasm. Yet, these aggregates are not unique to those carrying

TARDBP variants and TDP-43 inclusions are now considered a hallmark of a majority of ALS cases (Mejzini et al., 2019). In contrast, ALS patients carrying pathogenic *FUS* variants do not typically present with TDP-43 inclusions, rather they harbour unique FUS aggregation within ubiquitinated neuronal cytoplasmic inclusions throughout the central nervous system. FUS shares many gene expression roles with TDP-43, although with different targets. *FUS* variants have also been implicated in cases of concurrent ALS and FTD (Mejzini et al., 2019; Saberi, Stauffer, Schulte, & Ravits, 2015; Snowden et al., 2011).

The *C9orf72* gene contains a hexanucleotide repeat expansion (GGGGCC) in its non-coding region, and an expansion on at least one allele of >30 copies — although more often the pathogenic repeats are present in hundreds to thousands of copies — accounts for one of the most commonly inherited forms of ALS in Europeans, including cases of sALS. The repeat expansion also increases risk of FTD development, both with or without ALS (DeJesus-Hernandez et al., 2011; Renton et al., 2011; Shatunov et al., 2010). Again, exact mechanisms are unclear, but patients present with ubiquitinated TDP-43 aggregates in neuronal cells; however, repeat expansion carriers are unique as the TDP-43 positive inclusions are outnumbered by nucleoporin 62 ubiquitinated neuronal inclusions, as well as dipeptide repeats protein aggregates most commonly found within the frontal, occipital, temporal, and motor cortices (Saberi et al., 2015).

As previously described, sALS cases remain mostly genetically undefined, yet heritability estimates are ~60%; therefore, there is a large amount of missing heritability for this phenotype. GWAS analyses have been used to account for some of this missing heritability, but many have only identified genes associated with fALS or rare subtypes of the disease (Rich, Roggenbuck, & Kolb, 2020). RVAAs have also been used to identify genes harbouring rare variation that may be contributing to the phenotype (S. M. K. Farhan et al., 2019; Nicolas et al., 2018). Yet replication of associations remains difficult, and relatively small sample sizes of ALS cohorts are likely limiting the ability to identify novel genetic associations, such as ultra-rare SNVs or structural variants in novel disease genes (Theunissen et al., 2020).

1.3.1.1.1.2 Polygenic contributors to ALS

The aforementioned GWAS analyses of ALS have also successfully identified a selection of common SNPs with small contributions to genetic risk (Nicolas et al., 2018). Unfortunately, there are relatively few of these studies, as obtaining large cohorts of ALS patients, along with appropriate replication cohorts, remains difficult. In time, enough associated SNPs may be identified that the generation of an accurate PRS score is possible, particularly for the sporadic forms of disease with a large amount of remaining missing heritability. With that said, an interesting pattern has emerged in the literature with regard to the identification of polygenic contributors to phenotypic modifiers of ALS, rather than the disease state itself. For example, a SNP in the gene Unc-13 Homolog A (*UNC13A*) was associated with cognitive decline, prefrontal and temporal cortex degeneration, and burden of hyperphosphorylated TDP-43 in patients with ALS and FTD (Placek et al., 2019), while another recent study created a PRS for cognitive dysfunction in ALS that also predicted atrophy severity in the frontal and temporal lobes (Placek et al., 2021). Overall, these studies suggest that in addition to polygenic contributors to overall ALS diagnosis, variants of small to moderate effect — both common and rare — may contribute to the heterogeneous presentation between ALS patients as well.

1.3.1.2 Parkinson's disease (PD)

PD is the most common movement disorder, with a prevalence of 4.2 per 1,000 individuals over the age of 40 in Ontario (Ng et al., 2015). The phenotype is characterized by loss of dopaminergic neurons in the substantia nigra pars compacta, resulting in motor features, such as bradykinesia, rigidity, resting tremor, and postural instability (Kalia & Lang, 2015). Presentation between PD patients is highly heterogeneous and sub-typing is not uncommon — for example, differentiating tremor-dominant PD from non-tremor dominant PD. Non-motor features are also not uncommon in PD patients and may present before or after the onset of motor features. Cognitive impairment and dementia, psychiatric disorders, sleep disruptions, and olfactory dysfunction can all appear up to 12–14 years prior to the motor symptoms of PD, yet often remain unexplained until diagnosis (Kalia & Lang, 2015). Additionally, although

there are therapies that can manage disease symptoms, such as dopamine replacement therapy, patients' conditions continue to worsen. With an average disease duration of 17 years, late-stage patients present with severe, debilitating treatment-resistant symptoms, including freezing of gait and falls, dysphagia, speech impairment, urinary incontinence, and dementia (Kalia & Lang, 2015).

As with other neurodegenerative diseases, PD displays a hallmark of protein aggregates, most notably aggregates of α-synuclein that form inclusions within neurons called "Lewy bodies" (Spillantini et al., 1997). Lewy bodies do not stay centralized to the brain, but can also be found in both the spinal cord and peripheral nervous system, and it is hypothesized that these inclusions induce the onset of motor symptoms. It is also well established that Lewy bodies contribute to late-stage cognitive phenotypes within PD patients, and the inclusions are often used to define "Lewy body diseases" under the PD spectrum, including Lewy body dementia and PD dementia (Gomperts, 2016). In addition to α-synuclein, other proteins are also known to aggregate in the neurons of PD patients, including amyloid-β plaques and neurofibrillary tau tangles, both of which contribute to an earlier onset of dementia symptoms. Aside from the protein aggregates, PD pathology has been found to involve dysfunction in the ubiquitin-proteosome system, mitochondrial function, the lysosome-autophagy pathway, protein trafficking, synaptic response, and, of course, dopamine neurotransmission (Kalia & Lang, 2015).

1.3.1.2.1   Genetic determinants of PD

Approximately 20% of PD cases are considered familial, although only 5–10% of all PD cases have a known monogenic cause, leaving most of the remaining familial cases genetically undefined (Ohnmacht, May, Sinkkonen, & Kruger, 2020). Genetic risk factors of low penetrance and polygenic risk factors have also been associated with a small fraction of sporadic PD cases. Including all PD cases, both seemingly familial and sporadic, heritability estimates are approximately 30%, which may increase as understanding of the complex nature of PD continues to develop (Keller et al., 2012; Nalls et al., 2014). Therefore, a relatively large amount of heritability remains unaccounted for. Importantly, PD associated genetic variants often display a large

amount of variable expressivity, including variation in age of onset, features of disease, and rates of progression.

1.3.1.2.1.1 Monogenic determinants of PD

More than 20 years ago, the first gene to be associated with PD was identified as synuclein alpha (*SNCA*), which encodes the α-synuclein protein and has a role in synaptic vesicle trafficking (Polymeropoulos et al., 1997). Rare pathogenic variants in the gene result in autosomal dominant PD; however, common non-coding variants have also been identified at the locus attributable to sporadic PD (Maraganore et al., 2006; Simon-Sanchez et al., 2009). All PD patients harbouring monogenic *SNCA* variation display hallmarks of Lewy bodies, Lewy neurites, and dopaminergic neuronal loss, yet clinical manifestation of disease can vary widely depending on the variant. Although many of the variants within *SNCA* previously associated with PD are nonsynonymous SNVs, duplications and triplications of the entire gene are also causative of the disease (Lunati, Lesage, & Brice, 2018). Autosomal dominant variants in the genes leucine-rich repeat kinase 2 (*LRRK2*), VPS35 retromer complex component (*VPS35*), GTP cyclohydrolase 1 (*GCH1*), and Ataxin 2 (*ATXN2*) also display well-established associations with PD.

Other genes have been identified that display an autosomal recessive inheritance pattern for PD; listed in order of variant frequency in PD, the genes include parkin RBR E3 ubiquitin protein ligase (*PARK2*), PTEN induced kinase 1 (*PINK1*), and parkinsonism associated deglycase (*DJ1*) (Lunati et al., 2018). PD patients carrying homozygous variants in these genes typically present with an early-onset form of disease (<40 years). Additionally, all three genes are involved in mitochondrial quality control, and pathogenic variants within the genes result in mitochondrial and mitophagy dysfunction (Ryan, Hoek, Fon, & Wade-Martins, 2015). Many other genes have been identified through both family studies and GWAS as potentially associated with either autosomal dominant or autosomal recessive PD, yet a large portion have yet to be fully validated.

Generally accepted as the most common genetic risk factor for PD is variation in the glucocerebrosidase gene (*GBA*). Originally, the gene was associated with autosomal recessive Gaucher's disease; however, patients were identified that presented with

generalized parkinsonism, and, upon further review, family members of patients with Gaucher's disease displayed increased incidence of PD, and many were identified as heterozygous *GBA* variant carriers (Halperin, Elstein, & Zimran, 2006; Tayebi et al., 2001). Aside from a lower frequency of resting tremor and a much greater risk of cognitive impairment and motor dysfunction, *GBA* variant carriers display relatively "typical" sporadic PD presentation (Sidransky et al., 2009). The variants themselves have been associated with lysosomal dysfunction, ultimately resulting in increased α-synuclein aggregation. Neuroinflammation and mitochondrial dysfunction have also been observed in these patients (Gegg & Schapira, 2018). The association between *GBA* and PD will be explored further in Chapter 6.

1.3.1.2.1.2 Polygenic contributors to PD

Although there are a small portion of PD patients that have monogenic forms of disease, largely patients are considered complex, with likely polygenic inheritance models. Recent estimates indicate that heritability of PD attributable to common variants may be between 16–36% (Goldman et al., 2019; Nalls et al., 2019). Specifically, 90 SNPs have now been associated with PD; however, the PRS that was created using these SNPs was predicted to lead to a high rate of false positive associations if applied to a general population, creating concerns for the clinical application of a PRS in PD (Nalls et al., 2019). Yet, further analysis of polygenic contributors to PD may allow for the development of a more accurate PRS.

In a similar manner to ALS, polygenic contributors have also been identified in PD to modify classical disease presentation. Shared genetic loci have been identified to associate with both PD and schizophrenia, suggesting common mechanisms underlying clinical features of the diseases (Smeland et al., 2021). Further, a weighted PRS of 23 GWAS identified SNPs was associated with faster cognitive and motor decline in patients with PD (Paul, Schulz, Bronstein, Lill, & Ritz, 2018). Again, these studies suggest that genetic variants with small to moderate phenotypic impact may not only contribute to risk of PD, but to the differential presentation between PD patients.

1.3.2   Cognitive and/or behavioural disorders

1.3.2.1   Alzheimer's disease (AD)

As the most common neurodegenerative disease, AD has a prevalence across Ontario of 13.8 per 1,000 individuals over the age of 40 years and up to 50.4 per 1,000 individuals over the age of 65 years ("2020 Alzheimer's disease facts and figures," 2020; Ng et al., 2015). The disease is characterized by progressive cognitive decline and severe memory impairment that disrupts daily functioning. As patients progress through the course of AD, they become more dependent on external care and may experience behavioural changes, mobility impairment, and/or psychosis. The mean age of diagnosis is ~75 years — with early-onset (EOAD) cases defined as onset < 65 years of age and late-onset (LOAD) defined as onset ≥ 65 years of age — and the disease has an average duration of ~8.5 years, although duration is often longer in early-onset and/or familial cases (Jost & Grossberg, 1995). Similar to most neurodegenerative diseases, available treatments and interventions are only able to marginally slow disease progression and mitigate symptoms (Weller & Budson, 2018).

AD neuropathology is largely defined by amyloid-β plaques and neurofibrillary tau tangles. Amyloid-β plaques are extracellular inclusions of amyloid-β-40 and amyloid-β-42, which are the by-products of aberrant amyloid precursor protein (APP) metabolism, and the latter of which is of higher abundance in resulting plaque deposits (Tiwari, Atluri, Kaushik, Yndart, & Nair, 2019). The amyloid-β first forms fibrils that oligomerize and diffuse into the extracellular space. There, the fibrils interrupt synaptic signaling and polymerize further to form plaques. Interestingly, the accumulation of amyloid-β plaques activates the hyperphosphorylation of microtubule-associated tau proteins, resulting in the formation of neurofibrillary tau tangles (Tiwari et al., 2019). While amyloid-β plaques have long been associated with AD, deposition of amyloid-β has also been observed in cognitively normal elderly individuals, suggesting amyloid-β cannot drive disease presentation in singularity; in contrast, neurofibrillary tau tangle load has been directly associated with degree of cognitive impairment (Lane, Hardy, & Schott, 2018). Although it is well established that amyloid-β and neurofibrillary tau tangles have a large role in AD pathology, precise mechanisms are yet to be elucidated. Neuropathological

markers such as neuropil threads, dystrophic neurites, and CAA, or co-pathologies such as vascular disease, Lewy bodies, or TDP-43 inclusions are also not uncommon in AD patients (Lane et al., 2018).

1.3.2.1.1 Genetic determinants of AD

The dichotomy between EOAD and LOAD loosely translates to whether the disease is considered familial or not, with most EOAD cases considered as familial, and most LOAD cases considered sporadic. Therefore, it is unsurprising that EOAD displays heritability estimates of 90–100%. Yet, LOAD heritability remains unexpectedly high at 70–80% (Cacace, Sleegers, & Van Broeckhoven, 2016). However, even with such high heritability, ~90% of EOAD cases remain genetically unexplained, as do a large majority of LOAD cases.

1.3.2.1.1.1 Monogenic determinants of AD

There are three main genes — *APP*, presenilin 1 (*PSEN1*), and presenilin 2 (*PSEN2*) — that in aggregate are known to account for 5–10% of EOAD cases, all of which were identified by linkage analysis to contribute to the disease in an autosomal dominant manner (Cacace et al., 2016). *APP* was the first gene associated with AD and the protein product is thought to be involved in synaptic development (Priller et al., 2006). As previously described, APP is also cleaved by secretases into by-products that eventually contribute to amyloid-β plaque formation. Both rare SNVs and structural variants have been identified as pathogenic in *APP* for EOAD, likely contributing to the phenotype through the overproduction of amyloid-β-42 and resulting in increased amyloid plaque pathology (Shao, Peng, & Wang, 2017). *APP* has also been associated with CAA, characterized by cerebrovascular amyloid deposition and resultant dementia and/or stroke. *PSEN1* and *PSEN2* encode proteins involved in the secretase in the amyloidogenic pathway implicated in cleaving APP (Cacace et al., 2016). *PSEN1* is the most frequently mutated out of the three main EAOD genes, whereas variants within *PSEN2* are considered rarer. Within *PSEN1* there have been a wide range of variants identified, including SNVs and structural variants, and almost all are considered highly penetrant; however, while both SNVs and structural variants have been associated with

EOAD in *PSEN2*, penetrance is more variable and variable expressivity has been observed (Cacace et al., 2016).

Of high importance, genetic risk of AD could not be discussed without describing *APOE*, the strongest genetic risk factor for sporadic LOAD. The association was originally discovered through linkage analysis and has been further replicated in many case-control studies. Interestingly, *APOE* is associated with increased risk of EOAD, particularly in individuals with a family history of AD (van Duijn et al., 1994). *APOE* encodes the apo E glycoprotein, which is responsible for carrying and distributing cholesterol, both in the periphery and the brain, and has roles in neuronal growth, nerve regeneration, and immunoregulation. Notably, there are two variants within the gene, located at amino acid positions 130 and 176 — or 112 and 158, respectively, if the pro-peptide sequence is excluded — that give rise to three apo E isoforms: apo E2, apo E3, and apo E4 (**Figure 1.8**) (Zannis et al., 1982). While the wild-type gene is encoded by the E3 genotype, the E4 genotype increases risk of AD in a dose dependent manner, with approximate increased risks of three times in heterozygous carriers and 12 times in homozygous individuals. The E4 genotype is relatively common in the general population, with an MAF of ~0.14, although frequencies vary between ancestral populations (Heffernan, Chidgey, Peng, Masters, & Roberts, 2016). In contrast, the less common E2 genotype (MAF ~ 0.07) has been associated with protective effects against AD development.

Although the exact mechanism by which *APOE* contributes to AD risk is unclear, it is seemingly involved in many potentially pathogenic pathways (Munoz, Garner, & Ooi, 2019). Not only has apo E been observed within amyloid-β plaques, but it may have a direct role in amyloid-β fibril formation and amyloid plaque clearing within the brain. The E4 isoform is less effective at preventing the fibril formation and potentially less effective at clearing amyloid plaques (Munoz et al., 2019). However, apo E is also observed within neurofibrillary tangles, and overexpression of apo E increases tau hyperphosphorylation. Apo E4 also prevents synaptic protein expression and is associated with increased α-synuclein pathology (Zhao et al., 2020; Zhao, Liu, Qiao, & Bu, 2018).
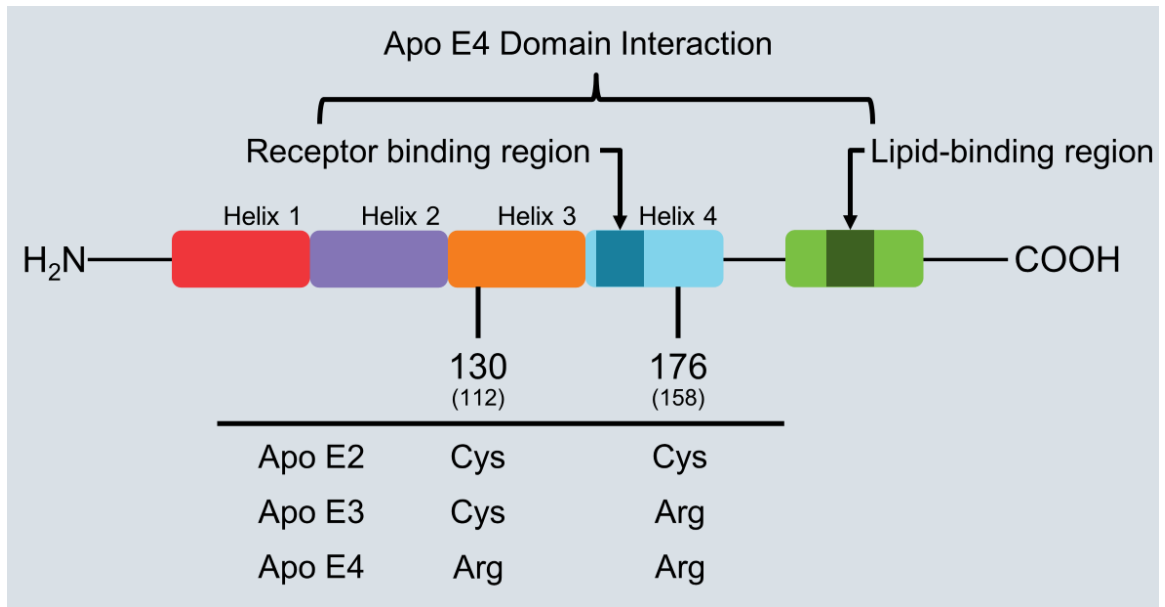
Figure 1.8 Schematic of the apolipoprotein (apo) E protein, demonstrating the location of the variants resulting in the differential apo E isoforms.
APOE encodes the apo E glycoprotein, which is responsible for carrying and distributing cholesterol. Two variants within the gene, located at amino acid positions 130 and 176, give rise to three apo E isoforms: apo E2, apo E3, and apo E4. E3 is considered the wild-type isoform, and E2 and E4 each have differential receptor binding affinity. Specifically, the apo E4 isoform results in a domain interaction that reduces the protein's binding affinity, and is the greatest genetic risk factor for the development of Alzheimer's disease. Adapted from (Yu, Tan, & Hardy, 2014).

1.3.2.1.1.2 Polygenic contributors to AD

Aside from *APOE*, there have been over 20 common variants identified through GWAS associated with LOAD. Current efforts are being put forth to create an accurate, replicable polygenic risk score to account for the potentially polygenic risk of AD; however, NGS association studies have also since identified rare variants in GWAS identified loci, as well as within other genes of potential interest, that may contribute moderate risk of disease (Raghavan & Tosto, 2017).

Interestingly, PRSs have been created for AD with relatively large predictive power for disease (Escott-Price, Myers, Huentelman, & Hardy, 2017; Escott-Price et al., 2015), but it has since been determined that a large amount of the predictive power from these scores is encompassed by the inclusion of the *APOE* E4 allele(s) (Escott-Price, Myers, Huentelman, Shoai, & Hardy, 2019). Yet, recent analyses have demonstrated that even when E4 carriers are excluded from PRS analyses, predictive power remains promising, although only in cohorts of pathologically confirmed AD cases. In clinical cohorts of AD, likely with an admixture of misdiagnosed individuals, the predictive power of the PRSs in non E4 carriers is reduced (Escott-Price et al., 2019; Escott-Price et al., 2015). The results are concerning, as the clinical utility of the currently available PRSs may not be generally applicable in AD clinics or in the offices of general practitioners, which are highly accessed by neurodegenerative disease patients living in rural areas.

1.3.2.2 Frontotemporal dementia (FTD)

FTD defines a collection of conditions, all characterized by atrophy of the frontal and anterior temporal lobes, and resulting in progressive behavioural changes and/or language dysfunction (Devenney, Ahmed, & Hodges, 2019). Overall, the disease displays a prevalence in Ontario of 0.591 per 1,000 individuals over the age of 40 and 1.944 per 1,000 individuals over the age of 65 (Hogan et al., 2016; Ng et al., 2015). FTD subtypes include: behavioural variant FTD (bvFTD), corticobasal syndrome (CBS), progressive supranuclear palsy (PSP), and primary progressive aphasia (PPA). BvFTD displays an average age of onset of ~58 years and is characterized by severe changes in social conduct and personality, including apathy, lack of inhibition and empathy, mental

rigidity, and deficits in executive function (Devenney et al., 2019). In contrast, CBS involves asymmetric motor features, such as limb rigidity, dystonia, or myoclonus and can sometimes present alongside other forms of FTD or dementia (Olney, Spina, & Miller, 2017). PSP is largely defined by vertical supranuclear gaze palsy or slow vertical saccades, as well as postural instability leading to falls, with onset after the age of 40 years. Patients may also present with the cognitive features common to many FTD sub-types, including personality changes, executive dysfunction, and reduced mental speed. Similar to CBS, PSP can co-exist with the other FTD sub-types (Olney et al., 2017). Finally, PPA is generally characterized by language impairment, but can be further subdivided into non-fluent and semantic variants. The non-fluent variant encompasses patients that present with severe difficulties with speech, whereas semantic variant patients present with language-based difficulties and issues with semantic knowledge (Olney et al., 2017).

It is also important to recognize the considerable overlap between FTD and ALS, specifically in regards to neuropathology and genetic associations. As previously described, between 40–50% of patients with ALS present with clinical features of FTD, such as behavioural changes, language impairment, and impaired executive function (Karch et al., 2018). It is also estimated that ~40% of FTD patients display a degree of motor dysfunction. If certain diagnostic criteria are met, patients may actually be diagnosed with concurrent ALS and FTD (ALS-FTD), accounting for ~15% of all FTD cases (Bennion Callister & Pickering-Brown, 2014; Strong et al., 2017).

Over 90% of FTD cases present with protein inclusions composed of hyperphosphorylated tau protein or ubiquitinated TDP-43; however, FUS positive inclusions and ubiquitin proteosome pathology are also observed in rare cases (Devenney et al., 2019). To some extent, neuropathology is dependent on FTD subtype, as well as any genetic contribution to disease, although correlations are not perfect. The majority of bvFTD cases present with heterogeneous inclusions containing at least one of: hyperphosphorylated tau, ubiquitinated TDP-43, or FUS. In cases of CBS and PSP, hyperphosphorylated tau is the predominant neuropathology, but in CBS, accumulation is in the cortex and basal ganglia, and in PSP, accumulation is observed in the basal ganglia,

brainstem, and cerebellum. Semantic variant PPA cases almost exclusively display TDP-43 inclusion pathology, whereas non-fluent variant PPA may present with either hyperphosphorylated tau or ubiquitinated TDP-43 (Devenney et al., 2019).

1.3.2.2.1 Genetic determinants of FTD

As previously stated, the genetic contributors to FTD influence patients' differential neuropathology and, in turn, the differential sub-types. Between 20–50% of FTD cases are considered familial, although frequencies differ dependent upon the FTD sub-type (Greaves & Rohrer, 2019). The remaining FTD cases are considered sporadic. Overall, FTD is considered highly heritable, yet a genetic cause has been identified in <20% of cases (Sieben et al., 2012).

1.3.2.2.1.1 Monogenic determinants of FTD

Three genes are able to account for the genetic predisposition to 60% of familial FTD cases, including microtubule associated protein tau (*MAPT*), granulin precursor (*GRN*), and *C9orf72*; the last of which refers to the hexanucleotide repeat expansion described in Section 1.2.1.1.1. In fact, the *C9orf72* expansion accounts for up to 25% of familial FTD cases alone and has also been observed in cases of seemingly sporadic FTD and ALS-FTD (Bennion Callister & Pickering-Brown, 2014).

MAPT was the first gene associated with familial FTD and was discovered through linkage analysis (Hutton et al., 1998; Poorkaj et al., 1998; Spillantini, Crowther, Kamphorst, Heutink, & van Swieten, 1998). The gene encodes the protein tau — the main component of hyperphosphorylated tau inclusions within neuron and glial cells in FTD. Normally, tau is involved in the stability of microtubules; however, when mutated, tau may become more abundant resulting in greater aggregation in the cytoplasm, may have an increased rate of phosphorylation, or may undergo alternative splicing resulting in imbalanced isoform ratios (Fenoglio, Scarpini, Serpente, & Galimberti, 2018). Carriers of *MAPT* pathogenic, rare variants with FTD most commonly present with the bvFTD, CBS, or PSP subtypes, and, unsurprisingly, carriers present with neuropathological hyperphosphorylated tau inclusions (Olszewska, Lonergan, Fallon, & Lynch, 2016). CBS and PSP have also been associated with common variation within *MAPT*, specifically in

reference to the *MAPT* haplotype. *MAPT* falls within the largest region of LD across the human genome, spanning ~1.8 Mb, inside of which 900 kb have been inverted, including the entire *MAPT* gene resulting in differential haplotypes, namely H1 and H2 (**Figure 1.9**) (Caffrey & Wade-Martins, 2007). Due to the large amount of LD throughout the region, the haplotype can be defined with only single SNPs that can be genotyped through methods such as TaqMan allelic discrimination assay, as will be applied in Chapter 2. Although the H1 haplotype is rather common in the general population, it has been associated with both CBS and PSP. Within the H1 haplotype, duplicated regions have also been discovered that can be defined by specific tagging SNPs, of which the H1C sub-haplotype has been associated with PSP (Caffrey & Wade-Martins, 2007).

GRN encodes progranulin, a protein involved in growth regulation, wound repair, and inflammation. Although variants within *GRN* are largely associated with familial FTD cases, there have been pathogenic rare variants identified in the gene in seemingly sporadic cases as well (Fenoglio et al., 2018). The variants result in haploinsufficiency, as the majority of known pathogenic variants result in premature termination of the protein and result in non-sense mediated decay. Missense variants have also been observed in *GRN* in FTD patients, yet their mechanism of pathogenicity remains unclear. *GRN* variants are associated with ubiquitinated TDP-43 inclusions and patients tend to have the bvFTD or non-fluent PPA sub-types (Miller & Llibre Guerra, 2019).

There are many other genes that have been associated with FTD, and its subtypes, in more rare instances, including *TARDBP* and *FUS*, indicating a large amount of genetic overlap between FTD and ALS (Bennion Callister & Pickering-Brown, 2014). Although monogenic inheritance of FTD is considered relatively uncommon, known Mendelian variants sufficiently account for disease risk in many cases and are, therefore, important markers to consider targeting in the development of novel therapeutics.
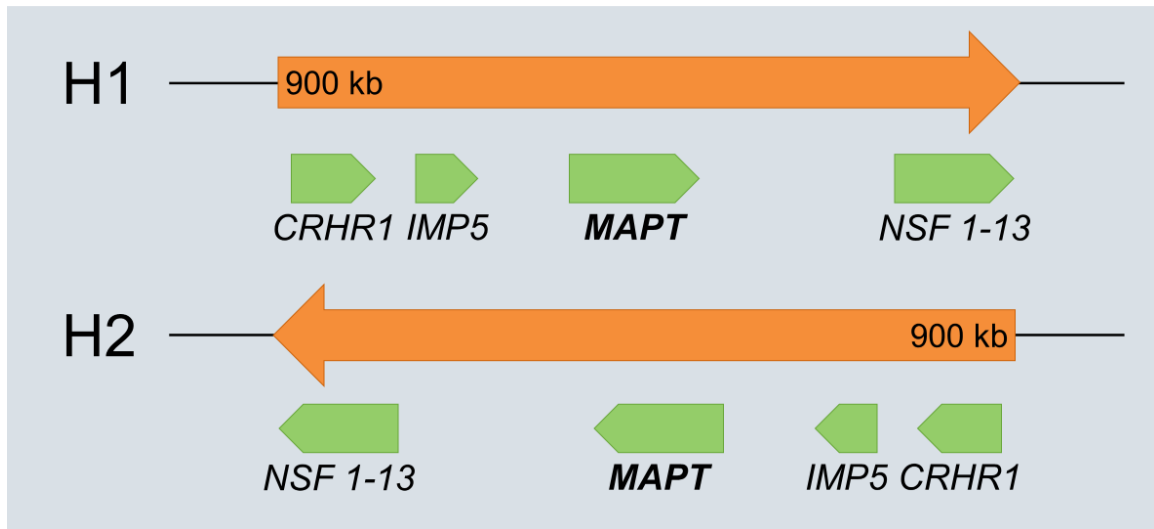
Figure 1.9 The 900 kilobase (kb) inversion on chromosome 17 (17q21) that results in two differential microtubule associated protein tau (*MAPT*) haplotypes.
MAPT falls within a ~1.8 megabase (Mb) region of linkage disequilibrium that includes 900 kb of inverted sequence, including the entire *MAPT* gene. The inversion results in differential haplotypes, namely H1 and H2. The H1 haplotype has been previously associated with subtypes of frontotemporal dementia. Adapted from (Caffrey & Wade-Martins, 2007).

1.3.2.2.1.2 Polygenic contributors to FTD

GWAS analyses are beginning to identify loci of interest associated with FTD, and attempts are being made to produce accurate polygenic risk scores to account for a portion of the remaining missing heritability (Ferrari et al., 2014; Hagenaars et al., 2018). Unfortunately, similar to GWAS of ALS, the large sample sizes required for these analyses have likely resulted in lower yield thus far, as FTD is considered a rarer type of dementia (Manzoni & Ferrari, 2021). While recent GWASs of AD have included >90,000 cases, the largest FTD GWAS to date has only included ~3,500 cases (Hagenaars et al., 2018; Kunkle et al., 2019). Further research into the genetic contributors to FTD with small to moderate phenotypic impact is warranted to account for the obvious missing heritability of the disease.

1.3.2.3 Mild cognitive impairment

MCI is a prodromal stage of dementia, characterized by memory impairment that does not disrupt daily functioning. Although some patients may not progress beyond the MCI diagnosis, over 50% will continue on to develop AD, vascular dementia, or other forms of dementia (Mitchell & Shiri-Feshki, 2009). As expected, MCI is highly heterogenous, and cases can be largely divided into two subtypes: 1) amnestic and 2) non-amnestic. Amnestic MCI is grossly characterized by memory impairment, whereas the non-amnestic form may present with intact memory, but impaired attention, language, or executive functioning. The former is more likely to progress to a classical AD presentation, while the latter may progress to other forms of neurodegeneration, such as FTD or PD dementia (Giau, Bagyinszky, & An, 2019).

1.3.2.3.1 Genetic determinants of MCI

Due to the large amount of heterogeneity within the MCI diagnosis, neuropathology varies. However, it is generally accepted that amyloid-β plaques and neurofibrillary tau tangle loads correlate with the likelihood that MCI will progress to AD or other dementias (Anderson, 2019). Similarly, the *APOE* E4 allele has been associated with both increased risk of MCI development, as well as increased risk of progression from

MCI to AD, yet any remaining genetic risk factors for MCI are undefined (Elcoroaristizabal Martin et al., 2011; Fleisher et al., 2007).

1.3.3   Cerebrovascular disease (CVD)

Not only does the brain consume 20% of cardiac output and require 20% of the body's available oxygen and glucose, but it is not able to store energy for the long-term and, therefore, requires a constant and dependable blood flow from its vascular system (Lendahl, Nilsson, & Betsholtz, 2019). It is unsurprising that when issues arise within the brain's intrinsic vasculature, or when blood flow to the brain is restricted, damage can be catastrophic. CVD is defined as any condition that affects the blood vessels within or the blood flow to the brain and encompasses both acute events, such as ischemic or hemorrhagic strokes, or CVD neuropathology, such as white matter hyperintensities, CAA, and enlarged perivascular spaces. Although CVD may solely drive neurodegenerative disease, as is the case with vascular dementia (VaD), it may also be a concurrent pathology with existing neurodegenerative disease, or act as one of many contributors to a neurodegenerative pathology, such as in AD or vascular PD (Seidel, Giovannetti, & Libon, 2012).

VCI broadly defines all forms of dementia with a vascular component, ranging from vascular MCI to VaD. The term can also encompass mixed pathologies, such as concurrent vascular disease and AD pathology. Generally, any individual displaying cognitive dysfunction along with the presence of CVD can be included under the VCI umbrella (Dichgans & Leys, 2017). While the prevalence of CVD in individuals over the age of 65 across Ontario is 37.8 per 1,000 individuals (Ng et al., 2015), it remains difficult to define prevalence of VCI, as inclusion criteria may vary across clinical and research cohorts. Additionally, it is not uncommon for cerebrovascular pathology and acute events to go unnoticed by patients and clinicians until brain imaging is employed. Yet, it is suggested that over 50% of dementia cases present with CVD features (Schneider, Arvanitakis, Bang, & Bennett, 2007). Further, CVD is not uncommon when considering motor/movement disorders (Chondrogianni et al., 2018; Mehanna & Jankovic, 2013).

Overall, it is continuously more accepted that neurodegenerative diseases are not only characterized by neuronal cell loss, but also by cerebrovascular injury. In the context of AD, approximately 40% of patients present with vascular changes and cerebral blood flow is largely reduced in many patients, suggesting a mixed dementia pathology (Lendahl et al., 2019). Further, blood-brain-barrier breakdown is common in AD, as well as in MCI, resulting in greater levels of immune cells in the brain and neuroinflammation, which is known to promote both neurodegenerative disease and CVD pathology. Amyloid-β has also been observed to accumulate within the cerebrovasculature resulting in CAA in more than 80% of AD cases (Lendahl et al., 2019). In contrast, there is much less evidence for a role of CVD in FTD, although white matter changes in the frontal and temporal lobes of FTD patients have been observed (Thal et al., 2015). And while dementia is the first consideration when discussing CVD and neurodegenerative disease co-morbidities, as stated above, motor/movement disorders are not excluded. Not only have CVD risk factors been associated with subsequent diagnosis of PD, but vascular PD has been estimated to account for between 3–12% of parkinsonism cases (Kummer et al., 2019; Mehanna & Jankovic, 2013). Additionally, a few cases of ALS have now been associated with cerebral arteriovenous malformations, and aspects of CVD, such as atherosclerosis and ischemic heart disease, may be associated with greater risk of ALS (Chondrogianni et al., 2018; Kioumourtzoglou et al., 2016).

When considering VaD, large and small cerebral vessel disease are considered causal, rather than a co-pathology. Pathology of VaD is largely defined by diminished cerebral blood flow, resulting in hypoxia, blood brain barrier dysfunction, and ultimately neurotoxic effects and/or amyloid deposition. VaD can be further subclassified into: 1) multi-infarct dementia, 2) strategic infarct dementia, 3) subcortical ischemic dementia, hemorrhagic dementia, 4) hypoperfusion dementia, and 5) other arteriopathies, such as CAA or CADASIL (Dichgans & Leys, 2017).

1.3.3.1   Genetic determinants of CVD and resultant neurodegenerative disease

Compared to the specific neurodegenerative diseases outlined throughout Sections 1.2.1 and 1.2.2, few genetic determinants of VCI have been identified. Although, those that

have been discovered have been found to either contribute to specific VaD diagnoses or to co-pathologies encompassed by the VCI umbrella term.

The most frequent form of monogenic cerebral small vessel disease is CADASIL, caused by heterozygous pathogenic variants in the Notch receptor 3 gene (*NOTCH3*) (Joutel et al., 1996). Although clinically heterogeneous, most CADASIL patients present with recurrent ischemic attacks, migraines with aura, cognitive decline and dementia, and, in some cases, mood disorders. Identification of white matter hyperintensities (WMHs) upon magnetic resonance imaging (MRI) is also a hallmark of the disease (Tikka et al., 2014). The Notch3 protein — a transmembrane receptor in smooth muscle cells — contains 34 extracellular epidermal growth factor (EGF)–like repeats, each containing cysteine residues integral to the protein structure. Variants affecting these cysteine residues are pathogenic for CADASIL, as they result in Notch3 protein misfolding and aggregation in the smooth muscle cells, including those within the brain's vasculature (Joutel et al., 1997; Opherk et al., 2009). Cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy (CARASIL), although much rarer, presents in a very similar manner to CADASIL; however, it is caused by homozygous pathogenic variants in the high-temperature requirement A serine peptidase 1 gene (*HTRA1*) (Hara et al., 2009; Tikka et al., 2014). Interestingly, heterozygous variants within *HTRA1* have also been associated with cerebral small vessel disease, although presentation is less severe than that of CARASIL (Bianchi et al., 2014; Y. Chen et al., 2013; Mendioroz et al., 2010).

In addition to genes contributing to monogenic forms of VaD, there are also genes that increase risk of stroke or other cerebrovascular accidents, such as collagen type IV alpha 1 (*COL4A1*) and collagen type IV alpha 2 (*COL4A2*). While these genes are pleiotropic for multiple conditions, including glaucoma and myopathy, pathogenic variants in *COL4A1* and *COL4A2* increase risk of hemorrhagic stroke. Other vascular conditions, such as pseudoxanthoma elasticum (PXE) caused by autosomal recessive variants in the ATP Binding Cassette Subfamily C Member 6 gene (*ABCC6*), can present with features of CVD as well and are important to continue to investigate in respect to how they may influence neurodegenerative disease pathology (Sunmonu, 2021). In fact, heterozygous

variants in *ABCC6* have also been observed in rare cases of PXE (Hu et al., 2003). Based on the previous associations between CVD and neurodegeneration, it is possible that variants within these genes may also contribute to increased risk of neurodegenerative disease presentation.

There are also genes previously associated with neurodegeneration that have overlapping influence with CVD. The most prominent example of this is *APOE*. Along with its strong association with AD, the E4 allele has been associated with ischemic CVD, as well as increased neuronal damage from CVD events (Laskowitz et al., 1997; McCarron, Delong, & Alberts, 1999). Further, the apo E4 isoform contributes to the accumulation of amyloid-β in the capillaries, resulting in increased rates of CAA in cases of AD, and the risk variant has been associated with increased WMH volumes in AD patients (Love & Miners, 2016). Recent analyses also suggest that *APOE* E4 contributes to blood-brain-barrier breakdown, resulting in cognitive decline independent of AD pathology (Montagne et al., 2020). Interestingly, although the *APOE* E2 allele is typically considered protective for AD, the allele is associated with increased WMH volume in CADASIL patients, as well as increased CAA in parenchymal and meningeal cerebrovasculature (Gesierich et al., 2016; Nelson et al., 2013).

1.3.4   Diagnosis and neurodegenerative disease heterogeneity

As can be gathered from the descriptions of different neurodegenerative diseases above, presentation across the diagnoses is highly heterogeneous. Not only can patients within one disease type, such as AD, present quite differently from one another, but patients with different diagnoses often have overlapping features. Particularly, early in disease presentation misdiagnosis can be quite common.

When considering motor/movement disorders, it is clearly recognized that patients also often present with features of cognitive or behavioural dysfunction. Over 40% of patients with ALS present with features of FTD, and common early features of PD include non-motor phenotypes, such as cognitive impairment, psychiatric disorders, and sleep disruptions (Kalia & Lang, 2015; Karch et al., 2018). In the early stages of disease, when motor features are mild or not yet manifested, these cognitive features can lead to

alternative diagnoses and mislead clinical care (Bicchi et al., 2015). Further, the motor/movement disorders themselves are known to have overlapping features, as parkinsonism has been observed in patients with ALS (Calvo et al., 2019). Although both conditions have clearly defined diagnostic criteria — based on clinical examination, nerve conduction studies, electromyography, and laboratory tests — if early signs of disease are considered abnormal, as is quite common, appropriate tests may not be ordered. Similarly, cognitive and behavioural disorders may present with features of movement dysfunction, or overlapping cognitive features between the differential diagnoses, perhaps to an even greater extent than the motor/movement disorders. Misdiagnosis as AD is particularly common, possibly due to its high prevalence — if a patient is displaying features of dementia that include memory impairment, AD is often first assumed, rather than considering the possibility of FTD with memory dysfunction or that the patient might have experienced silent CVD resulting in VCI (Beber & Chaves, 2013; Doran et al., 2007; Selvackadunco et al., 2019).

Unfortunately, current diagnostic processes are rather slow across neurodegenerative diseases, allowing for further progression of pathology and associated symptoms before potential interventions can be applied. For example, the average ALS patient will wait 9–12 months for a diagnosis following initial onset of symptoms, which is quite striking for a disease with a survival period of only 2–4 years (Hulisz, 2018; Mehta et al., 2017). Additionally, by the time even early symptom onset occurs, irreversible neuropathology changes may have begun 10–20 years before. In the case of PD, by the onset of motor symptoms, it is estimated that more than 50% of dopaminergic neurons have already been lost (Katsuno et al., 2018). As currently available treatments and interventions are only able to marginally slow disease progression and manage symptoms, rather than preventing progression or reversing damage, it is imperative that these are initiated as soon as possible to preserve patient's function ("The Need for Early Detection and Treatment in Alzheimer's Disease," 2016).

Although there are many emerging methods for neurodegenerative disease diagnosis, such as MRI algorithms and biomarker testing of cerebral spinal fluid or plasma (Sancesario & Bernardini, 2018; Shen et al., 2020; X. Y. Zhang, Yang, Lu, Yang, &

Zhang, 2017), definitive diagnosis requires postmortem neuropathological analysis. Genetics may offer another avenue for definitive diagnosis, but current clinical genetic tests for neurodegeneration screen only for variants known to cause the familial forms of the diseases, limiting their utility. Comprehensibly, the basis for this is the lack of understanding regarding the many genes that have now been linked to neurodegenerative diseases and the high occurrence of uninterpretable variation during genetic screening (J. S. Roberts, Patterson, & Uhlmann, 2020). However, it cannot be denied the profound effect wide-spread, accurate genetic diagnostics would have in the field of neurodegenerative disease. Not only would it offer definitive diagnosis for patients, but it would also offer the ability for pre-emptive screening in family members interested in knowing their risk of disease development.

1.3.5   Non-genetic risk factors for neurodegenerative disease

The most well-established risk factor for all neurodegenerative diseases is age. The root of this association is likely multifaceted, as many cellular mechanisms are known to breakdown with age. Suggested pathways include, but are not limited to: genomic instability resulting in DNA damage, epigenetic changes such as increased DNA methylation, mitochondrial dysfunction, cellular senescence, defects in telomere maintenance, and a lack of stem cell rejuvenation (Hou et al., 2019). Further, there are often prominent differences in risk between sexes for neurodegenerative disease development, but the differences tend to be specific to the particular diagnosis. For example, males have a greater risk of ALS, PD, and potentially FTD, while females have a greater risk of AD. Presentation of the diseases can also differ between sexes (Hanamsagar & Bilbo, 2016; Miller & Llibre Guerra, 2019).

There are a large number of environmental risk factors for neurodegenerative diseases that may vary between diagnoses, as well. Across most diagnoses common risk factors include traumatic brain injury, smoking, and chemical or heavy metal exposure (Delamarre & Meissner, 2017; Hulisz, 2018; Killin, Starr, Shiue, & Russ, 2016). Additionally, as described in Section 1.2.3, CVD is an imperative risk factor across all neurogenerative diseases. Of course, it is important to have an understanding of how environmental factors may influence risk of neurodegenerative disease development, as

lifestyle modification may be able to mitigate risk to a certain degree. Yet, the large amount of heritability attributable to neurodegeneration still suggests a dire need for effective therapeutics to slow, cease, and reverse disease progression.

1.4   The Ontario Neurodegenerative Disease Research Initiative

To date, many large consortia have been formed to study the factors contributing to and progression of neurodegenerative diseases, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Parkinson Progression Marker Initiative (PPMI), and Project MinE; yet these studies remain in their silos of specific diagnoses, studying AD, PD, and ALS, respectively. Without consistent assessments, enrollment criteria, and data processing, analyzing the participants across these studies remains difficult and they cannot account for the large amounts of heterogeneity and overlap between neurodegenerative diagnoses.

To gain a better understanding of the occurrence and progression of various neurodegenerative diseases and their respective similarities and differences, the Ontario Neurodegenerative Disease Research Initiative (ONDRI) was formed, funded through the Ontario Brain Institute (OBI) (ondri.ca). The longitudinal, province-wide, observational cohort study aims to characterize multiple neurodegenerative diseases, including: 1) AD; 2) ALS; 3) FTD (bvFTD, CBS, PSP, and PPA [non-fluent and semantic variants]); 4) amnestic MCI; and 5) PD, as well as a cohort of individuals with CVD with or without cognitive impairment (CVD ± CI). ONDRI takes a multimodal approach with multiple assessment platforms and its novel study design provides a longitudinal nature of evaluation, including follow-up and simultaneous analysis of patients across the five phenotypes using the same phenotypic and genotypic markers. Goals of the study included identifying markers that could be applied to early and accurate prediction tools for neurodegeneration, as well as thoroughly analyzing the contribution of cerebral small-vessel pathology to neurodegenerative phenotypes. For the purposes of brevity, only relevant details regarding the general experimental design, enrollment criteria, data management, and assessment platforms will be outlined herein. Full details have been

previously described by Farhan *et al*. and Sunderland *et al*. (S. M. K. Farhan et al., 2017; Sunderland et al., 2020).

Participant recruitment took place between June 2014 and March 2017. Although ONDRI originally aimed to enroll 600 participants across the various disease cohorts, final enrollment was capped at 520 participants, including 41 AD participants, 40 ALS participants, 161 CVD ± CI participants, 53 FTD participants, 85 MCI participants, and 140 PD participants. Upon original study design, the AD and MCI participants were binned into a single AD/MCI cohort, but were later divided, which is reflected in certain analyses in this Dissertation. AD participants included those with either amnestic and non-amnestic presentation and MCI participants may have had either single- or multi-domain amnestic presentation. Importantly, all AD/MCI participant MRI scans were assessed by a research neurologist to confirm absence of significant cerebrovascular pathology that may suggest non-Alzheimer cognitive impairment; if evidence of small vessel disease was observed, participants were instead enrolled into the CVD ± CI cohort. ALS participants may have had possible, probable, or definite ALS. Of the FTD participants, 21 were diagnosed with bvFTD (39.6%), 3 with CBS (5.7%), 16 with PSP (30.2%), 8 with non-fluent PPA (15.1%), and 5 with semantic variant PPA (9.4%). All PD participants were considered idiopathic and may or may not have had cognitive impairment. Full enrollment criteria and enrollment deviations were outlined by Sunderland *et al*. (Sunderland et al., 2020).

All participants were assessed across seven platforms, including: 1) clinical; 2) genomics; 3) neuropsychology; 4) gait and balance; 5) eye tracking; 6) neuroimaging; and 7) retinal imaging. There is also a neuropathology platform that is conducting neuropathological analysis postmortem on select participants and a neuroinformatics platform that acts as a structural backbone providing data processing, data structuring, and statistical support. All platforms' data are deposited in Brain-CODE (braincode.ca), allowing for a secure and central location to collect, store, and share data. The use of this portal is imperative as, in accordance with ONDRI's agreement with the OBI and the Ontario Government, all non-identifying data will be made available to the greater scientific community on

request. It is the hope that this open-science effort will allow for eventual collaboration with other consortium-based efforts to obtain larger sample sizes and to replicate results.

1.4.1 Genetic analysis of neurodegenerative disease patients

Largely, this Dissertation encompasses the work of the genomics platform of ONDRI, specifically focusing on the data produced using the ONDRISeq targeted NGS panel. However, multiple methods were employed within ONDRI to genetically characterize participants, including the aforementioned ONDRISeq panel, the NeuroX array, repeat-primed PCR, TaqMan allelic discrimination assay, and Sanger sequencing.

ONDRISeq was custom-designed for the ONDRI study, and was used to sequence the protein-coding regions of 80 genes that were previously associated with the neurodegenerative diseases encompassed by ONDRI's mandate (**Table 1.2**; **Appendix D**). With this methodology, we are able to harness high-quality NGS data in a focused and efficient manner. The design and validation of the ONDRISeq panel with multiple concordance studies was previously described, for which the ONDRISeq panel was able to identify novel, rare variants of possible clinical significance in 72.2% of 216 ONDRI participants used for panel validation (S.M.K. Farhan et al., 2016). As described in Section 1.1.3.2, WGS evaluates the entire genomic content of an individual, while WES involves sequencing only the protein-coding regions of the genome. Targeted sequencing, in contrast, focuses on specific regions of the genome based on relatively few specific loci linked by common pathological mechanisms or known clinical phenotype. Therefore, targeted sequencing can be an excellent approach when there is already a foundation of candidate genes known to be associated with the disease of interest, as is the case with neurodegenerative disease (Dilliott et al., 2018). Targeting specific regions of the genome also allows for elimination of superfluous and irrelevant genetic variation that can cloud or distract from data interpretation. While WGS and WES both produce high-quality data, the large amount of data can be overwhelming, and at the time of the ONDRI experimental design, cost for WGS and WES remained prohibitive.

To design ONDRISeq, 25 molecular genetics experts, including scientists and clinicians within ONDRI, were consulted to select the 80 genes included on the panel (S.M.K.

Farhan et al., 2016). Not only did the ONDRISeq panel include known monogenic genes, such as *SOD1*, *SNCA*, *APP*, and *NOTCH3*, but the method also targeted genes with less-established genetic associations, such as those within known neurodegeneration pathways and those identified through GWASs (**Table 1.2**). Detailed methodology of the bioinformatics pipeline to process the data from the ONDRISeq panel is included in **Appendix J**. Unfortunately, certain genes of interest had to be excluded from the ONDRISeq panel due to technological limitations. For example, the *GBA* gene — which has a pseudogene located downstream that causes NGS read misalignment — and the *C9orf72* repeat expansion could not be accurately assessed with ONDRISeq.

In addition to the ONDRISeq panel, all participants were genotyped using the NeuroX genotype array. NeuroX consists of both the Illumina Human Exome array, as well as a selection of variants located throughout the genome that have been previously associated with neurological disease. The majority of the variants are rare (~200,000), with fewer common variants (~25,000) (Nalls et al., 2015). Although the data produced from NeuroX were not used for the purposes of this Dissertation, it is important to highlight that the validation of the ONDRISeq panel relied heavily on this method, as it was able to validate the calls of all 122 non-synonymous variants identified as part of the NGS panel's proof of concept analysis (S.M.K. Farhan et al., 2016).

In addition to the ONDRISeq panel and NeuroX array, all ONDRI participants were genotyped for the two defining SNPs of the *APOE* genotype, rs429358(CT):p.Cys130Arg and rs7412(CT):p.Arg176Cys, using TaqMan allelic discrimination assay. Although the SNPs were also captured by the ONDRISeq panel and NeuroX array, the array's results were found to be of low quality, and it was necessary to confirm the allele calls obtained using NGS with another methodology. Further, all participants were Sanger sequenced for the protein-coding regions of the *GBA* gene.

Although 520 participants were enrolled in ONDRI, only 519 were included in the ONDRI genomics platform, as one PD participant's blood sample did not arrive at Robarts Research Institute in London, ON for DNA isolation, and the participant was not

able to provide a replacement sample upon follow up. Demographics of all participants included in the ONDRI genomics analyses are outline in **Table 1.3**.

Table 1.2 Genes included on the ONDRISeq next-generation sequencing gene panel, broken down based on clinical diagnosis and mode of inheritance.

Neurodegenerative disease	Clinical diagnosis	Mode of inheritance	Gene(s)
Alzheimer's disease (AD)	Early-onset AD	ADm	*APP, PSEN1, PRNP, PSEN2*
		AR	*APP*
	Late-onset AD	ACD	*APOE*
		ADm	*APOE, APP, SORL1*
		AR	*APOE, APP*
		RF	*ABCA7, BIN1, CD2AP, CD33, CLU, CR1, MAPT, MS4A4E, MS4A6A, PICALM, PLD3, TREM2*
	Early-onset dementia	ADm	*CSF1R, DNMT1, ITM2B*
	Juvenile-onset dementia	AR	*TYROBP*
	Late-onset dementia	ADm	*CSF1R, DNMT1, ITM2B*
		RF	*TREM2*
Amyotrophic lateral sclerosis (ALS)	Early-onset ALS	ADm	*HNRNPA2B1, OPTN, PFN1, SOD1, VAPB, VCP*
		AR	*OPTN, PNPLA6, SOD1*
	Juvenile-onset ALS	ADm	*SETX, SOD1*
		AR	*ALS2, SETX, SIGMAR1*
		X-Linked Dm	*UBQLN2*
	Late-onset ALS	ADm	*ANG, ARHGEF28, ATXN2, CHMP2B, DAO, DCTN1, FIG4, FUS, HNRNPA1, NEFH, PRPH, SOD1, SQSTM1, TAF15, TARDBP, VAPB*
		AR	*ARHGEF28, FUS, SOD1*
		RF	*CENPV, MAPT, UNC13A*
		X-Linked Dm	*UBQLN2*
	ALS-FTD	ADm	*CHMP2B, FUS, SQSTM1, TARDBP, UNC13A, VCP*
		RF	*MAPT*
		X-Linked Dm	*UBQLN2*
Cerebrovascular disease with or without cognitive impairment (CVD ± CI)[1]	Pseudoxanthoma elasticum	ADm	*ABCC6*
		AR	*ABCC6*
	Hemorrhage	RF	*COL4A1, COL4A2*
	Brain small vessel disease	ADm	*COL4A1, COL4A2*
	CADASIL	ADm	*HTRA1, NOTCH3*
	CARASIL	AR	*HTRA1*
	Aicardi-Goutieres syndrome 5	AR	*SAMHD1*
		ADm	*TREX1*

	Aicardi-Goutieres syndrome 1	AR	*TREX1*
	Vasculopathy, retinal, with cerebral leukoencephalopathy and systemic manifestations	ADm	*TREX1*
Frontotemporal dementia (FTD)	Early-onset FTD	ADm	*TARDBP*
	FTD	ADm	*CHMP2B, FUS, MAPT, OPTN, SIGMAR1, SQSTM1, TARDBP, VCP*
		AR	*OPTN*
		RF	*DCTN1*
		X-Linked Dm	*UBQLN2*
	IBMPFD2	RF	*HNRNPA2B1, VCP*
	Progressive supranuclear palsy	ADm	*MAPT*
	Ubiquitin-positive FTD	ADm	*GRN*
Parkinson's disease (PD)	Early-onset PD	ADm	*HTRA2, LRRK2, PINK1, SNCA, VPS35*
		AR	*FBX07, PANK2, PARK7, PINK1, PLA2G6*
		RF	*ATP13A2, GCH1, GIGYF2, HTRA2, LRRK2*
	Juvenile-onset PD	ADm	*UCHL1*
		AR	*PARK2, UCHL1*
	Late-onset PD	ADm	*DNAJC13, EIF4G1, HTRA2, LRRK2, PARK2, VPS35*
		AR	*PLA2G6*
		RF	*ADH1C, ATP13A2, ATXN2, GAK, GIGYF2, HTRA2, LRRK2, MAPT, MC1R, NR4A2, PM20D1, RAB7L1*
	-	BM	*PARL*

Abbreviations: ACD, autosomal co-dominant; ADm, autosomal dominant; AR, autosomal recessive; BM, biological mechanistic association; CADASIL, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy; CARASIL, cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy; IBMPFD2, inclusion body myopathy with Paget disease of the bone and frontotemporal dementia; RF, risk factor; X-Linked Dm, X-linked dominant.

Table 1.3 Demographics of the enrolled ONDRI participants that were included in the genomics analyses.

Cohort	Cases	Mean age (years ± sd)	Min. age (years)	Max. age (years)	Male:Female
ONDRI	519	68.9 ± 7.7	40.1	87.8	345:174
AD	41	71.8 ± 8.0	54.4	87.8	24:17
ALS	40	62.0 ± 8.7	40.1	77.2	24:16
CVD ± CI	161	69.2 ± 7.4	54.9	85.4	110:51
FTD	53	67.8 ± 7.0	49.7	80.9	34:19
bvFTD	21	65.9 ± 8.8	49.7	80.9	14:7
nfPPA	8	68.4 ± 5.9	59.6	75.1	5:3
PSP	16	69.8 ± 6.0	60.0	80.1	10:6
MCI	85	70.6 ± 8.3	53.4	87.2	45:40
PD	139	67.9 ± 6.3	55.1	85.9	108:31

Demographics of the FTD subtypes corticobasal syndrome and semantic variant primary progressive aphasia cohorts are not displayed due to the small sample sizes of the groups (n ≤ 5), as they may compromise confidentiality.

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; bvFTD, behavioural variant frontotemporal dementia; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; Max., maximum; MCI, mild cognitive impairment; Min., minimum; nfPPA, non-fluent primary progressive aphasia; PD, Parkinson's disease; PSP, progressive supranuclear palsy; sd, standard deviation.

1.5   Thesis outline

1.5.1   Overall aim and objectives

Genetic factors increase one's risk of developing neurodegeneration considerably and influence disease features. Yet, heretofore, many studies have aimed to identify the full spectrum of genetic variation contributing to each individual neurodegenerative disease, and they have failed to consider a holistic approach to studying neurodegeneration by not taking into account the vast amount of clinical and pathological overlap between the various diagnoses. It is imperative that we are able to gain a full understanding of the genetic determinants contributing not only to risk of the diseases themselves, but to the vast amount of heterogeneity observed within diagnoses, to gain a full understanding of how genetic determinants influence all neurodegenerative diseases and their presentation. By leveraging the unique study design of ONDRI and its rich dataset, efforts can be made to elucidate the overlapping genetic determinants between the various neurodegenerative diagnoses. Further, by focusing on the intrinsic goals of ONDRI we may establish an understanding of the genetic contribution to cerebral small-vessel pathology and its influence on neurodegenerative phenotypes.

The aim of my PhD research was to genetically characterize the full ONDRI cohort, and leverage the data to gain a greater understanding of the genetic overlap between the various neurodegenerative diseases, including AD, ALS, FTD, MCI, and PD, as well as CVD, largely using the data generated with the ONDRISeq panel.

The first objective of my work was to identify associations between common genetic variants of high phenotypic impact and the neurodegenerative diseases encompassed by ONDRI's mandate. Although the *APOE* genotype is a well-established risk factor for the development of AD, it was imperative to replicate the association within the ONDRI cohort to contribute to validation of the AD diagnoses, as well as assess whether the genotype was contributing to any of the other disease cohorts. *MAPT* haplotype of each ONDRI participant was also determined to assess contributions to disease presentation. The details of these analyses, which compared variant frequencies in each ONDRI cohort to a cognitively normal, elderly control cohort, are outlined in <u>Chapter 2</u>. The *APOE*

genotype has also controversially been associated with deficits in various cognitive domains within different individual neurodegenerative disease cohorts; therefore, I leveraged ONDRI's robust neuropsychology dataset to assess the contribution of the *APOE* genotype to cognitive deficits in five domains, across the diagnoses. The analysis and its results are included in Chapter 3.

The second objective of my PhD research was to assess the contribution of a spectrum of rare genetic variants to the neurodegenerative diseases encompassed by ONDRI's mandate. As previously described, there exists a gap in the literature accounting for the potential overlapping impact of genetic factors known to contribute to specific neurodegenerative diagnoses within other neurodegenerative phenotypes. To begin characterizing this heterogeneity, I identified all rare SNVs within the ONDRI samples in the 80 genes previously associated with neurodegenerative disease using the NGS targeted sequencing panel, ONDRISeq. I then performed rare variant association analyses on both binned gene sets, as well as individual genes to identify signals of variant enrichment and elucidate potential overlapping genetic factors across the neurodegenerative diagnoses, which is presented in Chapter 4. In Chapter 5, I leveraged the ONDRISeq NGS data to identify rare, large-scale CNVs within the ONDRI cohort using a recently developed depth-of-coverage approach. Finally, to account for variation that could not be captured using the ONDRISeq panel due to technical limitations, Chapter 6 outlines the Sanger sequencing analysis of all exons of *GBA* in the ONDRI participants and the identification of rare, nonsynonymous variants of interest within the gene and associations with the individual neurodegenerative disease cohorts.

Finally, the third objective of my work was to identify associations between rare genetic variants within ONDRI and the presentation and features of the neurodegenerative disease patients. Along with the sequencing of *GBA* presented in Chapter 6, I also outline a multivariate multiple regression analysis to assess the effects of *GBA* rare variants on age of onset, generalized cognition, and motor impairment across all neurodegenerative disease and CVD cohorts. Further, Chapter 7 highlights a novel association between *NOTCH3* rare variants and cerebrovascular disease burden in PD patients carrying variants in the gene. Although only focusing on two genes for this objective, the analyses

highlight a proof of concept that rare variants of possible moderate phenotypic impact may influence the presentation of neurodegenerative diseases not typically associated with the gene of interest, thereby accounting for some of the phenotypic heterogeneity observed in these disease cohorts.

1.5.2   Hypothesis

Susceptibility to neurodegenerative diseases and their heterogenous intermediate phenotypes display a complex genetic landscape with a combination of influence from rare and common genetic variation, including small-scale SNVs and large-scale CNVs.

1.5.3   Summary

This Dissertation describes my research elucidating the genetic determinants of patients with various neurodegenerative diseases, as well as patients with cerebrovascular disease. Not only does my work aim to characterize the genetic factors contributing to disease presentation overall, but it begins the important investigation into how genetics may contribute to the heterogenous features of neurodegenerative disease. Importantly, the research detailed herein also adds important data to ONDRI, which will become available to the greater scientific community in the coming years. To achieve these goals, I utilized the data generated using the ONDRISeq targeted NGS panel, which covers 80 genes previously associated with neurodegenerative and cerebrovascular phenotypes, along with other methodologies, such as Sanger sequencing, to account for the limitations of the NGS panel. Not only did my work validate previously identified genetic associations with the neurodegenerative diseases under study, such as the association between *APOE* E4 and AD, but it identified novel genetic associations between neurodegenerative disease associated genes and other neurodegenerative disease cohorts and patterns in clinical presentation. Importantly, the work presented highlights the need for a greater understanding of the complex genetic architecture contributing to neurodegeneration.

1.6   References

2020 Alzheimer's disease facts and figures. (2020). *Alzheimers Dement*. doi:10.1002/alz.12068

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Agrawal, M., & Biswas, A. (2015). Molecular diagnostics of neurodegenerative disorders. *Front Mol Biosci, 2*, 54. doi:10.3389/fmolb.2015.00054

Anderson, N. D. (2019). State of the science on mild cognitive impairment (MCI). *CNS Spectr, 24*(1), 78-87. doi:10.1017/S1092852918001347

Anna, A., & Monika, G. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet, 59*(3), 253-268. doi:10.1007/s13353-018-0444-7

Asimit, J., & Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annu Rev Genet, 44*, 293-308. doi:10.1146/annurev-genet-102209-163421

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Beber, B. C., & Chaves, M. L. F. (2013). Evaluation of patients with behavioral and cognitive complaints: misdiagnosis in frontotemporal dementia and Alzheimer's disease. *Dement Neuropsychol, 7*(1), 60-65. doi:10.1590/S1980-57642013DN70100010

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Arch Dis Child Educ Pract Ed, 98*(6), 236-238. doi:10.1136/archdischild-2013-304340

Bellou, E., Stevenson-Hoare, J., & Escott-Price, V. (2020). Polygenic risk and pleiotropy in neurodegenerative diseases. *Neurobiol Dis, 142*, 104953. doi:10.1016/j.nbd.2020.104953

Bennion Callister, J., & Pickering-Brown, S. M. (2014). Pathogenesis/genetics of frontotemporal dementia and how it relates to ALS. *Exp Neurol, 262 Pt B*, 84-90. doi:10.1016/j.expneurol.2014.06.001

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A, 74(8), 3171-3175. doi:10.1073/pnas.74.8.3171

Bianchi, S., Di Palma, C., Gallus, G. N., Taglia, I., Poggiani, A., Rosini, F., . . . Federico, A. (2014). Two novel HTRA1 mutations in a European CARASIL patient. *Neurology, 82*(10), 898-900. doi:10.1212/WNL.0000000000000202

Bicchi, I., Emiliani, C., Vescovi, A., & Martino, S. (2015). The Big Bluff of Amyotrophic Lateral Sclerosis Diagnosis: The Role of Neurodegenerative Disease Mimics. *Neurodegener Dis, 15*(6), 313-321. doi:10.1159/000435917

Bocchetta, M., Mega, A., Bernardi, L., Di Maria, E., Benussi, L., Binetti, G., . . . Bruni, A. C. (2016). Genetic Counseling and Testing for Alzheimer's Disease and Frontotemporal Lobar Degeneration: An Italian Consensus Protocol. *J Alzheimers Dis, 51*(1), 277-291. doi:10.3233/JAD-150849

Bosco, D. A., Morfini, G., Karabacak, N. M., Song, Y., Gros-Louis, F., Pasinelli, P., . . . Brown, R. H., Jr. (2010). Wild-type and mutant SOD1 share an aberrant conformation and a common pathogenic pathway in ALS. *Nat Neurosci, 13*(11), 1396-1403. doi:10.1038/nn.2660

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell, 169*(7), 1177-1186. doi:10.1016/j.cell.2017.05.038

Bush, W. S., & Haines, J. (2010). Overview of linkage analysis in complex traits. *Curr Protoc Hum Genet, Chapter 1*, Unit 1 9 1-18. doi:10.1002/0471142905.hg0109s64

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput Biol, 8*(12), e1002822. doi:10.1371/journal.pcbi.1002822

Cacace, R., Sleegers, K., & Van Broeckhoven, C. (2016). Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimers Dement, 12*(6), 733-748. doi:10.1016/j.jalz.2016.01.012

Caffrey, T. M., & Wade-Martins, R. (2007). Functional MAPT haplotypes: bridging the gap between genotype and neuropathology. *Neurobiol Dis, 27*(1), 1-10. doi:10.1016/j.nbd.2007.04.006

Calvo, A., Chio, A., Pagani, M., Cammarosano, S., Dematteis, F., Moglia, C., . . . Rizzone, M. G. (2019). Parkinsonian traits in amyotrophic lateral sclerosis (ALS): a prospective population-based study. *J Neurol, 266*(7), 1633-1642. doi:10.1007/s00415-019-09305-0

Cardon, L. R., & Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet, 361*(9357), 598-604. doi:10.1016/S0140-6736(03)12520-2

Chaisson, M. J., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., . . . Eichler, E. E. (2015). Resolving the complexity of the human genome

using single-molecule sequencing. *Nature, 517*(7536), 608-611. doi:10.1038/nature13907

Chakravarti, A. (1999). Population genetics--making sense out of sequence. *Nat Genet, 21*(1 Suppl), 56-60. doi:10.1038/4482

Chambers, L. W., Bancej, C., & McDowell, I. (2016). *Prevalence and montary costs of dementia in Canada*. In A. S. o. Canada (Ed.), *Population Health Expert Panel*.

Chapter 3: Mapping Connections: An understanding of neurological conditions in Canada – Scope (prevalence and incidence). (2014). Retrieved from https://www.canada.ca/en/public-health/services/reports-publications/mapping-connections-understanding-neurological-conditions/mapping-connections-understanding-neurological-conditions-canada-13.html

Chen, Y., He, Z., Meng, S., Li, L., Yang, H., & Zhang, X. (2013). A novel mutation of the high-temperature requirement A serine peptidase 1 (HTRA1) gene in a Chinese family with cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy (CARASIL). *J Int Med Res, 41*(5), 1445-1455. doi:10.1177/0300060513480926

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One, 8*(4), e62856. doi:10.1371/journal.pone.0062856

Choi, S. W., Mak, T. S., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc, 15*(9), 2759-2772. doi:10.1038/s41596-020-0353-1

Chondrogianni, M., Bregianni, M., Frantzeskaki, F., Giamarellos-Bourboulis, E., Anagnostou, E., Kararizou, E., . . . Stefanis, L. (2018). Three new case reports of Arteriovenous malformation-related Amyotrophic Lateral Sclerosis. *J Neurol Sci, 393*, 58-62. doi:10.1016/j.jns.2018.08.011

Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell, 12(1), 1-8. doi:10.1016/0092-8674(77)90180-5

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., & Pritchard, J. K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet, 38*(1), 75-81. doi:10.1038/ng1697

Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., & Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet, 132*(10), 1077-1130. doi:10.1007/s00439-013-1331-2

Coughlin, C. R., 2nd, Scharer, G. H., & Shaikh, T. H. (2012). Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome Med, 4*(10), 80. doi:10.1186/gm381

Crouch, D. J. M., & Bodmer, W. F. (2020). Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc Natl Acad Sci U S A, 117*(32), 18924-18933. doi:10.1073/pnas.2005634117

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., . . . Rademakers, R. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron, 72*(2), 245-256. doi:10.1016/j.neuron.2011.09.011

Delamarre, A., & Meissner, W. G. (2017). Epidemiology, environmental risk factors and genetics of Parkinson's disease. *Presse Med, 46*(2 Pt 1), 175-181. doi:10.1016/j.lpm.2017.01.001

Devenney, E. M., Ahmed, R. M., & Hodges, J. R. (2019). Frontotemporal dementia. *Handb Clin Neurol, 167*, 279-299. doi:10.1016/B978-0-12-804766-8.00015-7

Dichgans, M., & Leys, D. (2017). Vascular Cognitive Impairment. *Circ Res, 120*(3), 573-591. doi:10.1161/CIRCRESAHA.116.308426

Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2018). Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J Vis Exp*(134). doi:10.3791/57266

Dolgin, E. (2009). Human genomics: The genome finishers. *Nature, 462*(7275), 843-845. doi:10.1038/462843a

Doran, M., du Plessis, D. G., Ghadiali, E. J., Mann, D. M., Pickering-Brown, S., & Larner, A. J. (2007). Familial early-onset dementia with tau intron 10 + 16 mutation with clinical features similar to those of Alzheimer disease. *Arch Neurol, 64*(10), 1535-1539. doi:10.1001/archneur.64.10.1535

Dron, J. S., & Hegele, R. A. (2018). Polygenic influences on dyslipidemias. *Curr Opin Lipidol, 29*(2), 133-143. doi:10.1097/MOL.0000000000000482

Dudgeon, S. (2010). *Rising tide: the impact of dementia on Canadian society*. Retrieved from

Dugger, B. N., Adler, C. H., Shill, H. A., Caviness, J., Jacobson, S., Driver-Dunckley, E., . . . Arizona Parkinson's Disease, C. (2014). Concomitant pathologies among a spectrum of parkinsonian disorders. *Parkinsonism Relat Disord, 20*(5), 525-529. doi:10.1016/j.parkreldis.2014.02.012

Dugger, B. N., & Dickson, D. W. (2017). Pathology of Neurodegenerative Diseases. *Cold Spring Harb Perspect Biol, 9*(7). doi:10.1101/cshperspect.a028035

Duraes, F., Pinto, M., & Sousa, E. (2018). Old Drugs as New Treatments for Neurodegenerative Diseases. *Pharmaceuticals (Basel), 11*(2). doi:10.3390/ph11020044

Elcoroaristizabal Martin, X., Fernandez Martinez, M., Galdos Alcelay, L., Molano Salazar, A., Bereincua Gandarias, R., Ingles Borda, S., . . . de Pancorbo, M. M. (2011). Progression from amnesic mild cognitive impairment to Alzheimer's disease: ESR1 and ESR2 polymorphisms and APOE gene. *Dement Geriatr Cogn Disord, 32*(5), 332-341. doi:10.1159/000335541

Elston, R. C. (1998). Methods of linkage analysis--and the assumptions underlying them [see comment]. *Am J Hum Genet, 63*(4), 931-934. doi:10.1086/302073

Escott-Price, V., Myers, A., Huentelman, M., Shoai, M., & Hardy, J. (2019). Polygenic Risk Score Analysis of Alzheimer's Disease in Cases without APOE4 or APOE2 Alleles. *J Prev Alzheimers Dis, 6*(1), 16-19. doi:10.14283/jpad.2018.46

Escott-Price, V., Myers, A. J., Huentelman, M., & Hardy, J. (2017). Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann Neurol, 82*(2), 311-314. doi:10.1002/ana.24999

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., . . . Williams, J. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain, 138*(Pt 12), 3673-3684. doi:10.1093/brain/awv268

Fan, H., & Chu, J. Y. (2007). A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics, 5*(1), 7-14. doi:10.1016/S1672-0229(07)60009-6

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., . . . Strong, M. J. (2017). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci, 44*(2), 196-202. doi:10.1017/cjn.2016.415

Farhan, S. M. K., Dilliott, A. A., Ghani, G., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*(16032), 1-11.

Farhan, S. M. K., Howrigan, D. P., Abbott, L. E., Klim, J. R., Topp, S. D., Byrnes, A. E., . . . Neale, B. M. (2019). Exome sequencing in amyotrophic lateral sclerosis implicates a novel gene, DNAJC7, encoding a heat-shock protein. *Nat Neurosci, 22*(12), 1966-1974. doi:10.1038/s41593-019-0530-0

Fenoglio, C., Scarpini, E., Serpente, M., & Galimberti, D. (2018). Role of Genetics and Epigenetics in the Pathogenesis of Alzheimer's Disease and Frontotemporal Dementia. *J Alzheimers Dis, 62*(3), 913-932. doi:10.3233/JAD-170702

Ferrari, R., Hernandez, D. G., Nalls, M. A., Rohrer, J. D., Ramasamy, A., Kwok, J. B., . . . Momeni, P. (2014). Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurol, 13*(7), 686-699. doi:10.1016/S1474-4422(14)70065-1

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nat Rev Genet, 7*(2), 85-97. doi:10.1038/nrg1767

Feuk, L., Marshall, C. R., Wintle, R. F., & Scherer, S. W. (2006). Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet, 15 Spec No 1*, R57-66. doi:10.1093/hmg/ddl057

Fleisher, A. S., Sowell, B. B., Taylor, C., Gamst, A. C., Petersen, R. C., & Thal, L. J. (2007). Clinical predictors of progression to Alzheimer disease in amnestic mild cognitive impairment. *Neurology, 68*(19), 1588-1595. doi:10.1212/01.wnl.0000258542.58725.4c

Gammon, K. (2014). Neurodegenerative disease: brain windfall. *Nature, 515*(7526), 299-300. doi:10.1038/nj7526-299a

Gegg, M. E., & Schapira, A. H. V. (2018). The role of glucocerebrosidase in Parkinson disease pathogenesis. *FEBS J, 285*(19), 3591-3603. doi:10.1111/febs.14393

Georgadaki, K., Khoury, N., Spandidos, D. A., & Zoumpourlis, V. (2016). The molecular basis of fertilization (Review). *Int J Mol Med, 38*(4), 979-986. doi:10.3892/ijmm.2016.2723

Gesierich, B., Opherk, C., Rosand, J., Gonik, M., Malik, R., Jouvent, E., . . . Ewers, M. (2016). APOE varepsilon2 is associated with white matter hyperintensity volume in CADASIL. *J Cereb Blood Flow Metab, 36*(1), 199-203. doi:10.1038/jcbfm.2015.85

Giau, V. V., Bagyinszky, E., & An, S. S. A. (2019). Potential Fluid Biomarkers for the Diagnosis of Mild Cognitive Impairment. *Int J Mol Sci, 20*(17). doi:10.3390/ijms20174149

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat Rev Genet, 13*(2), 135-145. doi:10.1038/nrg3118

Goldman, S. M., Marek, K., Ottman, R., Meng, C., Comyns, K., Chan, P., . . . Tanner, C. M. (2019). Concordance for Parkinson's disease in twins: A 20-year update. *Ann Neurol, 85*(4), 600-605. doi:10.1002/ana.25441

Goldmann, J. M., Veltman, J. A., & Gilissen, C. (2019). De Novo Mutations Reflect Development and Aging of the Human Germline. *Trends Genet, 35*(11), 828-839. doi:10.1016/j.tig.2019.08.005

Gomperts, S. N. (2016). Lewy Body Dementias: Dementia With Lewy Bodies and Parkinson Disease Dementia. *Continuum (Minneap Minn), 22*(2 Dementia), 435-463. doi:10.1212/CON.0000000000000309

Grad, L. I., Rouleau, G. A., Ravits, J., & Cashman, N. R. (2017). Clinical Spectrum of Amyotrophic Lateral Sclerosis (ALS). *Cold Spring Harb Perspect Med, 7*(8). doi:10.1101/cshperspect.a024117

Greaves, C. V., & Rohrer, J. D. (2019). An update on genetic frontotemporal dementia. *J Neurol, 266*(8), 2075-2086. doi:10.1007/s00415-019-09363-4

Guerreiro, R., Sassi, C., Gibbs, J. R., Edsall, C., Hernandez, D., Brown, K., . . . Bras, J. (2018). A comprehensive assessment of benign genetic variability for neurodegenerative disorders. *bioRxiv*. doi:10.1101/270686

Hagenaars, S. P., Radakovic, R., Crockford, C., Fawns-Ritchie, C., International, F. T. D. G. C., Harris, S. E., . . . Deary, I. J. (2018). Genetic risk for neurodegenerative disorders, and its overlap with cognitive ability and physical function. *PLoS One, 13*(6), e0198187. doi:10.1371/journal.pone.0198187

Halperin, A., Elstein, D., & Zimran, A. (2006). Increased incidence of Parkinson disease among relatives of patients with Gaucher disease. *Blood Cells Mol Dis, 36*(3), 426-428. doi:10.1016/j.bcmd.2006.02.004

Hanamsagar, R., & Bilbo, S. D. (2016). Sex differences in neurodevelopmental and neurodegenerative disorders: Focus on microglial function and neuroinflammation during development. *J Steroid Biochem Mol Biol, 160*, 127-133. doi:10.1016/j.jsbmb.2015.09.039

Hannan, A. J. (2018). Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet, 19*(5), 286-298. doi:10.1038/nrg.2017.115

Hara, K., Shiga, A., Fukutake, T., Nozaki, H., Miyashita, A., Yokoseki, A., . . . Onodera, O. (2009). Association of HTRA1 mutations and familial ischemic cerebral small-vessel disease. *N Engl J Med, 360*(17), 1729-1739. doi:10.1056/NEJMoa0801560

Hassold, T., & Hunt, P. (2001). To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet, 2*(4), 280-291. doi:10.1038/35066065

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics, 107*(1), 1-8. doi:10.1016/j.ygeno.2015.11.003

Heffernan, A. L., Chidgey, C., Peng, P., Masters, C. L., & Roberts, B. R. (2016). The Neurobiology and Age-Related Prevalence of the epsilon4 Allele of Apolipoprotein E in Alzheimer's Disease Cohorts. *J Mol Neurosci, 60*(3), 316-324. doi:10.1007/s12031-016-0804-x

Hellwege, J. N., Keaton, J. M., Giri, A., Gao, X., Velez Edwards, D. R., & Edwards, T. L. (2017). Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet, 95*, 1 22 21-21 22 23. doi:10.1002/cphg.48

Hemani, G., Bowden, J., & Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet, 27*(R2), R195-R208. doi:10.1093/hmg/ddy163

Hogan, D. B., Jette, N., Fiest, K. M., Roberts, J. I., Pearson, D., Smith, E. E., . . . Maxwell, C. J. (2016). The Prevalence and Incidence of Frontotemporal Dementia: a Systematic Review. *Can J Neurol Sci, 43 Suppl 1*, S96-S109. doi:10.1017/cjn.2016.25

Hou, Y., Dan, X., Babbar, M., Wei, Y., Hasselbalch, S. G., Croteau, D. L., & Bohr, V. A. (2019). Ageing as a risk factor for neurodegenerative disease. *Nat Rev Neurol, 15*(10), 565-581. doi:10.1038/s41582-019-0244-7

Hu, X., Peek, R., Plomp, A., ten Brink, J., Scheffer, G., van Soest, S., . . . Bergen, A. A. (2003). Analysis of the frequent R1141X mutation in the ABCC6 gene in pseudoxanthoma elasticum. *Invest Ophthalmol Vis Sci, 44*(5), 1824-1829. doi:10.1167/iovs.02-0981

Hulisz, D. (2018). Amyotrophic lateral sclerosis: disease state overview. *Am J Manag Care, 24*(15 Suppl), S320-S326. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30207670

Hurles, M. E., Dermitzakis, E. T., & Tyler-Smith, C. (2008). The functional impact of structural variation in humans. *Trends Genet, 24*(5), 238-245. doi:10.1016/j.tig.2008.03.001

Hutton, M., Lendon, C. L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., . . . Heutink, P. (1998). Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature, 393*(6686), 702-705. doi:10.1038/31508

Iacocca, M. A., Wang, J., Dron, J. S., Robinson, J. F., McIntyre, A. D., Cao, H., & Hegele, R. A. (2017). Use of next-generation sequencing to detect LDLR gene copy number variation in familial hypercholesterolemia. *J Lipid Res, 58*(11), 2202-2209. doi:10.1194/jlr.D079301

International HapMap, C. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi:10.1038/nature04226

International Human Genome Sequencing, C. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931-945. doi:10.1038/nature03001

Jost, B. C., & Grossberg, G. T. (1995). The natural history of Alzheimer's disease: a brain bank study. *J Am Geriatr Soc, 43*(11), 1248-1255. doi:10.1111/j.1532-5415.1995.tb07401.x

Joutel, A., Corpechot, C., Ducros, A., Vahedi, K., Chabriat, H., Mouton, P., . . . Tournier-Lasserve, E. (1996). Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature, 383*(6602), 707-710. doi:10.1038/383707a0

Joutel, A., Vahedi, K., Corpechot, C., Troesch, A., Chabriat, H., Vayssiere, C., . . . Tournier-Lasserve, E. (1997). Strong clustering and stereotyped nature of Notch3 mutations in CADASIL patients. *Lancet, 350*(9090), 1511-1515. doi:10.1016/S0140-6736(97)08083-5

Kalia, L. V., & Lang, A. E. (2015). Parkinson's disease. *Lancet, 386*(9996), 896-912. doi:10.1016/S0140-6736(14)61393-3

Karch, C. M., Wen, N., Fan, C. C., Yokoyama, J. S., Kouri, N., Ross, O. A., . . . International Parkinson's Disease Genomics, C. (2018). Selective Genetic Overlap Between Amyotrophic Lateral Sclerosis and Diseases of the Frontotemporal Dementia Spectrum. *JAMA Neurol, 75*(7), 860-875. doi:10.1001/jamaneurol.2018.0372

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434-443. doi:10.1038/s41586-020-2308-7

Katsuno, M., Sahashi, K., Iguchi, Y., & Hashizume, A. (2018). Preclinical progression of neurodegenerative diseases. *Nagoya J Med Sci, 80*(3), 289-298. doi:10.18999/nagjms.80.3.289

Keller, M. F., Saad, M., Bras, J., Bettella, F., Nicolaou, N., Simon-Sanchez, J., . . . Wellcome Trust Case Control, C. (2012). Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Hum Mol Genet, 21*(22), 4996-5009. doi:10.1093/hmg/dds335

Killin, L. O., Starr, J. M., Shiue, I. J., & Russ, T. C. (2016). Environmental risk factors for dementia: a systematic review. *BMC Geriatr, 16*(1), 175. doi:10.1186/s12877-016-0342-y

Kioumourtzoglou, M. A., Seals, R. M., Gredal, O., Mittleman, M. A., Hansen, J., & Weisskopf, M. G. (2016). Cardiovascular disease and diagnosis of amyotrophic lateral sclerosis: A population based study. *Amyotroph Lateral Scler Frontotemporal Degener, 17*(7-8), 548-554. doi:10.1080/21678421.2016.1208247

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi:10.1038/ng.2892

Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y., & Chia, K. S. (2010). The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet, 55*(7), 403-415. doi:10.1038/jhg.2010.55

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc, 4*(7), 1073-1081. doi:10.1038/nprot.2009.86

Kummer, B. R., Diaz, I., Wu, X., Aaroe, A. E., Chen, M. L., Iadecola, C., . . . Navi, B. B. (2019). Associations between cerebrovascular risk factors and parkinson disease. *Ann Neurol, 86*(4), 572-581. doi:10.1002/ana.25564

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., . . . Environmental Risk for Alzheimer's Disease, C. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet, 51*(3), 414-430. doi:10.1038/s41588-019-0358-2

Kuntzer, J., Eggle, D., Klostermann, S., & Burtscher, H. (2010). Human variation databases. *Database (Oxford), 2010*, baq015. doi:10.1093/database/baq015

LaDuca, H., Farwell, K. D., Vuong, H., Lu, H. M., Mu, W., Shahmirzadi, L., . . . Chao, E. C. (2017). Exome sequencing covers >98% of mutations identified on targeted next generation sequencing panels. *PLoS One, 12*(2), e0170843. doi:10.1371/journal.pone.0170843

Lander, E. S. (1996). The new genomics: global views of biology. *Science, 274*(5287), 536-539. doi:10.1126/science.274.5287.536

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res, 42*(Database issue), D980-985. doi:10.1093/nar/gkt1113

Lane, C. A., Hardy, J., & Schott, J. M. (2018). Alzheimer's disease. *Eur J Neurol, 25*(1), 59-70. doi:10.1111/ene.13439

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. *Cell, 177*(1), 70-84. doi:10.1016/j.cell.2019.02.032

Laskowitz, D. T., Sheng, H., Bart, R. D., Joyner, K. A., Roses, A. D., & Warner, D. S. (1997). Apolipoprotein E-deficient mice have increased susceptibility to focal cerebral ischemia. *J Cereb Blood Flow Metab, 17*(7), 753-758. doi:10.1097/00004647-199707000-00005

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., . . . Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet, 91*(2), 224-237. doi:10.1016/j.ajhg.2012.06.007

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics, 13*(4), 762-775. doi:10.1093/biostatistics/kxs014

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. doi:10.1038/nature19057

Lendahl, U., Nilsson, P., & Betsholtz, C. (2019). Emerging links between cerebrovascular and neurodegenerative diseases-a special role for pericytes. *EMBO Rep, 20*(11), e48070. doi:10.15252/embr.201948070

Lewontin, R.C., & Kojima, K. (1960). THE EVOLUTIONARY DYNAMICS OF COMPLEX POLYMORPHISMS. *Evolution, 14*: 458-472.

Liang, K. C., Tseng, J. T., Tsai, S. J., & Sun, H. S. (2015). Characterization and distribution of repetitive elements in association with genes in the human genome. *Comput Biol Chem, 57*, 29-38. doi:10.1016/j.compbiolchem.2015.02.007

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., & Guo, J. T. (2017). Effects of short indels on protein structure and function in human genomes. *Sci Rep, 7*(1), 9313. doi:10.1038/s41598-017-09287-x

Liu, Q., Zhang, P., Wang, D., Gu, W., & Wang, K. (2017). Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med, 9*(1), 65. doi:10.1186/s13073-017-0456-7

Love, S., & Miners, J. S. (2016). Cerebrovascular disease in ageing and Alzheimer's disease. *Acta Neuropathol, 131*(5), 645-658. doi:10.1007/s00401-015-1522-0

Lunati, A., Lesage, S., & Brice, A. (2018). The genetic landscape of Parkinson's disease. *Rev Neurol (Paris), 174*(9), 628-643. doi:10.1016/j.neurol.2018.08.004

Mani, A. (2017). Pathogenicity of De Novo Rare Variants: Challenges and Opportunities. *Circ Cardiovasc Genet, 10*(6). doi:10.1161/CIRCGENETICS.117.002013

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi:10.1038/nature08494

Manzoni, C., & Ferrari, R. (2021). Mendelian and Sporadic FTD: Disease Risk and Avenues from Genetics to Disease Pathways Through In Silico Modelling. *Adv Exp Med Biol, 1281*, 283-296. doi:10.1007/978-3-030-51140-1_17

Maraganore, D. M., de Andrade, M., Elbaz, A., Farrer, M. J., Ioannidis, J. P., Kruger, R., . . . Genetic Epidemiology of Parkinson's Disease, C. (2006). Collaborative analysis of alpha-synuclein gene promoter variability and Parkinson disease. *JAMA, 296*(6), 661-670. doi:10.1001/jama.296.6.661

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380. doi:10.1038/nature03959

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet, 51*(4), 584-591. doi:10.1038/s41588-019-0379-x

McCarron, M. O., Delong, D., & Alberts, M. J. (1999). APOE genotype as a risk factor for ischemic cerebrovascular disease: a meta-analysis. *Neurology, 53*(6), 1308-1311. doi:10.1212/wnl.53.6.1308

Mehanna, R., & Jankovic, J. (2013). Movement disorders in cerebrovascular disease. *Lancet Neurol, 12*(6), 597-608. doi:10.1016/S1474-4422(13)70057-7

Mehta, P., Horton, D. K., Kasarskis, E. J., Tessaro, E., Eisenberg, M. S., Laird, S., & Iskander, J. (2017). CDC Grand Rounds: National Amyotrophic Lateral Sclerosis (ALS) Registry Impact, Challenges, and Future Directions. *MMWR Morb Mortal Wkly Rep, 66*(50), 1379-1382. doi:10.15585/mmwr.mm6650a3

Mejzini, R., Flynn, L. L., Pitout, I. L., Fletcher, S., Wilton, S. D., & Akkari, P. A. (2019). ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Front Neurosci, 13*, 1310. doi:10.3389/fnins.2019.01310

Mendel, J.G. (1865). Versuche über Pflanzenhybriden. In: Verhandlungen des Naturforschenden Vereins Brünn. Band 4, S. 3-47.

Mendioroz, M., Fernandez-Cadenas, I., Del Rio-Espinola, A., Rovira, A., Sole, E., Fernandez-Figueras, M. T., . . . Montaner, J. (2010). A missense HTRA1 mutation expands CARASIL syndrome to the Caucasian population. *Neurology, 75*(22), 2033-2035. doi:10.1212/WNL.0b013e3181ff96ac

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., . . . Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature, 585*(7823), 79-84. doi:10.1038/s41586-020-2547-7

Miller, B., & Llibre Guerra, J. J. (2019). Frontotemporal dementia. *Handb Clin Neurol, 165*, 33-45. doi:10.1016/B978-0-444-64012-3.00003-4

Mitchell, A. J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia--meta-analysis of 41 robust inception cohort studies. *Acta Psychiatr Scand, 119*(4), 252-265. doi:10.1111/j.1600-0447.2008.01326.x

Moldovan, V., Banescu, C., & Dobreanu, M. (2020). Molecular diagnosis methods in familial hypercholesterolemia. *Anatol J Cardiol, 23*(3), 120-127. doi:10.14744/AnatolJCardiol.2019.95038

Montagne, A., Nation, D. A., Sagare, A. P., Barisano, G., Sweeney, M. D., Chakhoyan, A., . . . Zlokovic, B. V. (2020). APOE4 leads to blood-brain barrier dysfunction predicting cognitive decline. *Nature, 581*(7806), 71-76. doi:10.1038/s41586-020-2247-3

Moraes, F., & Goes, A. (2016). A decade of human genome project conclusion: Scientific diffusion about our genome knowledge. *Biochem Mol Biol Educ, 44*(3), 215-223. doi:10.1002/bmb.20952

Morin, S. J., Eccles, J., Iturriaga, A., & Zimmerman, R. S. (2017). Translocations, inversions and other chromosome rearrangements. *Fertil Steril, 107*(1), 19-26. doi:10.1016/j.fertnstert.2016.10.013

Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am J Hum Genet, 7*(3), 277-318. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/13258560

Munoz, S. S., Garner, B., & Ooi, L. (2019). Understanding the Role of ApoE Fragments in Alzheimer's Disease. *Neurochem Res, 44*(6), 1297-1305. doi:10.1007/s11064-018-2629-1

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., . . . International Parkinson's Disease Genomics, C. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol, 18*(12), 1091-1102. doi:10.1016/S1474-4422(19)30320-5

Nalls, M. A., Bras, J., Hernandez, D. G., Keller, M. F., Majounie, E., Renton, A. E., . . . Singleton, A. B. (2015). NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiol Aging, 36*(3), 1605 e1607-1612. doi:10.1016/j.neurobiolaging.2014.07.028

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., . . . Singleton, A. B. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet, 46*(9), 989-993. doi:10.1038/ng.3043

The Need for Early Detection and Treatment in Alzheimer's Disease. (2016). *EBioMedicine, 9*, 1-2. doi:10.1016/j.ebiom.2016.07.001

Nelson, P. T., Pious, N. M., Jicha, G. A., Wilcock, D. M., Fardo, D. W., Estus, S., & Rebeck, G. W. (2013). APOE-epsilon2 and APOE-epsilon4 correlate with increased amyloid accumulation in cerebral vasculature. *J Neuropathol Exp Neurol, 72*(7), 708-715. doi:10.1097/NEN.0b013e31829a25b9

Ng, R., Maxwell, C. J., Yates, E. A., Nylen, K., Antflick, J., Jette, N., & Bronskill, S. E. (2015). *Brain Disorders in Ontario: Prevalence, Incidence and Costs from Health Administrative Data*. Retrieved from Toronto, ON:

Nicolas, A., Kenna, K. P., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., . . . Landers, J. E. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron, 97*(6), 1268-1283 e1266. doi:10.1016/j.neuron.2018.02.027

Nik, S., & Bowman, T. V. (2019). Splicing and neurodegeneration: Insights and mechanisms. *Wiley Interdiscip Rev RNA, 10*(4), e1532. doi:10.1002/wrna.1532

Nirenberg, M. W., & Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A, 47, 1588-1602. doi:10.1073/pnas.47.10.1588

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., . . . Phillippy, A. M. (2021). The complete sequence of a human genome. *bioRxiv*.

Ohnmacht, J., May, P., Sinkkonen, L., & Kruger, R. (2020). Missing heritability in Parkinson's disease: the emerging role of non-coding genetic variation. *J Neural Transm (Vienna), 127*(5), 729-748. doi:10.1007/s00702-020-02184-0

Olney, N. T., Spina, S., & Miller, B. L. (2017). Frontotemporal Dementia. *Neurol Clin, 35*(2), 339-374. doi:10.1016/j.ncl.2017.01.008

Olszewska, D. A., Lonergan, R., Fallon, E. M., & Lynch, T. (2016). Genetics of Frontotemporal Dementia. *Curr Neurol Neurosci Rep, 16*(12), 107. doi:10.1007/s11910-016-0707-9

Online Mendelian Inheritance in Man, OMIM®. Retrieved from https://omim.org/

Opherk, C., Duering, M., Peters, N., Karpinska, A., Rosner, S., Schneider, E., . . . Dichgans, M. (2009). CADASIL mutations enhance spontaneous multimerization of NOTCH3. *Hum Mol Genet, 18*(15), 2761-2767. doi:10.1093/hmg/ddp211

Ott, J., Wang, J., & Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet, 16*(5), 275-284. doi:10.1038/nrg3908

Pagani, F., & Baralle, F. E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet, 5*(5), 389-396. doi:10.1038/nrg1327

Page, G. P., Zakharkin, S. O., Kim, K., Mehta, T., Chen, L., & Zhang, K. (2007). Microarray analysis. *Methods Mol Biol, 404*, 409-430. doi:10.1007/978-1-59745-530-5_20

Pang, S. Y., Teo, K. C., Hsu, J. S., Chang, R. S., Li, M., Sham, P. C., & Ho, S. L. (2017). The role of gene variants in the pathogenesis of neurodegenerative disorders as revealed by next generation sequencing studies: a review. *Transl Neurodegener, 6*, 27. doi:10.1186/s40035-017-0098-0

Paul, K. C., Schulz, J., Bronstein, J. M., Lill, C. M., & Ritz, B. R. (2018). Association of Polygenic Risk Score With Cognitive Decline and Motor Progression in Parkinson Disease. *JAMA Neurol*. doi:10.1001/jamaneurol.2017.4206

Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and GC content of the human genome. *BMC Res Notes, 12*(1), 106. doi:10.1186/s13104-019-4137-z

Placek, K., Baer, G. M., Elman, L., McCluskey, L., Hennessy, L., Ferraro, P. M., . . . McMillan, C. T. (2019). UNC13A polymorphism contributes to frontotemporal disease in sporadic amyotrophic lateral sclerosis. *Neurobiol Aging, 73*, 190-199. doi:10.1016/j.neurobiolaging.2018.09.031

Placek, K., Benatar, M., Wuu, J., Rampersaud, E., Hennessy, L., Van Deerlin, V. M., . . . McMillan, C. T. (2021). Machine learning suggests polygenic risk for cognitive dysfunction in amyotrophic lateral sclerosis. *EMBO Mol Med, 13*(1), e12595. doi:10.15252/emmm.202012595

Poduri, A., Evrony, G. D., Cai, X., & Walsh, C. A. (2013). Somatic mutation, genomic variation, and neurological disease. *Science, 341*(6141), 1237758. doi:10.1126/science.1237758

Polymeropoulos, M. H., Lavedan, C., Leroy, E., Ide, S. E., Dehejia, A., Dutra, A., . . . Nussbaum, R. L. (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science, 276*(5321), 2045-2047. doi:10.1126/science.276.5321.2045

Poorkaj, P., Bird, T. D., Wijsman, E., Nemens, E., Garruto, R. M., Anderson, L., . . . Schellenberg, G. D. (1998). Tau is a candidate gene for chromosome 17 frontotemporal dementia. *Ann Neurol, 43*(6), 815-825. doi:10.1002/ana.410430617

Postuma, R. B., Berg, D., Adler, C. H., Bloem, B. R., Chan, P., Deuschl, G., . . . Stern, M. (2016). The new definition and diagnostic criteria of Parkinson's disease. *Lancet Neurol, 15*(6), 546-548. doi:10.1016/S1474-4422(16)00116-2

Priller, C., Bauer, T., Mitteregger, G., Krebs, B., Kretzschmar, H. A., & Herms, J. (2006). Synapse formation and function is modulated by the amyloid precursor protein. *J Neurosci, 26*(27), 7212-7221. doi:10.1523/JNEUROSCI.1450-06.2006

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *J Hum Genet, 59*(1), 5-15. doi:10.1038/jhg.2013.114

Raghavan, N., & Tosto, G. (2017). Genetics of Alzheimer's Disease: the Importance of Polygenic and Epistatic Components. *Curr Neurol Neurosci Rep, 17*(10), 78. doi:10.1007/s11910-017-0787-1

Ragoussis, J. (2009). Genotyping technologies for genetic research. *Annu Rev Genomics Hum Genet, 10*, 117-133. doi:10.1146/annurev-genom-082908-150116

Ramirez, J., Berezuk, C., McNeely, A. A., Scott, C. J., Gao, F., & Black, S. E. (2015). Visible Virchow-Robin spaces on magnetic resonance imaging of Alzheimer's disease patients and normal elderly from the Sunnybrook Dementia Study. *J Alzheimers Dis, 43*(2), 415-424. doi:10.3233/JAD-132528

Raz, L., Knoefel, J., & Bhaskar, K. (2016). The neuropathology and cerebrovascular mechanisms of dementia. *J Cereb Blood Flow Metab, 36*(1), 172-186. doi:10.1038/jcbfm.2015.164

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., . . . Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444-454. doi:10.1038/nature05329

Renton, A. E., Majounie, E., Waite, A., Simon-Sanchez, J., Rollinson, S., Gibbs, J. R., . . . Traynor, B. J. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron, 72*(2), 257-268. doi:10.1016/j.neuron.2011.09.010

Rich, K. A., Roggenbuck, J., & Kolb, S. J. (2020). Searching Far and Genome-Wide: The Relevance of Association Studies in Amyotrophic Lateral Sclerosis. *Front Neurosci, 14*, 603023. doi:10.3389/fnins.2020.603023

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med, 17*(5), 405-424. doi:10.1038/gim.2015.30

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273*(5281), 1516-1517. doi:10.1126/science.273.5281.1516

Roberts, J. S., Patterson, A. K., & Uhlmann, W. R. (2020). Genetic testing for neurodegenerative diseases: Ethical and health communication challenges. *Neurobiol Dis, 141*, 104871. doi:10.1016/j.nbd.2020.104871

Roberts, R., & Knopman, D. S. (2013). Classification and epidemiology of MCI. *Clin Geriatr Med, 29*(4), 753-772. doi:10.1016/j.cger.2013.07.003

Rodriguez-Murillo, L., & Greenberg, D. A. (2008). Genetic association analysis: a primer on how it works, its strengths and its weaknesses. *Int J Androl, 31*(6), 546-556. doi:10.1111/j.1365-2605.2008.00896.x

Ross, J. P., Dion, P. A., & Rouleau, G. A. (2020). Exome sequencing in genetic disease: recent advances and considerations. *F1000Res, 9*. doi:10.12688/f1000research.19444.1

Ryan, B. J., Hoek, S., Fon, E. A., & Wade-Martins, R. (2015). Mitochondrial dysfunction and mitophagy in Parkinson's: from familial to sporadic disease. *Trends Biochem Sci, 40*(4), 200-210. doi:10.1016/j.tibs.2015.02.003

Saberi, S., Stauffer, J. E., Schulte, D. J., & Ravits, J. (2015). Neuropathology of Amyotrophic Lateral Sclerosis and Its Variants. *Neurol Clin, 33*(4), 855-876. doi:10.1016/j.ncl.2015.07.012

Saier, M. H., Jr. (2019). Understanding the Genetic Code. *J Bacteriol, 201*(15). doi:10.1128/JB.00091-19

Saini, N., & Gordenin, D. A. (2018). Somatic mutation load and spectra: A record of DNA damage and repair in healthy human cells. *Environ Mol Mutagen, 59*(8), 672-686. doi:10.1002/em.22215

Sancesario, G. M., & Bernardini, S. (2018). Diagnosis of neurodegenerative dementia: where do we stand, now? *Ann Transl Med, 6*(17), 340. doi:10.21037/atm.2018.08.04

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A, 74*(12), 5463-5467. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/271968

Schacherer, J. (2016). Beyond the simplicity of Mendelian inheritance. *C R Biol, 339*(7-8), 284-288. doi:10.1016/j.crvi.2016.04.006

Schneider, J. A., Arvanitakis, Z., Bang, W., & Bennett, D. A. (2007). Mixed brain pathologies account for most dementia cases in community-dwelling older persons. *Neurology, 69*(24), 2197-2204. doi:10.1212/01.wnl.0000271090.28148.24

Seidel, G. A., Giovannetti, T., & Libon, D. J. (2012). Cerebrovascular disease and cognition in older adults. *Curr Top Behav Neurosci, 10*, 213-241. doi:10.1007/7854_2011_140

Selvackadunco, S., Langford, K., Shah, Z., Hurley, S., Bodi, I., King, A., . . . Al-Sarraj, S. (2019). Comparison of clinical and neuropathological diagnoses of neurodegenerative diseases in two centres from the Brains for Dementia Research (BDR) cohort. *J Neural Transm (Vienna), 126*(3), 327-337. doi:10.1007/s00702-018-01967-w

Shao, W., Peng, D., & Wang, X. (2017). Genetics of Alzheimer's disease: From pathogenesis to clinical usage. *J Clin Neurosci, 45*, 1-8. doi:10.1016/j.jocn.2017.06.074

Shastry, B. S. (2009a). Copy number variation and susceptibility to human disorders (Review). *Mol Med Rep, 2*(2), 143-147. doi:10.3892/mmr_00000074

Shastry, B. S. (2009b). SNPs: impact on gene function and phenotype. *Methods Mol Biol, 578*, 3-22. doi:10.1007/978-1-60327-411-1_1

Shatunov, A., Mok, K., Newhouse, S., Weale, M. E., Smith, B., Vance, C., . . . Al-Chalabi, A. (2010). Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *Lancet Neurol, 9*(10), 986-994. doi:10.1016/S1474-4422(10)70197-6

Shen, X. N., Li, J. Q., Wang, H. F., Li, H. Q., Huang, Y. Y., Yang, Y. X., . . . Alzheimer's Disease Neuroimaging, I. (2020). Plasma amyloid, tau, and neurodegeneration biomarker profiles predict Alzheimer's disease pathology and clinical progression in older adults without dementia. *Alzheimers Dement (Amst), 12*(1), e12104. doi:10.1002/dad2.12104

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., . . . Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science, 309*(5741), 1728-1732. doi:10.1126/science.1117389

Sidransky, E., Nalls, M. A., Aasly, J. O., Aharon-Peretz, J., Annesi, G., Barbosa, E. R., . . . Ziegler, S. G. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N Engl J Med, 361*(17), 1651-1661. doi:10.1056/NEJMoa0901281

Sieben, A., Van Langenhove, T., Engelborghs, S., Martin, J. J., Boon, P., Cras, P., . . . Cruts, M. (2012). The genetics and neuropathology of frontotemporal lobar degeneration. *Acta Neuropathol, 124*(3), 353-372. doi:10.1007/s00401-012-1029-x

Simon-Sanchez, J., Schulte, C., Bras, J. M., Sharma, M., Gibbs, J. R., Berg, D., . . . Gasser, T. (2009). Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet, 41*(12), 1308-1312. doi:10.1038/ng.487

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet, 15*(2), 121-132. doi:10.1038/nrg3642

Smeland, O. B., Shadrin, A., Bahrami, S., Broce, I., Tesli, M., Frei, O., . . . Andreassen, O. A. (2021). Genome-wide Association Analysis of Parkinson's Disease and Schizophrenia Reveals Shared Genetic Architecture and Identifies Novel Risk Loci. *Biol Psychiatry, 89*(3), 227-235. doi:10.1016/j.biopsych.2020.01.026

Snowden, J. S., Hu, Q., Rollinson, S., Halliwell, N., Robinson, A., Davidson, Y. S., . . . Mann, D. M. (2011). The most common type of FTLD-FUS (aFTLD-U) is associated with a distinct clinical form of frontotemporal dementia but is not related to mutations in the FUS gene. *Acta Neuropathol, 122*(1), 99-110. doi:10.1007/s00401-011-0816-0

Spillantini, M. G., Crowther, R. A., Kamphorst, W., Heutink, P., & van Swieten, J. C. (1998). Tau pathology in two Dutch families with mutations in the microtubule-binding region of tau. *Am J Pathol, 153*(5), 1359-1363. doi:10.1016/S0002-9440(10)65721-5

Spillantini, M. G., Schmidt, M. L., Lee, V. M., Trojanowski, J. Q., Jakes, R., & Goedert, M. (1997). Alpha-synuclein in Lewy bodies. *Nature, 388*(6645), 839-840. doi:10.1038/42166

Strande, N. T., Riggs, E. R., Buchanan, A. H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S. S., . . . Berg, J. S. (2017). Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am J Hum Genet, 100*(6), 895-906. doi:10.1016/j.ajhg.2017.04.015

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., . . . Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science, 315*(5813), 848-853. doi:10.1126/science.1136678

Strong, M. J., Abrahams, S., Goldstein, L. H., Woolley, S., McLaughlin, P., Snowden, J., . . . Turner, M. R. (2017). Amyotrophic lateral sclerosis - frontotemporal spectrum disorder (ALS-FTSD): Revised diagnostic criteria. *Amyotroph Lateral Scler Frontotemporal Degener, 18*(3-4), 153-174. doi:10.1080/21678421.2016.1267768

Subramanian, S., & Kumar, S. (2006). Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics, 7*, 306. doi:10.1186/1471-2164-7-306

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., . . . Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-81. doi:10.1038/nature15394

Sunderland, K. M., Beaton, D., Arnott, S. R., Kleinstiver, P. W., Kwan, D., Lawrence-Dewar, J. M., . . . Binns, M. A. (2020). The Ontario Neurodegenerative Disease Research Initiative. *medRxiv*, 2020.2007.2030.20165456. doi:10.1101/2020.07.30.20165456

Sunmonu, N. A. (2021). Multiple Cerebrovascular Insults in Pseudoxanthoma Elasticum. *J Stroke Cerebrovasc Dis, 30*(3), 105524. doi:10.1016/j.jstrokecerebrovasdis.2020.105524

Taioli, E., Pedotti, P., & Garte, S. (2004). Importance of allele frequency estimates in epidemiological studies. *Mutat Res, 567*(1), 63-70. doi:10.1016/j.mrrev.2004.06.001

Tayebi, N., Callahan, M., Madike, V., Stubblefield, B. K., Orvisky, E., Krasnewich, D., . . . Sidransky, E. (2001). Gaucher disease and parkinsonism: a phenotypic and genotypic characterization. *Mol Genet Metab, 73*(4), 313-321. doi:10.1006/mgme.2001.3201

Taylor, M. S., Ponting, C. P., & Copley, R. R. (2004). Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Res, 14*(4), 555-566. doi:10.1101/gr.1977804

Thal, D. R., von Arnim, C. A., Griffin, W. S., Mrak, R. E., Walker, L., Attems, J., & Arzberger, T. (2015). Frontotemporal lobar degeneration FTLD-tau: preclinical lesions, vascular, and Alzheimer-related co-pathologies. *J Neural Transm (Vienna), 122*(7), 1007-1018. doi:10.1007/s00702-014-1360-6

Theunissen, F., Flynn, L. L., Anderton, R. S., Mastaglia, F., Pytte, J., Jiang, L., . . . Akkari, P. A. (2020). Structural Variants May Be a Source of Missing Heritability in sALS. *Front Neurosci, 14*, 47. doi:10.3389/fnins.2020.00047

Thompson, C. T., & Gray, J. W. (1993). Cytogenetic profiling using fluorescence in situ hybridization (FISH) and comparative genomic hybridization (CGH). *J Cell Biochem Suppl, 17G*, 139-143. doi:10.1002/jcb.240531127

Tikka, S., Baumann, M., Siitonen, M., Pasanen, P., Poyhonen, M., Myllykangas, L., . . . Kalimo, H. (2014). CADASIL and CARASIL. *Brain Pathol, 24*(5), 525-544. doi:10.1111/bpa.12181

Tishkoff, S. A., & Verrelli, B. C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet, 4*, 293-340. doi:10.1146/annurev.genom.4.070802.110226

Tiwari, S., Atluri, V., Kaushik, A., Yndart, A., & Nair, M. (2019). Alzheimer's disease: pathogenesis, diagnostics, and therapeutics. *Int J Nanomedicine, 14*, 5541-5554. doi:10.2147/IJN.S200490

Trojsi, F., Christidi, F., Migliaccio, R., Santamaria-Garcia, H., & Santangelo, G. (2018). Behavioural and Cognitive Changes in Neurodegenerative Diseases and Brain Injury. *Behav Neurol, 2018*, 4935915. doi:10.1155/2018/4935915

Uricchio, L. H. (2020). Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. *Hum Genet, 139*(1), 5-21. doi:10.1007/s00439-019-02040-6

van Duijn, C. M., de Knijff, P., Cruts, M., Wehnert, A., Havekes, L. M., Hofman, A., & Van Broeckhoven, C. (1994). Apolipoprotein E4 allele in a population-based

study of early-onset Alzheimer's disease. *Nat Genet, 7*(1), 74-78. doi:10.1038/ng0594-74

Veitia, R. A., Caburet, S., & Birchler, J. A. (2018). Mechanisms of Mendelian dominance. *Clin Genet, 93*(3), 419-428. doi:10.1111/cge.13107

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *Am J Hum Genet, 90*(1), 7-24. doi:10.1016/j.ajhg.2011.11.029

Wall, J. D., & Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet, 4*(8), 587-597. doi:10.1038/nrg1123

Warner, J. P., Barron, L. H., Goudie, D., Kelly, K., Dow, D., Fitzpatrick, D. R., & Brock, D. J. (1996). A general method for the detection of large CAG repeat expansions by fluorescent PCR. *J Med Genet, 33*(12), 1022-1026. doi:10.1136/jmg.33.12.1022

Weller, J., & Budson, A. (2018). Current understanding of Alzheimer's disease diagnosis and treatment. *F1000Res, 7*. doi:10.12688/f1000research.14506.1

Wetterstrand, K. A. (2020, December 7, 2020). The Cost of Sequencing a Human Genome. Retrieved from https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Wijesekera, L. C., & Leigh, P. N. (2009). Amyotrophic lateral sclerosis. *Orphanet J Rare Dis, 4*, 3. doi:10.1186/1750-1172-4-3

Winsor, E. (1988). Mendelian genetics. *Can Fam Physician, 34*, 859-862. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21253091

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet, 89*(1), 82-93. doi:10.1016/j.ajhg.2011.05.029

Xuan, J., Yu, Y., Qing, T., Guo, L., & Shi, L. (2013). Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett, 340*(2), 284-295. doi:10.1016/j.canlet.2012.11.025

Youssef, N., Budd, A., & Bielawski, J. P. (2019). Introduction to Genome Biology and Diversity. *Methods Mol Biol, 1910*, 3-31. doi:10.1007/978-1-4939-9074-0_1

Yu, J. T., Tan, L., & Hardy, J. (2014). Apolipoprotein E in Alzheimer's disease: an update. *Annu Rev Neurosci, 37*, 79-100. doi:10.1146/annurev-neuro-071013-014300

Zannis, V. I., Breslow, J. L., Utermann, G., Mahley, R. W., Weisgraber, K. H., Havel, R. J., . . . Blum, C. (1982). Proposed nomenclature of apoE isoproteins, apoE

genotypes, and phenotypes. *J Lipid Res, 23*(6), 911-914. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7130859

Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Hum Mol Genet, 24*(R1), R102-110. doi:10.1093/hmg/ddv259

Zhang, H., Chen, W., & Sun, K. (2017). Mendelism: New Insights from Gregor Mendel's Lectures in Brno. *Genetics, 207*(1), 1-8. doi:10.1534/genetics.117.201434

Zhang, X. Y., Yang, Z. L., Lu, G. M., Yang, G. F., & Zhang, L. J. (2017). PET/MR Imaging: New Frontier in Alzheimer's Disease and Other Dementias. *Frontiers in Molecular Neuroscience, 10*, 343. doi:10.3389/fnmol.2017.00343

Zhao, N., Attrebi, O. N., Ren, Y., Qiao, W., Sonustun, B., Martens, Y. A., . . . Bu, G. (2020). APOE4 exacerbates alpha-synuclein pathology and related toxicity independent of amyloid. *Sci Transl Med, 12*(529). doi:10.1126/scitranslmed.aay1809

Zhao, N., Liu, C. C., Qiao, W., & Bu, G. (2018). Apolipoprotein E, Receptors, and Modulation of Alzheimer's Disease. *Biol Psychiatry, 83*(4), 347-357. doi:10.1016/j.biopsych.2017.03.003

Zlotogora, J. (1994). High frequencies of human genetic diseases: founder effect with genetic drift or selection? *Am J Med Genet, 49*(1), 10-13. doi:10.1002/ajmg.1320490104

Zou, H., Wu, L. X., Tan, L., Shang, F. F., & Zhou, H. H. (2020). Significance of Single-Nucleotide Variants in Long Intergenic Non-protein Coding RNAs. *Front Cell Dev Biol, 8*, 347. doi:10.3389/fcell.2020.00347

Chapter 2 – The influence of *APOE* and *MAPT* to a spectrum of neurodegenerative disease phenotypes.

The work presented in <u>Chapter 2</u> has been edited from the original publication in the *Canadian Journal of Neurological Sciences* for brevity and consistency throughout the entire Dissertation.
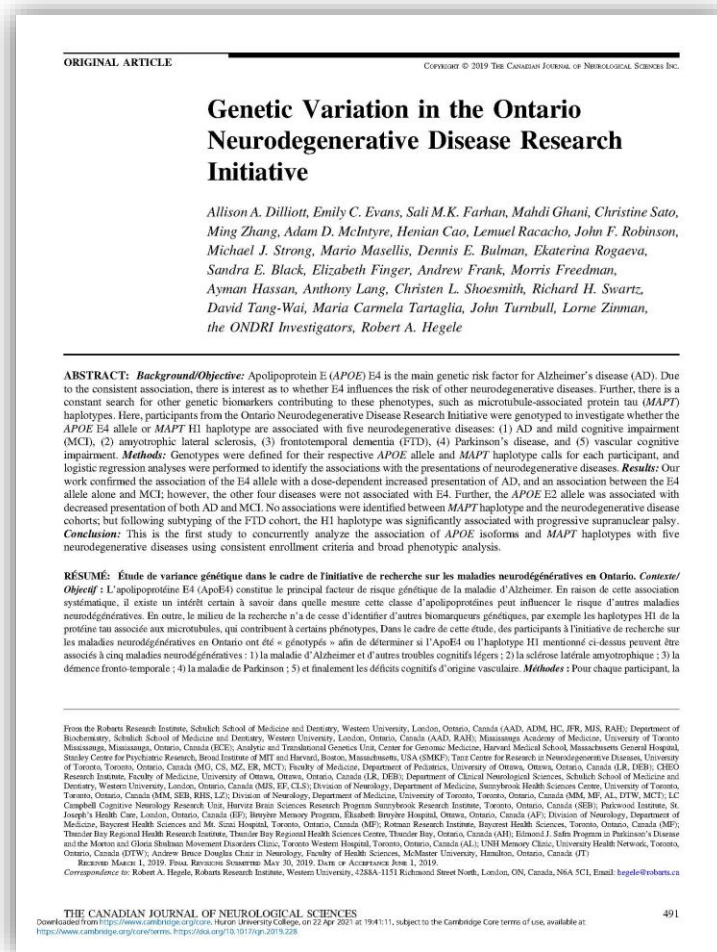
Dilliott A.A., Evans E.C., Farhan S.M.K., Ghani M., Sato C., Zhang M., McIntyre A.D., Cao H., Racacho L., Robinson J.F., Strong M.J., Masellis M., Bulman D.E., Rogaeva E., Black S., Finger E., Frank A., Freedman M., Hassan A., Lang A., Shoesmith C., Swartz R., Tang-Wai D., Tartaglia C., Turnbull J., Zinman L., the ONDRI Investigators, and Hegele R.A. (2019). Genetic variation in the Ontario Neurodegenerative Disease Research Initiative. *Can J Neurol Sci,* 46(5):491-498

Can J Neurol Sci Editor's Choice Award

2.1.   Abstract

Background/Objective: Apolipoprotein E (*APOE*) E4 is the main genetic risk factor for Alzheimer's disease (AD). Due to the consistent association, there is interest as to whether E4 influences the risk of other neurodegenerative diseases. Further, there is a constant search for other genetic biomarkers contributing to these phenotypes, such as microtubule associated protein tau (*MAPT*) haplotypes. Here, participants from the Ontario Neurodegenerative Disease Research Initiative were genotyped to investigate whether the *APOE* E4 allele or *MAPT* H1 haplotype are associated with multiple neurodegenerative diseases: 1) AD; 2) mild cognitive impairment (MCI); 3) amyotrophic lateral sclerosis; 4) frontotemporal dementia (FTD); and 5) Parkinson's disease, as well as cerebrovascular disease (CVD) with or without cognitive impairment.

Methods and Results: Genotypes were mapped to their respective *APOE* allele and *MAPT* haplotype calls for each participant and logistic regressions were performed to identify associations with the disease cohorts. Our work confirmed the association of the E4 allele with a dose-dependent increased presentation of AD, and an association between the E4 allele alone and MCI; however, the other four diseases were not associated with E4. Further, the *APOE* E2 allele was associated with decreased presentation of both AD and MCI. No associations were identified between *MAPT* haplotype and the disease cohorts, but following subtyping of the FTD cohort, the H1 allele was significantly associated with progressive supranuclear palsy.

Conclusion: This is the first study to concurrently analyze association of *APOE* isoforms and *MAPT* haplotypes with five neurodegenerative diseases and CVD using consistent enrollment criteria and broad phenotypic analysis.

2.2.   Introduction

With the aging of populations, the burden of neurodegenerative diseases is increasing, and substantial effort is directed towards identification of genetic biomarkers with the objective of improved disease prediction and the long-term goal of discovering therapeutic targets. In particular, molecular genetics efforts have focused on identifying common single nucleotide polymorphisms (SNPs) that contribute to disease risk (Lambert et al., 2013; Nalls et al., 2014; Simon-Sanchez et al., 2009). Although these types of markers usually account for only a small proportion of disease risk, two closely linked common SNPs on chromosome 19 have been identified to jointly impart a relatively large effect on the risk of a particular neurodegenerative phenotype, namely the apolipoprotein E (*APOE*) E4 allele and Alzheimer's disease (AD) (Bertram, McQueen, Mullin, Blacker, & Tanzi, 2007).

Apo E is found in chylomicrons, very low-density lipoproteins, intermediate-density lipoproteins and high-density lipoproteins; it provides structural support to these particles and also governs the catabolism of triglyceride rich lipoproteins through its role as a receptor ligand. Importantly, apo E is the principal cholesterol carrier in the brain (Mahley & Rall, 2000). There are three common protein isoforms of apo E — E2, E3, and E4 — historically designated based on protein mobility in isoelectrophoretic focusing gels (Kane & Gowland, 1986). At the DNA level, these three isoforms are encoded by two nonsynonymous SNPs within the *APOE* gene, occurring at amino acid positions 130 and 176 (also numbered as 112 and 158, respectively, if the pro-peptide sequence is excluded), and each involving cysteine or arginine as alternate residues (Zannis et al., 1982). The E4 allele, which has arginine at both positions 130 and 176, is the most common genetic risk factor for the development of late-onset AD and contributes to disease risk in a dose-dependent manner (Ward et al., 2012). Meta-analyses show that one and two copies of the E4 allele raise AD risk by ~3- to 4- and ~12-fold, respectively (Corder et al., 1993; Saunders et al., 1993). Due to the replicated high risk association from several meta-analyses of AD and *APOE* (Farrer et al., 1997; Liu et al., 2015; Ward et al., 2012), researchers have attempted to determine whether the E4 allele is also associated with other neurodegenerative diseases, including amyotrophic lateral sclerosis

(ALS) (Y. J. Li et al., 2004; Mui, Rebeck, McKenna-Yasek, Hyman, & Brown, 1995), frontotemporal dementia (FTD) (Agosta et al., 2009; Geschwind, Karrim, Nelson, & Miller, 1998), and Parkinson's disease (PD) (Ezquerra et al., 2008; Huang, Chen, & Poole, 2004; Pankratz et al., 2006), as well as cerebrovascular disease with or without cognitive impairment (CVD ± CI) (Baum et al., 2006; Chuang et al., 2010; Davidson et al., 2006), so far with mixed and inconsistent results.

Another gene less consistently associated with AD risk is the microtubule associated protein tau gene (*MAPT*), which encodes the protein tau (Myers et al., 2005). It remains to be established whether the *MAPT* is associated with other neurodegenerative diseases. Within the *MAPT* gene an ancestral inversion of ~900kb has resulted in two distinct haplotypes, H1 and H2, and creates a large region of linkage disequilibrium. Apart from the few associations found between H1 *MAPT* and AD, there has been debate as to whether the haplotype is associated with PD (Seto-Salvia et al., 2011) and with the FTD subtype progressive supranuclear palsy (Baker et al., 1999; Ferrari et al., 2017).

The Ontario Neurodegenerative Disease Research Initiative (ONDRI) is a multi-platform, provincial-wide, observational cohort study aiming to characterize multiple attributes of five neurodegenerative diseases, namely: 1) AD; 2) mild cognitive impairment (MCI); 3) ALS; 4) FTD; and 5) PD, as well as CVD ± CI (S. M. K. Farhan et al., 2017). In addition to genomic analysis, ONDRI incorporates a comprehensive phenotypic assessment on each participant. The large dataset, combined with the consistent enrollment criteria allows for the unique opportunity to assess the association of *APOE* genotype and *MAPT* haplotype across the respective neurodegeneration phenotypes. Here, we aim to replicate the known associations of the *APOE* E4 allele, *APOE* E4/4 genotype, and *MAPT* H1 haplotype with AD, in addition to assessing whether *APOE* E4 and *MAPT* H1 confer risk to ALS, FTD, PD, and CVD ± CI within the ONDRI cohort.

2.3.   Methods

2.3.1.   Study participants

Blood samples were collected from 519 ONDRI participants after informed consent was obtained, in accordance with the Research Ethics Boards at Hamilton General Hospital

(Hamilton, Ontario, Canada); McMaster (Hamilton, Ontario, Canada); Parkwood Hospital (London, Ontario, Canada); London Health Sciences Centre (London, Ontario, Canada); The Ottawa Hospital (Ottawa, Ontario, Canada); University Health Network-Elizabeth Bruyère Hospital (Ottawa, Ontario, Canada); Baycrest Centre for Geriatric Care (Toronto, Ontario, Canada); Centre for Addiction and Mental Health (Toronto, Ontario, Canada); St Michael's Hospital (Toronto, Ontario, Canada); Sunnybrook Health Sciences Centre (Toronto, Ontario, Canada); and Toronto Western Hospital (Toronto, Ontario, Canada). Formal diagnoses and demographic data were obtained by participants' clinicians upon enrollment in the study, in accordance with ONDRI standard operating protocols (S. M. K. Farhan et al., 2017).

2.3.2.   DNA preparation and genotyping

Genomic DNA was isolated from blood samples collected from each participant as described previously (S.M.K. Farhan et al., 2016). DNA samples were also obtained from 189 cognitively normal controls from the GenADA study (H. Li et al., 2008). All samples underwent targeted next-generation sequencing using the ONDRISeq neurodegenerative disease gene panel. Full methodology of DNA isolation, sequencing with the ONDRISeq panel (S.M.K. Farhan et al., 2016), and raw sequencing data processing were previously described (Dilliott et al., 2018).

Allele calls for the *APOE* risk alleles rs429358(CT):p.Cys130Arg and rs7412(CT):p.Arg176Cys were extracted from the ONDRISeq data files and mapped to their respective *APOE* genotype for each participant using a customized Annotate Variation (Wang, Li, & Hakonarson, 2010) script. Allele calls and mapped genotypes were validated with TaqMan allelic discrimination assay (Koch et al., 2002), as previously described (S.M.K. Farhan et al., 2016).

TaqMan was also used to determine the *MAPT* haplotype of the ONDRI participants and control samples. DNA samples were genotyped for the intronic SNP rs1800547, which is not covered by the ONDRISeq panel. Based on a region of linkage disequilibrium, allele calls were mapped to their respective *MAPT* haplotype (Lai et al., 2017).

2.3.3.   Statistical analysis

Statistical analyses were performed using SAS v9.4 (SAS Institute, Cary, NC). The Wilcoxon Mann-Whitney U test was utilized to determine the difference between the ages of the control cohort compared to the five disease cohorts of interest. Chi-squared analyses were used to determine the difference between the control cohort and disease cohorts' male:female ratios. Odds ratios and confidence intervals were obtained using logistic regression, adjusting for participants' age and sex.

2.4.   Results

2.4.1.   Study participants

Table 2.1 displays the demographics of the 519 ONDRI participants included in this study, as well as the cognitively normal controls. Of the ONDRI participants, 83.0% self-reported their ethnicity as Caucasian. The ALS cohort had the lowest mean age (62.0 ± 8.7 years), and the control cohort had the highest mean age (74.0 ± 8.2 years), which was significantly different from the mean age of the five ONDRI disease cohorts ($p < 1.0 \times 10^{-4}$). Additionally, the male:female ratio of the control cohort was significantly different from that of the overall ONDRI cohort ($p < 1.0 \times 10^{-4}$).

Table 2.1 Demographics of the 519 ONDRI participants and 189 controls genotyped for *APOE* and haplotyped for *MAPT*.

	ONDRI	AD/MCI	ALS	FTD	PD	CVD ± CI	Controls
Cases	519	126	40	53	139	161	189
Mean Age (years ± sd)	68.6 ± 7.7	71.0 ± 8.2	62.0 ± 8.7	67.8 ± 7.0	67.9 ± 6.3	69.2 ± 7.4	74.0 ± 8.2
Min. Age (years)	40.1	53.4	40.1	49.7	55.1	54.9	48.0
Max. Age (years)	87.8	87.8	77.2	80.9	85.9	85.4	92.0
Male:Female	345:174	69:57	24:16	34:19	108:31	110:51	77:112

Abbreviations: AD/MCI, Alzheimer's disease/mild cognitive impairment; ALS, amyotrophic lateral sclerosis; *APOE*, Apolipoprotein E gene; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; *MAPT*, microtubule associated protein tau; Max, maximum; Min, minimum; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease; sd, standard deviation.

2.4.2. *APOE* genotype associations

Calls of the *APOE* alleles were obtained using the ONDRISeq panel and validated using the TaqMan allelic discrimination assay with 100% concordance. Allele and genotype frequencies were calculated for the disease cohorts and the controls (**Table 2.2**). As expected, the AD/MCI cohort displayed the highest *APOE* E4 allele frequency (31.7%) and E4/4 genotype frequency (14.3%), compared to 14.6% and 3.7%, respectively in controls. The AD/MCI cohort also displayed the lowest *APOE* E2 allele frequency (2.4%), compared to 10.3% in controls. The lowest *APOE* E4 allele and E4/4 genotype frequencies were observed in the PD cohort (12.6% and 1.4%, respectively), differing marginally from the respective frequencies in controls.

Allele and genotype calls were compared between each ONDRI disease cohort and the control cohort, while adjusting for both age and sex of participants (**Figure 2.1**). The E4 allele was significantly associated with increased presentation of AD/MCI compared to controls (OR = 2.76, 95% CI = 1.85–4.11, p < 1.0×10^{-4}). Similarly, the E4/4 genotype significantly increased the presentation of AD/MCI when compared to controls (OR = 4.13, 95% CI = 1.64–10.37, p = 2.5×10^{-3}). As expected, the E2 allele was associated with a significantly decreased presentation of AD/MCI when compared to controls after adjusting for age and sex (OR = 0.21, 95% CI = 0.08–0.50, p = 5.0×10^{-4}; **Figure 2.2**). No association with *APOE* was found with the other four phenotypes in the ONDRI dataset.

The AD/MCI cohort was split into participants presenting with AD (n=41) and those presenting with MCI (n = 85) and *APOE* analyses were repeated. The AD and MCI sub cohorts displayed E4 allele frequencies of 46.3% and 24.7% and E4/4 genotype frequencies of 26.8% and 8.2%, respectively. Indeed, the E4 allele was significantly associated with both AD and MCI compared to controls (OR = 5.24, 95% CI = 3.07–8.92, p < 1.0×10^{-4} and OR = 1.94, 95% CI = 1.22–3.07, p = 4.9×10^{-3}, respectively) and the E2 allele was significantly associated with decreased presentation of both AD and MCI compared to controls (OR = 0.10, 95% CI = 0.01–0.77, p = 0.0268 and OR = 0.26, 95% CI = 0.10–0.68, p = 5.8×10^{-3}, respectively). The E4/4 genotype was also significantly associated with increased presentation of AD (OR = 10.36, 95% CI = 3.55–

30.19, p < 1.0x10^{-4}); however, the genotype did not significantly increase presentation of MCI.

Table 2.2 *APOE* allele and genotype frequencies in 519 ONDRI participants and 189 controls.

	APOE Genotype [n (%)]						*APOE* Alleles [n (%)]		
	E2/2	**E3/2**	**E4/2**	**E3/3**	**E4/3**	**E4/4**	**E2**	**E3**	**E4**
AD/MCI	0	5 (4.0)	1 (0.8)	59 (46.8)	43 (34.1)	18 (14.3)	6 (2.4)	166 (65.9)	80 (31.7)
ALS	0	6 (15.0)	0	21 (52.5)	12 (30.0)	1 (2.5)	6 (7.5)	60 (75.0)	14 (17.5)
FTD	1 (1.9)	6 (11.3)	0	26 (49.0)	18 (34.0)	2 (3.8)	8 (7.5)	76 (71.7)	22 (20.8)
PD	1 (0.7)	19 (13.7)	3 (2.2)	86 (67.9)	28 (20.1)	2 (1.4)	24 (8.6)	219 (78.8)	35 (12.6)
CVD ± CI	0	21 (13.0)	2 (1.2)	100 (62.1)	34 (21.1)	4 (2.5)	23 (7.1)	255 (79.2)	44 (13.7)
Controls	2 (1.1)	29 (15.3)	6 (3.2)	110 (58.2)	35 (18.6)	7 (3.7)	39 (10.3)	284 (75.1)	55 (14.6)

All study participants were genotyped using both the ONDRISeq panel and TaqMan allelic discrimination assay.

Abbreviations: AD/MCI, Alzheimer's disease/mild cognitive impairment; ALS, amyotrophic lateral sclerosis; APOE, Apolipoprotein E gene; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease.
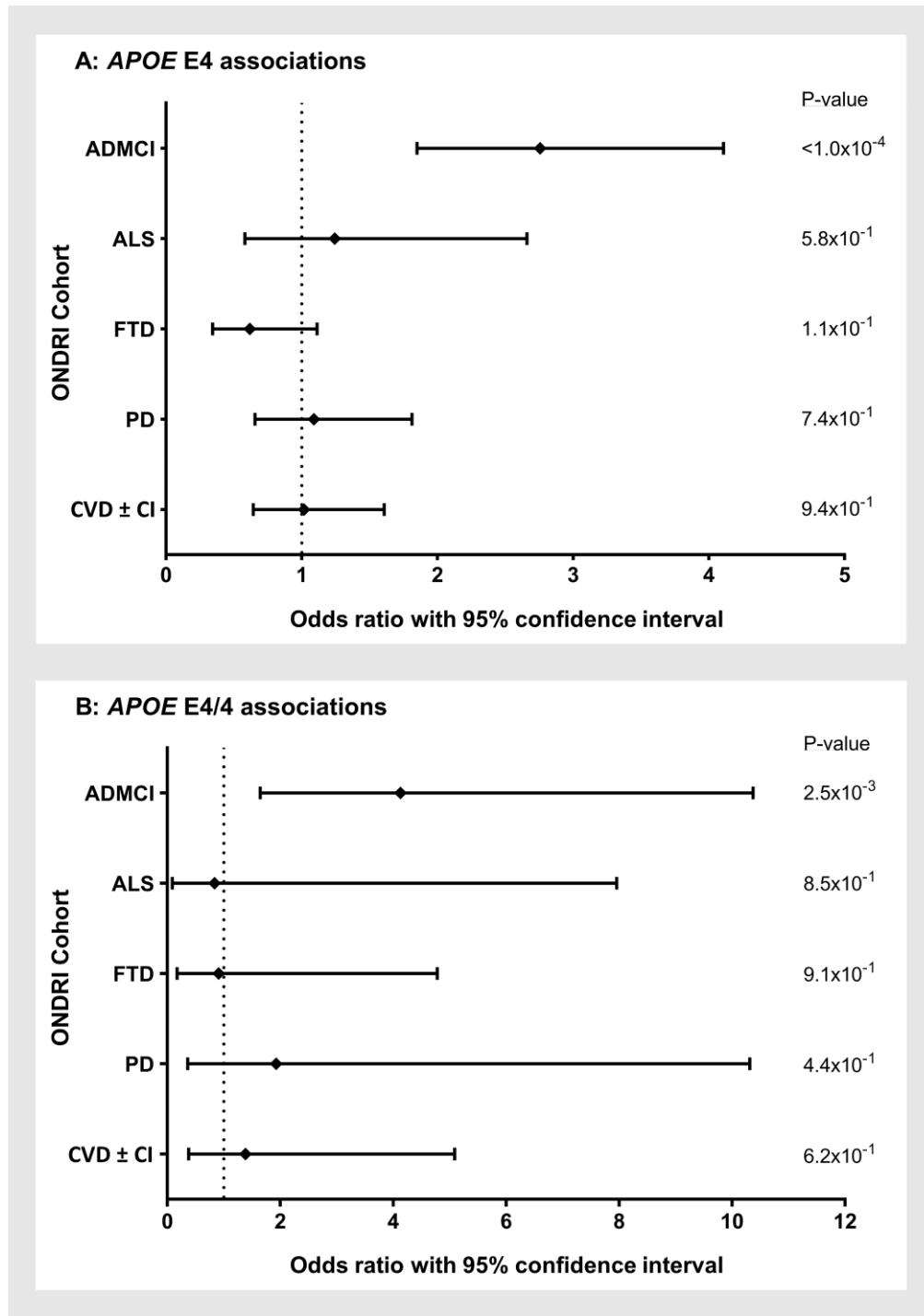
Figure 2.1 Forest plots of the relationship between *APOE* and risk of each of the diseases encompassed by ONDRI.
Logistic regressions adjusting for participant age and sex analyzed the *APOE* E4 allele and E4/4 genotype status of the ONDRI cohorts when compared to controls. A. Forest plot of the *APOE* E4 allele and associated risk of each ONDRI disease cohort. B. Forest plot of the *APOE* E4/4 genotype and associated risk of each ONDRI disease cohort.
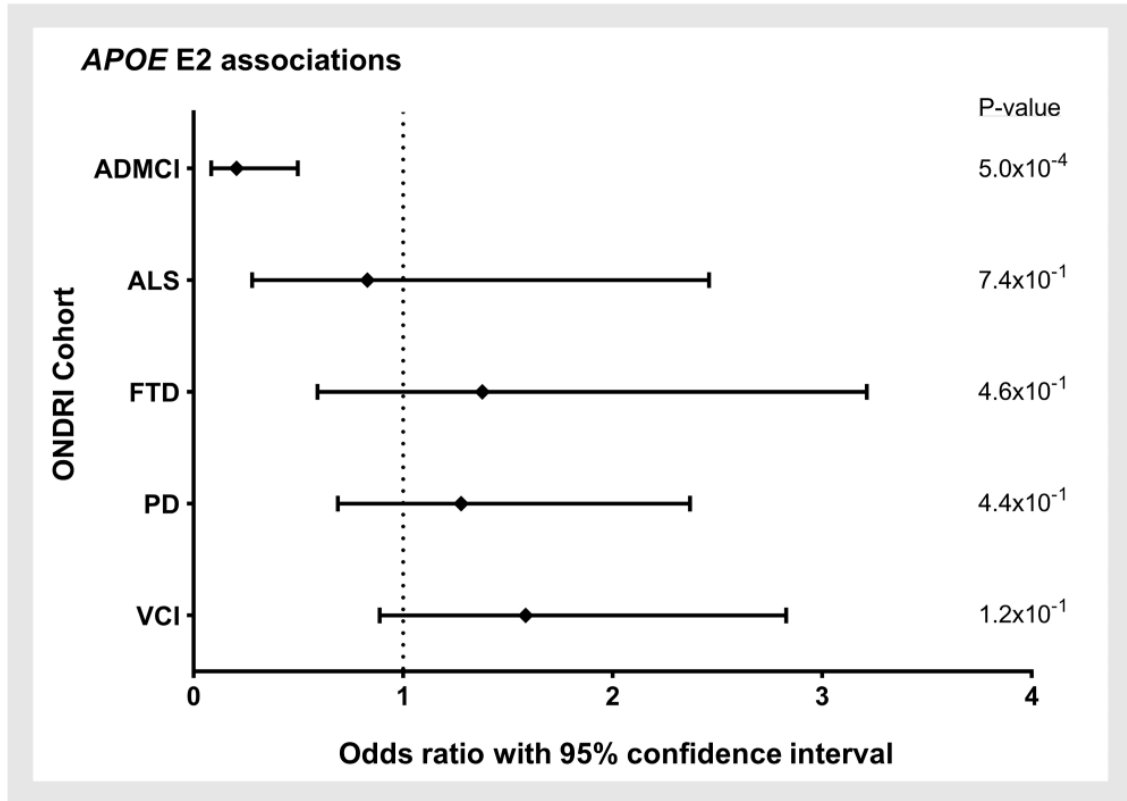
Figure 2.2 Forest plot of the relationship between the *APOE* E2 allele and risk of each of the diseases encompassed by ONDRI.
Logistic regressions adjusting for participant age and sex analyzed the *APOE* E2 allele status of the ONDRI cohorts when compared to controls.

2.4.3.   *MAPT* haplotype associations

Allele calls of the intronic *MAPT* variant, rs1800547, were mapped to their respective *MAPT* haplotype for each DNA sample. The ALS cohort had the highest frequencies of H1 haplotype and H1/H1 diplotype (87.5% and 75.0%, respectively), whereas the FTD cohort displayed the lowest frequencies (75.5% and 60.4%, respectively; **Table 2.3**). There were no significant associations found between *MAPT* and any of the disease phenotypes in ONDRI when compared to controls following adjustment for both age and sex.

Due to its previous associations with the PSP subtype of FTD, the FTD cohort was split into its respective subtypes, including behavioural variant FTD (bvFTD; n = 22), corticobasal syndrome (CBS; n = 3), progressive non-fluent aphasia (PNFA; n = 8), PSP (n=15), and semantic dementia (SD; n = 5), and *MAPT* analyses were repeated. Interestingly, the SD sub cohort displayed the greatest *MAPT* H1 haplotype frequency and the CBS sub cohort displayed the lowest, at 90.0% and 50.0% respectively. Similarly, the SD sub cohort, along with the PSP sub cohort, displayed the greatest H1H1 diplotype frequency of 80.0%, while the CBS sub cohort displayed the lowest of 33.3%. We also observed that the H1 haplotype was significantly associated with increased PSP prevalence (OR = 7.46, 95% CI = 2.39–23.29, p = 5.0×10^{-4}) following adjustment for age and sex; however, the H1H1 diplotype did not display significant associations with PSP presentation. In addition, there were no significant associations between *MAPT* and any of the other FTD subtypes.

Table 2.3 *MAPT* haplotype and diplotype frequencies in 519 ONDRI participants and 189 controls.

	MAPT Haplotype [n (%)]		*MAPT* Diplotype [n (%)]		
	H1	**H2**	**H1/H1**	**H1/H2**	**H2/H2**
AD/MCI	200 (79.4)	52 (20.6)	75 (59.5)	50 (39.7)	1 (0.8)
ALS	70 (87.5)	10 (12.5)	30 (75.0)	10 (25.0)	0
FTD	80 (75.5)	26 (24.5)	32 (60.4)	16 (30.2)	5 (9.4)
PD	227 (81.7)	51 (18.3)	93 (66.9)	41 (29.5)	5 (3.6)
VCI	264 (82.0)	58 (18.0)	108 (67.1)	48 (29.8)	5 (3.1)
Controls	293 (77.5)	85 (22.5)	111 (58.7)	71 (37.6)	7 (3.7)

All study participants were genotyped for the intronic SNP rs1800547 using TaqMan allelic discrimination assay and results were mapped to their respective haplotype.

Abbreviations: AD/MCI, Alzheimer's disease/mild cognitive impairment; ALS, amyotrophic lateral sclerosis; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; MAPT, microtubule associated protein tau gene; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease.

2.5.   Discussion

This is the first genetic characterization of the ONDRI cohort, which is important for upcoming multimodal, multi-year, prospective observational studies of the five phenotypes (e.g. *APOE*/*MAPT*-based stratification). The principal findings from the current study are: 1) a dose-dependent association of the *APOE* E4 allele with AD and an association between E4 and MCI; 2) an inverse association of the *APOE* E2 allele with AD and MCI presentation; 3) a lack of associations between the *APOE* alleles and other diseases included in the ONDRI mandate (ALS, FTD, PD or CVD ± CI); and 4) no associations between any of the disease cohorts and the *MAPT* H1 haplotype, but an association between H1 and the PSP subtype of FTD.

Our study design offers a unique opportunity to analyze individuals each with one of five neurodegenerative diseases or CVD ± CI enrolled with strict inclusion criteria and evaluated across a wide-range of platforms (S. M. K. Farhan et al., 2017). Because of this robust workflow, we can investigate the effect of the *APOE* alleles and genotypes and *MAPT* haplotypes across multiple diseases with common assessment. The control cohort had a significantly older mean age than the ONDRI disease cohorts, as well as a significantly different male:female ratio. For this reason, logistic regression was applied to obtain odds ratios adjusted for both the age and sex of participants.

The E4 allele frequency in previously reported AD patients is 28–37%, while in controls it is 8–14% (Farrer et al., 1997; Heffernan, Chidgey, Peng, Masters, & Roberts, 2016). The results presented here are comparable to these literature values, with E4 allele frequencies of 31.7% and 14.5% in the AD/MCI and control cohorts, respectively. More specifically, we observed an *APOE* E4 allele frequency of 46.3% in individuals with AD and 24.7% in individuals with MCI. Although the E4 allele frequency was significantly increased in cohorts of MCI compared to controls, the increase is not as great as that seen in those with AD only. Interestingly, the E4 allele has been shown to be a predictive risk factor for the clinical conversion from MCI to AD (Elcoroaristizabal Martin et al., 2011; Fleisher et al., 2007; Petersen et al., 2005), which, coupled with the increased E4 allele frequency in the MCI sub cohort, may indicate that a portion of the individuals enrolled in ONDRI with MCI will experience disease progression to AD. The longitudinal nature

of the ONDRI study will permit follow up of the individuals with MCI to determine whether their *APOE* status predicts possible progression to AD, and to evaluate the phenotypic measures most severely affected by their status.

In accordance with previous literature (Corder et al., 1993), we identified that *APOE* E4 was associated with AD in a dose-dependent manner following adjustment for age and sex. The presence of the E4 allele produced an approximately 5-fold increased risk of having AD/MCI, which was marginally greater than estimates previously found. However, we identified the E4/4 genotype to increase risk just over 10-fold, marginally lower than the commonly reported 12-fold (Corder et al., 1993; Saunders et al., 1993). The slight discrepancies are likely due to the modest number of individuals enrolled in ONDRI with AD.

Previous studies have also suggested that the E2 allele decreases risk of AD, which we also observed with our AD/MCI cohort (Corder et al., 1994; Farrer et al., 1997). It is hypothesized that the stability provided by the cysteine-to-arginine variant at amino acid 176 may be contributing to this protective effect (Zhong & Weisgraber, 2009) and allows the isoform to more effectively clear amyloid-β (Yang, Smith, Zhou, Gandy, & Martins, 1997), protect against synaptic degeneration (Dumanis et al., 2009), and facilitate anti-oxidant activity (Miyata & Smith, 1996). However, because of the small sample sizes within ONDRI, no individuals in the AD/MCI cohort harboured the E2/2 genotype and precise genotypic risk associations could not be evaluated. Due to this absence of individuals with the E2/2 genotype in the AD/MCI cohort, it is expected that two copies of the E2 allele would incrementally decrease the risk of the disease in a dose-dependent manner, particularly in those with AD, but larger cohorts would be needed to validate this hypothesis.

Although an association was not observed between *MAPT* H1 and the total FTD cohort, we did observe an association between the haplotype and increased presentation of PSP, as has been previously identified (Baker et al., 1999). Yet, we were not able to replicate the previously observed increased prevalence of PSP associated with the H1H1 diplotype, again possibly as a result of the modest number of individuals with PSP. Interestingly, the

SD sub cohort of FTD displayed the highest *MAPT* H1 haplotype frequency and the same H1H1 diplotype frequency as the PSP sub cohort, but was not significantly associated with *MAPT*. We expect that the very low number of individuals enrolled in ONDRI with SD may be driving factor in the lack of association observed here and believe that further analysis into the association between *MAPT* H1 and SD is warranted with larger sample sizes.

Due to the strong association between both the E4 allele and E4/4 genotype and AD status, many studies have attempted to identify associations with other neurodegenerative disorders (Verghese, Castellano, & Holtzman, 2011). However, these studies have reported inconsistent results regarding the risk associated with *APOE* E4 and either onset and/or progression of the other four diseases studied in ONDRI, namely ALS (Y. J. Li et al., 2004; Mui et al., 1995), FTD (Agosta et al., 2009; Geschwind et al., 1998), PD (Ezquerra et al., 2008; Huang et al., 2004; Pankratz et al., 2006), and CVD ± CI (Baum et al., 2006; Chuang et al., 2010; Davidson et al., 2006). Within the ONDRI cohort, no associations were identified for the other disease phenotypes. Similarly, no associations were identified between the *MAPT* H1 haplotype or H1/H1 diplotype and any of the five complete neurodegenerative disease cohorts or the CVD ± CI cohort. Absence of associations could have been due to small sample sizes, and thus false negative inferences, or to the true lack of a biological effect of E4 and H1 in these conditions. Associations previously reported may have been due to the diagnostic challenges associated with neurodegenerative diseases. Admixture of AD pathology in individuals with other neurodegenerative diseases, including CVD associated neurodegenerative diseases such as vascular dementia, may produce false positive associations with E4, and co-pathologies within neurodegenerative diseases are far more common than previously appreciated (Robinson et al., 2018). Due to the spectrum of overlapping features that can be observed across neurodegenerative phenotypes it will be important to identify those that are associated with the *APOE* E4 allele and *MAPT* H1 haplotype to better understand patient prognosis. Future analyses will utilize ONDRI's robust assessment of structural and cognitive measures to identify whether common phenotypes across the various diseases are influenced by *APOE* and *MAPT*.

While ONDRI is unique in terms of the number of different clinical conditions evaluated simultaneously within the process, there is still a limitation due to modest sample sizes. Larger cohorts may produce results that align more closely with those previously reported. Additionally, an important limitation to this study is the lack of correlation with cognitive status within the disease cohorts. Assessments of cognitive impairment are ongoing, and future studies will incorporate these measures from the participants in order to assess the effects of the *APOE* E4 allele and *MAPT* haplotype on cognitive status within all five disease cohorts.

2.6.   Conclusion

In conclusion, E4 allele carriers in the ONDRI study displayed a dose-dependent increased risk of AD/MCI, specifically in those diagnosed with AD, which is consistent with current *APOE* literature. Similarly, this study was concordant with recent evidence that the *APOE* E2 allele decreases risk of AD/MCI. Further, the *MAPT* H1 haplotype was significantly associated with the PSP subtype of FTD. The work also confirmed that risks of the other four diseases evaluated within ONDRI, namely ALS, FTD, PD and CVD ± CI are not associated with the E4 allele or E4/4 genotype and that none of the complete disease cohorts are associated with the *MAPT* H1 haplotype or H1/H1 diplotype. To our knowledge this is the first study to analyze *APOE* genotypes and *MAPT* haplotypes across these five neurodegenerative diseases and CVD using common enrollment criteria and comprehensive phenotypic analysis. Future studies will investigate the structural and cognitive symptoms of neurodegeneration influenced by the E4 allele and H1 haplotype and the contributions of other genetic factors to these phenotypes.

2.7.   References

Agosta, F., Vossel, K. A., Miller, B. L., Migliaccio, R., Bonasera, S. J., Filippi, M., . . . Gorno-Tempini, M. L. (2009). Apolipoprotein E epsilon4 is associated with disease-specific effects on brain atrophy in Alzheimer's disease and frontotemporal dementia. *Proc Natl Acad Sci U S A, 106*(6), 2018-2022. doi:10.1073/pnas.0812697106

Baker, M., Litvan, I., Houlden, H., Adamson, J., Dickson, D., Perez-Tur, J., . . . Hutton, M. (1999). Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum Mol Genet, 8*(4), 711-715. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10072441

Baum, L., Lam, L. C., Kwok, T., Lee, J., Chiu, H. F., Mok, V. C., . . . Ng, H. K. (2006). Apolipoprotein E epsilon4 allele is associated with vascular dementia. *Dement Geriatr Cogn Disord, 22*(4), 301-305. doi:10.1159/000095246

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet, 39*(1), 17-23. doi:ng1934 [pii] 10.1038/ng1934

Chuang, Y. F., Hayden, K. M., Norton, M. C., Tschanz, J., Breitner, J. C., Welsh-Bohmer, K. A., & Zandi, P. P. (2010). Association between APOE epsilon4 allele and vascular dementia: The Cache County study. *Dement Geriatr Cogn Disord, 29*(3), 248-253. doi:10.1159/000285166000285166 [pii]

Corder, E. H., Saunders, A. M., Risch, N. J., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Jr., . . . et al. (1994). Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat Genet, 7*(2), 180-184. doi:10.1038/ng0694-180

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., . . . Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science, 261*(5123), 921-923. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8346443

Davidson, Y., Gibbons, L., Purandare, N., Byrne, J., Hardicre, J., Wren, J., . . . Mann, D. M. (2006). Apolipoprotein E epsilon4 allele frequency in vascular dementia. *Dement Geriatr Cogn Disord, 22*(1), 15-19. doi:10.1159/000092960

Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2018). Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J Vis Exp*(134). doi:10.3791/57266

Dumanis, S. B., Tesoriero, J. A., Babus, L. W., Nguyen, M. T., Trotter, J. H., Ladu, M. J., . . . Hoe, H. S. (2009). ApoE4 decreases spine density and dendritic complexity in cortical neurons in vivo. *J Neurosci, 29*(48), 15317-15322. doi:10.1523/JNEUROSCI.4026-09.2009

Elcoroaristizabal Martin, X., Fernandez Martinez, M., Galdos Alcelay, L., Molano Salazar, A., Bereincua Gandarias, R., Ingles Borda, S., . . . de Pancorbo, M. M. (2011). Progression from amnesic mild cognitive impairment to Alzheimer's disease: ESR1 and ESR2 polymorphisms and APOE gene. *Dement Geriatr Cogn Disord, 32*(5), 332-341. doi:10.1159/000335541000335541 [pii]

Ezquerra, M., Campdelacreu, J., Gaig, C., Compta, Y., Munoz, E., Marti, M. J., . . . Tolosa, E. (2008). Lack of association of APOE and tau polymorphisms with dementia in Parkinson's disease. *Neurosci Lett, 448*(1), 20-23. doi:10.1016/j.neulet.2008.10.018S0304-3940(08)01401-8 [pii]

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., . . . Strong, M. J. (2017). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci, 44*(2), 196-202. doi:10.1017/cjn.2016.415 S0317167116004157 [pii]

Farhan, S. M. K., Dilliott, A. A., Ghani, G., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*(16032), 1-11.

Farrer, L. A., Cupples, L. A., Haines, J. L., Hyman, B., Kukull, W. A., Mayeux, R., . . . van Duijn, C. M. (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA, 278*(16), 1349-1356. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9343467

Ferrari, R., Wang, Y., Vandrovcova, J., Guelfi, S., Witeolar, A., Karch, C. M., . . . Desikan, R. S. (2017). Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer's and Parkinson's diseases. *J Neurol Neurosurg Psychiatry, 88*(2), 152-164. doi:10.1136/jnnp-2016-314411

Fleisher, A. S., Sowell, B. B., Taylor, C., Gamst, A. C., Petersen, R. C., & Thal, L. J. (2007). Clinical predictors of progression to Alzheimer disease in amnestic mild cognitive impairment. *Neurology, 68*(19), 1588-1595. doi:01.wnl.0000258542.58725.4c [pii]10.1212/01.wnl.0000258542.58725.4c

Geschwind, D., Karrim, J., Nelson, S. F., & Miller, B. (1998). The apolipoprotein E epsilon4 allele is not a significant risk factor for frontotemporal dementia. *Ann Neurol, 44*(1), 134-138. doi:10.1002/ana.410440122

Heffernan, A. L., Chidgey, C., Peng, P., Masters, C. L., & Roberts, B. R. (2016). The Neurobiology and Age-Related Prevalence of the epsilon4 Allele of

Apolipoprotein E in Alzheimer's Disease Cohorts. *J Mol Neurosci, 60*(3), 316-324. doi:10.1007/s12031-016-0804-x

Huang, X., Chen, P. C., & Poole, C. (2004). APOE-[epsilon]2 allele associated with higher prevalence of sporadic Parkinson disease. *Neurology, 62*(12), 2198-2202. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/15210882

Kane, J. W., & Gowland, E. (1986). A method for the identification of apolipoprotein E isoforms employing chemical precipitation and flat bed isoelectric focusing in agarose. *Ann Clin Biochem, 23 (Pt 5)*, 509-513. doi:10.1177/000456328602300504

Koch, W., Ehrenhaft, A., Griesser, K., Pfeufer, A., Muller, J., Schomig, A., & Kastrati, A. (2002). TaqMan systems for genotyping of disease-related polymorphisms present in the gene encoding apolipoprotein E. *Clin Chem Lab Med, 40*(11), 1123-1131. doi:10.1515/CCLM.2002.197

Lai, M. C., Bechy, A. L., Denk, F., Collins, E., Gavriliouk, M., Zaugg, J. B., . . . Caffrey, T. M. (2017). Haplotype-specific MAPT exon 3 expression regulated by common intronic polymorphisms associated with Parkinsonian disorders. *Mol Neurodegener, 12*(1), 79. doi:10.1186/s13024-017-0224-6

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., . . . Amouyel, P. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet, 45*(12), 1452-1458. doi:10.1038/ng.2802ng.2802 [pii]

Li, H., Wetten, S., Li, L., St Jean, P. L., Upmanyu, R., Surh, L., . . . Roses, A. D. (2008). Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol, 65*(1), 45-53. doi:2007.3 [pii]10.1001/archneurol.2007.3

Li, Y. J., Pericak-Vance, M. A., Haines, J. L., Siddique, N., McKenna-Yasek, D., Hung, W. Y., . . . Siddique, T. (2004). Apolipoprotein E is associated with age at onset of amyotrophic lateral sclerosis. *Neurogenetics, 5*(4), 209-213. doi:10.1007/s10048-004-0193-0

Liu, Y., Yu, J. T., Wang, H. F., Han, P. R., Tan, C. C., Wang, C., . . . Tan, L. (2015). APOE genotype and neuroimaging markers of Alzheimer's disease: systematic review and meta-analysis. *J Neurol Neurosurg Psychiatry, 86*(2), 127-134. doi:10.1136/jnnp-2014-307719jnnp-2014-307719 [pii]

Mahley, R. W., & Rall, S. C., Jr. (2000). Apolipoprotein E: far more than a lipid transport protein. *Annu Rev Genomics Hum Genet, 1*, 507-537. doi:1/1/507 [pii]10.1146/annurev.genom.1.1.507

Miyata, M., & Smith, J. D. (1996). Apolipoprotein E allele-specific antioxidant activity and effects on cytotoxicity by oxidative insults and beta-amyloid peptides. *Nat Genet, 14*(1), 55-61. doi:10.1038/ng0996-55

Mui, S., Rebeck, G. W., McKenna-Yasek, D., Hyman, B. T., & Brown, R. H., Jr. (1995). Apolipoprotein E epsilon 4 allele is not associated with earlier age at onset in amyotrophic lateral sclerosis. *Ann Neurol, 38*(3), 460-463. doi:10.1002/ana.410380318

Myers, A. J., Kaleem, M., Marlowe, L., Pittman, A. M., Lees, A. J., Fung, H. C., . . . Hardy, J. (2005). The H1c haplotype at the MAPT locus is associated with Alzheimer's disease. *Hum Mol Genet, 14*(16), 2399-2404. doi:ddi241 [pii]10.1093/hmg/ddi241

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., . . . Singleton, A. B. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet, 46*(9), 989-993. doi:10.1038/ng.3043ng.3043 [pii]

Pankratz, N., Byder, L., Halter, C., Rudolph, A., Shults, C. W., Conneally, P. M., . . . Nichols, W. C. (2006). Presence of an APOE4 allele results in significantly earlier onset of Parkinson's disease and a higher risk with dementia. *Mov Disord, 21*(1), 45-49. doi:10.1002/mds.20663

Petersen, R. C., Thomas, R. G., Grundman, M., Bennett, D., Doody, R., Ferris, S., . . . Thal, L. J. (2005). Vitamin E and donepezil for the treatment of mild cognitive impairment. *N Engl J Med, 352*(23), 2379-2388. doi:NEJMoa050151 [pii]10.1056/NEJMoa050151

Robinson, J. L., Lee, E. B., Xie, S. X., Rennert, L., Suh, E., Bredenberg, C., . . . Trojanowski, J. Q. (2018). Neurodegenerative disease concomitant proteinopathies are prevalent, age-related and APOE4-associated. *Brain, 141*(7), 2181-2193. doi:10.1093/brain/awy146

Saunders, A. M., Strittmatter, W. J., Schmechel, D., George-Hyslop, P. H., Pericak-Vance, M. A., Joo, S. H., . . . et al. (1993). Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology, 43*(8), 1467-1472. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8350998

Seto-Salvia, N., Clarimon, J., Pagonabarraga, J., Pascual-Sedano, B., Campolongo, A., Combarros, O., . . . Kulisevsky, J. (2011). Dementia risk in Parkinson disease: disentangling the role of MAPT haplotypes. *Arch Neurol, 68*(3), 359-364. doi:10.1001/archneurol.2011.1768/3/359 [pii]

Simon-Sanchez, J., Schulte, C., Bras, J. M., Sharma, M., Gibbs, J. R., Berg, D., . . . Gasser, T. (2009). Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet, 41*(12), 1308-1312. doi:10.1038/ng.487ng.487 [pii]

Verghese, P. B., Castellano, J. M., & Holtzman, D. M. (2011). Apolipoprotein E in Alzheimer's disease and other neurological disorders. *Lancet Neurol, 10*(3), 241-252. doi:10.1016/S1474-4422(10)70325-2S1474-4422(10)70325-2 [pii]

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res, 38*(16), e164. doi:10.1093/nar/gkq603gkq603 [pii]

Ward, A., Crean, S., Mercaldi, C. J., Collins, J. M., Boyd, D., Cook, M. N., & Arrighi, H. M. (2012). Prevalence of apolipoprotein E4 genotype and homozygotes (APOE e4/4) among patients diagnosed with Alzheimer's disease: a systematic review and meta-analysis. *Neuroepidemiology, 38*(1), 1-17. doi:10.1159/000334607000334607 [pii]

Yang, D. S., Smith, J. D., Zhou, Z., Gandy, S. E., & Martins, R. N. (1997). Characterization of the binding of amyloid-beta peptide to cell culture-derived native apolipoprotein E2, E3, and E4 isoforms and to isoforms from human plasma. *J Neurochem, 68*(2), 721-725. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9003062

Zannis, V. I., Breslow, J. L., Utermann, G., Mahley, R. W., Weisgraber, K. H., Havel, R. J., . . . Blum, C. (1982). Proposed nomenclature of apoE isoproteins, apoE genotypes, and phenotypes. *J Lipid Res, 23*(6), 911-914. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7130859

Zhong, N., & Weisgraber, K. H. (2009). Understanding the association of apolipoprotein E4 with Alzheimer disease: clues from its structure. *J Biol Chem, 284*(10), 6027-6031. doi:10.1074/jbc.R800009200

Chapter 3 – Association of apolipoprotein E variation with cognitive impairment across multiple neurodegenerative diagnoses.

The work presented in <u>Chapter 3</u> has been edited from the original publication in *Neurobiology of Aging* for brevity and consistency throughout the entire Dissertation.

Dilliott A.A., Sunderland K.M., McLaughlin P.M., Roberts A.C., Evans E.C., Abrahao A., Binns M.A., Black S.E., Borrie M., Casaubon L.K., Dowlatshahi D., Finger E., Fischer C.E., Frank A., Freedman M., Grimes D., Hassan A., Jog M., Kumar S., Kwan D., Lang A.E., Mandzia J., Marras C., Masellis M., McIntyre A.D., Pasternak S., Pollock D.G., Rajji T.K., Robinson J.F., Rogaeva E., Sahlas D.J., Saposnik G., Sato C., Seitz D., Shoesmith C., Steeves T., Strother S.C., Swartz R.H., Tan B., Tang-Wai D., Tartaglia M.C., Troyer A.K., Turnbull J., Zinman L., ONDRI Investigators, and Hegele R.A. (2021). Association of apolipoprotein E variation with cognitive impairment across multiple neurodegenerative diagnoses. *Neurobiol Aging*, 105:378.e1-378.e9

3.1.   Abstract

Background/Objective: For many years, there has been uncertainty regarding how apolipoprotein E (*APOE*) E2 and E4 variants may influence overlapping features of neurodegeneration, such as cognitive impairment. We aimed to identify whether the APOE variants are associated with cognitive function across various neurodegenerative and cerebrovascular diagnoses (n=513).

Methods and Results: Utilizing a comprehensive neuropsychology battery, multivariate multiple regression was used to assess the influence of *APOE* carrier status and disease cohort on performance across five cognitive domains. Irrespective of disease cohort, E4 carriers had significantly lower performance in verbal memory and visuospatial domains than those with E3/3, while E2 carriers' cognitive performance was not significantly different. However, E2 carriers with frontotemporal dementia (FTD) performed significantly worse than those with E3/3 in the attention/working memory, executive function, and visuospatial domains.

Conclusion: Our results highlight that the influence of *APOE* variation on cognition is complex, in some cases varying based on diagnosis and possibly underlying disease pathology.

3.2. Introduction

Apolipoprotein E (*APOE*) is located on chromosome 19q13 and encodes a lipoprotein component responsible for the transportation of cholesterol, both in plasma and within the central nervous system. The formed apo E protein exists as one of three isoforms — E3, otherwise considered the wild-type isoform, E2, and E4 — defined by two single nucleotide polymorphisms within the gene at amino acid positions 130 and 176 (Zannis et al., 1982). Importantly, the *APOE* E4 allele displays a high-risk association with sporadic Alzheimer's disease (AD), in a dose dependent manner (Ward et al., 2012). Previously, we analyzed the contribution of the variant to multiple neurodegenerative phenotypes (Dilliott et al., 2019). Consistent with previous studies, we reported an association with AD. We also observed an association between E4 and amnestic mild cognitive impairment (MCI) presentation; however, we did not observe associations between *APOE* and the other neurodegenerative diseases under study, including amyotrophic lateral sclerosis (ALS), cerebrovascular disease (CVD; previously referred to as vascular cognitive impairment), frontotemporal dementia (FTD), or Parkinson's disease (PD).

While the association between *APOE* and sporadic AD is well established, the exact mechanism of pathology is not fully understood. Apo E is an amyloid-β (Aβ) chaperone and modulates its metabolism, aggregation, and deposition by directly binding to the Aβ peptide (T. C. Dickson, Saunders, & Vickers, 1997; Kanekiyo, Xu, & Bu, 2014), and controlling its production by regulating neuronal cholesterol levels (Osenkowski, Ye, Wang, Wolfe, & Selkoe, 2008). However, these functions are influenced by the isoform of apo E. Apo E4 has been found to promote Aβ oligomer aggregation and prevent its clearance, thereby resulting in Aβ accumulation into amyloid plaques (Deane et al., 2008; T. Hashimoto et al., 2012; Liu et al., 2017). In addition to its effects on Aβ, apo E4 has also been proposed to contribute to neurodegenerative disease pathology by increasing tau aggregates in brains with existing Aβ pathology (Farfel, Yu, De Jager, Schneider, & Bennett, 2016), as well as tau phosphorylation in mouse models (Brecht et al., 2004; Shi et al., 2017), indicating that the isoform may also play a role in the presentation of tauopathies. Similarly, in individuals with Lewy body disease, apo E4 was associated

with increased α-synuclein pathology, irrespective of tau and Aβ, suggesting a role in synucleinopathies as well (D. W. Dickson et al., 2018). In contrast, although apo E2 is well known for its reduced binding affinity for the low-density lipoprotein receptor, causing hyperlipoproteinemia type III when inherited in the homozygous state (Rall, Weisgraber, Innerarity, & Mahley, 1982), the isoform was also associated with increased tau pathology in mouse models of FTD (Zhao et al., 2018).

Although a large amount of effort has been put forth in determining the contribution of *APOE* genotypes to neurodegenerative disease development and pathology, how the variants influence features of disease has long been a source of uncertainty (Swartz, Black, & St George-Hyslop, 1999), and still remains to be fully elucidated. In particular, how *APOE* may contribute to overlapping features of neurodegeneration — such as cognitive impairment — remains unclear with previous analyses reporting conflicting results. Specifically, researchers have reported that the E4 allele is associated with reduced performance on measures of verbal and episodic memory, executive function, and speed of processing in healthy elderly cohorts (O'Donoghue, Murphy, Zamboni, Nobre, & Mackay, 2018); memory and attention in probable AD patients (M. Hashimoto et al., 2001; Lehtovirta et al., 1996; Marra et al., 2004; van der Vlies et al., 2007); executive function in amnestic MCI patients (Seo et al., 2016); and memory, attention, and executive function in patients with PD (Mata et al., 2014). However, other studies have reported no difference in cognition between *APOE* E4 carriers and non-carriers in healthy elderly cohorts (O'Donoghue et al., 2018), or AD (Lehtovirta et al., 1996; van der Vlies et al., 2007; Wolk, Dickerson, & Alzheimer's Disease Neuroimaging, 2010) and PD (Mengel et al., 2016) patients.

In contrast, the *APOE* E2 allele — commonly accepted as protective against the development of AD (Dilliott et al., 2019) — has been associated with increased risk of FTD, specifically tauopathies, such as progressive supranuclear palsy (PSP) and corticobasal degeneration (CBD) (Verpillat et al., 2002; Zhao et al., 2018). It is also associated with an increased co-occurrence of FTD with a primary diagnosis of ALS (ALS-FTSD) (Chio et al., 2016). To date, studies have also presented inconsistent findings regarding the relationship between *APOE* status and ischaemic stroke outcomes

(Klages, Fisk, & Rockwood, 2005; Pendlebury et al., 2020; Verghese, Castellano, & Holtzman, 2011), including emerging evidence of a potentially complex interaction between *APOE* E4, cerebrovascular burden, and cognition that may offer a new lens through which to examine the effect of *APOE* status in these groups (Mirza et al., 2019).

The Ontario Neurodegenerative Disease Research Initiative (ONDRI) is a prospective longitudinal cohort study with the aim of developing deep phenotypes of 1) AD; 2) ALS; 3) CVD; 4) FTD; 5) amnestic MCI; and 6) PD, and understanding potential cerebrovascular disease contributors to neurodegeneration (Farhan et al., 2017; Sunderland et al., 2020). The ONDRI protocol includes a comprehensive neuropsychological battery (McLaughlin et al., 2020) and well-established clinical diagnostic criteria, affording the opportunity to examine *APOE* variants cross-sectionally in rigorously characterized cohorts using converging information across genetics, clinical, and neuropsychological platforms. In doing so, we overcome limitations of other studies that report *APOE* associations using a) a singular disease cohort that are unable to identify overlapping features of neurodegeneration and b) cognitive characterizations based on a restricted set of domains and sub-clinical standards for determining impairment (i.e., use of a single test to label a domain as impaired) that may underestimate the nature of and degree of cognitive impairment.

In this paper, we aimed to identify whether carriers of the *APOE* E2 or E4 variants display differing cognitive function across multiple domains compared to those with the wild-type *APOE* E3/3 genotype across neurodegenerative and cerebrovascular diseases.

3.3.   Methods

3.3.1.   Study participants

In total, 520 participants previously clinically diagnosed with a neurodegenerative or cerebrovascular disease were recruited into ONDRI from 14 tertiary care centers across Ontario. Diseases included: 1) AD; 2) ALS; 3) CVD; 4) FTD (including behavioural variant FTD [bvFTD], corticobasal syndrome [CBS], PSP, and primary progressive aphasia (PPA, including non-fluent and semantic variant); 5) amnestic MCI; and 6) PD.

Based on the Montreal Cognitive Assessment (MoCA) administered during an initial study screening visit, the CVD and PD cohorts aimed to include a mix of individuals with (<26) and without (≥26) cognitive impairment. MoCA scores across disease cohorts are outlined in **Table 3.1**. Ethics approval was obtained from each of the participating sites. Descriptions of the inclusion/exclusion criteria of ONDRI participants were previously reported (Farhan et al., 2017; Sunderland et al., 2020). All participants provided informed written consent.

Table 3.1 Distribution of general cognition in all enrolled ONDRI participants, as determined by MoCA score.

Cohort	Total number of participants	Mean MoCA (score ± sd)	Max. MoCA	Min. MoCA
ONDRI	520[I]	24.4 ± 3.4	30	13
AD	41	20.9 ± 2.8	26	15
ALS	40[I]	25.5 ± 2.8	30	19
CVD	161	25.2 ± 3.0	30	18
FTD	53	21.5 ± 3.9	29	13
MCI	85	23.5 ± 2.7	30	18
PD	140[II]	25.8 ± 2.6	30	18

[I]One participant did not complete MoCA testing.

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; Max., maximum; MCI, mild cognitive impairment; Min., minimum; MoCA, Montreal Cognitive Assessment; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease; sd, standard deviation.

3.3.2. *APOE* genotyping

Participant blood samples were obtained and genomic DNA was isolated from blood and subsequently sequenced using ONDRISeq, a custom-designed next-generation sequencing gene panel that targets 80 genes previously associated with the disease cohorts in ONDRI (Dilliott et al., 2018; Farhan et al., 2016).

APOE genotypes were obtained from the ONDRISeq data using a customized Annotate Variation (ANNOVAR) script to extract allele calls for the *APOE* variants rs429358(CT):p.Cys130Arg and rs7412(CT):p.Arg176Cys and map to the respective genotype for each participant (Dilliott et al., 2019; Dilliott et al., 2018). All *APOE* genotypes were subsequently validated using a TaqMan allelic discrimination assay. Genotypes were used to bin participants into three groups based on carrier status: 1) E2 carriers (E2/2, E3/2); 2) E3/3; and 3) E4 carriers (E4/3, E4/4). All participants harbouring an E4/2 genotype were excluded from the study.

3.3.3. Neuropsychology assessment

Participants completed a comprehensive neuropsychological battery (McLaughlin et al., 2020) that included 23 measures from 14 neuropsychological tests across five cognitive domains: 1) attention/working memory; 2) executive function; 3) language; 4) verbal memory; and 5) visuospatial abilities (**Table 3.2**). Cognitive domains and associated test measures were grouped together based on general neuropsychology conventions (Lezak, Howieson, Loring, Hannay, & Fischer, 2004) and consensus agreement among the ONDRI Clinical Neuropsychologists.

Domain scores were created by averaging the standardized residuals of a linear model for each raw test score with age, sex, years of education, and, where necessary, assessment version as main effects, across variables within a domain. All timed scores were multiplied by -1 such that faster response times (higher scores) reflected better performance.

Prior to creating domain scores, missing neuropsychology data were imputed using a regularized iterative principal component analysis (PCA) process from the missMDA package in R (Josse & Husson, 2016). Briefly, using the standardized residuals starting values were imputed and a PCA was performed. From the PCA, a set of fitted values was derived, and a new PCA was performed using the fitted values as updated imputed values. The number of components retained for calculating the fitted values was determined based on the minimum mean square predicted error of the observed values. The process was repeated until the overall change in fitted values was negligible.

Starting values were the observed variable mean (adjusted for age, sex, and years of education of the participant) for values missing for reasons unrelated to the participant, and the extreme score when the participant was unable to complete the task because of difficulty understanding the task instructions or performing the task as intended.

Both the *APOE* and Neuropsychology datasets underwent rigorous quality assurance and quality control processes to ensure accuracy (Dilliott et al., 2018; McLaughlin et al., 2020; Sunderland et al., 2019).

Table 3.2 Binning strategy of the neuropsychological tests into cognitive domains.

Cognitive Domain	Attention and Working Memory	Executive Function	Language	Verbal Memory	Visuospatial Awareness
Neuropsychological Tests	Symbol Digit Modality Test (coding)	Trail Making Test – Part B (time)	Boston Naming – 15 Item (pro-rated)	RAVLT: Immediate	Judgement of Line Orientation
	Trail Making Test – Part A (time)	DKEFS: Interference (time)	TAWF: Verb Naming	RAVLT: Long-delay	VOSP: Incomplete Letters
	WAIS-III: Digit Span Forward (longest span)	DKEFS: Inhibition/ Switching (time)	BDAE: Semantic Probe	RAVLT: Recognition Discrimination	BVMT-R: Copy Trial (raw)
	WAIS-III: Digit Span Backward (longest span)	DKEFS: Letter Fluency	WASI-II: Vocabulary		
	WAIS-III: Digit Span Total	DKEFS: Category Fluency			
	DKEFS: Color naming (time)	WASI-II: Matrix			
	DKEFS: Word reading (time)				

Abbreviations: BDAE, Boston Diagnostic Aphasia Examination; BVMT-R, Brief Visuospatial Memory Test-Revised; DKEFS, Delis-Kaplan Executive Function System; RAVLT, Rey Auditory Verbal Learning Test; VOSP, Visual Object and Space Perception Battery; WAIS-III, Wechsler Adult Intelligence Scale – Third Edition; WASI-II, Wechsler Abbreviated Scale Intelligence – Second Edition.

3.3.4.   Statistical analysis

A multivariate multiple regression model was used to estimate effects across the five cognitive domains as a function of *APOE* carrier status and disease cohort, while accounting for interactions between the two predictor variables (Johnson & Wichern, 2001). Wilks' lambda assessed predictor contributions of the five parallel equations, where significance indicated that the predictor contribution was non-zero for at least one of the five cognitive domains. To determine the extent to which domain(s) a predictor contributed, the individual coefficients were examined.

Participant *APOE* carrier statuses were transformed into dummy variables with E3/3 as the reference category. Participant disease cohorts were transformed with weighted-effect coding, using the *wec* R package (v0.4-1) (Nieuwenhuis et al., 2017). With weighted-effect coding, the point of reference is the sample mean of all participants, regardless of cohort. While the parameter for one cohort (in this case, CVD was chosen) must be excluded from the model to avoid statistical redundancy, the coefficient estimate can still be calculated (te Grotenhuis et al., 2017a, 2017b). Significance for the multivariate multiple regression model was measured at an alpha-level of 5.0e-2, although non-significant trends were reported in the results at $p < 7.5e\text{-}2$.

Statistical analyses were performed using the R statistical software 3.6.0 (R Core Team, 2014) in R Studio 1.1.463 and data visualization was performed using the *ggplot2* R package (v3.3.s) (Wickham, 2009).

3.4.   Results

3.4.1.   Study participants

In total, 519 of the 520 participants enrolled in ONDRI were included in the genetic analysis, as one participant did not complete the genomics assessment. Of the 519, six with the *APOE* E4/2 genotype were excluded. Of the 513 participants remaining, we identified that 292 (56.9%) participants harboured the E3/3 genotype, 59 (11.5%)

participants were *APOE* E2 carriers, and 162 (31.6%) participants were *APOE* E4 carriers (**Table 3.3**). Neuropsychology data were imputed using a regularized iterative PCA for 176 missing values (1.5% of all cells) across 100 participants (19.5%) that were distributed amongst the six cohorts.

Table 3.3 Demographics and *APOE* carrier status of the ONDRI participants included in the *APOE* cognitive impairment analysis.

Cohort	Participants included in *APOE* analysis	Mean age (years ± sd)	Male:Female	*APOE* E3/3	*APOE* E2 carriers	*APOE* E4 carriers
ONDRI	513	68.6 ± 7.6	341:172	292 (56.9%)	59 (11.5%)	162 (31.6%)
AD	40	71.9 ± 8.1	23:17	14 (35.0%)	0	26 (65.0%)
ALS	40	62.0 ± 8.7	24:16	21 (52.5%)	6 (15.0%)	13 (32.5%)
CVD	159	69.2 ± 7.4	109:50	100 (62.9%)	21 (13.2%)	38 (23.9%)
FTD	53	67.8 ± 7.1	34:19	26 (49.1%)	7 (13.2%)	20 (37.7%)
bvFTD	21	65.9 ± 8.8	14:7	11	0	10
nfPPA	8	68.4 ± 5.9	5:3	4	1	3
PSP	15	69.7 ± 6.2	10:6	6	6	4
MCI	85	70.6 ± 8.3	45:40	45 (52.9%)	5 (5.9 %)	35 (41.2%)
PD	136	67.8 ± 6.4	106:30	86 (63.2%)	20 (14.7%)	30 (22.1%)

APOE E2 carriers are those harbouring the E3/2 or E2/2 genotypes, whereas *APOE* E4 carriers are those harbouring the E4/4 or E4/3 genotypes. Individuals harbouring the *APOE* E4/2 genotype were excluded from the cognitive analysis. The patient diagnosed with bvFTD and PSP was included in the PSP cohort. Demographics and APOE carrier status of the CBS and svPPA cohorts are not displayed due to the small sample sizes of the groups (n ≤ 5).

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; *APOE*, Apolipoprotein E gene; bvFTD, behavioural variant frontotemporal dementia; CBS, corticobasal syndrome; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; MCI, mild cognitive impairment; nfPPA, non-fluent primary progressive aphasia; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease; PSP, progressive supranuclear palsy; sd, standard deviation; svPPA, semantic variant primary progressive aphasia.

3.4.2. Influence of *APOE* across all neurodegenerative diseases

All coefficient estimates and standard errors obtained by the multivariate multiple regression model are summarized in **Table 3.4**. A Wilks' lambda test on the multivariate multiple regression analysis confirmed contribution to the model from *APOE* carrier status ($\Lambda = 0.937$, p = 3.58e-4), disease cohort ($\Lambda = 0.705$, p < 2.20e-16), and their interaction ($\Lambda = 0.874$, p = 1.81e-2).

Figure 3.1 shows the magnitude of difference in cognition of participants carrying E4 or E2 in relation to E3 homozygotes across ONDRI as a whole and all disease cohorts individually for each cognitive domain. To better illustrate these changes, disease effects were removed.

Combining all participants in a single group revealed that *APOE* E4 carriers had significantly lower performance in verbal memory (p = 9.22e-3; **Figure 3.1I**) and visuospatial abilities (p = 2.20e-2; **Figure 3.1J**) compared to those with the *APOE* E3/3 genotype. In contrast, *APOE* E2 carriers displayed similar cognitive performance across domains to those with the *APOE* E3/3 genotype.

Including an interaction term in the model also allowed us to compare cognitive performance of the two *APOE* variant carrier groups (E4 and E2) to the *APOE* E3/3 carriers in each individual disease cohort to identify differences between *APOE* carrier status and disease group.

Table 3.4 Coefficient estimates (SE) of z-score transformed cognitive domain scores for participants in each neurodegenerative disease cohort, and in the total cohort, based on *APOE* carrier status obtained using multivariate multiple regression.

Cognitive Domain	Attention and Working Memory	Executive Function	Language	Verbal Memory	Visuospatial Abilities
Intercept	0.013 (0.040)	-0.012 (0.044)	-0.020 (0.050)	0.057 (0.047)*	0.056 (0.043)
APOE E2	-0.041 (0.097)	-0.0801 (0.107)	0.081 (0.121)	0.014 (0.114)	-0.147 (0.104)
APOE E4	-0.036 (0.070)	-0.029 (0.077)	-0.024 (0.087)	-0.215 (0.082)***	-0.173 (0.0752)*
AD	-0.222 (0.170)	-0.340 (0.188)	-0.730 (0.211)***	-0.475 (0.199)**	-0.070 (0.182)
AD x *APOE* E2	NA	NA	NA	NA	NA
AD x *APOE* E4	0.056 (0.196)	0.049 (0.217)	0.384 (0.245)	-0.419 (0.230)	-0.158 (0.211)
ALS	0.076 (0.140)	0.181 (0.155)	-0.004 (0.175)	0.062 (0.165)	0.080 (0.151)
ALS x *APOE* E2	0.482 (0.294)	0.623 (0.325)	0.447 (0.366)	0.054 (0.345)	0.279 (0.315)
ALS x *APOE* E4	-0.116 (0.228)	-0.199 (0.252)	0.058 (0.284)	0.219 (0.238)	0.166 (0.245)
FTD	-0.642 (0.124)***	-0.767 (0.137)***	-0.854 (0.154)***	-0.649 (0.145)***	-0.213 (0.133)
FTD x *APOE* E2	-0.827 (0.268)**	-0.605 (0.296)*	0.779 (0.334)*	0.260 (0.314)	-1.051 (0.287)***
FTD x *APOE* E4	-0.100 (0.187)	-0.076 (0.207)	0.190 (0.233)	0.486 (0.219)*	0.145 (0.201)
MCI	0.050 (0.091)	0.005 (0.101)	0.135 (0.113)	-0.190 (0.107)	-0.085 (0.098)
MCI x *APOE* E2	0.328 (0.304)	0.175 (0.336)	0.036 (0.379)	0.574 (0.357)	0.385 (0.327)
MCI x *APOE* E4	0.052 (0.134)	0.132 (0.148)	-0.101 (0.167)	-0.093 (0.157)	0.081 (0.144)
PD	0.230 (0.064)	0.238 (0.070)***	0.184 (0.079)*	0.254 (0.075)***	0.086 (0.068)
PD x *APOE* E2	-0.062 (0.135)	-0.238 (0.150)	-0.224 (0.169)	-0.265 (0.159)	0.033 (0.145)
PD x *APOE* E4	-0.021 (0.133)	-0.080 (0.147)	-0.155 (0.166)	0.113 (0.156)	-0.227 (0.143)
CVD +/- CI	0.027 (0.057)	0.090 (0.064)	0.239 (0.072)	0.205 (0.067)	0.040 (0.062)
CVD +/- CI x *APOE* E2	0.119 (0.130)	0.209 (0.143)	-0.183 (0.162)	0.014 (0.152)	0.148 (0.139)
CVD +/- CI x *APOE* E4	0.023 (0.117)	0.017 (0.129)	-0.166 (0.146)	-0.048 (0.137)	0.080 (0.125)

APOE carrier statuses were transformed into dummy variables with E3/3 as the reference. Participant disease cohorts were transformed using weighted-effect coding using the CVD ± CI cohort as the reference. The model was rerun using the ALS cohort as the reference in the weighted-effect coding to obtain coefficient estimates for the CVD ± CI cohort, as displayed below the line. *p-value < 0.05; **p-value < 0.005; ***p-value < 0.0005.

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; *APOE*, Apolipoprotein E gene; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; MCI mild cognitive impairment; PD, Parkinson's disease; SE, standard error.

3.4.3. Influence of *APOE* in the individual disease cohorts

Considering individual disease cohorts, direction of association based on the *APOE* E4 allele was not uniform across disease cohorts. E4 carrier status was significantly associated with *better* verbal memory performance in the FTD cohort relative to those harbouring the E3/3 genotype (p = 2.72e-2; **Figure 3.1I**). In contrast, in the AD cohort poorer verbal memory performance by *APOE* E4 carriers approached significance (p = 6.97e-2; **Figure 3.1I**). No other differences of interest were observed between *APOE* E4 and E3/3 across disease cohorts.

The only significant interaction effect of *APOE* E2 was observed in the FTD cohort. FTD patients harbouring the E2 allele performed significantly better in the language domain (p = 2.00e-2; **Figure 3.1C**), yet significantly worse in the attention/working memory (p = 2.12e-3; **Figure 3.1A**), executive function (p = 4.17e-2; **Figure 3.1B**), and visuospatial domains (p = 2.37e-4; **Figure 1E**), compared to patients harbouring the E3/3 genotype. In the ALS cohort, better executive function performance by *APOE* E2 carriers approached significance (p = 5.58e-2; **Figure 3.1B**).
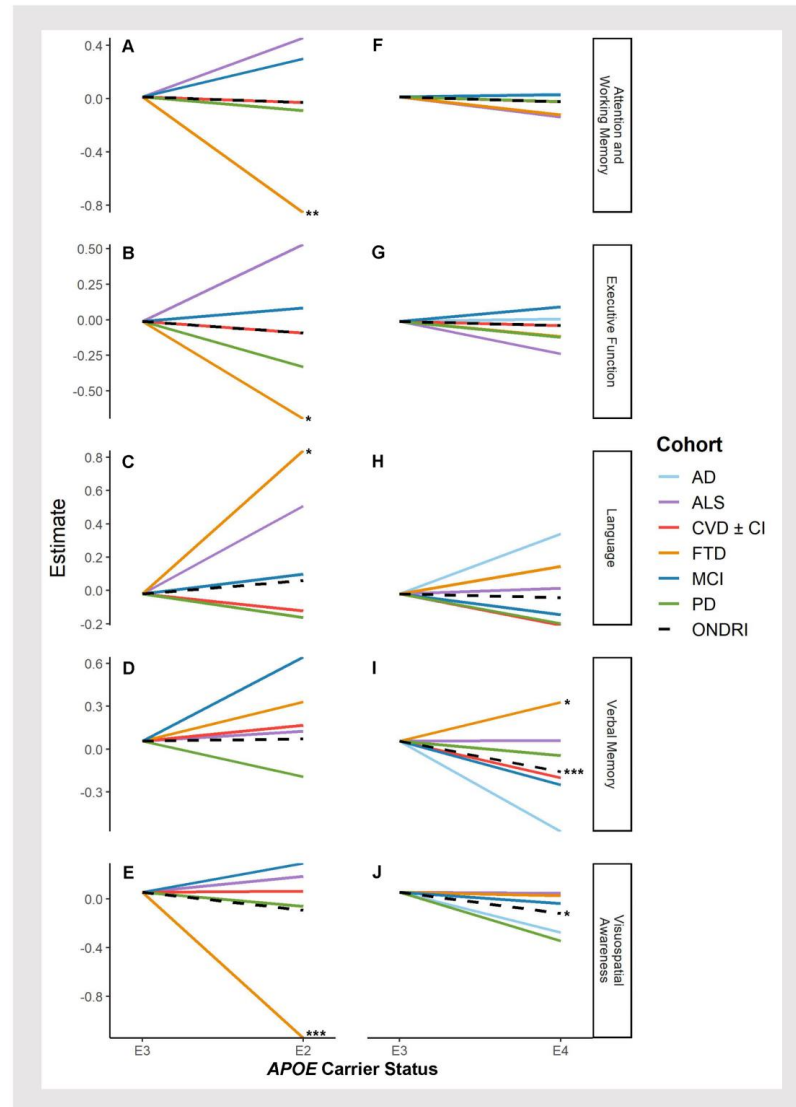
Figure 3.1 Estimates based *APOE* variant carrier status in the neurodegenerative disease cohorts encompassed by ONDRI with respect to performance on neuropsychology tests binned into cognitive domains.
From the neuropsychology core dataset, 23 raw scores were binned into five cognitive domains, raw scores were standardized into z-scores, and mean z-scores were derived for each cognitive domain per participant. Multivariate multiple regression analysis was used to model performance across the cognitive domains as a function of *APOE* carrier status, disease cohort, and interactions between the predictor variables. Estimates for differences between groups were calculated based on the intercept, coefficient estimates from *APOE* carrier status, and coefficient estimates from the *APOE* carrier status and disease cohort interaction effect. *p-value < 0.05; **p-value < 0.005; ***p-value < 0.0005.
Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; *APOE*, Apolipoprotein E gene; CVD ± CI, cerebrovascular disease with or without cognitive impairment; E2, E2 carriers; E3, E3/3 carriers; E4, E4 carriers; FTD, frontotemporal dementia; MCI mild cognitive impairment; ONDRI, Ontario Neurodegenerative Disease Research Initiative; PD, Parkinson's disease.

3.5.   Discussion

Here, we investigated the influence of the *APOE* E4 and E2 alleles on cognitive functioning in multiple cognitive domains across neurodegenerative and cerebrovascular disease cohorts.

Generally, carriers of the E4 allele had an estimated significantly lower performance in the verbal memory and visuospatial domains than those with E3/3. Multiple studies have reported associations between the *APOE* E4 allele and poor memory performance, including adults who were considered cognitively normal for their age (Baxter, Caselli, Johnson, Reiman, & Osborne, 2003; Caselli et al., 2009; Caselli et al., 2007). Combined with the results that we observed across all participants regardless of cohort in the verbal memory domain, this suggests that *APOE* E4 may influence memory performance regardless of disease status, independent of age and sex. Two hypotheses have been put forth in the literature to account for these findings (O'Donoghue et al., 2018). The first suggests that *APOE* E4 carriers are in a prodromal stage of AD, with the Apo E4 isoform increasing the deposition of early-AD pathology, such as Aβ and tau, thereby indirectly influencing memory function. The second hypothesis suggests that the Apo E4 isoform directly influences cognition, in a mechanism unrelated to AD pathology – possibly via a role of Apo E in synaptic plasticity and neurotoxicity. Perhaps both mechanisms contribute, but this must be further elucidated (O'Donoghue et al., 2018). Our data showed an intriguing exception in the FTD cohort, where participants with E4 performed better than those with the E3/3 genotype; a reverse effect than what was observed in the E4 group effect. Yet, FTD patients carrying an E4 allele still performed below the group average for *APOE* E4 carriers across all ONDRI samples in the verbal memory domain (**Table 3.4**). In addition, while we observed poorer performance in E4 carriers in comparison to E3/3 AD patients, the interaction term did not reach statistical significance, as we expected it to (M. Hashimoto et al., 2001; Lehtovirta et al., 1996; Marra et al., 2004; van der Vlies et al., 2007). Due to the high proportion of E4 carriers in the AD cohort (65%) and the large influence from an AD diagnosis and E4 individually (β = -0.475, p = 1.73e-2 and β = -0.215, p = 9.22e-3, respectively), there may not have

been enough power to achieve statistical significance despite the relatively large interaction estimate (β = -0.419).

In the FTD cohort, carriers of the *APOE* E2 allele performed significantly worse in multiple cognitive domains, including attention/working memory, executive function, and visuospatial abilities. Zhao *et al.* identified an association between *APOE* E2 and increased neurofibrillary tau pathology (Zhao et al., 2018), which is a hallmark of certain subtypes of FTD — including PSP, CBD, and FTD with tau pathology — and is a risk factor for increased cognitive impairment (Koga et al., 2017). Further, the *APOE* E2/2 genotype was previously directly associated with increased risk of PSP and CBD (Zhao et al., 2018) and the E2 allele with increased risk of ALS-FTSD (Chio et al., 2016). Within ONDRI, of the seven FTD patients carrying an E2 allele, five were diagnosed with PSP (the Richardson phenotype, PSP-RS) and one was diagnosed with both bvFTD and PSP (**Table 3.3**). Individuals with PSP often present with a high degree of cognitive impairment, particularly in executive function and speed of processing, verbal fluency, and visuospatial abilities (Bak, Crawford, Hearn, Mathuranath, & Hodges, 2005). Therefore, it is possible that we observed an increased presence of the PSP subtype — and cognitive impairment — as a result of increased neurofibrillary tau pathology in individuals carrying *APOE* E2. Unfortunately, neuropathology data are currently unavailable for ONDRI patients to assess tau pathology; however, the PSP-RS phenotype is highly predictive of underlying tau pathology (Kovacs et al., 2020). In contrast, E2 carriers in the FTD cohort performed significantly better than patients with *APOE* E3/3 on memory measures. No other disease cohorts displayed significant interaction effects on cognitive function between the E2 allele and disease diagnosis; however, the effects could not be assessed in the AD cohort due to a lack of E2 carriers. This observation was expected as the E2 allele is commonly accepted as protective against AD (Dilliott et al., 2019).

We recognize this study's limitations, including the relatively modest sample sizes of the individual disease cohorts, which limited the ability to further analyze the subtypes of heterogenous diseases, as was observed in the FTD cohort. Despite this, a goal of the analysis was to evaluate across all neurodegenerative diagnoses and account for

overlapping disease features. The cross-cohort analysis was made possible by ONDRI's rigorous and consistent enrollment criteria and cross-platform assessment. It is important for this study to be replicated with larger sample sizes, particularly of FTD patients. Limitations were also introduced by not accounting for progression of the ONDRI patient's cognitive impairment over time and the lack of access to post-mortem neuropathology to confirm diagnoses, although these analyses may be revisited in the future if data are available. Finally, we utilized a theory-based framework derived from published guidelines to reduce our neuropsychological measures into five cognitive domains. This theory-driven data reduction approach has inherent limitations, such that some tests may not contribute solely to one domain and there is debate as to how tests should be binned; however, the use of a comprehensive assessment was prioritized as previous analyses on *APOE* have mostly utilized only a limited number of tests accounting for fewer cognitive domains per investigation or introducing single test biases.

3.6.   Conclusion

In summary, our study allowed for simultaneous analysis of the influence of *APOE* on multiple cognitive domains across various neurodegenerative diseases and CVD with rigorous data collection and a comprehensive neuropsychological assessment protocol. Considering all neurodegenerative and cerebrovascular disease patients as a group, we identified an association between the *APOE* E4 allele and worse performance on measures of verbal memory and visuospatial processing. Further, we identified a potential association between the *APOE* E2 allele and poor performance on measures of attention/working memory, executive function, and visuospatial abilities in patients with FTD, which may add to the previously observed associations between the variant and PSP and tau pathology. Future work incorporating the longitudinal ONDRI neuropsychology data will be imperative for understanding how *APOE* influences disease progression. Additionally, including other biomarkers obtained in ONDRI, such as neuroimaging data, may further elucidate the mechanisms by which *APOE* influences cognition across the disease cohorts. Gaining a greater understanding of the contributions of genetic risk factors, such as *APOE* variation to features of neurodegenerative disease,

may present the opportunity for genetic testing to become a progression prediction tool for these typically heterogenous phenotypes.

3.7.   References

Bak, T. H., Crawford, L. M., Hearn, V. C., Mathuranath, P. S., & Hodges, J. R. (2005). Subcortical dementia revisited: similarities and differences in cognitive function between progressive supranuclear palsy (PSP), corticobasal degeneration (CBD) and multiple system atrophy (MSA). *Neurocase, 11*(4), 268-273. doi:10.1080/13554790590962997

Baxter, L. C., Caselli, R. J., Johnson, S. C., Reiman, E., & Osborne, D. (2003). Apolipoprotein E epsilon 4 affects new learning in cognitively normal individuals at risk for Alzheimer's disease. *Neurobiol Aging, 24*(7), 947-952. doi:10.1016/s0197-4580(03)00006-x

Brecht, W. J., Harris, F. M., Chang, S., Tesseur, I., Yu, G. Q., Xu, Q., . . . Huang, Y. (2004). Neuron-specific apolipoprotein e4 proteolysis is associated with increased tau phosphorylation in brains of transgenic mice. *J Neurosci, 24*(10), 2527-2534. doi:10.1523/JNEUROSCI.4315-03.2004

Caselli, R. J., Dueck, A. C., Osborne, D., Sabbagh, M. N., Connor, D. J., Ahern, G. L., . . . Reiman, E. M. (2009). Longitudinal modeling of age-related memory decline and the APOE epsilon4 effect. *N Engl J Med, 361*(3), 255-263. doi:10.1056/NEJMoa0809437

Caselli, R. J., Reiman, E. M., Locke, D. E., Hutton, M. L., Hentz, J. G., Hoffman-Snyder, C., . . . Osborne, D. (2007). Cognitive domain decline in healthy apolipoprotein E epsilon4 homozygotes before the diagnosis of mild cognitive impairment. *Arch Neurol, 64*(9), 1306-1311. doi:10.1001/archneur.64.9.1306

Chio, A., Brunetti, M., Barberis, M., Iazzolino, B., Montuschi, A., Ilardi, A., . . . Calvo, A. (2016). The Role of APOE in the Occurrence of Frontotemporal Dementia in Amyotrophic Lateral Sclerosis. *JAMA Neurol, 73*(4), 425-430. doi:10.1001/jamaneurol.2015.4773

Deane, R., Sagare, A., Hamm, K., Parisi, M., Lane, S., Finn, M. B., . . . Zlokovic, B. V. (2008). apoE isoform-specific disruption of amyloid beta peptide clearance from mouse brain. *J Clin Invest, 118*(12), 4002-4013. doi:10.1172/JCI36663

Dickson, D. W., Heckman, M. G., Murray, M. E., Soto, A. I., Walton, R. L., Diehl, N. N., . . . Ross, O. A. (2018). APOE epsilon4 is associated with severity of Lewy body pathology independent of Alzheimer pathology. *Neurology, 91*(12), e1182-e1195. doi:10.1212/WNL.0000000000006212

Dickson, T. C., Saunders, H. L., & Vickers, J. C. (1997). Relationship between apolipoprotein E and the amyloid deposits and dystrophic neurites of Alzheimer's disease. *Neuropathol Appl Neurobiol, 23*(6), 483-491. doi:10.1111/j.1365-2990.1997.tb01325.x

Dilliott, A. A., Evans, E. C., Farhan, S. M. K., Ghani, M., Sato, C., Zhang, M., . . . Hegele, R. A. (2019). Genetic Variation in the Ontario Neurodegenerative Disease Research Initiative. *Can J Neurol Sci, 46*(5), 491-498. doi:10.1017/cjn.2019.228

Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2018). Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J Vis Exp*(134). doi:10.3791/57266

Farfel, J. M., Yu, L., De Jager, P. L., Schneider, J. A., & Bennett, D. A. (2016). Association of APOE with tau-tangle pathology with and without beta-amyloid. *Neurobiol Aging, 37*, 19-25. doi:10.1016/j.neurobiolaging.2015.09.011

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., . . . Strong, M. J. (2017). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci, 44*(2), 196-202. doi:10.1017/cjn.2016.415

Farhan, S. M. K., Dilliott, A. A., Ghani, M., Sato, C., Liang, E., Zhang, M., . . . Investigators, O. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med, 1*, 16032. doi:10.1038/npjgenmed.2016.32

Hashimoto, M., Yasuda, M., Tanimukai, S., Matsui, M., Hirono, N., Kazui, H., & Mori, E. (2001). Apolipoprotein E epsilon 4 and the pattern of regional brain atrophy in Alzheimer's disease. *Neurology, 57*(8), 1461-1466. doi:10.1212/wnl.57.8.1461

Hashimoto, T., Serrano-Pozo, A., Hori, Y., Adams, K. W., Takeda, S., Banerji, A. O., . . . Hyman, B. T. (2012). Apolipoprotein E, especially apolipoprotein E4, increases the oligomerization of amyloid beta peptide. *J Neurosci, 32*(43), 15181-15192. doi:10.1523/JNEUROSCI.1542-12.2012

Johnson, R. A., & Wichern, D. W. (2001). Applied Multivariate Statistical Analysis. In. New York, USA: Prentice Hall.

Josse, J., & Husson, F. (2016). missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software, 70*(1), 1-31. doi:10.18637/jss.v070.i01

Kanekiyo, T., Xu, H., & Bu, G. (2014). ApoE and Abeta in Alzheimer's disease: accidental encounters or partners? *Neuron, 81*(4), 740-754. doi:10.1016/j.neuron.2014.01.045

Klages, J. D., Fisk, J. D., & Rockwood, K. (2005). APOE genotype, vascular risk factors, memory test performance and the five-year risk of vascular cognitive impairment or Alzheimer's disease. *Dement Geriatr Cogn Disord, 20*(5), 292-297. doi:10.1159/000088317

Koga, S., Parks, A., Kasanuki, K., Sanchez-Contreras, M., Baker, M. C., Josephs, K. A., . . . Dickson, D. W. (2017). Cognitive impairment in progressive supranuclear palsy is associated with tau burden. *Mov Disord, 32*(12), 1772-1779. doi:10.1002/mds.27198

Kovacs, G. G., Lukic, M. J., Irwin, D. J., Arzberger, T., Respondek, G., Lee, E. B., . . . Hoglinger, G. U. (2020). Distribution patterns of tau pathology in progressive supranuclear palsy. *Acta Neuropathol, 140*(2), 99-119. doi:10.1007/s00401-020-02158-2

Lehtovirta, M., Soininen, H., Helisalmi, S., Mannermaa, A., Helkala, E. L., Hartikainen, P., . . . Riekkinen, P. J. (1996). Clinical and neuropsychological characteristics in familial and sporadic Alzheimer's disease: relation to apolipoprotein E polymorphism. *Neurology, 46*(2), 413-419. doi:10.1212/wnl.46.2.413

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). *Neuropsychological assessment* (4th ed.): Oxford University Press.

Liu, C. C., Zhao, N., Fu, Y., Wang, N., Linares, C., Tsai, C. W., & Bu, G. (2017). ApoE4 Accelerates Early Seeding of Amyloid Pathology. *Neuron, 96*(5), 1024-1032 e1023. doi:10.1016/j.neuron.2017.11.013

Marra, C., Bizzarro, A., Daniele, A., De Luca, L., Ferraccioli, M., Valenza, A., . . . Masullo, C. (2004). Apolipoprotein E epsilon4 allele differently affects the patterns of neuropsychological presentation in early- and late-onset Alzheimer's disease patients. *Dement Geriatr Cogn Disord, 18*(2), 125-131. doi:10.1159/000079191

Mata, I. F., Leverenz, J. B., Weintraub, D., Trojanowski, J. Q., Hurtig, H. I., Van Deerlin, V. M., . . . Zabetian, C. P. (2014). APOE, MAPT, and SNCA genes and cognitive performance in Parkinson disease. *JAMA Neurol, 71*(11), 1405-1412. doi:10.1001/jamaneurol.2014.1455

McLaughlin, P. M., Sunderland, K. M., Beaton, D., Binns, M. A., Kwan, D., Levine, B., . . . Troyer, A. K. (2020). The Quality Assurance and Quality Control Protocol for Neuropsychological Data Collection and Curation in the Ontario Neurodegenerative Disease Research Initiative (ONDRI) Study. *Assessment*, 1073191120913933. doi:10.1177/1073191120913933

Mengel, D., Dams, J., Ziemek, J., Becker, J., Balzer-Geldsetzer, M., Hilker, R., . . . Dodel, R. (2016). Apolipoprotein E epsilon4 does not affect cognitive performance in patients with Parkinson's disease. *Parkinsonism Relat Disord, 29*, 112-116. doi:10.1016/j.parkreldis.2016.04.013

Mirza, S. S., Saeed, U., Knight, J., Ramirez, J., Stuss, D. T., Keith, J., . . . Alzheimer's Disease Neuroimaging, I. (2019). APOE epsilon4, white matter hyperintensities, and cognition in Alzheimer and Lewy body dementia. *Neurology, 93*(19), e1807-e1819. doi:10.1212/WNL.0000000000008377

Nieuwenhuis, R., te Grotenhuis, M., Pelzer, B., Schmidt, A., Konig, R., & Eisinga, R. (2017). Weighted Effect Coding (Version 0.4-1). Online: CRAN. Retrieved from https://cran.r-project.org/web/packages/wec/wec.pdf

O'Donoghue, M. C., Murphy, S. E., Zamboni, G., Nobre, A. C., & Mackay, C. E. (2018). APOE genotype and cognition in healthy individuals at risk of Alzheimer's disease: A review. *Cortex, 104*, 103-123. doi:10.1016/j.cortex.2018.03.025

Osenkowski, P., Ye, W., Wang, R., Wolfe, M. S., & Selkoe, D. J. (2008). Direct and potent regulation of gamma-secretase by its lipid microenvironment. *J Biol Chem, 283*(33), 22529-22540. doi:10.1074/jbc.M801925200

Pendlebury, S. T., Poole, D., Burgess, A., Duerden, J., Rothwell, P. M., & Oxford Vascular, S. (2020). APOE-epsilon4 Genotype and Dementia Before and After Transient Ischemic Attack and Stroke: Population-Based Cohort Study. *Stroke, 51*(3), 751-758. doi:10.1161/STROKEAHA.119.026927

R Core Team. (2014). R: A language and environment for statistical computing. . Retrieved from http://www.R-project.org/

Rall, S. C., Jr., Weisgraber, K. H., Innerarity, T. L., & Mahley, R. W. (1982). Structural basis for receptor binding heterogeneity of apolipoprotein E from type III hyperlipoproteinemic subjects. *Proc Natl Acad Sci U S A, 79*(15), 4696-4700. doi:10.1073/pnas.79.15.4696

Seo, E. H., Kim, S. H., Park, S. H., Kang, S. H., Choo, I. H., & Alzheimer's Disease Neuroimaging, I. (2016). Independent and Interactive Influences of the APOE Genotype and Beta-Amyloid Burden on Cognitive Function in Mild Cognitive Impairment. *J Korean Med Sci, 31*(2), 286-295. doi:10.3346/jkms.2016.31.2.286

Shi, Y., Yamada, K., Liddelow, S. A., Smith, S. T., Zhao, L., Luo, W., . . . Holtzman, D. M. (2017). ApoE4 markedly exacerbates tau-mediated neurodegeneration in a mouse model of tauopathy. *Nature, 549*(7673), 523-527. doi:10.1038/nature24016

Sunderland, K. M., Beaton, D., Arnott, S. R., Kleinstiver, P. W., Kwan, D., Lawrence-Dewar, J. M., . . . Binns, M. A. (2020). The Ontario Neurodegenerative Disease Research Initiative. *medRxiv*, 2020.2007.2030.20165456. doi:10.1101/2020.07.30.20165456

Sunderland, K. M., Beaton, D., Fraser, J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., . . . Binns, M. A. (2019). The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project. *BMC Med Res Methodol, 19*(1), 102. doi:10.1186/s12874-019-0737-5

Swartz, R. H., Black, S. E., & St George-Hyslop, P. (1999). Apolipoprotein E and Alzheimer's disease: a genetic, molecular and neuroimaging review. *Can J Neurol*

Sci, 26(2), 77-88. Retrieved from
https://www.ncbi.nlm.nih.gov/pubmed/10352866

te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., &
Konig, R. (2017a). A novel method for modelling interaction between categorical
variables. *International Journal of Public Health, 62*(3), 427-431.
doi:10.1007/s00038-016-0902-0

te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., &
Konig, R. (2017b). When size matters: advantages of weighted effect coding in
observational studies. *International Journal of Public Health, 62*(1), 163-167.
doi:10.1007/s00038-016-0901-1

van der Vlies, A. E., Pijnenburg, Y. A., Koene, T., Klein, M., Kok, A., Scheltens, P., &
van der Flier, W. M. (2007). Cognitive impairment in Alzheimer's disease is
modified by APOE genotype. *Dement Geriatr Cogn Disord, 24*(2), 98-103.
doi:10.1159/000104467

Verghese, P. B., Castellano, J. M., & Holtzman, D. M. (2011). Apolipoprotein E in
Alzheimer's disease and other neurological disorders. *Lancet Neurol, 10*(3), 241-
252. doi:10.1016/S1474-4422(10)70325-2

Verpillat, P., Camuzat, A., Hannequin, D., Thomas-Anterion, C., Puel, M., Belliard, S., . .
. Clerget-Darpoux, F. (2002). Apolipoprotein E gene in frontotemporal dementia:
an association study and meta-analysis. *Eur J Hum Genet, 10*(7), 399-405.
doi:10.1038/sj.ejhg.5200820

Ward, A., Crean, S., Mercaldi, C. J., Collins, J. M., Boyd, D., Cook, M. N., & Arrighi, H.
M. (2012). Prevalence of apolipoprotein E4 genotype and homozygotes (APOE
e4/4) among patients diagnosed with Alzheimer's disease: a systematic review
and meta-analysis. *Neuroepidemiology, 38*(1), 1-17. doi:10.1159/000334607

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York, USA:
Springer.

Wolk, D. A., Dickerson, B. C., & Alzheimer's Disease Neuroimaging, I. (2010).
Apolipoprotein E (APOE) genotype has dissociable effects on memory and
attentional-executive network function in Alzheimer's disease. *Proc Natl Acad Sci
U S A, 107*(22), 10256-10261. doi:10.1073/pnas.1001412107

Zannis, V. I., Breslow, J. L., Utermann, G., Mahley, R. W., Weisgraber, K. H., Havel, R.
J., . . . Blum, C. (1982). Proposed nomenclature of apoE isoproteins, apoE
genotypes, and phenotypes. *J Lipid Res, 23*(6), 911-914. Retrieved from
http://www.ncbi.nlm.nih.gov/pubmed/7130859

Zhao, N., Liu, C. C., Van Ingelgom, A. J., Linares, C., Kurti, A., Knight, J. A., . . . Bu, G.
(2018). APOE epsilon2 is associated with increased tau pathology in primary
tauopathy. *Nat Commun, 9*(1), 4388. doi:10.1038/s41467-018-06783-0

Chapter 4 – Contribution of rare variant associations to neurodegenerative disease presentation in the Ontario Neurodegenerative Disease Research Initiative.

The work presented in <u>Chapter 4</u> has been edited from the original manuscript currently under review at *npj Genomic Medicine* for brevity and consistency throughout the entire Dissertation.

4.1.   Abstract

Background/Objective: Genetic factors contribute to neurodegenerative disease, with high heritability estimates across diagnoses; however, a large portion of this genetic influence remains poorly understood. Many previous studies have attempted to fill the gaps by performing linkage analyses and association studies in individual disease cohorts but have failed to consider the clinical and pathological overlap observed across neurodegenerative diseases and the potential for genetic overlap between the phenotypes. Here, we leveraged rare variant association analyses (RVAA) to elucidate the genetic overlap among multiple neurodegenerative diagnoses, including Alzheimer's disease, amyotrophic lateral sclerosis, frontotemporal dementia (FTD), mild cognitive impairment, and Parkinson's disease (PD), as well as cerebrovascular disease, using the data generated with a custom-designed neurodegenerative disease gene panel in the Ontario Neurodegenerative Disease Research Initiative (ONDRI).

Methods and Results: As expected, only ~3% of ONDRI participants harboured a monogenic variant likely driving their disease presentation. Yet, when genes were binned based on previous disease associations, we observed an enrichment of putative loss-of-function variants in PD genes across all ONDRI cohorts. Further, individual gene-based RVAA identified significant enrichment of rare, non-synonymous variants in *PARK2* in the FTD cohort, and in *NOTCH3* in the PD cohort.

Conclusion: The results indicate that there may be greater heterogeneity in the genetic factors contributing to neurodegeneration than previously appreciated. Although the mechanisms by which these genes contribute to disease presentation must be further explored, we hypothesize they may be a result of rare variants of moderate phenotypic effect contributing to overlapping pathology and clinical features observed across neurodegenerative diagnoses.

4.2. Introduction

Neurodegenerative diseases are characterized by neuronal degeneration resulting in cognitive decline and/or motor dysfunction. Mainly manifesting in late adulthood, neurodegenerative diseases are often tightly correlated with the deposition of protein aggregates, such as amyloid-β and neurofibrillary tau tangles in Alzheimer's disease (AD) and α-synuclein in Parkinson's disease (PD) (Kovacs, Botond, & Budka, 2010). Although diagnoses are typically based on clinical presentation, definitive diagnosis requires post-mortem pathologic analysis to identify the pathogenic protein aggregates *in situ*. Further, neurodegenerative disease presentation is highly heterogeneous, and it is increasingly accepted that diagnoses exist along a spectrum, with a greater amount of mixed pathology — and overlapping clinical features — than previously thought (Kovacs et al., 2013).

Genetic factors are known to increase risk of neurodegeneration and influence expression of disease features (Bocchetta et al., 2016); however, only ~10% of neurodegenerative disease patients are considered to have familial forms of disease, a fraction of which are caused by known rare, highly penetrant genetic variants. Similarly, while genome-wide association studies (GWASs) have identified many common GWAS-significant single nucleotide polymorphisms (SNPs) in neurodegenerative disease cohorts and thus have advanced the field considerably (Lambert et al., 2013; Nalls et al., 2014; Simon-Sanchez et al., 2009), such variants account for only a small amount of heritable risk (Keller et al., 2014; Ridge, Mukherjee, Crane, & Kauwe, 2013; Singleton & Hardy, 2016). Even after considering the collective effects of both Mendelian large-effect rare mutations and common disease-associated SNPs, a considerable portion of heritability across neurodegenerative diseases remains unexplained (Ciani et al., 2019; Ridge et al., 2013; Van Damme, 2018).

Recent studies have reported enrichment of rare variants in genes typically considered only in early-onset, familial neurodegenerative disease cases in cohorts with sporadic forms of disease, likely constituting a moderate effect on disease risk. For example, rare variants have been identified in patients with late-onset sporadic AD in amyloid precursor protein (*APP*), presenilin 1 (*PSEN1*), and presenilin 2 (*PSEN2*) (Cruchaga et al., 2012);

in patients with both late- and early-onset sporadic PD in α-synuclein (*SNCA*), parkin RBR E3 ubiquitinated protein ligase (*PARK2*), leucine-rich repeat kinase 2 (*LRRK2*), and VPS35 retromer complex component (*VPS35*) (Lesage & Brice, 2012; Robak et al., 2017); and in patients with both familial and sporadic ALS in superoxide dismutase (*SOD1*), FUS RNA binding protein (*FUS*), and DnaJ heat shock protein family (Hsp40) member C7 (*DNAJC7*) (S. M. K. Farhan et al., 2019). In addition, the explanation for heterogeneity of phenotypic expression of neurodegeneration among individuals with identical rare variants is unclear, as is the potential influence of genetic factors on the overlapping clinical and pathological features of different neurodegenerative diagnoses. Such gaps in knowledge reinforce how much remains to be learned regarding genetic risk of neurodegeneration, even with respect to known neurodegenerative disease genes.

While rare variants likely account for at least a portion of the missing heritability of neurodegenerative diseases, as well as the phenotypic heterogeneity between diseases, they remain difficult to detect. Rare variants with large effect sizes are individually very uncommon and require large samples sizes to obtain the statistical power necessary for detection — even some of the largest GWASs, with sample sizes >100,000, are still unable to detect rare disease-associated variants. However, by binning variants into gene-based groupings of their original disease associations — or by analyzing each gene individually — rare variant association analyses (RVAAs) may identify gene-disease associations and explain additional disease risk even with modest sample sizes (Lee, Abecasis, Boehnke, & Lin, 2014).

Here, we utilize targeted next-generation sequencing (NGS) data coupled with a RVAA-based binning strategy to identify the contribution of rare genetic variants in participants from the Ontario Neurodegenerative Disease Research Initiative (ONDRI) to multiple neurodegenerative disease phenotypes, including: 1) AD; 2) amnestic MCI; 3) ALS; 4) frontotemporal dementia (FTD); and 5) PD, as well as cerebrovascular disease (CVD) (S. M. K. Farhan et al., 2017; Sunderland et al., 2020). By binning variants into disease-association-based gene groupings and individual gene-based groupings, and comparing variant enrichment to that of a cognitively normal, elderly control cohort, we seek to identify whether rare variants significantly contribute to disease presentation in ONDRI

participants. Furthermore, by studying six phenotypes concurrently, we can determine whether associations exist across disease phenotypes, and whether these might account for the overlapping features often observed across neurodegenerative diseases.

4.3.   Methods

4.3.1.   Study participants

In total, 520 individuals passed ONDRI preliminary screening (S. M. K. Farhan et al., 2017). Of those, 519 participants had a blood sample collected, from which genomic DNA was extracted. Study ethics approval was obtained from the Research Ethics Boards at Baycrest Centre for Geriatric Care (Toronto, Ontario, Canada); Centre for Addiction and Mental Health (Toronto, Ontario, Canada); Elizabeth Bruyère Hospital (Ottawa, Ontario, Canada); Hamilton General Hospital (Hamilton, Ontario, Canada); McMaster (Hamilton, Ontario, Canada); London Health Sciences Centre (London, Ontario, Canada); Parkwood Hospital (London, Ontario, Canada); St Michael's Hospital (Toronto, Ontario, Canada); Sunnybrook Health Sciences Centre (Toronto, Ontario, Canada); The Ottawa Hospital (Ottawa, Ontario, Canada); and University Health Network-Toronto Western Hospital (Toronto, Ontario, Canada). All participants provided written, informed consent in accordance with the Research Ethics Boards and regulatory requirements. DNA was also obtained from 189 cognitively normal control genomic DNA samples from the GenADA study (Li et al., 2008).

4.3.2.   DNA preparation and sequencing

All DNA samples were sequenced using the targeted NGS panel, ONDRISeq, on the Illumina MiSeq Personal Genome Sequencer (Illumina, San Diego, CA, United States) and raw sequencing data were processed with a custom bioinformatics workflow. Briefly, FASTQ files were imported into CLC Bio Genomics Workbench v10 (CLC Bio, Aarhus, Denmark) to perform pre-processing and variant annotation, which produced a variant calling format (VCF) file and binary alignment map (BAM) file for each participant. Detailed methodology outlining DNA isolation, DNA sequencing, and sequencing analysis has been previously described (Dilliott et al., 2018).

4.3.3. Identification of variants likely contributing to Mendelian disease

ONDRISeq VCF files of the ONDRI cases were imported into VarSeq® (Golden Helix, Bozeman, MT, United States) and variants were annotated with sequence ontologies. Minor allele frequencies (MAFs) were obtained from the Genome Aggregation Database (gnomAD v.2.0.1v3 non-neuro) (Karczewski et al., 2020). Rare (MAF < 0.01), nonsynonymous variants were prioritized. Further assessment of variants was performed to identify those in genes known to contribute to Mendelian forms of the patient's disease of diagnosis and those classified as pathogenic or likely pathogenic in ClinVar (Landrum et al., 2014), Online Mendelian Inheritance in Man (OMIM) ("Online Mendelian Inheritance in Man, OMIM®,"), and/or the Alzforum Mutation Database ("Alzforum Mutations,"). All identified variants were considered those likely to be contributing to Mendelian forms of disease.

All samples were genotyped for chromosome 9 open reading frame 72 (*C9orf72*) using both amplicon length analysis and repeat-primed polymerase chain reaction (PCR), as previously described (Xi et al., 2012). Harbouring > 30 repeats is a commonly accepted genetic cause of ALS and FTD (Xi et al., 2015; Xi et al., 2012), and therefore was the cutoff used to determine those with pathogenic repeat expansions.

4.3.4. Ancestry matching and estimation

The ONDRISeq VCF files of all cases and controls were merged and filtered to include only SNPs within exonic and splicing regions with a MAF > 0.005 in the Genome Aggregation Database (gnomAD v.2.0.1v3) using VarSeq®. Variants that were located on the sex chromosomes or within the *MAPT* gene were excluded, due to potential influence from the cohort's sex distribution and a common haplotype variation found across the gene, respectively. The filtered, merged VCF was processed with a bash-based tool that contains a collection of scripts necessary to run region-based RVAA, "Exautomate" (Davis, Dron, Robinson, Hegele, & Lizotte, 2019), to produce PLINK compatible MAP and PED files. SNP & Variation Suite v8.8.3 (SVS; Golden Helix Inc.)

was used to perform linkage disequilibrium (LD) pruning (threshold = 0.5) and a
principal component analysis (PCA) was performed to identify the genetic ancestry.

In accordance with standard quality control in genomic studies, a logistic regression
analysis was performed within R on the generated principal components to identify
individuals with divergent ancestries to minimize false discoveries due to population
stratification. A multidimensional outlier analysis (multiplier = 1.5) was performed
within SVS using the significant components to identify outlier samples based on
ancestral variation and batch effects, which were not included in the RVAAs described
below.

To predict the genomic ancestry of the samples, we used the whole genome sequences
from the 1000 Genomes Project (1000G; N = 2693), which are binned into ancestral
groups, including: African, Admixed American, East Asian, European, and South Asian
(Auton et al., 2015). The 1000G VCFs were merged and filtered to include only SNPs
within the exonic and splicing regions captured by the ONDRISeq panel with an MAF >
0.005 in gnomAD. The resulting filtered merged VCF was processed using "Exautomate"
to produce MAP and PED files and a PCA was performed using the SNPRelate
Bioconductor R package (v1.22.0; LD pruning threshold = 0.5) (Zheng, Gogarten,
Laurie, & Weir, 2015). The SNP loadings from this PCA and the PED file of the ONDRI
cases and controls were used to project the ONDRI cases and controls onto the
components of the 1000G PCA (Zheng et al., 2012).

4.3.5.   Rare variant association analysis

The VCF files of all ancestry matched ONDRI cases and controls were imported into the
variant annotation software, VarSeq®. Variants were annotated with sequence
ontologies, MAFs from gnomAD, and *in silico* prediction scores from Combined
Annotation Dependent Depletion (CADD; v1.3) (Kircher et al., 2014), Sorting Intolerant
from Tolerant (SIFT) (Kumar, Henikoff, & Ng, 2009), and Polymorphism Phenotyping
v2 (PolyPhen-2) (Adzhubei et al., 2010). Variants were prioritized and variants with a
sequence ontology of nonsense, stop-loss, splicing acceptor, splicing donor, frameshift,
or missense, and a MAF < 0.01 in the Exome Aggregation Consortium (ExAC v1.0) were

included in subsequent analyses. Both heterozygous and homozygous variants were retained for RVAAs. Variants were binned into three groups: 1) putative loss-of-function variants (LOF; including nonsense, stop-loss, frameshift, splice acceptor, and splice donor sequence ontologies); 2) missense variants; and 3) possibly deleterious missense variants (including missense variants with either a CADD Phred ≥ 20 or a likely damaging/damaging prediction from both SIFT and PolyPhen-2). Carriers of these variants were considered 'variant positive' and 'variant negative', respectively.

Variants were also binned into groups based on the previous disease association of the gene in which the variant was located. In total, the 80 genes encompassed by the ONDRISeq panel were binned into four disease-association groups: 1) AD/MCI-associated genes; 2) ALS/FTD-associated genes; 3) PD-associated genes; and 4) CVD-associated genes based on the most well-established previous disease association, as determined by Farhan *et al.* (**Figure 4.1**) (S.M.K. Farhan et al., 2016).

RVAAs were performed using multinomial logistic regression models. A model was produced for each variant subgroup (putative LOF, missense, and possibly deleterious missense) to compare the number of variant-positive individuals in each of the ONDRI disease cohorts to the cognitively normal control cohort, while correcting for age and sex. In addition, participants were weighted to better reflect disease prevalence in the general elderly population, accounting for potential inference bias as a result of the non-probability sampling mechanism ("2020 Alzheimer's disease facts and figures," 2020; "Chapter 3: Mapping Connections: An understanding of neurological conditions in Canada – Scope (prevalence and incidence)," 2014; Hogan et al., 2016; Mehta et al., 2018; Ng et al., 2015; Roberts & Knopman, 2013). The brglm2 R package (v0.6.2) (Kosmidis, 2020) was used to fit the regression models and apply a mean bias reduction (Kosmidis, Kenne Pagui, & Sartori, 2020) that accounts for the low variant positive counts.

A gene-based RVAA, the optimal unified Sequence Kernel Association Test (SKAT-O), was also performed using the script package "Exautomate" (Davis et al., 2019). This method identified specific genes covered by ONDRISeq with an increased frequency of

nonsynonymous, rare variants (MAF < 0.01, ExAC) in the disease cohorts compared to controls, and in the disease cohorts compared to each other. To maximize sample sizes, the AD and amnestic MCI cohorts were combined for SKAT-O analyses. As SKAT-O was not able to account for multi-nucleotide variants, follow-up analyses were performed on SKAT-O results with a detected signal using Firth logistic regression, adjusting for age and sex, using the brglm2 R package. Genes that had total rare variant counts between the two cohorts of <5, or with zero rare variants in one of the cohorts were excluded from analyses.

Analyses were performed using R statistical software 3.6.0 (Team, 2014) in R Studio 1.1.463 and data visualization was performed using the ggplot2 R package (v3.3.s) (Wickham, 2009). Significance for all regression analyses was measured at an alpha-level of p < 0.050, although regression results with p < 0.075 were still reported.

Figure 4.1 Binning of genes included on the ONDRISeq next-generation sequencing panel based on their previous disease associations for use in the rare variant association analyses.
Genes were binned based on the most well-established previous disease association, as determined by Farhan *et al.* (S.M.K. Farhan et al., 2016).

4.4.   Results

4.4.1.   Variants likely contributing to Mendelian disease

In the total cohort of 519 ONDRI cases (**Table 4.1**), seven participants harboured non-synonymous rare variants likely contributing to a Mendelian form of disease with each harbouring a unique variant of interest (**Table 4.2**), including one participant with AD, two participants with CVD, one participant with MCI, and three participants with PD, each harbouring a unique variant of interest (**Table 4.3**). Further, seven participants carried pathogenic repeat expansions within *C9orf72* (**Table 4.2**), including four participants with ALS, two participants with FTD, and one participant with AD. Overall, monogenic variants were observed at a frequency of ~3% both before and after ancestral outlier analysis (0.027 [0.015–0.045] and 0.030 [0.016–0.052] in the total ONDRI cohort and ancestry matched cohort, respectively). All participants were retained for subsequent analyses.

Table 4.1 Demographics of the total ONDRI cohorts and cognitively normal control cohort at baseline and the demographics of the cohorts following multivariate outlier analysis.

	Total Participants			**Ancestry Matched Participants**		
Cohort	**Samples**	**Mean age (years ± sd)**	**Male:Female**	**Samples**	**Mean age (years ± sd)**	**Male:Female**
ONDRI	519	68.6 ± 7.6	341:172	396	68.7 ± 7.9	268:128
AD	41	71.8 ± 8.0	24:17	33	71.4 ± 7.9	19:14
ALS	40	62.0 ± 8.7	24:16	32	61.9 ± 9.2	23:9
CVD	161	69.2 ± 7.4	109:50	124	69.6 ± 7.6	87:37
FTD	53	67.8 ± 7.1	34:19	39	67.5 ± 7.3	25:14
MCI	85	70.6 ± 8.3	45:40	59	71.9 ± 8.3	29:30
PD	139	67.8 ± 6.4	106:30	109	67.5 ± 6.3	85:24
Controls	189	74.0 ± 8.2	77:112	164	74.0 ± 7.9	68:96

The common variation (MAF > 0.005) captured by the ONDRISeq next-generation sequencing panel was used to perform principal component analysis on the ONDRI cases and controls accounting for variance introduced by differential ancestry and batch effects. Multivariate outlier analysis was performed using the first 8 principal components. Ancestry matched refers to all samples that were not removed by the outlier analysis.

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CVD, cerebrovascular disease; FTD, frontotemporal dementia; MCI, mild cognitive impairment; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease; sd, standard deviation.

Table 4.2 ONDRI participants harbouring rare variants likely contributing to Mendelian forms of neurodegenerative disease and cerebrovascular disease.

Cohort	Total Participants			Ancestry Matched Participants		
	Samples	Monogenic rare variants	*C9orf72* expansion	Samples	Monogenic rare variants	*C9orf72* expansion
ONDRI	519	7	7	396	5	7
AD	41	1	1	33	1	1
ALS	40	0	4	32	0	4
CVD	161	2	0	124	1	0
FTD	53	0	2	39	0	2
MCI	85	1	0	59	1	0
PD	139	3	0	109	2	0

Ancestry matched refers to all samples that were not removed by the outlier analysis following principal component analysis on common variation (MAF > 0.005) captured by the ONDRISeq next-generation sequencing panel. Monogenic rare variants refers to individuals harbouring variants with a MAF < 0.01 in gnomAD v.2.1.1 (non-neuro) in a gene known to contribute to Mendelian forms of the disease of patient diagnosis and that was classified as likely pathogenic/pathogenic in ClinVar, OMIM, and/or the AlzForum mutations database. *C9orf72* expansions refers to individuals harbouring the G_4C_2-expansion, which were genotyped using amplicon length analysis and repeat primed PCR; all expansions were > 60 repeats in length.

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; *C9orf72*, chromosome 9 open reading frame 72 gene; CVD, cerebrovascular disease; FTD, frontotemporal dementia; MCI, mild cognitive impairment; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease.

Table 4.3 Nonsynonymous rare variants likely contributing to Mendelian forms of neurodegenerative disease and cerebrovascular disease identified in the total ONDRI cohort.

Gene	cDNA alteration	Protein alteration	Reference SNP identifier	Sequence ontology	MAF (gnomAD)	CADD Phred	Previous disease association	Participants harbouring variant (cohort)
APP	c.2137G>A	p.Ala713Thr	rs63750066	Missense	2.88e-5	34.0	AD[1]	1 (AD)
GCH1	c.671A>G	p.Lys224Arg	rs41298442	Missense	4.70e-5	13.9	PD[2]	1 (PD)
LRRK2	c.6055G>A	p.Gly2019Ser	rs34637584	Missense	2.26e-4	35	PD[3]	1 (PD)
NOTCH3	c.544C>T	p.Arg182Cys	rs28933697	Missense	4.70e-5	31.0	CADASIL[4]	1 (CVD)
NOTCH3	c.580T>C	p.Cys194Arg	rs1568361818	Missense	NA	23.5	CADASIL[5]	1 (CVD)
PSEN1	c.118_120delGAC	p.Asp40del	rs759538127	In-frame deletion	1.40e-4	14.1	AD[6]	1 (MCI)
SNCA	c.150T>G	p.His50Gln	rs201106962	Missense	8.28e-5	4.9	PD[7]	1 (PD)

[1](Armstrong, Boada, Rey, Vidal, & Ferrer, 2004; Carter et al., 1992; Pera et al., 2013); [2](Leuzzi et al., 2002; Saunders-Pullman et al., 2004; Trender-Gerhard et al., 2009); [3](Di Fonzo et al., 2005; Zabetian et al., 2006); [4](Joutel et al., 2001; Joutel et al., 1997); [5](Kalimo, Ruchoux, Viitanen, & Kalaria, 2002); [6](Nygaard, Lippa, Mehdi, & Baehring, 2014); [7](Khalaf et al., 2014)

Nonsynonymous rare variants likely contributing to Mendelian forms of neurodegenerative disease were considered variants with a MAF < 0.01 in gnomAD v.2.1.1 (non-neuro) in a gene known to contribute to Mendelian forms of the disease of patient diagnosis and that was classified as likely pathogenic/pathogenic in ClinVar, OMIM, and/or the AlzForum mutations database.

Abbreviations: AD, Alzheimer's disease; *APP*, amyloid precursor protein gene; CADASIL, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy; CADD Phred, combined annotation dependent depletion Phred; cDNA, coding DNA; *COL4A2*, collagen type IV alpha 2 chain protein; CVD, cerebrovascular disease; *GCH1*, GTP cyclohydrolase 1 gene; *LRRK2*, leucine rich repeat kinase 2 gene; MAF, minor allele frequency; MCI, mild cognitive impairment; NA, not applicable; *NOTCH3*, notch receptor 3 gene; PD, Parkinson's disease; *PSEN1*, presenilin 1 gene; *SNCA*, α-synuclein gene; SNP, single nucleotide polymorphism.

4.4.2. Principal component analysis

Ancestry of the ONDRI cases and cognitively normal control samples was estimated by projecting their SNP loadings onto a PCA of the 1000G population (**Figure 4.2**). The large degree of overlap observed between the ONDRI cases, as well as the cognitively normal controls, and the European cohort of the 1000G population suggests that the participants within our study were largely of European descent. To produce a homogenous genetic dataset, which minimizes any false discoveries due to population stratification in accordance with standard genomics quality control best practices, we performed a logistic regression of the ONDRI case and control principal components and identified the first eight components as significantly contributing to variance in the samples. Following multidimensional ancestral outlier analysis and outlier removal, the data consisted of 396 ONDRI cases and 164 cognitively normal controls (**Table 4.1**).
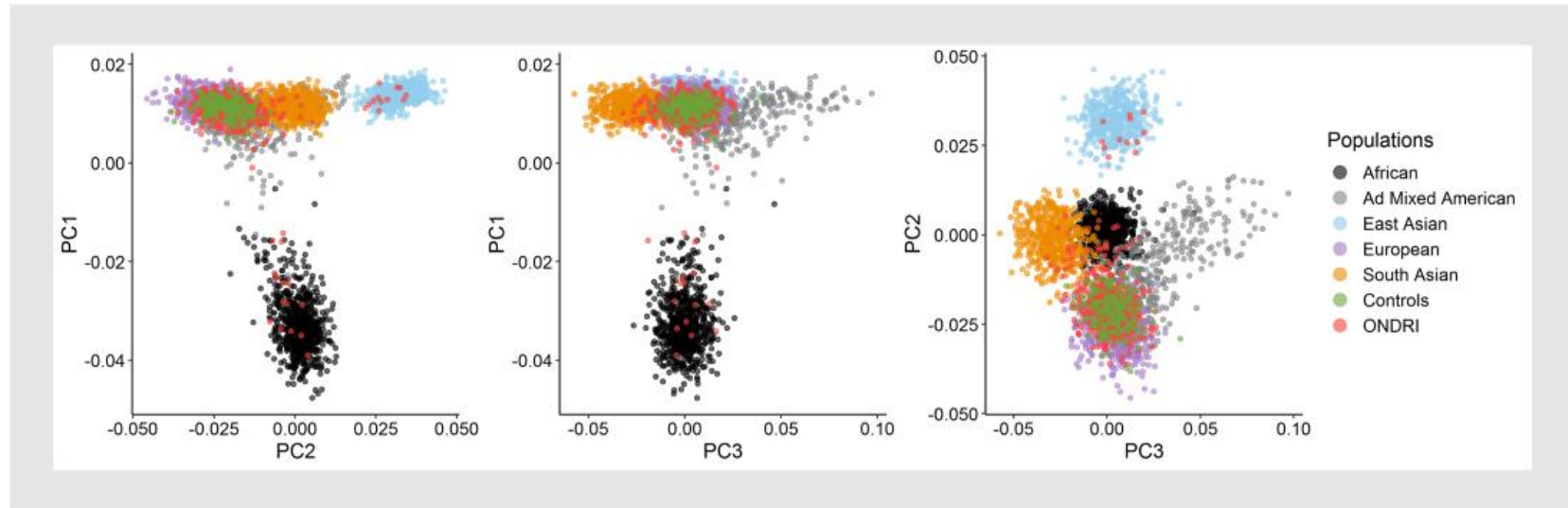
Figure 4.2 Ancestry estimate of the ONDRI cases and cognitively normal control cohort using a principal component analysis (PCA) of data from the 1000 Genomes Project.
Whole genome sequencing VCFs from 1000G were merged and filtered to include SNPs (MAF > 0.005, gnomAD) within the exonic and splicing regions captured by the ONDRISeq panel. A PCA was run using the *SNPRelate* Bioconductor R package (LD pruning threshold = 0.5). SNP loadings from the PCA were used to project the ONDRI cases and controls onto the components to estimate the participant's ancestries. Abbreviations: ONDRI, Ontario Neurodegenerative Disease Research Initiative; PC, principal component.

4.4.3. Disease association based RVAA

All regression coefficients and standard errors obtained by the multinomial logistic regression models are summarized in **Appendix E**. Combined analysis of neurodegenerative disease cohorts revealed that ONDRI participants were significantly more likely to carry a putative LOF variant in PD-associated genes in comparison to the normal controls (OR = 7.322, p = 0.031; **Figure 4.3**). Interestingly, similar significant associations were observed within many individual disease cohorts when compared to controls, including for the AD (OR = 12.307, p = 0.023), ALS (OR = 127.744, p = 0.013), and FTD (OR = 51.832, p = 0.031) cohorts. Although not significant, the remaining disease cohorts displayed trends towards association in the same manner, including the CVD (OR = 10.698, p = 0.071), MCI (OR = 6.273, p = 0.053), and PD (OR = 30.821, p = 0.061) cohorts (**Figure 4.3**).

In addition, ALS and MCI cases were significantly more likely to carry rare putative LOF variants in ALS/FTD-associated genes, compared to the control cohort (OR = 33.169, p = 0.045 and OR = 2.905, p = 0.044, respectively; **Figure 4.3**). The ALS cases were also more likely to carry rare putative LOF variants in AD- and CVD-associated genes, although results were not significant (OR = 25.572, p = 0.072 and OR = 57.857, p = 0.074, respectively; **Figure 4.3**).

No differences in odds of carrying rare missense variants or possibly deleterious missense variants were observed between the participants in the neurodegenerative disease cohorts and the controls (**Appendix F**).
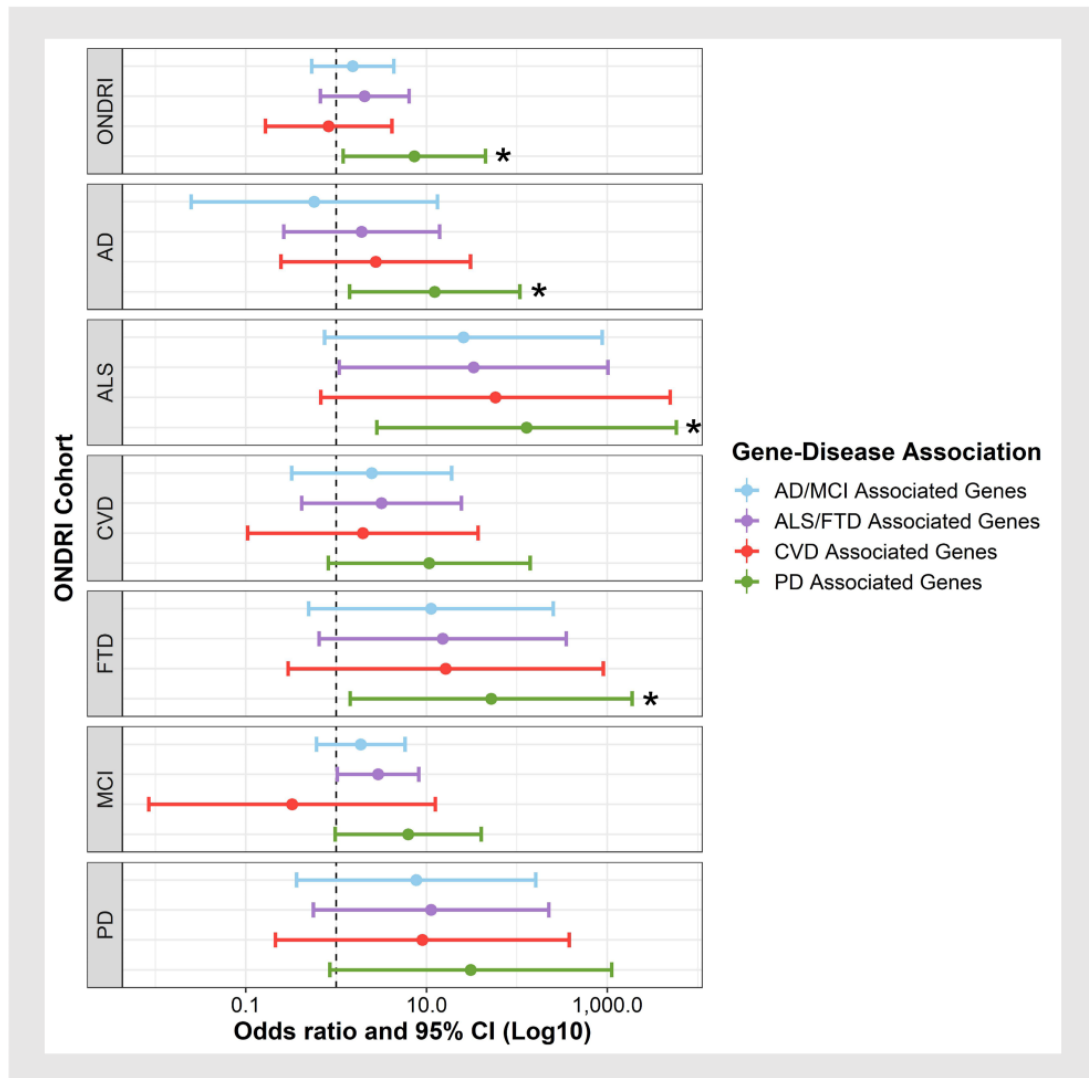
Figure 4.3 Rare variant association analysis comparing the enrichment of putative loss-of-function (LOF) rare variants in four disease-associated gene groupings in the ONDRI cohorts compared to cognitively normal controls.
Multinomial logistic regressions adjusted for age, sex, and disease prevalence were performed to analyze enrichment of putative LOF variants (including nonsense, stop loss, frameshift, splice acceptor, and splice donor sequence ontologies) identified in the 80 genes encompassed by the ONDRISeq panel, which were binned into four disease-associated gene groupings: AD associated genes, ALS/FTD associated genes, CVD associated genes, and PD associated genes, across the ONDRI cohorts compared to the control cohort. Only ancestry matched participants were included in the analyses. The *brglm2* R package was used to fit the regression model and apply a mean bias reduction accounting for the low variant positive counts. *p-value < 0.05. Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CVD, cerebrovascular disease; FTD, frontotemporal dementia; LOF, loss-of-function; MCI mild cognitive impairment; ONDRI, Ontario Neurodegenerative Disease Research Initiative; PD, Parkinson's disease.

4.4.4. Gene-based RVAA using SKAT-O

Following gene-based RVAA using SKAT-O, 11 genes were identified to be likely enriched in nonsynonymous rare variants in the ONDRI cohorts compared to the controls that also had sufficient total rare variant counts for subsequent analysis. Firth logistic regression, which was used to accommodate for the limitations of SKAT-O, revealed significant differences in nonsynonymous rare variant enrichment in three genes in the ONDRI cohorts compared to the controls (**Table 4.4**), including charged multivesicular body protein 2B (*CHMP2B*) across the combined neurodegenerative disease ONDRI cohort (OR = 0.080, p = 0.0008), neurofilament heavy chain (*NEFH*) in the CVD cohort (OR = 0.360, p = 0.036), and *PARK2* in the FTD cohort (OR = 11.602, p = 0.023).

Similarly, SKAT-O revealed 6 genes likely enriched in nonsynonymous rare variants in the individual ONDRI disease cohorts compared to each other that also had sufficient total rare variant counts for subsequent analysis. Firth logistic regression identified two genes with a significantly different enrichment of nonsynonymous rare variants in one cohort when compared to another (**Table 4.5**), including an enrichment of variants in Notch receptor 3 (*NOTCH3*) in the PD cohort compared to the combined AD and MCI cohort (OR = 2.986, p = 0.009), and an enrichment of variants in *NEFH* in the combined AD and MCI cohort compared to the CVD cohort (OR = 0.272, p = 0.011).

Table 4.4 Gene-based rare variant association analyses using Firth logistic regression comparing rare variant enrichment in the ONDRI cohorts compared to the control cohort, in genes identified as having differing enrichment by SKAT-O.

Cohort	Gene	Previous disease association	Odds ratio	Lower CI	Upper CI	P-value	
ONDRI	*ABCA7*	AD	1.630	0.768	3.460	0.204	
ONDRI	*CHMP2B*	ALS/FTD	0.080	0.018	0.351	0.0008	**
AD/MCI	*ABCA7*	AD	2.042	0.862	4.838	0.105	
AD/MCI	*SETX*	ALS/FTD	2.085	0.744	5.843	0.162	
CVD	*NEFH*	ALS/FTD	0.360	0.138	0.936	0.036	*
CVD	*PARK2*	PD	3.129	0.397	24.669	0.279	
CVD	*PINK1*	PD	2.278	0.382	13.600	0.366	
CVD	*SETX*	ALS/FTD	0.567	0.213	1.512	0.257	
FTD	*ALS2*	ALS/FTD	3.543	0.708	17.721	0.123	
FTD	*PARK2*	PD	11.602	1.395	96.519	0.023	*
PD	*ABCA7*	AD	2.155	0.831	5.590	0.114	

Odds ratio and 95% CI (Log10): 0.01 0.1 1 10 100

Due to limitations of SKAT-O and to account for the effects of age and sex, a Firth logistic regression was performed on each gene identified by SKAT-O to compare the enrichment of rare variants in the respective ONDRI cohorts in comparison to a cognitively normal control cohort. Only ancestry matched participants were included in the analyses. Genes that had total rare variant counts <5, or with zero rare variants in one of the cohorts were excluded from analyses. Statistical analyses were performed using the R statistical software 3.6.0 in R Studio 1.1.463. *p-value < 0.05; **p-value < 0.005; ***p-value < 0.0005. Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CI, confidence interval; CVD, cerebrovascular disease; FTD, frontotemporal dementia; MCI, mild cognitive impairment; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease; SKAT-O, the optimal unified Sequence Kernel Association Test.

Table 4.5 Gene-based rare variant association analyses using Firth logistic regression comparing rare variant enrichment between the individual ONDRI cohorts, in genes identified as having differing enrichment by SKAT-O.

Reference Cohort	Comparison Cohort	Gene	Previous disease association	Odds ratio	Lower CI	Upper CI	P-value	
AD/MCI	PD	*NOTCH3*	CVD	2.986	1.310	6.806	0.009	
AD/MCI	CVD	*NEFH*	ALS/FTD	0.272	0.100	0.739	0.011	
CVD	FTD	*NEFH*	ALS/FTD	2.392	0.630	9.092	0.200	
PD	CVD	*ALS2*	ALS/FTD	3.092	0.442	21.630	0.256	
PD	FTD	*PARK2*	PD	2.563	0.592	11.085	0.208	

Odds ratio and 95% CI (Log10)

Due to limitations of SKAT-O and to account for the effects of age and sex, a Firth logistic regression was performed on each gene identified by SKAT-O to compare the enrichment of rare variants between the individual ONDRI disease cohorts. Only ancestry matched participants were included in the analyses. Genes that had total rare variant counts <5, or with zero rare variants in one of the cohorts were excluded from analyses. Statistical analyses were performed using the R statistical software 3.6.0 in R Studio 1.1.463. *p-value < 0.05; **p-value < 0.005; ***p-value < 0.0005.

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CI, confidence interval; CVD, cerebrovascular disease; FTD, frontotemporal dementia; MCI, mild cognitive impairment; ONDRI, Ontario neurodegenerative disease research initiative; PD, Parkinson's disease; SKAT-O, the optimal unified Sequence Kernel Association Test.

4.5.   Discussion

As previously described, a large amount of missing heritability remains across neurodegenerative diagnoses and little is known regarding the contribution of rare genetic factors to the heterogenous presentation of these diseases. Due to the established infrequency of Mendelian forms of neurodegenerative phenotypes (Ghasemi & Brown, 2018; Reed, Bandres-Ciga, Blauwendraat, & Cookson, 2019; Takada, 2015; Van Cauwenberghe, Van Broeckhoven, & Sleegers, 2016), it was not surprising that few ONDRI participants harboured monogenic variants likely driving their disease presentation, including seven carriers of likely monogenic non-synonymous, rare variants — defined as variants previously reported as pathogenic within relevant mutations databases and the literature in respect to the participant's diagnosis — and seven carriers of the pathogenic *C9orf72* repeat expansion. Yet, the low frequency of monogenic variants observed in our cohorts has highlighted the need to investigate the contribution of rare variants in genes previously associated with neurodegeneration to the presentation of the entire spectrum of neurodegenerative and CVD diagnoses utilizing RVAA. Associations between specific neurodegenerative diagnoses and known neurodegeneration genes were identified with SKAT-O and subsequent logistic regression, as well as with disease-association based RVAA. Our principal findings included associations between: 1) nonsynonymous rare variants in *PARK2* and the FTD cohort; 2) nonsynonymous rare variants in *NOTCH3* and the PD cohort; 3) rare, putative LOF variants in PD-associated genes across the entire ONDRI cohort; and 4) rare, putative LOF variants in ALS/FTD-associated genes in the ALS and MCI cohorts.

Nonsynonymous rare variants in *PARK2* were enriched in the FTD cohort. While *PARK2* is well-established to be associated with autosomal recessive familial PD (Abbas et al., 1999) and potentially with autosomal dominant sporadic PD (Jeon, Kim, Lee, Hattori, & Mizuno, 2001; Tan et al., 2005), it has not been previously associated with FTD. However, both PD and FTD are influenced by lysosome dysfunction, which can be exacerbated by mutated *PARK2* (Wallings, Humble, Ward, & Wade-Martins, 2019). The variants identified in the FTD cohort were all of heterozygous zygosity, and two had been previously reported as variants of uncertain significance in ClinVar (i.e. p.Arg402Cys

and p.Pro437Leu). Although the variants may have contributed to the FTD diagnoses in our cohort, it remains possible that the variants had a moderate phenotypic effect and/or decreased penetrance. If so, our result may be consistent with some of the heterogeneity and overlap often seen across neurodegenerative disease presentations, therefore highlighting the potential impact of unexpected rare variation to features of disease, which is an area of neurogenetics that must be further explored.

One example of how rare variants may contribute to intermediate phenotypes of neurodegeneration, rather than a diagnosis itself, is demonstrated for rare variants in *NOTCH3* among participants with PD. Typically, rare variants within *NOTCH3* are associated with a monogenic subtype of CVD called cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL). Leukoencephalopathy associated with CADASIL can be manifested as white matter hyperintensities seen on T2-weighted MRI scans. We previously observed that PD participants harbouring rare *NOTCH3* variants had double the volume of white matter hyperintensities than those that did not (Dilliott et al., 2020). Herein, we did not observe an association between PD and *NOTCH3* when compared to the controls, but the history of CVD in the control cohort was unknown. An association was observed between PD and *NOTCH3* when compared to the combined AD and MCI cohort, which was of particular interest as ONDRI excluded participants from the AD and MCI cohorts who had significant evidence of vascular pathology (Sunderland et al., 2020). Therefore, this result seemingly supports the hypothesis that rare *NOTCH3* variation may be contributing to cerebrovascular features within patients with PD (Dilliott et al., 2020).

The gene-based RVAA also identified *CHMP2B* as having significantly fewer variants in the entire ONDRI cohort compared to controls, and *NEFH* as having significantly fewer variants in the CVD cohort compared to the controls or the combined AD/MCI cohort. These results could be interpreted as protection against neurodegenerative diseases and cerebrovascular phenotypes from rare variants in *CHMP2B* and *NEFH*, respectively. However, the association within *CHMP2B* was likely driven by a single splicing variant (c.34+8C>T) harboured by the only three ONDRI participants with rare *CHMP2B* variants and five of the eight controls with rare *CHMP2B* variants. The variant had a

MAF in ExAC of 2.80E-3 and was previously reported in ClinVar as benign. It is possible that the variants in *CHMP2B* and *NEFH* may have gain-of-function protective effects, explaining the unexpected signal; however, functional assays are needed for confirmation. Based on the large amount of influence from single variants in these potentially protective results, specifically in the case of *CHMP2B*, and no previously established protective effects for either gene within the literature, further interpretation remains unclear.

No 'expected' rare variant associations were observed between the individual neurodegenerative disease cohorts and genes previously associated with the disease cohorts in the gene-based analysis. For example, there were no associations between rare variants in *APP*, *PSEN1*, or *PSEN2* with the combined AD and MCI cohort. Although this was unsurprising due to the low frequency of monogenic variants we identified in the ONDRI cohorts (**Table 1.2**), it suggests there may be other genetic determinants driving disease presentation and progression. So, to maximize analytic power, we also assessed rare variant frequency in groups of genes, based on the disease with which the genes have been most commonly associated. Across all neurodegenerative diagnoses, rare, putative LOF variants in PD-associated genes were enriched when compared to the control cohort. Although unsurprising in the PD cohort, this interesting trend was observed in all five remaining neurodegenerative disease cohorts individually, as well as in the combined ONDRI cohort.

When we examined the individual genes that contributed to the association, eight of the 13 putative LOF PD-associated variant positive participants (61.5% of variant positive participants) harboured variants in melanocortin 1 receptor (*MC1R*). Specifically, the putative LOF variants in *MC1R* were identified in the CVD, FTD, MCI, and PD cohorts, as well as in one control participant. *MC1R* on chromosome 16 encodes a receptor typically involved in the regulation of melanin pigment within the skin, but is also expressed in the periaqueductal gray matter of the brain (Xia, Wikberg, & Chhajlani, 1995). The gene was originally associated with PD in a study by Tell-Marti *et al.* (Tell-Marti et al., 2015), in which a common missense variant (p.Arg160Trp) was associated with the disease in a Spanish population. Previous research has also suggested an

association between both red hair and melanoma — for which *MC1R* variants are a risk factor — and PD (Chen, Feng, Schwarzschild, & Gao, 2017; Liu, Gao, Lu, & Chen, 2011). The MC1R protein was also found to be neuroprotective in dopaminergic neurons, which are integral to PD pathology. Unfortunately, the association between *MC1R* and PD is controversial, with multiple studies unable to replicate the finding (Gan-Or et al., 2016; Lorenzo-Betancor, Wszolek, & Ross, 2016), and to date, no strong evidence linking *MC1R* variation to any other neurodegenerative disease exists.

We also observed a significant enrichment of rare, putative LOF variants in ALS/FTD-associated genes in the ALS and MCI cohorts. No single gene stood out in the analysis and it is important to recognize that the number of participants in each cohort harbouring rare putative LOF variants was low, so we are cautious to not draw conclusions from these results given the small sample sizes. However, it cannot be discounted that the enrichment signal within the MCI cohort may suggest the participants' potential to progress to FTD, rather than AD. Typically, we anticipate that amnestic MCI patients will progress to an AD phenotype or will not progress at all, yet the possibility remains that presentation will develop into a variant of FTD and follow up of the MCI participants in ONDRI remains imperative.

Our study did have limitations that deserve comment. The analysis was largely limited by modest sample sizes, particularly after accounting for variance resulting from differential ancestry and batch effects. Combined with the inherent rarity of the variants, the number of variant-positive participants in each cohort remained small. Yet, the study still identified interesting signals that are reasonable contenders for replication within larger cohorts and hypothesis generating for further analyses. Further, apart from basic demographic information and Montreal Cognitive Assessment scores to define the control cohort as cognitively normal upon enrollment, no further data were available. Therefore, it is unclear whether control participants had any history of CVD without cognitive impairment and analyses may not have been sensitive to signals from the CVD-associated genes on the ONDRISeq panel, as highlighted by the association between *NOTCH3* and PD when compared to the AD/MCI cohort, but not the controls. Our analyses were also limited to individuals of probable European ancestry and replication in

other populations is necessary. Finally, our results were restricted to the 80 genes covered by our custom, targeted NGS panel (S.M.K. Farhan et al., 2016) and identification of novel genes associated with neurodegenerative disease was not possible. Despite this, we still identified associations between specific neurodegenerative diagnoses and known neurodegeneration genes, such as the enrichment of *PARK2* variants in FTD.

4.6. Conclusion

Our analyses allowed us to observe considerable heterogeneity in the genetic contribution underlying neurodegenerative disease presentation. While we could not conclude that the rare variants observed were driving diagnoses in all instances, it is reasonable to assume that some of the variability observed in neurodegenerative disease presentation may be driven by the rare variants in 'atypical' neurodegenerative disease genes. Future analyses are required to replicate our findings; however, our study demonstrates the potential for RVAA as an approach to identify genes in which rare variants may have moderate and somewhat unanticipated phenotypic effects in certain neurodegenerative disease cohorts, either by directly influencing disease pathology or by potentially contributing to the overlapping features across neurodegenerative disease. Overall, this may suggest a more complex genetic architecture of neurodegeneration than the familiar simple monogenic model of inheritance in which a variant fully explains a clinical phenotype.

4.7. References

2020 Alzheimer's disease facts and figures. (2020). *Alzheimers Dement*. doi:10.1002/alz.12068

Abbas, N., Lucking, C. B., Ricard, S., Durr, A., Bonifati, V., De Michele, G., . . . Brice, A. (1999). A wide variety of mutations in the parkin gene are responsible for autosomal recessive parkinsonism in Europe. French Parkinson's Disease Genetics Study Group and the European Consortium on Genetic Susceptibility in Parkinson's Disease. *Hum Mol Genet, 8*(4), 567-574. doi:10.1093/hmg/8.4.567

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Alzforum Mutations. Retrieved from https://www.alzforum.org/mutations

Armstrong, J., Boada, M., Rey, M. J., Vidal, N., & Ferrer, I. (2004). Familial Alzheimer disease associated with A713T mutation in APP. *Neurosci Lett, 370*(2-3), 241-243. doi:10.1016/j.neulet.2004.08.026

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Bocchetta, M., Mega, A., Bernardi, L., Di Maria, E., Benussi, L., Binetti, G., . . . Bruni, A. C. (2016). Genetic Counseling and Testing for Alzheimer's Disease and Frontotemporal Lobar Degeneration: An Italian Consensus Protocol. *J Alzheimers Dis, 51*(1), 277-291. doi:10.3233/JAD-150849

Carter, D. A., Desmarais, E., Bellis, M., Campion, D., Clerget-Darpoux, F., Brice, A., . . . Mallet, J. (1992). More missense in amyloid gene. *Nat Genet, 2*(4), 255-256. doi:10.1038/ng1292-255

Chapter 3: Mapping Connections: An understanding of neurological conditions in Canada – Scope (prevalence and incidence). (2014). Retrieved from https://www.canada.ca/en/public-health/services/reports-publications/mapping-connections-understanding-neurological-conditions/mapping-connections-understanding-neurological-conditions-canada-13.html

Chen, X., Feng, D., Schwarzschild, M. A., & Gao, X. (2017). Red hair, MC1R variants, and risk for Parkinson's disease - a meta-analysis. *Ann Clin Transl Neurol, 4*(3), 212-216. doi:10.1002/acn3.381

Ciani, M., Bonvicini, C., Scassellati, C., Carrara, M., Maj, C., Fostinelli, S., . . . Benussi, L. (2019). The Missing Heritability of Sporadic Frontotemporal Dementia: New Insights from Rare Variants in Neurodegenerative Candidate Genes. *Int J Mol Sci, 20*(16). doi:10.3390/ijms20163903

Cruchaga, C., Haller, G., Chakraverty, S., Mayo, K., Vallania, F. L., Mitra, R. D., . . . Consortium, N.-L. N. F. S. (2012). Rare variants in APP, PSEN1 and PSEN2 increase risk for AD in late-onset Alzheimer's disease families. *PLoS One, 7*(2), e31039. doi:10.1371/journal.pone.0031039

Davis, B. D., Dron, J. S., Robinson, J. F., Hegele, R. A., & Lizotte, D. J. (2019). Exautomate: A user-friendly tool for region-based rare variant association analysis (RVAA). *bioRxiv, 649368*. doi:https://doi.org/10.1101/649368

Di Fonzo, A., Rohe, C. F., Ferreira, J., Chien, H. F., Vacca, L., Stocchi, F., . . . Italian Parkinson Genetics, N. (2005). A frequent LRRK2 gene mutation associated with autosomal dominant Parkinson's disease. *Lancet, 365*(9457), 412-415. doi:10.1016/S0140-6736(05)17829-5

Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2018). Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J Vis Exp*(134). doi:10.3791/57266

Dilliott, A. A., Ramirez, J., Binns, M. A., Breen, D. P., Evans, E. C., Beaton, D., . . . Ontario Neurodegenerative Disease Research Initiative, I. (2020). Parkinson's Disease, NOTCH3 Genetic Variants, and White Matter Hyperintensities. *Mov Disord, 35*(11), 2090-2095. doi:10.1002/mds.28171

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., . . . Strong, M. J. (2017). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci, 44*(2), 196-202. doi:10.1017/cjn.2016.415

Farhan, S. M. K., Dilliott, A. A., Ghani, G., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*(16032), 1-11.

Farhan, S. M. K., Howrigan, D. P., Abbott, L. E., Klim, J. R., Topp, S. D., Byrnes, A. E., . . . Neale, B. M. (2019). Exome sequencing in amyotrophic lateral sclerosis implicates a novel gene, DNAJC7, encoding a heat-shock protein. *Nat Neurosci, 22*(12), 1966-1974. doi:10.1038/s41593-019-0530-0

Gan-Or, Z., Mohsin, N., Girard, S. L., Montplaisir, J. Y., Ambalavanan, A., Strong, S., . . . Rouleau, G. A. (2016). The role of the melanoma gene MC1R in Parkinson disease and REM sleep behavior disorder. *Neurobiol Aging, 43*, 180 e187-180 e113. doi:10.1016/j.neurobiolaging.2016.03.029

Ghasemi, M., & Brown, R. H., Jr. (2018). Genetics of Amyotrophic Lateral Sclerosis. *Cold Spring Harb Perspect Med, 8*(5). doi:10.1101/cshperspect.a024125

Hogan, D. B., Jette, N., Fiest, K. M., Roberts, J. I., Pearson, D., Smith, E. E., . . . Maxwell, C. J. (2016). The Prevalence and Incidence of Frontotemporal

Dementia: a Systematic Review. *Can J Neurol Sci, 43 Suppl 1*, S96-S109. doi:10.1017/cjn.2016.25

Jeon, B. S., Kim, J. M., Lee, D. S., Hattori, N., & Mizuno, Y. (2001). An apparently sporadic case with parkin gene mutation in a Korean woman. *Arch Neurol, 58*(6), 988-989. doi:10.1001/archneur.58.6.988

Joutel, A., Favrole, P., Labauge, P., Chabriat, H., Lescoat, C., Andreux, F., . . . Tournier-Lasserve, E. (2001). Skin biopsy immunostaining with a Notch3 monoclonal antibody for CADASIL diagnosis. *Lancet, 358*(9298), 2049-2051. doi:10.1016/S0140-6736(01)07142-2

Joutel, A., Vahedi, K., Corpechot, C., Troesch, A., Chabriat, H., Vayssiere, C., . . . Tournier-Lasserve, E. (1997). Strong clustering and stereotyped nature of Notch3 mutations in CADASIL patients. *Lancet, 350*(9090), 1511-1515. doi:10.1016/S0140-6736(97)08083-5

Kalimo, H., Ruchoux, M. M., Viitanen, M., & Kalaria, R. N. (2002). CADASIL: a common form of hereditary arteriopathy causing brain infarcts and dementia. *Brain Pathol, 12*(3), 371-384. doi:10.1111/j.1750-3639.2002.tb00451.x

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434-443. doi:10.1038/s41586-020-2308-7

Keller, M. F., Ferrucci, L., Singleton, A. B., Tienari, P. J., Laaksovirta, H., Restagno, G., . . . Nalls, M. A. (2014). Genome-wide analysis of the heritability of amyotrophic lateral sclerosis. *JAMA Neurol, 71*(9), 1123-1134. doi:10.1001/jamaneurol.2014.1184

Khalaf, O., Fauvet, B., Oueslati, A., Dikiy, I., Mahul-Mellier, A. L., Ruggeri, F. S., . . . Lashuel, H. A. (2014). The H50Q mutation enhances alpha-synuclein aggregation, secretion, and toxicity. *J Biol Chem, 289*(32), 21856-21876. doi:10.1074/jbc.M114.553297

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi:10.1038/ng.2892

Kosmidis, I. (2020). brglm2: Bias Reduction in Generalized Linear Models (Version 0.6.2). Online: CRAN. Retrieved from https://cran.r-project.org/package=brglm2

Kosmidis, I., Kenne Pagui, E. C., & Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Stat Comput, 30*, 43-59. doi:https://doi.org/10.1007/s11222-019-09860-6

Kovacs, G. G., Botond, G., & Budka, H. (2010). Protein coding of neurodegenerative dementias: the neuropathological basis of biomarker diagnostics. *Acta Neuropathol, 119*(4), 389-408. doi:10.1007/s00401-010-0658-1

Kovacs, G. G., Milenkovic, I., Wohrer, A., Hoftberger, R., Gelpi, E., Haberler, C., . . . Budka, H. (2013). Non-Alzheimer neurodegenerative pathologies and their combinations are more frequent than commonly believed in the elderly brain: a community-based autopsy series. *Acta Neuropathol, 126*(3), 365-384. doi:10.1007/s00401-013-1157-y

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc, 4*(7), 1073-1081. doi:10.1038/nprot.2009.86

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., . . . Amouyel, P. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet, 45*(12), 1452-1458. doi:10.1038/ng.2802

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res, 42*(Database issue), D980-985. doi:10.1093/nar/gkt1113

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet, 95*(1), 5-23. doi:10.1016/j.ajhg.2014.06.009

Lesage, S., & Brice, A. (2012). Role of mendelian genes in "sporadic" Parkinson's disease. *Parkinsonism Relat Disord, 18 Suppl 1*, S66-70. doi:10.1016/S1353-8020(11)70022-0

Leuzzi, V., Carducci, C., Carducci, C., Cardona, F., Artiola, C., & Antonozzi, I. (2002). Autosomal dominant GTP-CH deficiency presenting as a dopa-responsive myoclonus-dystonia syndrome. *Neurology, 59*(8), 1241-1243. doi:10.1212/wnl.59.8.1241

Li, H., Wetten, S., Li, L., St Jean, P. L., Upmanyu, R., Surh, L., . . . Roses, A. D. (2008). Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol, 65*(1), 45-53. doi:10.1001/archneurol.2007.3

Liu, R., Gao, X., Lu, Y., & Chen, H. (2011). Meta-analysis of the relationship between Parkinson disease and melanoma. *Neurology, 76*(23), 2002-2009. doi:10.1212/WNL.0b013e31821e554e

Lorenzo-Betancor, O., Wszolek, Z. K., & Ross, O. A. (2016). Rare variants in MC1R/TUBB3 exon 1 are not associated with Parkinson's disease. *Ann Neurol, 79*(2), 331. doi:10.1002/ana.24581

Mehta, P., Kaye, W., Raymond, J., Wu, R., Larson, T., Punjani, R., . . . Horton, K. (2018). Prevalence of Amyotrophic Lateral Sclerosis - United States, 2014. *MMWR Morb Mortal Wkly Rep, 67*(7), 216-218. doi:10.15585/mmwr.mm6707a3

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., . . . Singleton, A. B. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet, 46*(9), 989-993. doi:10.1038/ng.3043

Ng, R., Maxwell, C. J., Yates, E. A., Nylen, K., Antflick, J., Jette, N., & Bronskill, S. E. (2015). *Brain Disorders in Ontario: Prevalence, Incidence and Costs from Health Administrative Data*. Retrieved from Toronto, ON:

Nygaard, H. B., Lippa, C. F., Mehdi, D., & Baehring, J. M. (2014). A Novel Presenilin 1 Mutation in Early-Onset Alzheimer's Disease With Prominent Frontal Features. *Am J Alzheimers Dis Other Demen, 29*(5), 433-435. doi:10.1177/1533317513518653

Online Mendelian Inheritance in Man, OMIM®. Retrieved from https://omim.org/

Pera, M., Alcolea, D., Sanchez-Valle, R., Guardia-Laguarta, C., Colom-Cadena, M., Badiola, N., . . . Lleo, A. (2013). Distinct patterns of APP processing in the CNS in autosomal-dominant and sporadic Alzheimer disease. *Acta Neuropathol, 125*(2), 201-213. doi:10.1007/s00401-012-1062-9

Reed, X., Bandres-Ciga, S., Blauwendraat, C., & Cookson, M. R. (2019). The role of monogenic genes in idiopathic Parkinson's disease. *Neurobiol Dis, 124*, 230-239. doi:10.1016/j.nbd.2018.11.012

Ridge, P. G., Mukherjee, S., Crane, P. K., & Kauwe, J. S. (2013). Alzheimer's disease: analyzing the missing heritability. *PLoS One, 8*(11), e79771. doi:10.1371/journal.pone.0079771

Robak, L. A., Jansen, I. E., van Rooij, J., Uitterlinden, A. G., Kraaij, R., Jankovic, J., . . . Shulman, J. M. (2017). Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain, 140*(12), 3191-3203. doi:10.1093/brain/awx285

Roberts, R., & Knopman, D. S. (2013). Classification and epidemiology of MCI. *Clin Geriatr Med, 29*(4), 753-772. doi:10.1016/j.cger.2013.07.003

Saunders-Pullman, R., Blau, N., Hyland, K., Zschocke, J., Nygaard, T., Raymond, D., . . . Bressman, S. B. (2004). Phenylalanine loading as a diagnostic test for DRD:

interpreting the utility of the test. *Mol Genet Metab, 83*(3), 207-212. doi:10.1016/j.ymgme.2004.07.010

Simon-Sanchez, J., Schulte, C., Bras, J. M., Sharma, M., Gibbs, J. R., Berg, D., . . . Gasser, T. (2009). Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet, 41*(12), 1308-1312. doi:10.1038/ng.487

Singleton, A., & Hardy, J. (2016). The Evolution of Genetics: Alzheimer's and Parkinson's Diseases. *Neuron, 90*(6), 1154-1163. doi:10.1016/j.neuron.2016.05.040

Sunderland, K. M., Beaton, D., Arnott, S. R., Kleinstiver, P. W., Kwan, D., Lawrence-Dewar, J. M., . . . Binns, M. A. (2020). The Ontario Neurodegenerative Disease Research Initiative. *medRxiv*, 2020.2007.2030.20165456. doi:10.1101/2020.07.30.20165456

Takada, L. T. (2015). The Genetics of Monogenic Frontotemporal Dementia. *Dement Neuropsychol, 9*(3), 219-229. doi:10.1590/1980-57642015DN93000003

Tan, E. K., Shen, H., Tan, J. M., Lim, K. L., Fook-Chong, S., Hu, W. P., . . . Zhao, Y. (2005). Differential expression of splice variant and wild-type parkin in sporadic Parkinson's disease. *Neurogenetics, 6*(4), 179-184. doi:10.1007/s10048-005-0001-5

Team, R. C. (2014). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org/

Tell-Marti, G., Puig-Butille, J. A., Potrony, M., Badenas, C., Mila, M., Malvehy, J., . . . Puig, S. (2015). The MC1R melanoma risk variant p.R160W is associated with Parkinson disease. *Ann Neurol, 77*(5), 889-894. doi:10.1002/ana.24373

Trender-Gerhard, I., Sweeney, M. G., Schwingenschuh, P., Mir, P., Edwards, M. J., Gerhard, A., . . . Bhatia, K. P. (2009). Autosomal-dominant GTPCH1-deficient DRD: clinical characteristics and long-term outcome of 34 patients. *J Neurol Neurosurg Psychiatry, 80*(8), 839-845. doi:10.1136/jnnp.2008.155861

Van Cauwenberghe, C., Van Broeckhoven, C., & Sleegers, K. (2016). The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med, 18*(5), 421-430. doi:10.1038/gim.2015.117

Van Damme, P. (2018). How much of the missing heritability of ALS is hidden in known ALS genes? *J Neurol Neurosurg Psychiatry, 89*(8), 794. doi:10.1136/jnnp-2018-318354

Wallings, R. L., Humble, S. W., Ward, M. E., & Wade-Martins, R. (2019). Lysosomal Dysfunction at the Centre of Parkinson's Disease and Frontotemporal Dementia/Amyotrophic Lateral Sclerosis. *Trends Neurosci, 42*(12), 899-912. doi:10.1016/j.tins.2019.10.002

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. New York, USA: Springer.

Xi, Z., van Blitterswijk, M., Zhang, M., McGoldrick, P., McLean, J. R., Yunusova, Y., . . . Rogaeva, E. (2015). Jump from pre-mutation to pathologic expansion in C9orf72. *Am J Hum Genet, 96*(6), 962-970. doi:10.1016/j.ajhg.2015.04.016

Xi, Z., Zinman, L., Grinberg, Y., Moreno, D., Sato, C., Bilbao, J. M., . . . Rogaeva, E. (2012). Investigation of c9orf72 in 4 neurodegenerative disorders. *Arch Neurol, 69*(12), 1583-1590. doi:10.1001/archneurol.2012.2016

Xia, Y., Wikberg, J. E., & Chhajlani, V. (1995). Expression of melanocortin 1 receptor in periaqueductal gray matter. *Neuroreport, 6*(16), 2193-2196. doi:10.1097/00001756-199511000-00022

Zabetian, C. P., Hutter, C. M., Yearout, D., Lopez, A. N., Factor, S. A., Griffith, A., . . . Payami, H. (2006). LRRK2 G2019S in families with Parkinson disease who originated from Europe and the Middle East: evidence of two distinct founding events beginning two millennia ago. *Am J Hum Genet, 79*(4), 752-758. doi:10.1086/508025

Zheng, X., Gogarten, S., Laurie, C., & Weir, B. (2015). SNPRelate [R Package]: Bioconductor. Retrieved from http://github.com/zhengxwen/SNPRelate

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics, 28*(24), 3326-3328. doi:10.1093/bioinformatics/bts606

Chapter 5 – Targeted copy number variant identification across the neurodegenerative disease spectrum.

The work presented in <u>Chapter 5</u> has been edited from the original manuscript that is being prepared for submission for brevity and consistency throughout the entire Dissertation.

Dilliott A.A., Zhang K.K., Wang J., Abrahao A., Binns M.A., Black S.E., Borrie M., Dowlatshahi D., Finger E., Fischer C.E., Frank A., Freedman M., Grimes D., Hassan A., Jog M., Kumar S., Kwan D., Lang A.E., Mandzia J., Masellis M., Pasternak S., Pollock D.G., Rajji T.K., Rogaeva E., Sahlas D.J., Saposnik G., Sato C., Seitz D., Shoesmith C., Steeves T., Swartz R.H., Tan B., Tang-Wai D., Tartaglia M.C., Turnbull J., Zinman L., ONDRI Investigators, and Hegele R.A.

5.1.   Abstract

Background/Objective: Neurodegenerative diseases are a range of conditions affecting the aging population that are caused by neuronal loss and result in heterogenous presentations of motor impairment and cognitive and/or behavioural dysfunction. Although genetic factors are known to contribute to neurodegenerative disease susceptibility, there remains a large amount of heritability unaccounted for across the diagnoses. Copy number variants (CNVs) throughout the genome contribute to these phenotypes, but their presence and influence on disease state remains relatively understudied. Here, we addressed this gap by leveraging the recent improvements in the bioinformatics analysis of next-generation sequencing and applying a depth of coverage approach to detect CNVs in 80 genes previously associated with neurodegenerative disease within participants of the Ontario Neurodegenerative Disease Research Initiative (ONDRI).

Methods and Results: In total, we identified and validated four CNVs in the ONDRI cohort, including: 1) a heterozygous deletion of exon 5 in *OPTN* in an Alzheimer's disease participant; 2) a duplication of exons 1–5 in *PARK7* in an amyotrophic lateral sclerosis participant; 3) a duplication of >3 Mb, which encompassed *ABCC6*, in a cerebrovascular disease (CVD) participant; and 4) a duplication of exons 7–11 in *SAMHD1* in a mild cognitive impairment participant. We also identified two duplications with high confidence in the genes *ABCA7* and *ATP13A2*, each carried by an additional CVD participant. To our knowledge, all six CNVs were novel with respect to their association with the disease phenotype of the carrier.

Conclusion: The identification of CNVs across our study of neurodegenerative disease suggest that a portion of the apparent missing heritability of the phenotypes may be due to these structural variants, and thus their assessment is imperative for a thorough understanding of the genetic spectrum of neurodegeneration.

5.2.   Introduction

Neurodegenerative diseases are a collection of progressive conditions characterized by neuronal degeneration and protein aggregation within the brain. Although typically defining behavioural and/or cognitive phenotypes, such as Alzheimer's disease (AD), frontotemporal dementia (FTD), and mild cognitive impairment (MCI) or motor phenotypes, such as amyotrophic lateral sclerosis (ALS) and Parkinson's disease (PD), the umbrella term may also encompass neurodegenerative phenotypes that result from, or present alongside, cerebrovascular disease (CVD).

Age is the strongest risk factor for neurodegeneration; however, another important risk factor encompasses genetic variation that contributes to the diseases. It is known that genetic factors increase one's risk of developing neurodegenerative disease considerably, with relatively high heritability estimates across the various diagnoses (Cacace, Sleegers, & Van Broeckhoven, 2016; Greaves & Rohrer, 2019; Mejzini et al., 2019). However, our existing understanding of the genetic contributors to neurodegenerative disease fails to reach these estimates, leaving a large amount of missing heritability (Cacace et al., 2016; Hagenaars et al., 2018; Keller et al., 2012; Mejzini et al., 2019).

Copy number variants (CNVs) are large-scale deletions or duplications of DNA of at least 50 base pairs in length. While CNVs are generally common across the human genome (Redon et al., 2006), when encompassing genes, these structural variants often have large phenotypic impacts, including affecting gene expression, organization, and dosage (Stranger et al., 2007). CNVs have been shown to contribute to neurodegenerative disease presentation, including in individuals with AD (Cuccaro, De Marco, Cittadella, & Cavallaro, 2017; Ghani et al., 2012; Hooli et al., 2014); ALS (Morello et al., 2018); and PD (Nuytemans, Theuns, Cruts, & Van Broeckhoven, 2010). Yet studies identifying these variants in neurodegenerative disease cohorts are sparse, potentially due to the previous intricacies of accurately detecting CNVs. It is therefore hypothesized that some of the missing heritability of neurodegeneration may be accounted for by these large-scale variants.

Recently, the development of new bioinformatics algorithms have made the identification of CNVs more accessible, with the ability to detect variants using next-generation sequencing (NGS) and a depth of coverage (DOC) assessment (Iacocca et al., 2017). Here, we leveraged this approach to identify CNVs across the participants of the Ontario Neurodegenerative Disease Research Initiative (ONDRI), a multi-cohort study aiming to characterize a selection of neurodegenerative diseases, including AD, ALS, FTD, MCI, and PD, as well as CVD and its potential influence on neurodegeneration. Previously, the ONDRI cohort was genetically sequenced using the ONDRISeq NGS targeted panel, which covers 80 genes known to contribute to neurodegenerative disease (Dilliott et al., 2018; S.M.K. Farhan et al., 2016). Following the identification of CNVs using a DOC approach, we also aimed to validate a subset of CNVs to determine appropriate metrics by which DOC analysis can define CNVs of high confidence when evaluating ONDRISeq data.

5.3. Methods

5.3.1. Study participants

ONDRI enrolled 520 individuals that passed preliminary screening and were each clinically diagnosed with one of the following conditions: 1) AD; 2) ALS; 3) CVD; 4) FTD; 5) MCI; and 6) PD. Research ethics board approval was obtained from each participating site. Descriptions of the inclusion/exclusion criteria of ONDRI participants were previously reported (S. M. K. Farhan et al., 2017; Sunderland et al., 2020). All participants provided informed written consent. Clinical diagnoses and demographic data were obtained during participant screening and baseline assessment. When possible, participants provided clinical longitudinal follow-up assessment yearly, for up to three years (S. M. K. Farhan et al., 2017; Sunderland et al., 2020).

5.3.2. Next-generation targeted sequencing

Of the 520 enrolled participants, 519 participants had a blood sample collected, from which genomic DNA was extracted. DNA was also obtained from 189 cognitively normal elderly controls from the GenADA study (Li et al., 2008).

All ONDRI participant and control DNA samples were subjected to targeted NGS using the ONDRISeq neurodegenerative disease gene panel, as previously described (Dilliott et al., 2018). DNA samples were pooled and paired-end NGS was performed using the MiSeq Personal Genome Sequencer (Illumina, San Diego, CA, United States) and MiSeq Reagent Kit v3. Raw sequencing data FASTQ files were imported into CLC Bio Genomics Workbench v10 (CLC Bio, Aarhus, Denmark) to perform pre-processing and variant annotation, which produced a variant calling format (VCF) file and binary alignment map (BAM) file for each participant. Read mapping was performed using the human reference genome GRCh37/hg19.

5.3.3. Copy number variant detection

The CNV Caller tool, an application within VarSeq® (v1.4.3; Golden Helix, Bozeman, MT), was used to detect CNVs from ONDRISeq-generated data. The CNV Caller tool employs a normalized depth of coverage algorithm, such that an increase in sample coverage in comparison to a set of reference samples suggests a gain of copy number, and a decrease in coverage suggests a loss of copy number.

The ONDRISeq browser extensible data (BED) file was imported into VarSeq®, as well as the VCF and BAM files of the 189 control samples from which the algorithm selected 48 samples to use as a reference set with the lowest percent difference in coverage data compared to each ONDRI sample. The algorithm excluded control samples in the reference set if they displayed a >20% difference in coverage compared with the samples of interest. The matched reference sets also corrected for GC-content bias and accounted for any regions exhibiting inaccurate mapping. By comparing to the reference set, the CNV Caller tool was used to identify CNVs across the 519 ONDRI participants. A DOC ratio and z-score were computed for each target region covered by the NGS panel within each ONDRI sample. The DOC ratios measured the normalized DOC of the sample of interest compared to the normalized mean DOC of the reference set, whereas z-scores measured the number of standard deviations of each target region's DOC from the normalized mean DOC of the reference set. In addition to the DOC algorithm, the CNV Caller tool examined single nucleotide polymorphism (SNP) heterozygosity by

examining variant allele frequencies across target regions to provide further evidence for suspected CNVs, as previously described (Iacocca et al., 2017). The CNV Caller tool assigned each suspected CNV an average DOC ratio, average z-score, and a p-value. CNVs detected using ONDRISeq data are referred to as "potential CNVs".

5.3.4. Deletion confirmation using breakpoint analysis

To verify the presence of a partial optineurin (*OPTN*) gene deletion, primers were designed to flank regions surrounding putative deletion breakpoints and used for PCR amplification of the mutant allele. The Expand 20kb[plus] PCR system was used for DNA amplification (Roche, Basel, Switzerland). Forward (F3) and reverse (R1) primers flanking the deletion junction were: F3 5'-GTGACTCCATCACTCTGAACCTCC and R1 5'-CGAGTCTTCCTTCACATACGTGCC. Gel electrophoresis of the PCR product provided a visual confirmation of the mutant allele.

Once deletion breakpoints were identified, confirmation primers (P1:5'-TCCCTTGACATTTGCAGTGGAATC, P2: 5'-ACTGAGAGAACAGACAAGGTCAAC, P4: 5'-GGTCACTTAGGGAACAAGATAGTC) spanning proximal and distal breakpoints were designed for PCR and Sanger sequencing to verify the deletion breakpoint sequences for the wild type and mutant alleles. Thirty seconds of extension time for PCR cycles were used to achieve amplification of the normal allele using primer pair P1 and P2, while primer pair P1 and P4 amplified the mutant allele. Electropherograms were analyzed using the Applied Biosystems SeqScape Software version 2.6 (Thermo Fischer Scientific, Waltham, MA, USA) with the reference sequence obtained from the National Center for Biotechnology Information (NCBI) GenBank database (https://www.ncbi.nlm.nih.gov/genbank/).

5.3.5. Duplication confirmation using whole exome sequencing

To validate the presence of potential CNV duplications detected using the ONDRISeq data, six samples each with at least one potential duplication were selected for whole-exome sequencing (WES). DNA samples were sent to the McGill University and

Genome Quebec Innovation Centre (MUGQIC) for WES using the HiSeq 4000 instrument (Illumina) and Roche Nimblegen chemistry (Roche, Basel, Switzerland). FASTQ files were again imported into CLC Bio Genomics Workbench v10 (CLC Bio) to perform pre-processing and variant annotation to produce a VCF file and BAM file for each participant.

VCF and BAM files of the six ONDRI participants and the BED file that defined the Roche Nimblegen chemistry target regions were imported into VarSeq®, along with VCF and BAM files from WES of eight reference samples obtained from cognitively normal individuals diagnosed with atrial fibrillation and sequenced on the same HiSeq 4000 run at the MUGQIC as the six ONDRI cases. Five reference samples were selected by the algorithm based on similarity of the normalized coverage to the samples of interest, as described above. Again, the CNV Caller tool applied a DOC approach and computed a DOC ratio, z-score, and p-value for each detected CNV.

5.4.   Results

5.4.1.   Study participants and ONDRISeq copy number variant analysis

Using the VarSeq® CNV Caller tool, at least one potential CNV was detected in 44 of the total 519 ONDRI participants screened (8.5%). A total of 47 potential CNVs were detected among the 44 participants, including 37 duplications and ten heterozygous deletions (**Appendix G**). The CNVs ranged in size from 150 to 74,407 bp.

5.4.2.   Deletion confirmation using breakpoint analysis

Of the ten potential heterozygous deletions identified, one was chosen for breakpoint analysis based on our high confidence in the variant call, as determined by the metrics produced by the CNV Caller algorithm (DOC ratio = 0.487; z-score = -6.851; p = 1.10E-12). Specifically, the chosen heterozygous deletion encompassed exon 5 of *OPTN*. Sanger sequencing across the CNV breakpoints confirmed the presence of a 4,969 bp deletion in *OPTN* that encompassed all of exon five, with proximal and distal breakpoints

at chr10: 13,152,598 and chr10: 13,157,566, respectively (**Figure 5.1**). The deletion was carried by subject 1, who was diagnosed with AD (**Table 5.1**). The remaining nine heterozygous deletions did not undergo breakpoint analysis due to lower confidence in the reliability of the CNV calls, as determined by the CNV's individual metrics and/or relatively short span.
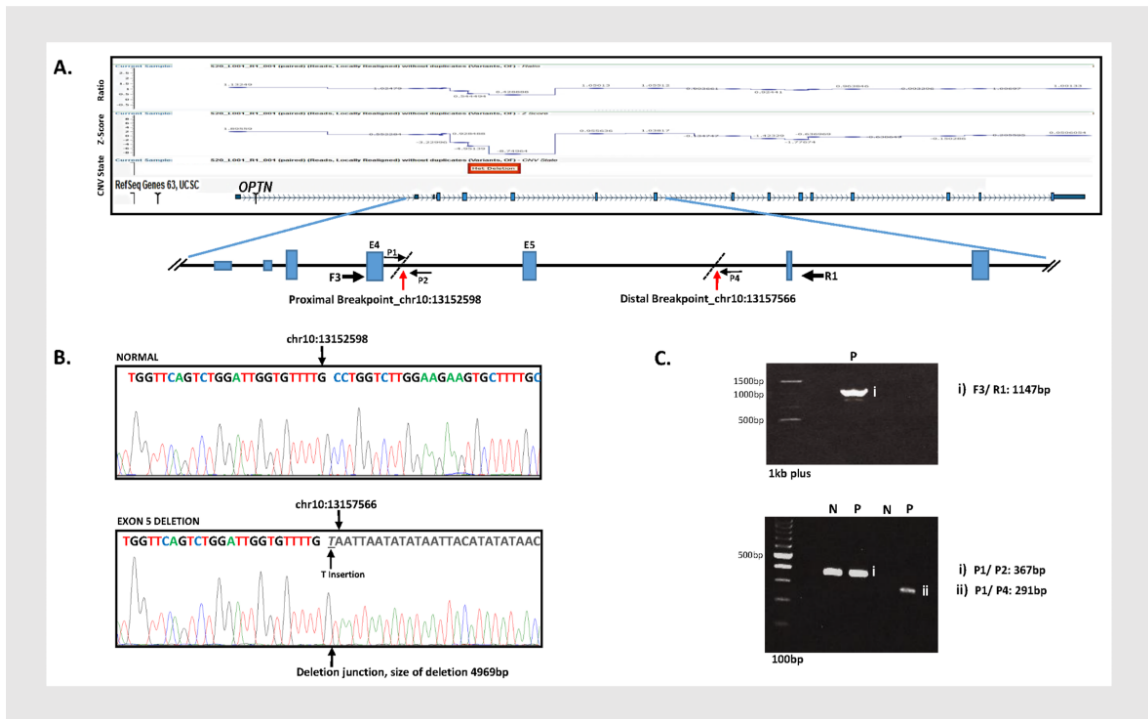
Figure 5.1 Validation of single-exon deletion in *OPTN* of subject 1 with Alzheimer's disease (AD).
(A) Screen capture of ONDRISeq-generated data from subject 1 processed by the VarSeq® v1.4.3 CNV Caller tool identified a potential heterozygous deletion, as indicated by a drop in DOC ratio. The bottom section shows the *OPTN* gene and location of primers used to confirm and sequence across the breakpoint. **(B)** Sanger sequencing results for the deletion junction. Results from a cognitively normal control are presented on the top, with results from subject 1 on the bottom. Internal sequence missing in the deleted allele is written in grey. **(C)** Gel electrophoresis of PCR products across the deletion breakpoint. The top gel shows amplification products generated using F3 and R1. The normal sequence distance between primer pair F3 and R1 generated a product size of 6116 bp; however, PCR amplification of subject 1's genomic DNA using F3 and R1 generated a product size of 1147 bp, suggesting a 4969 bp deletion. The bottom gel contains amplification products generated using primer pairs (i) P1, located in the proximal side of the suspected breakpoint, and P2, located within the deleted fragment, as well as (ii) P1 and P4, located on the distal side of the suspected breakpoint. Both the normal control (N) and proband (P) demonstrate amplification (367 bp) for the proximal primer pair. Amplification (291 bp) with P1 and P4 is seen in the proband, but not the normal control. For individuals without the deletion, the span between P1 and P4 would be too large to amplify under standard conditions; thus, if amplification occurred, it confirms the presence of a large deletion between the primer pair.

Table 5.1 Demographics and clinical data of the six ONDRI participants identified to have CNVs of high confidence in the 80 neurodegenerative disease genes covered by the ONDRISeq panel.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Diagnosis	AD	ALS	CVD	MCI	CVD	CVD
ASO (years)	73	55	70	57	73	63
Sex	Male	Male	Male	Female	Male	Male
MoCA	19	26	29	23	26	30
Other relevant clinical information	Exhibited cataracts in both eyes. No evidence of motor impairment reported.	Exhibited slight kinetic tremor; fasciculation of the torso, arms, and legs; brisk deep tendon reflex; and diffuse denervation of the lower motor neurons. No signs of parkinsonism were reported (H&Y = 0).	Experienced right-sided, anterior, large-artery atherosclerosis. Exhibited coronary artery disease, hypertension, and high cholesterol and previously had undergone coronary artery bypass graft surgery.	Although symptom onset was 15 years ago, the subject has not progressed to AD. No history of significant CVD.	Experienced right-sided, anterior, small-artery occlusion. Exhibited hypertension, atrial fibrillation, and diabetes.	Experienced right sided, posterior, large-artery atherosclerosis. Exhibited hypertension and high cholesterol. No signs of parkinsonism were reported (H&Y = 0) or other motor symptoms.
Relevant family history	N/A	N/A	Both parents and one sibling exhibited heart disease.	N/A	N/A	Mother experienced peripheral vascular disease, father experienced a stroke, and one half-sibling was diagnosed with ALS.
CNV Identified	Heterozygous deletion of exon 5 in *OPTN*	Duplication of exons 1−5 in *PARK7*	Duplication of all exons of *ABCC6* as well as 42 other genes[a]	Duplication of exons 7−11 in *SAMHD1*	Duplication of the latter part of exon 16 in *ABCA7*	Duplication of exon 16 in *ATP13A2*
Span (bp)	4,969[b]	9,810	3,233,228[c]	8,802	276	707
Genomic region	chr10: 13,152,598−13,157,566	~chr1:8,021,464−8,031,273	~chr16:15,185,138−18,418,365[c]	~chr20:35,539,371−35,548,172	~chr19:1,047,628−1,047,903	~chr1:17,319,874−17,320,580
Validation	Breakpoint	WES	WES	WES	N/A	N/A
DOC ratio	-6.851	6.816	6.371	5.872	9.287	7.175
z-score	0.487	1.455	1.482	1.474	3.221	1.430
p-value	1.10E-12	0	0	1.70E-21	3.14E-13	2.02E-8

Clinical data was obtained during participant screening and baseline analysis. Span, genomic region, DOC ratio, z-score, and p-value were based on the ONDRISeq CNV analysis, unless otherwise indicated. [a]Other genes encompassed by the duplication are outlined in **Table 5.2**. [b]The span presented is based on the breakpoint analysis of the heterozygous duplication. [c]The span and genomic region presented is based on the WES analysis of the duplication.
Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; ASO, age of symptom onset; bp, base pairs; chr, chromosome; CNV, copy number variant; CVD, cerebrovascular disease; DOC, depth of coverage; H&Y, Hoehn and Yahr; MCI, mild cognitive impairment; MoCA, Montreal Cognitive Assessment; N/A, not applicable; WES, whole exome sequencing.

5.4.3. Duplication confirmation using whole-exome sequencing

Of the participants harbouring the 37 potential duplications, six were identified as CNVs of relatively high confidence, as determined by their DOC ratios, z-scores, and p-values. Following WES of the samples and subsequent analysis with the CNV Caller tool, we identified and validated three participants as carriers of large-scale duplications (**Table 5.1**). Subject 2 was diagnosed with ALS and harboured a duplication spanning 9,810 bp that encompassed the first five exons of parkinsonism associated deglycase (*PARK7*). Subject 3 was diagnosed with CVD and harboured a duplication encompassing the entirety of ATP binding cassette subfamily C member 6 (*ABCC6*), which was detected using the ONDRISeq analysis; however, WES revealed the duplication also encompassed 42 other neighbouring genes, including 15 protein-coding genes, 12 pseudogenes, 11 microRNA encoding genes, and 4 non-coding RNA genes (**Table 5.2**). In total, the duplication spanned over 3 Mb. Finally, subject 4 was diagnosed with MCI and harboured a duplication of exons 7–11 of SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1 (*SAMHD1*).

One of the six samples sent for WES exhibited unmappable and incorrectly mapped reads, failing to pass the quality control standards of the CNV Caller tool algorithm. Validation of the duplication carried by this individual remains inconclusive.

Table 5.2 Genes encompassed by the >3 Mb duplication harboured by Subject 3.

Gene type	Genes
MicroRNA encoding genes	*MIR1972-1; MIR6511B2; MIR3180-4; MIR6506; MIR484; MIR3179-2; MIR3670-2; MIR3180-2; MIR6511A2; MIR6770-2; MIR6511A3;*
Non-coding RNA encoding genes	*LOC100505915; PKD1P6-NPIPP1; PKD1P1; LOC102723692*
Protein-coding genes	*PDXPC1; NTAN1; RRN3; NPIPA5; MPV17L; C16orf45; KIAA0430; NDE1; MYH11; FOPNL; ABCC1; ABCC6; NOMO3; NPIPA7; XYLT1; NPIPA8*
Pseudogenes	*LOC728138; NPIPP1; PKD1P6; RNU6-213P; RPL15P20; RPL17P40; PKD1P2; LOC100133127; LOC441750; LOC100421029; RPL7P47; LOC100133137*

Whole-exome sequencing confirmation of a full gene duplication of ABCC6 revealed that the duplication also encompassed 42 other neighbouring genes.

5.4.4. Copy number variants of high confidence

Based on the ONDRISeq analysis average DOC ratios, average z-scores, and p-values of the CNVs that we were both able and unable to validate, we identified two additional CNVs of high confidence (**Table 5.1**). The first duplication encompassed part of exon 16 of ATP binding cassette subfamily A member 7 (*ABCA7*) and had an average DOC ratio of 3.221 and average z-score of 9.961. It was found in subject 5, who was diagnosed with CVD. The second duplication encompassed all of exon 16 of ATPase cation transporting 13A2 (*ATP13A2*), with a DOC ratio of 1.430 and an average z-score of 7.175, and was found in subject 6, who was also diagnosed with CVD. Although these CNVs have not be formally validated, the CNV Caller tool metrics suggest that they are true positives.

Importantly, none of the confirmed CNVs, nor the CNVs of high confidence, were identified in any of the 189 cognitively normal elderly control samples. The six CNV carriers did not harbour additional pathogenic single nucleotide variants in the neurodegenerative disease associated genes encompassed by ONDRISeq relevant to their diagnoses. Clinical case information of subjects 1–6 can be found in **Table 5.1**.

5.5. Discussion

Of the 519 individuals diagnosed with neurodegenerative disease and/or CVD enrolled in the ONDRI study, we identified 44 individuals (8.5%) with potential CNVs in the 80 neurodegenerative disease associated genes covered by the ONDRISeq panel, of which six CNVs, each in a different participant (1.2%), were validated or determined to be of high confidence. Breakpoint analysis confirmed the presence of a heterozygous deletion in *OPTN* harboured by an individual with AD, and WES confirmed the presence of duplications in *PARK7*, *ABCC6*, and *SAMHD1*, in individuals diagnosed with AD, CVD, and MCI, respectively. We were also able to identify two additional CNVs from the 44 detected that were deemed to be CNVs of high confidence, which were duplications in *ABCA7* and *ATP13A2* also in individuals diagnosed with CVD. To our knowledge, all six CNVs were novel with respect to each carrier's diagnosis.

Among the validated CNVs, the heterozygous deletion of exon 5 in *OPTN* was the only CNV we confirmed using Sanger-based breakpoint analysis, and it was identified in a participant with AD. *OPTN* encodes optineurin, and pathogenic variants in the gene are associated with both autosomal dominant, adult-onset glaucoma and ALS (Maruyama et al., 2010; Schilter, Reis, Sorokina, & Semina, 2015). Interestingly, two studies have reported similar heterozygous deletions of exon 5 of *OPTN* in Japanese ALS cohorts (Iida et al., 2012; Maruyama et al., 2010). It has been hypothesized that there may be a relationship between *OPTN*, glaucoma, and AD, due to the high rate of co-morbidity between glaucoma and AD, as well as the observation of optineurin in neurofibrillary tau tangles — a hallmark of AD pathology (Liu & Tian, 2011). Yet, AD patients with pathogenic *OPTN* variants have not been previously observed. Herein, subject 1 exhibited gradual cognitive decline but did not demonstrate ALS-associated motor symptoms. Although no glaucoma diagnosis was documented and the participant's intraocular pressure was within the normal range for both eyes (10–21 mm Hg) at baseline and upon one-year follow-up, the participant did report vision loss and presented with cataracts at baseline assessment. It remains unclear whether the observation of the partial *OPTN* heterozygous deletion suggests a novel relationship between the gene and AD, or whether the variant may be contributing to the participant's ocular phenotypes.

We confirmed the presence of a duplication spanning exons 1–5 of *PARK7* in a participant diagnosed with ALS, although breakpoints could not be determined using WES. *PARK7*, otherwise referred to as *DJ-1*, encodes a conserved protein belonging to the peptidase C56 family and is thought to help inhibit aggregation of α-synuclein — a hallmark of PD pathology — as well as protect neurons against oxidative stress and cell death (Lev, Roncevic, Ickowicz, Melamed, & Offen, 2006). Mutations in *PARK7* cause autosomal recessive early-onset PD (EOPD) (Bonifati et al., 2003). Previously, duplications of the first five exons of *PARK7* have been associated with EOPD (Macedo et al., 2009), and small-scale variants in the gene have been associated with autosomal recessive Guamanian ALS/EOPD, which presents with a heterogeneous symptoms including muscular atrophy, cognitive decline, and tremor or rigidity (Annesi et al., 2005; Hanagasi et al., 2016). However, no previous associations of *PARK7* duplications have been reported in patients diagnosed with ALS alone. While the participant presented

herein did report slight tremors at baseline, which progressed to moderate tremors upon three-year follow-up based on the MDS Unified Parkinson's Disease Rating Scale (MDS-UPDRS), they had no further clinical signs of parkinsonism and had a Hoehn and Yahr score of zero both at baseline and follow-up. Yet the involvement of *PARK7* in ALS cannot be ruled out, as the DJ-1 protein is involved in sensing oxidative stress (Lev et al., 2006), and *PARK7* variants may therefore increase risk of oxidative stress, which is implicated as a major component in ALS pathologic mechanisms (Barber & Shaw, 2010). Our study is the first to report a case in which a structural variant affecting *PARK7* may have a role in a non-parkinsonism condition; however, functional analyses are required for further investigation of this relationship.

Furthermore, a CVD participant presented with a full-gene duplication of *ABCC6*. Confirmation of the duplication using WES found the CNV actually spanned over 3 Mb encompassing 43 total genes. Pathogenic variants within *ABCC6*, including CNVs, cause pseudoxanthoma elasticum (PXE), a rare autosomal recessive disorder characterized by elastic tissue fragmentation and arterial calcification (Bergen et al., 2000; Kringen et al., 2015; Ringpfeil, Lebwohl, Christiano, & Uitto, 2000). It is not uncommon for PXE patients to present with cerebral artery calcification, and studies have shown that ischemic CVD is highly prevalent in patients with PXE (Kauw et al., 2017; Pavlovic et al., 2005). Here, the participant carrying the *ABCC6* duplication exhibited symptoms characteristic of PXE, including hypertension, atherosclerosis, stroke, mood disorders, and ocular features such as cataracts. The participant also reported that all immediate family members had a history of heart disease, albeit segregation analysis of the duplication was not possible. Although the specific CNV we identified has not been previously reported, a full gene duplication of *ABCC6* was reported in gnomAD (v2.1.1 non-neuro; https://gnomad.broadinstitute.org/) (Karczewski et al., 2020). The duplication was identified in 18 gnomAD samples (allele frequency = 1.08E-3); however, all individuals were under the age of 60 years, and it remains unclear whether any of the individuals presented with features of CVD. To our knowledge, this is the first reporting of a large-scale duplication involving *ABCC6* in an individual with CVD. Although this CNV spanned 42 other genes (**Table 5.2**), there is currently no evidence suggesting that

structural variation of these other genes contributed to the participant's disease presentation.

We also identified and validated a duplication of exons 7–11 in *SAMHD1* in a subject with MCI, and identified duplications of high confidence in exon 16 of both *ABCA7* and *ATP13A2* in two different participants with CVD. Previous studies have associated large duplication events in *ABCA7* and *SAMHD1* with developmental delay and autism (Coe et al., 2014; Cooper et al., 2011); however, no reports of neurodegenerative symptoms have been made in patients demonstrating similar CNVs. Additionally, small-scale duplications within exon 16 of *ATP13A2* have been linked to Kufor-Rakeb syndrome, a juvenile-onset condition characterized by parkinsonism and dementia, although full duplications of exon 16 have not been previously described (Ramirez et al., 2006). Subject 6, who harboured the *ATP13A2* duplication, was diagnosed with CVD and did not present with clinical features of PD at baseline, nor upon two-year follow-up. Further evaluation of these CNVs will be needed to gain a better understanding of their contribution to neurodegeneration and CVD.

Although we have high confidence in the presence of six CNVs across the ONDRI participants, we were able to confirm the exact breakpoints of only one, namely the deletion of exon 5 in *OPTN*. Identification of CNVs using NGS is limited to only determine which NGS probes are affected by the structural variant, thereby requiring further analysis to determine CNV breakpoints. However, breakpoint analysis remains challenging for duplications, as it is unclear whether the duplicated sequence will appear in tandem with the original sequence, or will be inserted unpredictably into a distal region of the genome. Therefore, we were unable to determine the exact location of the identified duplications and whether they may be interrupting other important genomic sequences that could contribute to the neurodegenerative phenotypes. Further, 41 CNVs identified using DOC analysis of ONDRISeq data remain unvalidated with average DOC ratios, average z-scores, and p-values of unknown confidence. Confirmation of these CNVs using alternative methods will be required. Despite these limitations, DOC CNV detection with targeted NGS continues to produce comprehensive, high-quality data, while remaining more time- and cost-effective than the 'gold-standard' Sanger

sequencing or multiplex ligation-dependent probe amplification (MLPA) approaches (Iacocca et al., 2017).

5.6.   Conclusion

In summary, we were able to identify potentially pathogenic, novel CNVs with high confidence in 6 individuals who were diagnosed with neurodegenerative disease or CVD. Further, we present an additional 41 potential CNVs that will be candidates for future replication studies. Although functional analyses are still required to determine how the CNVs may contribute to pathologic mechanisms of disease, the results highlight the need for further investigation into structural variants and their impact on neurodegenerative and cerebrovascular phenotypes, as they may account for a portion of the missing heritability observed across the individual diagnoses. Assessing the full spectrum of potential variants that can contribute to the disease states is imperative for a complete understanding of the genetic etiology of these highly prevalent and progressive conditions, which, in due course, will contribute to more accurate genetic diagnostic screening and therapeutic targeting.

5.7.   References

Annesi, G., Savettieri, G., Pugliese, P., D'Amelio, M., Tarantino, P., Ragonese, P., . . . Quattrone, A. (2005). DJ-1 mutations and parkinsonism-dementia-amyotrophic lateral sclerosis complex. *Ann Neurol, 58*(5), 803-807. doi:10.1002/ana.20666

Barber, S. C., & Shaw, P. J. (2010). Oxidative stress in ALS: key role in motor neuron injury and therapeutic target. *Free Radic Biol Med, 48*(5), 629-641. doi:10.1016/j.freeradbiomed.2009.11.018

Bergen, A. A., Plomp, A. S., Schuurman, E. J., Terry, S., Breuning, M., Dauwerse, H., . . . de Jong, P. T. (2000). Mutations in ABCC6 cause pseudoxanthoma elasticum. *Nat Genet, 25*(2), 228-231. doi:10.1038/76109

Bonifati, V., Rizzu, P., Squitieri, F., Krieger, E., Vanacore, N., van Swieten, J. C., . . . Heutink, P. (2003). DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism. *Neurol Sci, 24*(3), 159-160. doi:10.1007/s10072-003-0108-0

Cacace, R., Sleegers, K., & Van Broeckhoven, C. (2016). Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimers Dement, 12*(6), 733-748. doi:10.1016/j.jalz.2016.01.012

Coe, B. P., Witherspoon, K., Rosenfeld, J. A., van Bon, B. W., Vulto-van Silfhout, A. T., Bosco, P., . . . Eichler, E. E. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet, 46*(10), 1063-1071. doi:10.1038/ng.3092

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., . . . Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nat Genet, 43*(9), 838-846. doi:10.1038/ng.909

Cuccaro, D., De Marco, E. V., Cittadella, R., & Cavallaro, S. (2017). Copy Number Variants in Alzheimer's Disease. *J Alzheimers Dis, 55*(1), 37-52. doi:10.3233/JAD-160469

Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2018). Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J Vis Exp*(134). doi:10.3791/57266

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., . . . Strong, M. J. (2017). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci, 44*(2), 196-202. doi:10.1017/cjn.2016.415

Farhan, S. M. K., Dilliott, A. A., Ghani, G., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*(16032), 1-11.

Ghani, M., Pinto, D., Lee, J. H., Grinberg, Y., Sato, C., Moreno, D., . . . Rogaeva, E. (2012). Genome-wide survey of large rare copy number variants in Alzheimer's disease among Caribbean hispanics. *G3 (Bethesda), 2*(1), 71-78. doi:10.1534/g3.111.000869

Greaves, C. V., & Rohrer, J. D. (2019). An update on genetic frontotemporal dementia. *J Neurol, 266*(8), 2075-2086. doi:10.1007/s00415-019-09363-4

Hagenaars, S. P., Radakovic, R., Crockford, C., Fawns-Ritchie, C., International, F. T. D. G. C., Harris, S. E., . . . Deary, I. J. (2018). Genetic risk for neurodegenerative disorders, and its overlap with cognitive ability and physical function. *PLoS One, 13*(6), e0198187. doi:10.1371/journal.pone.0198187

Hanagasi, H. A., Giri, A., Kartal, E., Guven, G., Bilgic, B., Hauser, A. K., . . . Lohmann, E. (2016). A novel homozygous DJ1 mutation causes parkinsonism and ALS in a Turkish family. *Parkinsonism Relat Disord, 29*, 117-120. doi:10.1016/j.parkreldis.2016.03.001

Hooli, B. V., Kovacs-Vajna, Z. M., Mullin, K., Blumenthal, M. A., Mattheisen, M., Zhang, C., . . . Tanzi, R. E. (2014). Rare autosomal copy number variations in early-onset familial Alzheimer's disease. *Mol Psychiatry, 19*(6), 676-681. doi:10.1038/mp.2013.77

Iacocca, M. A., Wang, J., Dron, J. S., Robinson, J. F., McIntyre, A. D., Cao, H., & Hegele, R. A. (2017). Use of next-generation sequencing to detect LDLR gene copy number variation in familial hypercholesterolemia. *J Lipid Res, 58*(11), 2202-2209. doi:10.1194/jlr.D079301

Iida, A., Hosono, N., Sano, M., Kamei, T., Oshima, S., Tokuda, T., . . . Ikegawa, S. (2012). Novel deletion mutations of OPTN in amyotrophic lateral sclerosis in Japanese. *Neurobiol Aging, 33*(8), 1843 e1819-1824. doi:10.1016/j.neurobiolaging.2011.12.037

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434-443. doi:10.1038/s41586-020-2308-7

Kauw, F., Kranenburg, G., Kappelle, L. J., Hendrikse, J., Koek, H. L., Visseren, F. L. J., . . . Spiering, W. (2017). Cerebral disease in a nationwide Dutch pseudoxanthoma elasticum cohort with a systematic review of the literature. *J Neurol Sci, 373*, 167-172. doi:10.1016/j.jns.2016.12.053

Keller, M. F., Saad, M., Bras, J., Bettella, F., Nicolaou, N., Simon-Sanchez, J., . . . Wellcome Trust Case Control, C. (2012). Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Hum Mol Genet, 21*(22), 4996-5009. doi:10.1093/hmg/dds335

Kringen, M. K., Stormo, C., Berg, J. P., Terry, S. F., Vocke, C. M., Rizvi, S., . . . Piehler, A. P. (2015). Copy number variation in the ATP-binding cassette transporter ABCC6 gene and ABCC6 pseudogenes in patients with pseudoxanthoma elasticum. *Mol Genet Genomic Med, 3*(3), 233-237. doi:10.1002/mgg3.137

Lev, N., Roncevic, D., Ickowicz, D., Melamed, E., & Offen, D. (2006). Role of DJ-1 in Parkinson's disease. *J Mol Neurosci, 29*(3), 215-225. doi:10.1385/jmn:29:3:215

Li, H., Wetten, S., Li, L., St Jean, P. L., Upmanyu, R., Surh, L., . . . Roses, A. D. (2008). Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol, 65*(1), 45-53. doi:10.1001/archneurol.2007.3

Liu, Y. H., & Tian, T. (2011). Hypothesis of optineurin as a new common risk factor in normal-tension glaucoma and Alzheimer's disease. *Med Hypotheses, 77*(4), 591-592. doi:10.1016/j.mehy.2011.06.040

Macedo, M. G., Verbaan, D., Fang, Y., van Rooden, S. M., Visser, M., Anar, B., . . . Heutink, P. (2009). Genotypic and phenotypic characteristics of Dutch patients with early onset Parkinson's disease. *Mov Disord, 24*(2), 196-203. doi:10.1002/mds.22287

Maruyama, H., Morino, H., Ito, H., Izumi, Y., Kato, H., Watanabe, Y., . . . Kawakami, H. (2010). Mutations of optineurin in amyotrophic lateral sclerosis. *Nature, 465*(7295), 223-226. doi:10.1038/nature08971

Mejzini, R., Flynn, L. L., Pitout, I. L., Fletcher, S., Wilton, S. D., & Akkari, P. A. (2019). ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Front Neurosci, 13*, 1310. doi:10.3389/fnins.2019.01310

Morello, G., Guarnaccia, M., Spampinato, A. G., La Cognata, V., D'Agata, V., & Cavallaro, S. (2018). Copy Number Variations in Amyotrophic Lateral Sclerosis: Piecing the Mosaic Tiles Together through a Systems Biology Approach. *Mol Neurobiol, 55*(2), 1299-1322. doi:10.1007/s12035-017-0393-x

Nuytemans, K., Theuns, J., Cruts, M., & Van Broeckhoven, C. (2010). Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update. *Hum Mutat, 31*(7), 763-780. doi:10.1002/humu.21277

Pavlovic, A. M., Zidverc-Trajkovic, J., Milovic, M. M., Pavlovic, D. M., Jovanovic, Z., Mijajlovic, M., . . . Sternic, N. (2005). Cerebral small vessel disease in pseudoxanthoma elasticum: three cases. *Can J Neurol Sci, 32*(1), 115-118. doi:10.1017/s0317167100016991

Ramirez, A., Heimbach, A., Grundemann, J., Stiller, B., Hampshire, D., Cid, L. P., . . . Kubisch, C. (2006). Hereditary parkinsonism with dementia is caused by

mutations in ATP13A2, encoding a lysosomal type 5 P-type ATPase. *Nat Genet, 38*(10), 1184-1191. doi:10.1038/ng1884

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., . . . Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444-454. doi:10.1038/nature05329

Ringpfeil, F., Lebwohl, M. G., Christiano, A. M., & Uitto, J. (2000). Pseudoxanthoma elasticum: mutations in the MRP6 gene encoding a transmembrane ATP-binding cassette (ABC) transporter. *Proc Natl Acad Sci U S A, 97*(11), 6001-6006. doi:10.1073/pnas.100041297

Schilter, K. F., Reis, L. M., Sorokina, E. A., & Semina, E. V. (2015). Identification of an Alu-repeat-mediated deletion of OPTN upstream region in a patient with a complex ocular phenotype. *Mol Genet Genomic Med, 3*(6), 490-499. doi:10.1002/mgg3.159

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., . . . Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science, 315*(5813), 848-853. doi:10.1126/science.1136678

Sunderland, K. M., Beaton, D., Arnott, S. R., Kleinstiver, P. W., Kwan, D., Lawrence-Dewar, J. M., . . . Binns, M. A. (2020). The Ontario Neurodegenerative Disease Research Initiative. *medRxiv*, 2020.2007.2030.20165456. doi:10.1101/2020.07.30.20165456

Chapter 6 – Rare variation within *GBA* and its associations across multiple neurodegenerative diseases.

The work presented in <u>Chapter 6</u> has been edited from the original manuscript currently under review at *Parkinsonism & Related Disorders* for brevity and consistency throughout the entire Dissertation.

Dilliott A.A., Vaillancourt A., Wang J., Beaulieu L., Evans E., Sunderland K.M., Abrahao A., Binns M.A., Black S.E., Borrie M., Breen D.P., Dowlatshahi D., Finger E., Fischer C.E., Frank A., Freedman M., Grimes D., Hassan A., Jog M., Kumar S., Lang A.E., Mandzia J., Marras C., Masellis M., Pasternak S., Pollock D.G., Rajji T.K., Rogaeva E., Sahlas D.J., Saposnik G., Sato C., Seitz D., Shoesmith C., Steeves T., Swartz R.H., Tan B., Tang-Wai D., Tartaglia M.C., Turnbull J., Zinman L., ONDRI Investigators, and Hegele R.A. (2021). Rare variation within *GBA* and its associations across multiple neurodegenerative diseases. *Parkinsonism Relat Disord*, [Under Review]. Ms. No. PARKRELDIS-D-21-00522.

6.1.   Abstract

Background/Objective: Pathogenic homozygous variants in *GBA* are well known to cause Gaucher's disease, while heterozygous variants are the most common genetic risk factor for Parkinson's disease (PD). However, it is increasingly recognized that some neurodegenerative diseases may show overlap in terms of both clinical expression and underlying pathology, suggesting overlap of underlying genetic risk factors. Yet *GBA* has not been assessed in cohorts with neurodegeneration other than classic synucleinopathies. Here, we aimed to assess rare, likely pathogenic *GBA* variants across the six disease cohorts encompassed by the Ontario Neurodegenerative Disease Research Initiative (ONDRI), and determine whether *GBA* variants were associated with disease presentation.

Methods and Results: We found that likely pathogenic *GBA* variants were significantly enriched in the PD cohort versus other ONDRI disease cohorts. However, we observed a high frequency of *GBA* variants across the entire sample, including among participants with Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD), mild cognitive impairment (MCI), and cerebrovascular disease with or without cognitive impairment (CVD ± CI), compared to previously reported control cohorts, potentially indicative of underlying Lewy body co-pathologies previously unrecognized in the ONDRI participants. We also identified novel collections of rare variants in *GBA* exons five and six across all ONDRI disease phenotypes. However, *GBA* variant status was not associated with clinical phenotypes across the cohorts.

Conclusion: Our findings indicate that diverse neurodegenerative diagnoses may have common underlying genetic risk factors — in this case, rare *GBA* variants — which could help with understanding the pathogenesis of these diseases and possible interventions.

6.2.   Introduction

The glucosylceramidase beta gene (*GBA*), located on chromosome 1q21, encodes the lysosomal enzyme glucocerebrosidase (GCase) that cleaves glucosylceramide within the cell membrane. Bi-allelic pathogenic *GBA* mutations cause Gaucher's disease, an autosomal recessive disorder characterized by a range of non-neurological and neurological features, including dementia and parkinsonism. Moreover, mono-allelic or heterozygous *GBA* variants are a common risk factor for developing Parkinson's disease (PD), with 5–25% of affected individuals carrying a pathogenic *GBA* variant (Sidransky et al., 2009). Pathogenic variants in *GBA* are also enriched in cohorts of patients with Lewy body disease (Mata et al., 2008; Shiner et al., 2016).

PD is a widely prevalent neurodegenerative disease characterized by both motor symptoms, such as bradykinesia, resting tremor, postural instability, and rigidity, and non-motor symptoms, such as cognitive impairment, depression, autonomic dysfunction, impaired sleep, and hyposmia, although presentation is highly heterogeneous (Poewe et al., 2017). A signature brain pathology of PD is the presence of Lewy bodies — intraneuronal inclusions containing aggregates of α-synuclein — the accumulation of which may lead to neurotoxicity and depending on distribution can contribute to dementia in PD patients (Burre, Sharma, & Sudhof, 2015; Poewe et al., 2017).

Although exactly how *GBA* variation contributes to the risk of PD development is unknown, a few hypotheses are being explored. First, GCase is a lysosomal hydrolase involved in the endolysosomal pathway, often implicated in PD pathology (Klein & Mazzulli, 2018), and its substrate, glucocerebroside, can accumulate in dopaminergic neurons, leading to cellular stress and possible interruption of dopamine production. Additionally, dysregulated GCase has also been associated with accumulation of α-synuclein (Mazzulli et al., 2011). Compared to patients with idiopathic PD, *GBA* PD patients tend to develop onset of clinical symptoms at an earlier age, have a greater motor decline, and are more likely to present with non-motor symptoms, such as cognitive impairment (Blauwendraat, Nalls, & Singleton, 2020; Gan-Or, Liong, & Alcalay, 2018).

GBA is currently considered to be an important target for PD and Lewy body disease therapeutics, with one small positive open label trial of a GCase pharmacological chaperone published and several phase 2 clinical trials underway (ClinicalTrials.gov: NCT02941822; NCT02914366; NCT04588285) (Mullin et al., 2020); yet it remains unclear whether variants within the gene contribute to other neurodegenerative diagnoses. Furthermore, the substantial overlap between features of non-PD neurodegenerative diseases and the frequent co-occurrence of neuropathology — including α-synuclein (Cairns et al., 2015) — suggests potentially common genetic mechanisms such as *GBA*. As *GBA* has not been systematically studied in neurodegenerative diseases other than synucleinopathies its involvement in the presentation of various neurodegenerative diseases is not known (Gan-Or et al., 2018).

To study the association of *GBA* variants with a range of neurodegenerative phenotypes, we used Sanger sequencing to identify all likely pathogenic non-synonymous rare variants in *GBA* harboured by individuals from the Ontario Neurodegenerative Disease Research Initiative (ONDRI) affected by: PD, Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD), mild cognitive impairment (MCI), and cerebrovascular disease with or without cognitive impairment (CVD ± CI). Furthermore, we investigated the relationship between *GBA* variants and clinical features in these patient cohorts.

6.3.   Methods

6.3.1.   Study participants

ONDRI enrolled 520 participants with neurodegenerative disease who provided informed consent, most of whom were in the early stages of disease progression (Farhan et al., 2017; Sunderland et al., 2020). Of those participants, 519 provided blood samples, from which genomic DNA was isolated as previously described (Dilliott et al., 2018). Study ethics approval was obtained from the Research Ethics Boards at all recruitment sites. All participants provided written, informed consent in accordance with the Research Ethics Boards and regulatory requirements. Clinical data were obtained upon participant enrollment including demographics, ages of symptom onset, Montreal Cognitive

Assessment (MoCA), and Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS).

Using the MDS-UPDRS, axial scores were calculated to represent motor severity across the disease cohorts. The score encompassed the summation of subscales, including those pertaining to difficulties with speech, swallowing, turning over in bed, freezing of gait, arising from chairs, gait, postural stability, and generalized posture (MDS-UPDRS subsections: 2.1, 2.3, 2.9, 2.13, 3.9, 3.10, 3.11, 3.12, and 3.13) (Bohnen et al., 2017; O'Gorman Tuura, Baumann, & Baumann-Vogel, 2018). Although there is no single scale fully appropriate for this purpose across diagnoses, the MDS-UPDRS was available for most participants and, aside from freezing of gait, all other metrics encompassed in the subscales can be observed in the full range of neurodegenerative diseases.

6.3.2. Sanger sequencing

All exons of *GBA* were Sanger sequenced in the ONDRI participants. Briefly, polymerase chain reaction (PCR) amplifications were carried out, yielding three amplicons that contained: 1) exons 1–4; 2) exons 5–7; and 3) exons 8–11; the most highly specific of previously published primers were used to provide specificity for isolating *GBA* (Mata et al., 2016). Sanger sequencing of each exon of *GBA* was performed using previously established primer sets (Mata et al., 2016). All 11 exons of *GBA* were sequenced, including 100 base pairs at each intron-exon boundary. Primer sequences are listed in **Appendix H**.

6.3.3. Variant annotation and prioritization

Variants were mapped to their genomic, mRNA, and protein coding positions using the NG_009783.1, NM_000157.4, and NP_000148.2 reference sequences, respectively (https://www.ncbi.nlm.nih.gov/gene/). Minor allele frequencies (MAFs) were obtained using the Genome Aggregation Database (gnomAD; v2.1 non-neuro; N = 114,704) (Karczewski et al., 2020). *In silico* predictions of variant deleteriousness were obtained from Combined Annotation Dependent Depletion Phred scoring (CADD Phred) (Kircher et al., 2014) and variant pathogenicity was investigated using ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/).

All non-synonymous and splicing region variants were prioritized to identify those most likely to be pathogenic. Variants were first prioritized to include only those that were rare, with a MAF < 0.01 in gnomAD. Variants were further prioritized to identify those with at least one of the following: a) a CADD Phred score >15; b) a classification of likely pathogenic or pathogenic in ClinVar; or c) a sequence ontology of putative loss-of-function (frameshift, splice-site donor, splice-site acceptor, and nonsense variants). All prioritized variants were referred to as "likely pathogenic variants." Individuals who did and did not harbour likely pathogenic variants were referred to as "variant positive" and "variant negative," respectively.

6.3.4.   Statistical analyses

To determine whether individuals with the *GBA* likely pathogenic variants were more likely to be diagnosed with PD than one of the other ONDRI diseases, we used logistic regression, adjusted for age and sex.

The contribution of *GBA* variant status to clinical outcomes was also compared between variant positive and variant negative participants across the ONDRI cohorts. A multivariate multiple regression model was used to estimate the influence of *GBA* variant status within the disease cohorts, adjusting for age and sex, on age of symptom onset, total MoCA score, and MDS-UPDRS axial score (Johnson & Wichern, 2001). The interaction terms between the two predictor variables of interest were included to identify any disease-specific effects of *GBA* carrier status. To use the sample mean of all participants, regardless of cohort, as a point of reference, disease cohorts were transformed using weighted-effect coding with the *wec* R package (v0.4-1) (Nieuwenhuis et al., 2017) as previously described (Dilliott et al., 2021; Te Grotenhuis et al., 2017). A multivariate regression, assessing the clinical outcomes between variant positive and variant negative participants in only the PD cohort confirmed results in the singular disease cohort. Wilks' lambda assessed predictor contributions across the three clinical outcomes, where significance indicated that the corresponding predictor contribution was non-zero for at least one of the three outcomes.

Significance for the regression analyses was evaluated at an alpha-level of 0.05. All statistical analyses were performed using the R statistical software 3.6.0 (R Core Team, 2014) in R Studio 1.1.463 and data visualization was performed using GraphPad Prism 9.

6.4. Results

6.4.1. Likely pathogenic *GBA* variants

GBA was successfully sequenced in 515 of the 519 participants (**Table 6.1**), although data from at least one exon was missing for 12 of the sequenced participants (**Appendix I**). Of the sequenced participants, 195 (38%) harboured at least one of 47 unique *GBA* variants. Following variant prioritization, 74 participants were found to harbour at least one of 18 unique *GBA* likely pathogenic heterozygous variants (**Table 6.2**; **Figure 6.1**).

Eight of the likely pathogenic variants were exclusively found within exons 5 and 6 and were each identified in a relatively high number of ONDRI participants (n = 4–46), resulting in ONDRI-wide MAFs ranging from 0.008–0.089, although each variant had an MAF in gnomAD $\leq$ 8.17E-05. Therefore, the variants were considered common among ONDRI participants. Further, 47 participants across all ONDRI cohorts harboured at least two of these variants together, hereafter referred to as complex variants (**Figure 6.1**). Of the seven identified complex variants, one consisted of two rare likely pathogenic variants (c.681T>G, p.N227K and c.689T>G, p.V230G), which was exclusively present in 15 PD participants. An additional two complex variants were each identified in a single PD patient. The remaining four complex variants were found within patients across the five other ONDRI cohorts, excluding PD (**Table 6.3**).

Aside from the variants present within exons five and six, the most common variants found within the ONDRI cohort were c.C1223T, p.T408M and c.A1226G, p.N409S, which were present in ten and nine participants, respectively (ONDRI-wide MAFs = 0.019 and 0.017, respectively; **Table 6.2**).

Although *GBA* likely pathogenic variants were identified in all ONDRI cohorts, logistic regression confirmed that participants carrying *GBA* likely pathogenic variants were more likely to be in the PD cohort compared to the AD, ALS, CVD, FTD, and MCI cohorts

combined (OR = 1.793 [1.061–3.031], p = 0.0293). However, *GBA* complex variants were not found to display the same associations (OR = 1.614 [0.852–3.057], p = 0.1416).

Table 6.1 Demographics of the 515 ONDRI participants successfully sequenced for *GBA* and their clinical outcomes.

	Participants [n (%)]	Percent male (%)	Age (mean ± sd)	Participants with clinical data (n)	Age of onset (mean ± sd)¹	MoCA (mean ± sd)¹	Axial score (mean ± sd)¹
AD							
GBA (+)	4 (9.8%)	100	72.8 ± 4.7	4	68.3 ± 7.2	20.0 ± 2.9	1.75 ± 2.9
GBA (-)	37 (90.2%)	54.1	71.7 ± 8.3	35	67.0 ± 8.6	21.0 ± 2.7	1.31 ± 2.0
ALS							
GBA (+)	6 (15.4%)	50	65.8 ± 6.1	4	65.1 ± 4.3	24.5 ± 2.9	3.75 ± 4.9*
GBA (-)	33 (84.6%)	60.6	61.3 ± 9.2	26	59.4 ± 9.5	25.7 ± 2.9	7.73 ± 5.7
CVD ± CI							
GBA (+)	15 (9.3%)	73.3	67.5 ± 6.9	15	65.2 ± 8.0	25.5 ± 2.1	1.07 ± 1.8
GBA (-)	146 (90.7%)	67.8	69.4 ± 7.5	145	67.3 ± 8.5	25.2 ± 3.1	1.77 ± 2.5
FTD							
GBA (+)	8 (15.7%)	62.5	69.6 ± 5.0	8	65.3 ± 5.7	23.1 ± 3.5	7.12 ± 7.8
GBA (-)	43 (84.3%)	62.8	67.4 ± 7.5	42	62.5 ± 7.7	21.1 ± 4.0	5.57 ± 7.1
MCI							
GBA (+)	12 (14.3%)	75	69.9 ± 7.1	10	68.9 ± 6.6	23.4 ± 3.6	1.00 ± 2.2
GBA (-)	72 (85.7%)	50	70.8 ± 8.5	63	67.0 ± 9.5	23.3 ± 2.7	0.86 ± 1.6
PD							
GBA (+)	28 (20.1%)	71.4	66.9 ± 6.4	28	61.3 ± 6.6	25.7 ± 2.7	4.89 ± 3.5
GBA (-)	111 (79.9%)	79.3	68.1 ± 6.3	111	61.6 ± 6.7	25.9 ± 2.5	4.45 ± 3.4

Clinical data were not available for all participants. ¹denotes clinical variables calculated based on the sample sizes outlined in "Participants with clinical data." Axial scores were derived from the Movement Disorder Society-Unified Parkinson's Disease Rating Scale part three (MDS-UPDRS) with lower scores referring to less motor impairment. Multivariate multiple linear regression was used to assess the association between *GBA* variant status and age of symptom onset, MoCA scores, and axial scores. *p-value < 0.05.

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; *GBA*, glucosylceramidase beta gene; MCI, mild cognitive impairment; MoCA, Montreal Cognitive Assessment; n, sample size; PD, Parkinson's disease; sd, standard deviation.

Table 6.2 Rare, likely deleterious variants identified within the *GBA* gene harboured by ONDRI participants.

Variant	Protein alteration	Reference SNP identifier	Exon	Sequence ontology	MAF (gnomAD)	CADD Phred	ClinVar (LP/P)	Number of carriers
c.115+1G>A	NA	rs104886460	2	Splice Donor	7.42E-05	32.0	GD and PD	1
c.C475T	p.R159W	rs439898	5	Missense	NA	25.1	GD	29
c.T667C	p.W223R	rs61748906	6	Missense	9.63E-06	22.0	GD	4
c.T681G	p.N227K	rs381418	6	Missense	8.17E-05	13.25	GD	19
c.680_681delinsGG	p.N227R	NA	6	Missense	NA	NA	Phenotype NP	28
c.T689G	p.V230G	rs381427	6	Missense	1.44E-05	13.40	GD*	46
c.T703C	p.S235P	rs1064644	6	Missense	9.61E-06	14.61	GD	31
c.G721A	p.G241R	rs409652	6	Missense	3.84E-05	23.5	GD	31
c.T754A	p.F252I	rs381737	6	Missense	4.36E-05	22.5	GD and PD*	32
c.G887A	p.R296Q	rs78973108	7	Missense	4.36E-05	25.90	LBD, PD, and GD	1
c.C1073T	p.P358L	NA	8	Missense	NA	26.9	NA	1
c.C1223T	p.T408M	rs75548401	8	Missense	5.89E-03	22.2	PD*	10
c.A1226G	p.N409S	rs76763715	9	Missense	1.86E-03	23.8	LBD, PD, GD	9
c.T1249G	p.W417G	rs1450426641	9	Missense	NA	27.80	NA	1
c.1265–1319del	p.L422Pfs3X	NA	9	Frameshift deletion	NA	NA	NA	1
c.T1448C	p.L483P	rs421016	10	Missense	1.28E-03	23.6	NA	2
c.G1483C	p.A495P	rs368060	10	Missense	1.44E-04	17.68	NA	1
c.G1495C	p.V499L	rs369068553	10	Missense	5.77E-05	18.57	GD	1

*Phenotype not explicitly listed as likely pathogenic or pathogenic in ClinVar, but the ClinVar entry includes many publications that indicate potential pathogenicity of the phenotype. "Phenotype NP" refers to the variant having a ClinVar designation of likely pathogenic or pathogenic, but no definitive phenotype being specified.

Abbreviations: CADD Phred, Combined annotation dependent depletion Phred score; *GBA*, glucosylceramidase beta gene; GD, Gaucher's disease; gnomAD, Genome Aggregation Database (v2.1 non-neuro); MAF, minor allele frequency; LBD, Lewy body dementia; LP, likely pathogenic; NA, not applicable; NP, not provided; ONDRI, Ontario Neurodegenerative Disease Research Initiative; PD, Parkinson's disease; P, pathogenic; SNP, Single nucleotide polymorphism.

Table 6.3 Complex variants identified within *GBA* exons five and six harboured by ONDRI participants according to disease cohort.

Complex Variant	Variant carriers in ONDRI					
	AD	**ALS**	**CVD ± CI**	**FTD**	**MCI**	**PD**
c.681T>G, p.N227K; c.689T>G, p.V230G	0	0	0	0	0	15
c.475C>T, p.R159W; c.680_681delinsGG, p.N227R; c.689T>G, p.V230G; c.703T>C, p.S235P; c.721G>A, p.G241R; c.754T>A, p.F252I	1	4	9	4	7	0
c.475C>T, p.R159W; c.680A>G, p.N227S; c.689T>G, p.V230G; c.703T>C, p.S235P; c.721G>A, p.G241R; c.754T>A, p.F252I	0	0	0	0	1	0
c.681T>G, p.N227K; c.689T>G, p.V230G; c.703T>C, p.S235P; c.721G>A, p.G241R; c.754T>A, p.F252I	0	0	0	0	0	1
c.681T>G, p.N227K; c.689T>G, p.V230G; c.754T>A, p.F252I	0	0	0	0	0	1
c.681T>G, p.N227K; c.703T>C, p.S235P; c.721G>A, p.G241R; c.754T>A, p.F252I	0	0	1	0	0	0
c.475C>T, p.R159W; c.667T>C, p.W223R; c.680_681delinsGG, p.N227R; c.689T>G, p.V230G; c.703T>C, p.S235P; c.721G>A, p.G241R; c.754T>A, p.F252I	0	0	1	0	2	0

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; *GBA*, glucosylceramidase beta gene; MCI, mild cognitive impairment; ONDRI, Ontario Neurodegenerative Disease Research Initiative; PD, Parkinson's disease.
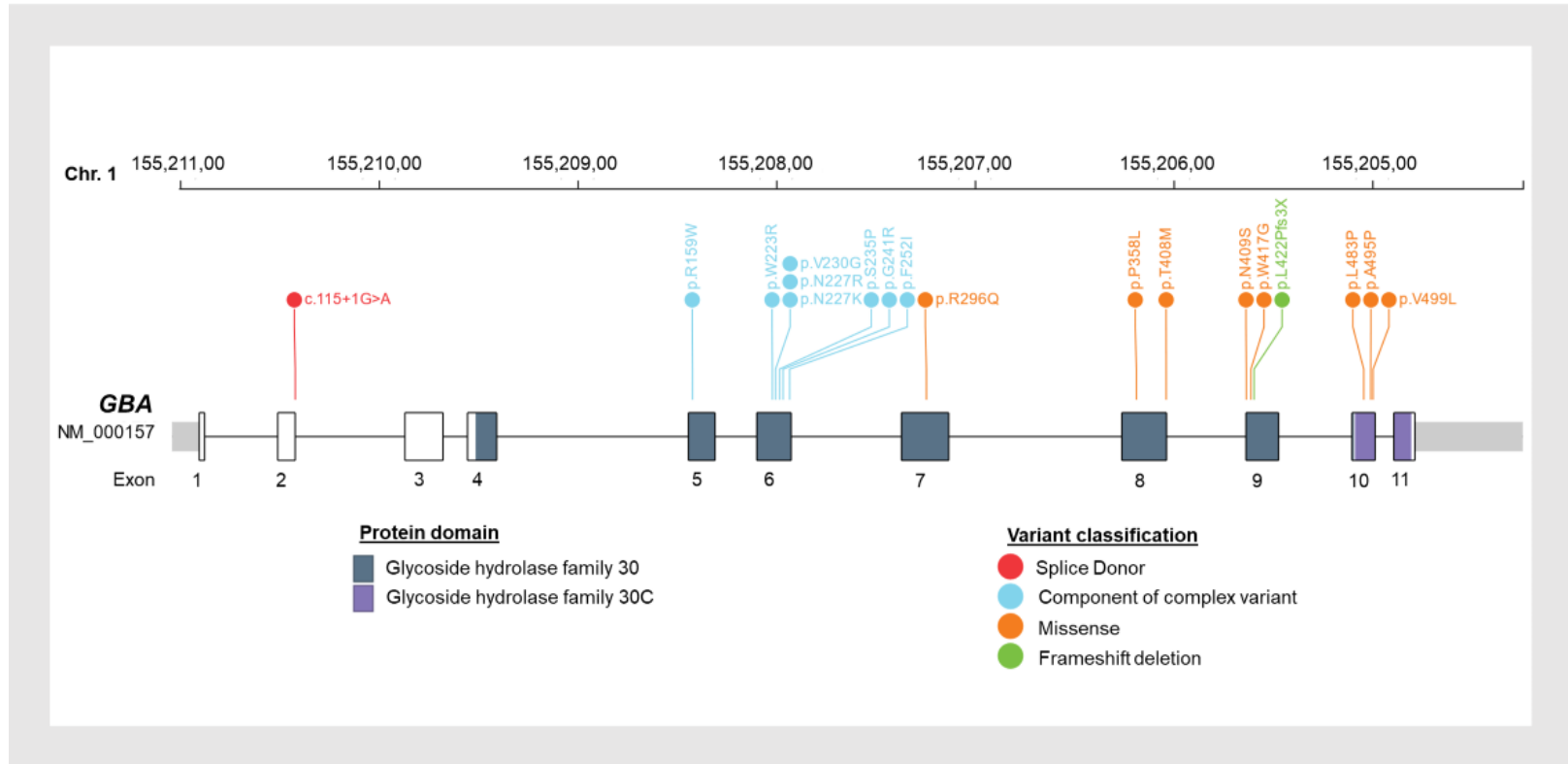
Figure 6.1 Schematic of the likely pathogenic variants identified in *GBA* across 515 ONDRI participants.
GBA variation was assessed using gold-standard Sanger sequencing. Variants were prioritized to identify those that were non-synonymous, rare (MAF < 0.01, gnomAD v2.1 non-neuro), and displayed at least one of the following: a) a CADD Phred score >15; b) a classification of likely pathogenic or pathogenic in ClinVar or Franklin; or c) a sequence ontology of putative loss-of-function (frameshift, splice-site donor, and nonsense variants). Each box in the schematic represents an exon of *GBA*. "Component of complex variant" refers to variants that were identified in exons five and six in various combinations with each other in a relatively high number of ONDRI participants. Figure was made using https://proteinpaint.stjude.org/.

6.4.2. Associations with clinical outcomes

Data for all three clinical outcomes were available for 491 of the ONDRI participants who had been sequenced for *GBA* (**Table 6.1**). The Wilks' lambda test confirmed that the contribution of the disease cohort and *GBA* variant status interaction was significant (Λ = 0.928, p = 0.0021); among participants in the ALS cohort, variant positive status negatively associated with MDS-UPDRS axial scores (β = -4.670, SE = 1.870, p = 0.0129). There were no further associations between *GBA* variant status and the clinical outcomes within the ONDRI cohorts, nor any associations between *GBA* complex variant status and the clinical outcomes.

6.5. Discussion

Here, we evaluated the contribution of *GBA* genetic variation across a spectrum of neurodegenerative disease and CVD phenotypes. We found that: 1) *GBA*-likely pathogenic variants were associated with diagnosis and enrollment in the PD cohort compared to the other ONDRI disease cohorts; 2) *GBA* likely pathogenic variants were still identified across all ONDRI disease cohorts; 3) there were intriguing clusters of *GBA* likely pathogenic variants in exons five and six — referred to as "complex variants" — across all ONDRI participants; and 4) there was not an obvious relationship between *GBA* and clinical phenotypes across the neurodegenerative cohorts.

Of the 18 likely pathogenic variants identified, eight were previously associated with only Gaucher's disease, one was previously associated with only PD, and four were previously associated with both Gaucher's disease and PD (**Table 6.2**). Individuals carrying *GBA* likely pathogenic variants were more likely to be in the PD cohort of ONDRI compared to the remaining five ONDRI disease cohorts. Nonetheless, likely pathogenic variants were identified across all study participants. Thus, the variants may represent novel genetic associations for neurodegenerative disease.

Although control samples sequenced with the same methodology were unavailable for these analyses, all variants were considered rare in the gnomAD non-neurological cohort (MAF < 0.01) and previous reports have suggested that the frequency of heterozygous

likely pathogenic/pathogenic *GBA* variants range from 0–4.0% in elderly control cohorts, dependent on ancestry and spectrum of variants screened (Sato et al., 2005; Sidransky et al., 2009). By comparison, the cohort with the lowest *GBA* variant frequency in ONDRI was the CVD ± CI cohort with a frequency of 9.3%, over double that of the highest frequency in previously reported controls (Sato et al., 2005; Sidransky et al., 2009), leaving us to hypothesize that the *GBA* likely pathogenic variants identified herein are likely contributing to the spectrum of neurodegenerative diseases.

The markedly high frequency of *GBA* likely pathogenic variants across all ONDRI disease cohorts suggests possible undetected pathologic connections between these disease cohorts. Importantly, when *GBA* is mutated, decreased function of the encoded GCase — a lysosomal hydrolase — causes lysosomal dysfunction, a pathway implicated in the pathologic mechanism of many neurodegenerative diseases (Root, Merino, Nuckols, Johnson, & Kukar, 2021), and suggests the variants we observed in *GBA* may be contributing to disease risk across the neurodegenerative diagnoses. Additionally, pathogenic *GBA* variants were previously associated with increased prominence of Lewy bodies in PD patients, as well as increased risk of Lewy body dementia (Mata et al., 2008; Nalls et al., 2013). Many cases of co-pathology between various neurodegenerative diagnoses and Lewy body dementia have been identified — including Lewy bodies and general α-synuclein aggregation in patients with AD (Cairns et al., 2015; Savica et al., 2019), ALS (Ayaki et al., 2018; Doherty, Bird, & Leverenz, 2004), and FTD (Forrest et al., 2019) — which can remain undetected until post-mortem pathologic analysis. Patients with confirmed AD and Lewy body dementia co-pathology are significantly more likely to carry *GBA* mutations than controls (Tsuang et al., 2012). Therefore, while it remains possible that the *GBA* variants increase vulnerability to multiple neurodegenerative diseases through common pathogenic mechanisms such as lysosomal dysfunction, the variants may also suggest a higher admixture of co-pathology than was previously appreciated. Post-mortem analyses will be necessary for validation of this hypothesis.

Of the *GBA* variants identified, a particularly interesting — and relatively novel — pattern arose in exons five and six, resulting in seven distinct complex variants identified

across 47 participants. One of the complex variants was observed in 15 patients with PD, namely the combination of p.N227K and p.V230G. While the variant p.N227K has previously only been associated with Gaucher's disease (Germain, Puech, Caillaud, Kahn, & Poenaru, 1998; Stone et al., 2000), p.V230G has been associated with Gaucher's disease and has also been observed in individuals with PD and with Lewy body dementia (Du, Ding, Chen, Guo, & Wang, 2018; Mitsui et al., 2009). Whether both variants or only one are contributing to PD development remains to be determined. Interestingly, both variants were also observed in combination with other *GBA* variants in ONDRI participants in other disease cohorts, as well as in the two other PD participants carrying complex variants.

Four complex variants were only identified in the AD, ALS, CVD, FTD, and MCI cohorts. Although the complex variants found among the ONDRI participants have not been previously reported in the specific combinations identified, similar patterns of complex variants in exons five and six have been observed on occasion in patients with both Gaucher's disease (Sheth et al., 2019) and PD (Mitsui et al., 2009). Yet, the high frequency and complex patterns of these variants in ONDRI subjects remains novel. Importantly, a pseudogene located downstream of *GBA* — glucosylceramidase beta pseudogene 1 (*GBAP*) — might affect primer alignment during sequencing, resulting in false positive variant calls. However, because exons 5 through 7 were amplified together in a single fragment with highly specific PCR primers, and no complex variants included likely pathogenic variants in exon 7, despite the exon's dissimilarity of sequence with *GBAP*, we are confident that the results are not due to artefacts related to sequencing of the pseudogene. The Sanger sequencing method employed was limited in its ability to discern whether the identified variants within each complex variant were in *cis* or *trans*, although variants in *trans* — i.e. compound heterozygous variants — might have been expected to result in Gaucher's disease. The self-reported ethnicities and family disease histories of all participants harbouring the complex variants were reviewed and no distinct patterns stood out. It has previously been proposed that other complex variants within *GBA*, particularly within exons 9 and 10, could be the result of small-scale rearrangements between *GBA* and *GBAP*, which may be a potential mechanism underlying the complex variants we observed herein (Eyal, Wilder, & Horowitz, 1990).

Ultimately, whether the complex variants indicate a rearrangement and how they may contribute to the spectrum of neurodegenerative diseases must be further explored.

Although we expected *GBA* variation might be associated with clinical phenotypes, such as earlier symptom onset, lower cognitive performance, and increased motor impairment, no definitive differences were observed between variant positive and variant negative ONDRI participants. Surprisingly, these results were similarly not observed within the PD cohort when assessed alone (Blauwendraat et al., 2020; Gan-Or et al., 2018), although follow up analyses with longitudinal data are necessary to determine whether variant positive participants will progress through their disease course differently than those variant negative. The absence of association was potentially a result of our relatively small sample sizes. The only significant result observed among quantitative intermediate traits was a lower axial score, corresponding to less motor impairment, in *GBA* variant positive participants with ALS compared with variant negative participants. However, only four ALS participants were variant-positive and standard deviation of the axial score was relatively large, so we are cautious not to draw conclusions from this result. Further analysis of the contribution of *GBA* variation to the clinical spectrum of various neurodegenerative diseases using larger sample sizes is required.

We recognize that this analysis has a few limitations. Control samples were not available for the analyses to compare *GBA* variant enrichment in the ONDRI cohorts to healthy elderly individuals. Yet all of the *GBA* likely pathogenic variants reported herein were rare in the general population and estimates of variant frequencies in other studies' control samples were readily available (Sato et al., 2005; Sidransky et al., 2009). Further, the entire *GBA* gene could not be sequenced in four ONDRI participants, and individual exons could not be sequenced in 12 additional participants. At least one clinical metric was missing for 24 participants who were subsequently excluded from the intermediate phenotype analysis. The majority of missing variables were due to lack of relevant variables from the MDS-UPDRS Part 3 (95.8% of missing data), and most of the individuals missing the clinical data were from the AD, MCI, or ALS cohorts (91.7% of participants missing data). Finally, a subset of likely pathogenic variants has not been previously associated with PD, or other neurodegenerative diseases, and functional

analyses are required to validate the variants' pathogenicity; however, a portion of these variants has been previously associated with Gaucher's disease, and the remaining variants were determined to be putative loss-of-function variants, or have an *in silico* prediction indicating they are likely deleterious (CADD Phred score > 15).

6.6. Conclusion

The analyses presented herein demonstrate that potentially deleterious *GBA* variation may be more prevalent among various neurodegenerative conditions than previously known. Although individuals with *GBA* likely pathogenic variants were significantly more likely to be in the PD cohort than the other ONDRI disease cohorts, variant frequencies in all ONDRI cohorts were remarkably high in contrast to previously reported elderly control cohorts (Sato et al., 2005; Sidransky et al., 2009). We hypothesize that this enrichment of *GBA* variation may indicate underlying Lewy body co-pathologies previously unrecognized among ONDRI participants or common pathogenic mechanisms such as lysosomal dysfunction. We also report a relatively high frequency of complex variants observed in *GBA* exons 5 and 6 across the neurodegenerative disease and CVD cohorts, which may indicate small-scale rearrangements with the *GBA* pseudogene and warrant further exploration. Our findings highlight the complexity and overlap of genetic factors contributing to various neurodegenerative diagnoses.

6.7.   References

Ayaki, T., Ito, H., Komure, O., Kamada, M., Nakamura, M., Wate, R., . . . Takahashi, R. (2018). Multiple Proteinopathies in Familial ALS Cases With Optineurin Mutations. *J Neuropathol Exp Neurol, 77*(2), 128-138. doi:10.1093/jnen/nlx109

Blauwendraat, C., Nalls, M. A., & Singleton, A. B. (2020). The genetic architecture of Parkinson's disease. *Lancet Neurol, 19*(2), 170-178. doi:10.1016/S1474-4422(19)30287-X

Bohnen, N. I., Haugen, J., Ridder, A., Kotagal, V., Albin, R. L., Frey, K. A., & Muller, M. (2017). Color discrimination errors associate with axial motor impairments in Parkinson's disease. *Mov Disord Clin Pract, 4*(6), 864-869. doi:10.1002/mdc3.12527

Burre, J., Sharma, M., & Sudhof, T. C. (2015). Definition of a molecular pathway mediating alpha-synuclein neurotoxicity. *J Neurosci, 35*(13), 5221-5232. doi:10.1523/JNEUROSCI.4650-14.2015

Cairns, N. J., Perrin, R. J., Franklin, E. E., Carter, D., Vincent, B., Xie, M., . . . Dominantly Inherited Alzheimer, N. (2015). Neuropathologic assessment of participants in two multi-center longitudinal observational studies: the Alzheimer Disease Neuroimaging Initiative (ADNI) and the Dominantly Inherited Alzheimer Network (DIAN). *Neuropathology, 35*(4), 390-400. doi:10.1111/neup.12205

Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2018). Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J Vis Exp*(134). doi:10.3791/57266

Dilliott, A. A., Sunderland, K. M., McLaughlin, P. M., Roberts, A. C., Evans, E. C., Abrahao, A., . . . Hegele, R. A. (2021). Association of apolipoprotein E variation with cognitive impairment across multiple neurodegenerative diagnoses. *Neurobiol Aging*([In Press]). doi:10.1016/j.neurobiolaging.2021.04.011

Doherty, M. J., Bird, T. D., & Leverenz, J. B. (2004). Alpha-synuclein in motor neuron disease: an immunohistologic study. *Acta Neuropathol, 107*(2), 169-175. doi:10.1007/s00401-003-0790-2

Du, X., Ding, Q., Chen, Q., Guo, P., & Wang, Q. (2018). Three mutations of adult type 1 Gaucher disease found in a Chinese patient: A case report. *Medicine (Baltimore), 97*(47), e13161. doi:10.1097/MD.0000000000013161

Eyal, N., Wilder, S., & Horowitz, M. (1990). Prevalent and rare mutations among Gaucher patients. *Gene, 96*(2), 277-283. doi:10.1016/0378-1119(90)90264-r

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., . . . Strong, M. J. (2017). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci, 44*(2), 196-202. doi:10.1017/cjn.2016.415

Forrest, S. L., Crockford, D. R., Sizemova, A., McCann, H., Shepherd, C. E., McGeachie, A. B., . . . Halliday, G. M. (2019). Coexisting Lewy body disease and clinical parkinsonism in frontotemporal lobar degeneration. *Neurology, 92*(21), e2472-e2482. doi:10.1212/WNL.0000000000007530

Gan-Or, Z., Liong, C., & Alcalay, R. N. (2018). GBA-Associated Parkinson's Disease and Other Synucleinopathies. *Curr Neurol Neurosci Rep, 18*(8), 44. doi:10.1007/s11910-018-0860-4

Germain, D. P., Puech, J. P., Caillaud, C., Kahn, A., & Poenaru, L. (1998). Exhaustive screening of the acid beta-glucosidase gene, by fluorescence-assisted mismatch analysis using universal primers: mutation profile and genotype/phenotype correlations in Gaucher disease. *Am J Hum Genet, 63*(2), 415-427. doi:10.1086/301969

Johnson, R. A., & Wichern, D. W. (2001). Applied Multivariate Statistical Analysis. In. New York, USA: Prentice Hall.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434-443. doi:10.1038/s41586-020-2308-7

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi:10.1038/ng.2892

Klein, A. D., & Mazzulli, J. R. (2018). Is Parkinson's disease a lysosomal disorder? *Brain, 141*(8), 2255-2262. doi:10.1093/brain/awy147

Mata, I. F., Leverenz, J. B., Weintraub, D., Trojanowski, J. Q., Chen-Plotkin, A., Van Deerlin, V. M., . . . Zabetian, C. P. (2016). GBA Variants are associated with a distinct pattern of cognitive deficits in Parkinson's disease. *Mov Disord, 31*(1), 95-102. doi:10.1002/mds.26359

Mata, I. F., Samii, A., Schneer, S. H., Roberts, J. W., Griffith, A., Leis, B. C., . . . Zabetian, C. P. (2008). Glucocerebrosidase gene mutations: a risk factor for Lewy body disorders. *Arch Neurol, 65*(3), 379-382. doi:10.1001/archneurol.2007.68

Mazzulli, J. R., Xu, Y. H., Sun, Y., Knight, A. L., McLean, P. J., Caldwell, G. A., . . . Krainc, D. (2011). Gaucher disease glucocerebrosidase and alpha-synuclein form a bidirectional pathogenic loop in synucleinopathies. *Cell, 146*(1), 37-52. doi:10.1016/j.cell.2011.06.001

Mitsui, J., Mizuta, I., Toyoda, A., Ashida, R., Takahashi, Y., Goto, J., . . . Tsuji, S. (2009). Mutations for Gaucher disease confer high susceptibility to Parkinson disease. *Arch Neurol, 66*(5), 571-576. doi:10.1001/archneurol.2009.72

Mullin, S., Smith, L., Lee, K., D'Souza, G., Woodgate, P., Elflein, J., . . . Schapira, A. H. V. (2020). Ambroxol for the Treatment of Patients With Parkinson Disease With and Without Glucocerebrosidase Gene Mutations: A Nonrandomized, Noncontrolled Trial. *JAMA Neurol, 77*(4), 427-434. doi:10.1001/jamaneurol.2019.4611

Nalls, M. A., Duran, R., Lopez, G., Kurzawa-Akanbi, M., McKeith, I. G., Chinnery, P. F., . . . Sidransky, E. (2013). A multicenter study of glucocerebrosidase mutations in dementia with Lewy bodies. *JAMA Neurol, 70*(6), 727-735. doi:10.1001/jamaneurol.2013.1925

Nieuwenhuis, R., te Grotenhuis, M., Pelzer, B., Schmidt, A., Konig, R., & Eisinga, R. (2017). Weighted Effect Coding (Version 0.4-1). Online: CRAN. Retrieved from https://cran.r-project.org/web/packages/wec/wec.pdf

O'Gorman Tuura, R. L., Baumann, C. R., & Baumann-Vogel, H. (2018). Beyond Dopamine: GABA, Glutamate, and the Axial Symptoms of Parkinson Disease. *Front Neurol, 9*, 806. doi:10.3389/fneur.2018.00806

Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., . . . Lang, A. E. (2017). Parkinson disease. *Nat Rev Dis Primers, 3*, 17013. doi:10.1038/nrdp.2017.13

R Core Team. (2014). R: A language and environment for statistical computing. . Retrieved from http://www.R-project.org/

Root, J., Merino, P., Nuckols, A., Johnson, M., & Kukar, T. (2021). Lysosome dysfunction as a cause of neurodegenerative diseases: Lessons from frontotemporal dementia and amyotrophic lateral sclerosis. *Neurobiol Dis, 154*, 105360. doi:10.1016/j.nbd.2021.105360

Sato, C., Morgan, A., Lang, A. E., Salehi-Rad, S., Kawarai, T., Meng, Y., . . . Rogaeva, E. (2005). Analysis of the glucocerebrosidase gene in Parkinson's disease. *Mov Disord, 20*(3), 367-370. doi:10.1002/mds.20319

Savica, R., Beach, T. G., Hentz, J. G., Sabbagh, M. N., Serrano, G. E., Sue, L. I., . . . Adler, C. H. (2019). Lewy body pathology in Alzheimer's disease: A clinicopathological prospective study. *Acta Neurol Scand, 139*(1), 76-81. doi:10.1111/ane.13028

Sheth, J., Bhavsar, R., Mistri, M., Pancholi, D., Bavdekar, A., Dalal, A., . . . Sheth, F. (2019). Gaucher disease: single gene molecular characterization of one-hundred Indian patients reveals novel variants and the most prevalent mutation. *BMC Med Genet, 20*(1), 31. doi:10.1186/s12881-019-0759-1

Shiner, T., Mirelman, A., Gana Weisz, M., Bar-Shira, A., Ash, E., Cialic, R., . . . Giladi, N. (2016). High Frequency of GBA Gene Mutations in Dementia With Lewy Bodies Among Ashkenazi Jews. *JAMA Neurol, 73*(12), 1448-1453. doi:10.1001/jamaneurol.2016.1593

Sidransky, E., Nalls, M. A., Aasly, J. O., Aharon-Peretz, J., Annesi, G., Barbosa, E. R., . . . Ziegler, S. G. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N Engl J Med, 361*(17), 1651-1661. doi:10.1056/NEJMoa0901281

Stone, D. L., Tayebi, N., Orvisky, E., Stubblefield, B., Madike, V., & Sidransky, E. (2000). Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. *Hum Mutat, 15*(2), 181-188. doi:10.1002/(SICI)1098-1004(200002)15:2<181::AID-HUMU7>3.0.CO;2-S

Sunderland, K. M., Beaton, D., Arnott, S. R., Kleinstiver, P. W., Kwan, D., Lawrence-Dewar, J. M., . . . Binns, M. A. (2020). The Ontario Neurodegenerative Disease Research Initiative. *medRxiv*, 2020.2007.2030.20165456. doi:10.1101/2020.07.30.20165456

Te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2017). A novel method for modelling interaction between categorical variables. *Int J Public Health, 62*(3), 427-431. doi:10.1007/s00038-016-0902-0

Tsuang, D., Leverenz, J. B., Lopez, O. L., Hamilton, R. L., Bennett, D. A., Schneider, J. A., . . . Zabetian, C. P. (2012). GBA mutations increase risk for Lewy body disease with and without Alzheimer disease pathology. *Neurology, 79*(19), 1944-1950. doi:10.1212/WNL.0b013e3182735e9a

Chapter 7 – Parkinson's Disease, *NOTCH3* Genetic Variants, and White Matter Hyperintensities.

The work presented in <u>Chapter 7</u> has been edited from the original publication in *Movement Disorders* for clarity and consistency throughout the entire Dissertation.

Dilliott A.A.*, Ramirez J.*, Binns M., Breen D., Evans E., Beaton D., McLaughlin P.M., Kwan D., Holmes M.F., Ozzoude M., Scott C.J.M., Strother S.C., Symons S., Swartz R.H., Grimes D., Jog M., Masellis M., Black S.E., Joutel A., Marras C., Rogaeva E., Hegele R.A., and Lang A.E., on behalf of the ONDRI Investigators (2020). Parkinson's disease, *NOTCH3* genetic variants, and white matter hyperintensities. *Mov Disord*, 35(11):2090-2095

    *Authors contributed equally to this work.

7.1.   Abstract

Background/Objective: White matter hyperintensities (WMH) on MRI may influence clinical presentation in patients with Parkinson's disease (PD), although their significance and pathophysiological origins remain unresolved. Studies examining WMH have identified pathogenic variants in *NOTCH3* as an underlying cause of inherited forms of cerebral small vessel disease.

Methods and Results: We examined *NOTCH3* variants, WMH volumes, and clinical correlates, in 139 PD participants in the Ontario Neurodegenerative Disease Research Initiative (ONDRI) cohort. We identified 13 PD participants (~9%) with rare (<1% of general population), non-synonymous *NOTCH3* variants. Bayesian linear modelling demonstrated a doubling of WMH between variant negative and positive participants (3.1 vs. 6.9mL), with large effect sizes for periventricular WMH (d=0.8) and lacunes (d=1.2). Negative correlations were observed between WMH and global cognition (r=-0.2).

Conclusion: The *NOTCH3* rare variants in PD may significantly contribute to increased WMH burden, which in turn may negatively influence cognition.

7.2.   Introduction

White matter hyperintensities (WMH), commonly associated with aging and cerebral small vessel disease (Wardlaw et al., 2013), have been observed on MRI of Parkinson's disease (PD) patients (Bohnen & Albin, 2011); however, their exact role in clinical presentation and disease progression remains controversial. Although some studies demonstrate WMH to be significantly associated with greater burden of motor dysfunction, mild cognitive impairment, and progression to dementia (Dunet et al., 2019; Foo et al., 2016; Kandiah et al., 2013; Slawek et al., 2013; Toda, Iijima, & Kitagawa, 2019), other studies have found no independent association beyond that of normal aging (Acharya, Bouchard, Emery, & Camicioli, 2007; Hanning et al., 2019; Mak et al., 2015; Pozorski et al., 2019; Ten Harmsen et al., 2018; Vesely, Antonini, & Rektor, 2016).

Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is a monogenic form of cerebral small vessel disease with hallmark signs of WMH. CADASIL is caused by heterozygous pathogenic mutations in the Notch Receptor 3 gene (*NOTCH3*) located on chromosome 19p13 (Joutel et al., 1996), which encodes a transmembrane receptor in smooth muscle cells and pericytes (Joutel et al., 2000). The protein's extracellular domain comprises 34 epidermal growth factor (EGF)-like repeats that each have six cysteines forming disulphide bonds integral to Notch3's tertiary structure and overall function (Sakamoto, Chao, Katsube, & Yamaguchi, 2005). Traditionally, pathogenic variants for CADASIL were considered those that occurred in an EGF-like repeat — most commonly, EGF-like repeats 1–6 — and contributed to the loss or gain of a cysteine residue (Joutel et al., 1997). These variants disrupt disulphide bridging, leading to protein misfolding, and ultimately Notch3 protein aggregation in smooth muscle cell walls, including within the brain's vasculature (Joutel et al., 2010; Opherk et al., 2009). However, *NOTCH3* variants pathogenic for diffuse WMH and the CADASIL phenotype have also been discovered within EGF-like repeats not affecting cysteine residues (Muino et al., 2017). In addition, variants outside the 34 EGF-like repeats may contribute to milder cerebrovascular disease (Schmidt et al., 2011). There is also evidence that some CADASIL patients can present with late-onset, slowly progressive parkinsonism. In general, these cases are not levodopa responsive,

have more falls, and display greater levels of cognitive impairment (Ragno et al., 2013). However, the contribution of *NOTCH3* variants to idiopathic PD has not been previously described.

Here we examine the presence of *NOTCH3* variants and WMH volumes, along with their clinical correlates, in PD participants from the Ontario Neurodegenerative Disease Research Initiative (ONDRI) cohort.

7.3. Methods

7.3.1. Study participants and clinical assessments

Participants were recruited from ONDRI (http://ondri.ca/) (S. M. K. Farhan et al., 2017), a multi-centre cohort study that included PD participants who met criteria for idiopathic PD defined by the United Kingdom's Parkinson's Disease Society Brain Bank clinical diagnostic criteria at the time of enrollment (Hughes, Daniel, Kilford, & Lees, 1992). Participants had a good and sustained response to dopaminergic therapy and had a Hoehn and Yahr stage 1–3 in the "on" medication state. Study participants gave written informed consent in accordance with each site's institutional Research Ethics Board.

Standard clinical and demographic assessments including the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) (Goetz et al., 2008), the Montreal Cognitive Assessment tool (MoCA) (Nasreddine et al., 2005), hemoglobin A1C (HbA1C), high- and low-density lipoproteins (HDL/LDL), seated/standing systolic and diastolic blood pressure (BP), hip/waist ratio, and clinical history of diabetes and hypertension, were all collected and underwent thorough quality control processes (Sunderland et al., 2019).

7.3.2. Genetic analysis

Blood-derived genomic DNA samples were sequenced using ONDRISeq, a custom-designed next-generation sequencing panel that targets 80 genes associated with Alzheimer's disease, amyotrophic lateral sclerosis, frontotemporal dementia, PD, and/or cerebrovascular disease (e.g. *NOTCH3*). Importantly, the ONDRISeq panel has been

previously validated (S.M.K. Farhan et al., 2016), and detailed sequencing and quality control methodology has been described (Dilliott et al., 2018).

Variants identified within *NOTCH3* with an allelic depth of at least 10x were prioritized if they were nonsynonymous in sequence ontology, were of either heterozygous or homozygous zygosity, and appeared at a minor allele frequency (MAF) of less than 1% in the Exome Aggregation Consortium database (ExAC) (Lek et al., 2016). Variants were further prioritized if they displayed evidence of being likely deleterious, either with an in silico Combined Annotation Dependent Depletion (CADD) Phred score of >20 (top 1% of deleterious variants in the human genome) or if previously associated with disease by the Human Gene Mutation Database (HGMD) (Stenson et al., 2014) or ClinVar (Landrum et al., 2016). Individuals that carried at least one prioritized *NOTCH3* variant are hereafter referred to as *NOTCH3* variant positive, whereas those who did not carry a variant are referred to as *NOTCH3* variant negative.

Due to the association between the apolipoprotein E (*APOE*) E2 allele and increased WMH volume in individuals with CADASIL (Gesierich et al., 2016), *APOE* genotypes were obtained from the ONDRISeq data. A customized ANNOVAR script was used to extract calls for the *APOE* variants rs429358(CT):p.Cys130Arg and rs7412(CT):p.Arg176Cys and map to the respective *APOE* genotype for each participant, as previously described (Dilliott et al., 2019).

7.3.3.   Neuroimaging

Harmonized with the Canadian Dementia Imaging Protocol (Duchesne et al., 2019), 3T MRI was performed at each ONDRI site and included the following sequences: 3D T1-weighted (T1), T2-weighted fluid attenuated inversion recovery (FLAIR), and an interleaved T2-weighted (T2) and proton density. MRI scans were fully evaluated by a neuroradiologist (S.S.) for incidental findings. MRI-based volumetrics were acquired using previously published methods, which included quantification of WMH volumes and an individualized measure of head size obtained from the supra-tentorial intracranial volume (ST-TIV) (Ramirez et al., 2020; Ramirez et al., 2014).

7.3.4.   Statistical analysis

Our primary aims were to determine the prevalence of *NOTCH3* variants in the ONDRI
PD cohort, and to test the hypothesis that WMH volumes were significantly higher in the
participants with *NOTCH3* variants, after controlling for hypertension, diabetes, age, sex
and the presence of the *APOE* E2 allele (Adib-Samii, Brice, Martin, & Markus, 2010).
Individual differences in head size and the highly skewed distribution of WMH volumes
were adjusted for with a natural log transform and ST-TIV volumes (DeCarli et al., 2005;
Ramirez et al., 2016). To account for conventional risk factors and presence of the *APOE*
E2 allele (Raz, Yang, Dahle, & Land, 2012), we found significant bivariate associations
with age, male sex, and clinical history of hypertension and diabetes (odds ratios ranging
from 0.72 to 1.41). Based on associations between these variables, we calculated an
aggregate of the presumed risk factors. In order to estimate the magnitude of association
between *NOTCH3* variation and WMH volume, we fit a Bayesian linear model (Sturtz,
Ligges, & Gelman, 2005). Graphical diagnostics were examined to ensure successful
convergence of the simulation. Thus, the outcome measure was log-transformed, head-
size adjusted total WMH volume. Regressor variables included an aggregate of the risk
factors, an *APOE* E2 allele indicator variable, and a *NOTCH3* variant indicator variable.
An uninformative prior normal distribution ($\mu = 0$, $\sigma = 1{,}000$) was specified for each of
the intercept and regression coefficients and a gamma distribution for residual variance.
Cohen's d and Cramér's ϕ (for categorical variables) effect sizes according to *NOTCH3*
status were estimated for MRI-derived whole brain volumetrics and clinical measures
(HbA1C, HDL/LDL, BP, hip/waist). Across the entire sample, Pearson r correlations
between WMH and tremor, postural instability/gait, and motor phenotype (tremor
dominant vs. postural instability/gait disorder) were also examined (Stebbins et al.,
2013). Statistical analyses were performed using the R package and *R2WinBUGS*
package (R Core Team, 2014; Sturtz et al., 2005). Data visualization was performed
using RStudio version 1.2.1335 (RStudio, Inc., Boston, MA) and ITKSnap (Yushkevich
et al., 2006).

7.4.   Results

7.4.1.   Study participants and *NOTCH3* variants

ONDRI enrolled 139 PD participants that had blood samples obtained for DNA isolation. Participant demographics, MRI-derived whole brain volumetrics, clinical measures, and effect sizes according to *NOTCH3* status are compared in **Table 7.1**.

In total, 13 of the 139 (9%) PD participants harboured rare, non-synonymous, likely deleterious *NOTCH3* variants. Twelve participants each harboured a single variant, and one participant harboured two variants (c.3704A>T: p.His1235Leu and c.6201dupC: p.Gly2068Argfs). In total, there were thirteen unique variants, with only one variant identified in two different individuals (**Table 7.2**). Eight of the prioritized variants were previously associated with disease according to HGMD or ClinVar. Importantly, three of the variants were cysteine modifying — one identified in EGF-like repeat 14 and two identified in EGF-like repeat 31.

Table 7.1 Study participant demographics, raw whole brain volumetrics obtained from MRI, and clinical measures commonly associated with vascular risk.

	NOTCH3 **Variant Negative**	*NOTCH3* **Variant Positive**	**Effect Size** [†]
Demographics			
N	126	13	--
Sex, M/F	99/27	9/4	0.07
Age	67.7 (6.4)	69.4 (6.3)	0.27
Education, years	15.5 (2.7)	15.5 (2.5)	0.02
MoCA	25.7 (2.6)	27.1 (1.9)	0.13
Modified Rankin Score	1.7 (0.7)	1.6 (0.7)	0.53
Whole brain volumetrics			
Total Intracranial Capacity, mL	1313.7 (126.1)	1335.1 (139.4)	0.17
White matter, mL	446.0 (61.3)	446.8 (65.0)	0.01
Gray matter, mL	573.6 (47.3)	583.0 (46.8)	0.20
Sulcal cerebrospinal fluid, mL	251.2 (53.1)	257.6 (54.5)	0.12
Ventricular cerebrospinal fluid, mL	38.2 (19.6)	38.2 (17.9)	0.00
Periventricular WMH, mm^3	4079.0 (5140.9)	8379.2 (8473.1)	**0.78***
Deep WMH, mm^3	484.4 (764.3)	657.8 (909.1)	0.22
Lacunes, mm^3	82.5 (183.1)	485.0 (947.2)	**1.22***
Enlarged PVS, mm^3	52.3 (52.9)	59.6 (80.8)	0.13
Clinical measures			
TD/PIGD Ratio (MDS-UPDRS)[a]	1.1 (0.2)	1.1 (0.2)	0.15
Blood glucose: HbA1C, %	5.7 (0.8)	5.7 (0.3)	0.07
Cholesterol, mmol/L	4.6 (1.1)	4.6 (1.1)	0.01
Low density lipoprotein, mmol/L	2.6 (0.9)[b]	2.5 (0.8)	0.12
High density lipoprotein, mmol/L	1.4 (0.4)[b]	1.6 (0.4)	0.45
Cholesterol/HDL ratio	3.4 (0.9)[b]	2.9 (0.4)	0.59
BP Systolic (seated)	131.1 (20.0)	136.6 (19.5)	0.28
BP Diastolic (seated)	78.0 (10.6)	78.0 (8.3)	0.01
BP Systolic (standing)	126.1 (20.4)	126.2 (15.3)	0.00
BP Diastolic (standing)	78.3 (11.6)	76.3 (8.3)	0.17
Hip/Waist Ratio	1.8 (3.9)[a]	2.2 (5.5)	0.04

Note that raw volumes are reported for transparency, statistical analyses were performed on risk adjusted, head-size adjusted, normalized values.

† Cohen's d for continuous variables, Cramér's ɸ for the sex variable. Large effect sizes are bold and highlighted with an asterisk (*).

[a] Tremor dominant (TD) vs postural instability/gait difficulty (PIGD) ratio.[34]

[b] Data available in 125/126 individuals.

Abbreviations: MoCA, Montreal Cognitive Assessment tool; PVS, perivascular spaces; MDS-UPDRS, Movement Disorder Society-Unified Parkinson's Disease Rating Scale; HbA1C, hemoglobin A1C; BP, blood pressure.

Table 7.2 Prioritized non-synonymous rare variants identified within the *NOTCH3* gene carried by 13 individuals with PD.

cDNA Alteration	Protein Alteration	Reference SNP Identifier	Exon	EGFr	Sequence Ontology	MAF (ExAC)	CADD Phred	Previous Disease Association	Participants Harbouring Variant
c.884T>G	p.Leu295Arg	rs143117018	6	7	Missense	1.89E-04	22.9	NA	1
c.1490C>T	p.Ser497Leu	rs114207045	9	12	Missense	5.69E-03	23.3	WML[1]	1
c.1732C>T	p.Arg578Cys	rs769773673	11	14	Missense	4.12E-05	30	CADASIL[2,3]	1
c.1931T>A	p.Val644Asp	rs148046938	12	16	Missense	7.00E-04	24.2	WML[4]	1
c.2978C>T	p.Thr993Met	rs371278091	18	25	Missense	8.28E-06	26.5	NA	1
c.3664T>G	p.Cys1222Gly	rs199638166	22	31	Missense	7.41E-05	25.6	CADASIL[5]	1
c.3691C>T	p.Arg1231Cys	rs201680145	22	31	Missense	9.88E-04	30	CADASIL[2,5]	1
c.3704A>T	p.His1235Leu	rs55882518	22	NA	Missense	3.95E-03	14.8	WML[1]	1
c.4552C>A	p.Leu1518Met	rs141320511	25	NA	Missense	3.17E-03	27.8	WML[1]	1
c.5510G>A	p.Arg1837His	rs138265894	30	NA	Missense	8.48E-04	35	NA	1
c.5854G>A	p.Val1952Met	rs115582213	32	NA	Missense	8.47E-03	34	AD modifier[6]	2
c.6025C>T	p.Arg2009Trp	rs151322770	33	NA	Missense	4.12E-05	33	NA	1
c.6201dupC	p.Gly2068Argfs	NA	33	NA	Frameshift insertion	NA	32	NA	1

[1](Schmidt et al., 2011); [2](Joutel et al., 1997); [3](Yoon et al., 2015); [4](Ungaro et al., 2009); [5](Rutten et al., 2016); [6](Guerreiro et al., 2012).
Abbreviations: AD, Alzheimer's disease; CADASIL, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy; cDNA, coding DNA; EGFr, epidermal growth factor-like repeat; ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; NA, not applicable; WML, white matter lesions. Previous disease associations were defined based on the Human Gene Mutation Database.

7.4.2.   Associations between *NOTCH3* and cerebrovascular burden

Periventricular WMH volumes were greater in the *NOTCH3* variant positive group (8379.2mm3 vs. 4079.0mm3, Cohen's d = 0.78). Lacunar volumes were also greater in the *NOTCH3* variant positive group (485.0 mm3 vs. 82.6mm3, Cohen's d = 1.22). The FLAIR MRI of a *NOTCH3* variant positive PD participant with diffuse WMH is shown in **Figure 7.1**.

There was an association between the aggregate risk factor and WMH volumes (**Figure 7.2**). There were two *NOTCH3* variant positive participants — one harbouring c.884T>G: p.Leu295Arg and one harbouring c.1931T>A: p.Val644Asp — with high aggregate risk factor and relatively low WMH. They might be expected to exert a negative bias on the estimate of the magnitude of association between *NOTCH3* and WMH. In order to estimate the strength of evidence for a conditional association between *NOTCH3* and log-adjusted WMH, we calculated Bayes factor using the *BayesFactor* package in R. The magnitude of the multiplicative effect of *NOTCH3* on WMH volume was about 2.2 (between 1.2 and 3.7, 95% probability). An approximate doubling of total WMH volume between *NOTCH3* variant negative (mean ~3.1 mL) and *NOTCH3* variant positive (mean ~6.9 mL) was determined at the midpoint of the aggregate risk factor.

Scatter plots showing log WMH volumes (head size corrected) versus clinical factors for the entire PD sample are shown in **Figure 7.3**. There was an inverse correlation between WMH volume and global cognition (r=-0.19, C.I: -0.32, -0.02), but no other significant associations were demonstrated.

Upon follow-up, comprehensive clinical assessment resulted in the revision of the original diagnosis of one *NOTCH3* variant negative participant who was subsequently removed from the study. Re-analysis of the data did not significantly impact the findings: epsilon changed from 0.8–1.3, to 0.8–1.4; medians and credible regions describing associations between *NOTCH3*, risk score, and *APOE* E2 and head-size adjusted total WMH volume were unchanged.

Figure 7.1 Axial view of fluid-attenuated inversion recovery (FLAIR) MRI showing multiple diffuse WMH (bright) in a *NOTCH3* variant positive Parkinson's disease participant from the Ontario Neurodegenerative Disease Research Initiative (ONDRI).
White matter hyperintensities (WMHs) are indicated by the red arrows.
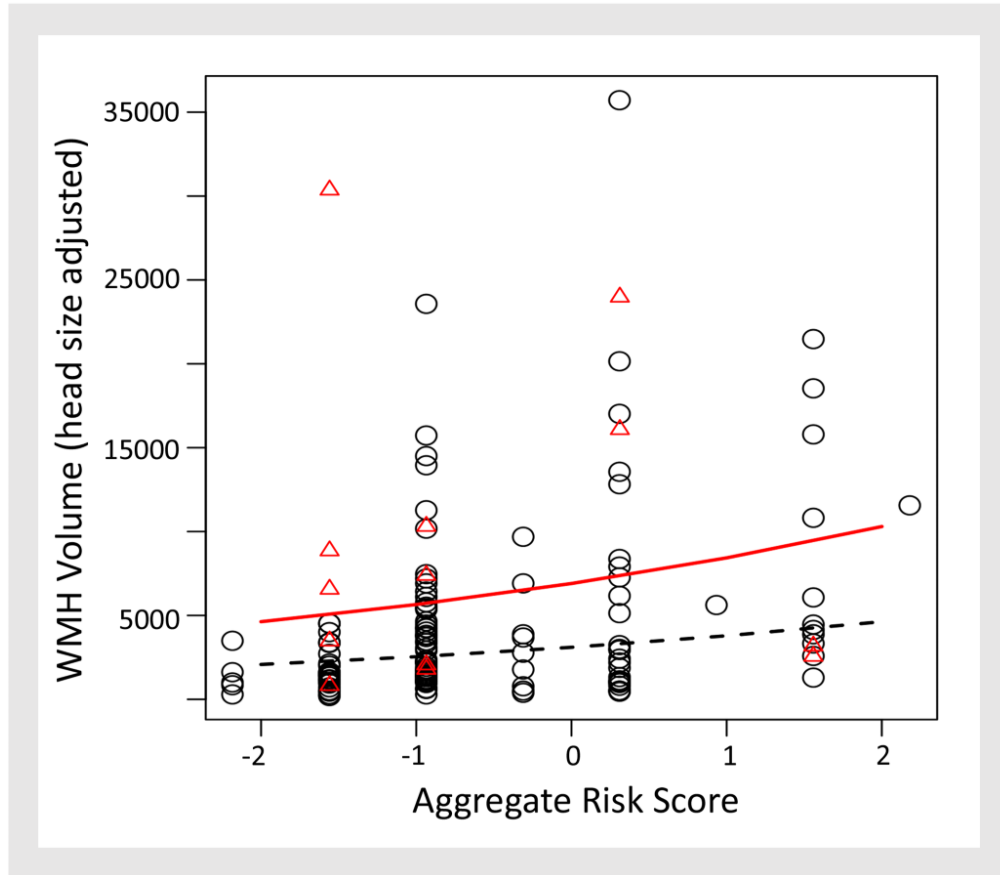
Figure 7.2 Scatter plot of WMH (head-size adjusted) versus aggregate risk score.
NOTCH3 variant positive participants are indicated in red; non-parametric smoothers (solid) and additive linear fit (dotted) are superimposed.
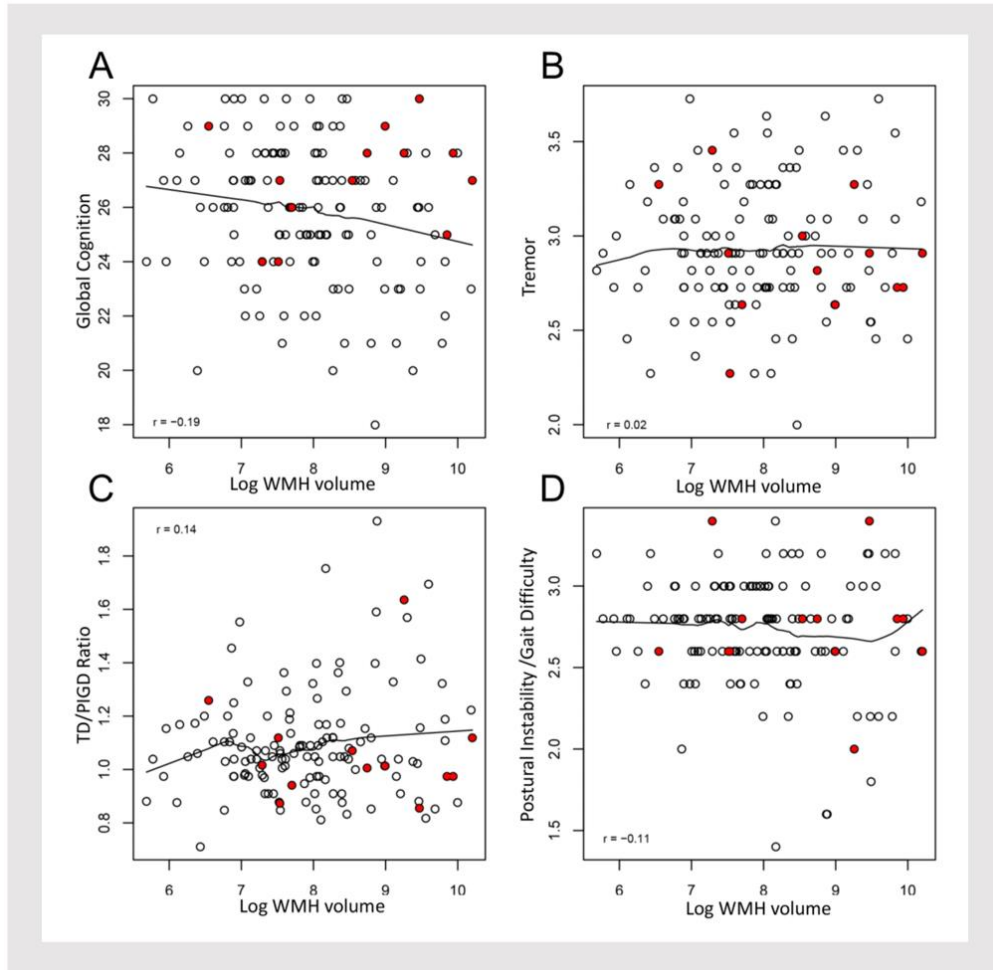
Figure 7.3 Scatter plots showing head-size corrected log WMH volume (x-axis) by global cognition (A), tremor (B), motor phenotype (tremor dominant vs. postural instability/gait disorder; C), and postural instability/gait (D).

7.5.   Discussion

The main finding of this study was that PD participants carrying prioritized *NOTCH3* variants had double the WMH volumes compared to *NOTCH3* variant negative participants. These findings were based on data adjusted for head size, age, vascular risk factors and the presence of the *APOE* E2 allele. To our knowledge, this is the first report suggesting that rare *NOTCH3* variants may be relevant in the neuropathology associated with idiopathic PD.

The 13 unique *NOTCH3* variants identified have not undergone functional analysis to confirm pathogenicity. However, using a non-synonymous, rare variant prioritization strategy, along with pathogenicity predictions based on *in silico* analysis or previous association with disease, we aimed to select variants most likely to contribute to disease presentation. Only one variant had a CADD Phred score below 20 (top 1% of deleterious variants) and eight of the variants were previously identified in individuals with CADASIL, WMH or Alzheimer's disease. Yet, ten of the variants were not cysteine-modifying and six were located outside of the EGF-like repeats, both of which would be atypical for CADASIL-related variants. Previously, it was hypothesized that these atypical *NOTCH3* variants may contribute to more mild presentations of CADASIL and cerebrovascular disease (Muino et al., 2017; Ragno et al., 2013; Ungaro et al., 2009). Rather than contributing to the patients' PD pathology directly, *NOTCH3* variants may modify disease presentation by influencing WMH burden, but further analysis is required.

Three individuals carried cysteine-modifying *NOTCH3* variants within an EGF-like repeat that have been previously associated with the CADASIL phenotype (Rutten et al., 2016; Yoon et al., 2015). These participants were individually analyzed and only one (carrier of c.3691C>T: p.Arg1231Cys) stood out as having more severe cerebrovascular burden, consistently appearing in the top 10 worst cases of total, periventricular and deep WMH, and with a higher burden of lacunes and enlarged perivascular spaces. Interestingly, the individual with the greatest amount of vascular burden carried a variant

in EGF-like repeat 25 that is not cysteine-altering (c.2978C>T: p.The993Met) and has not been previously associated with disease. Further, this individual harboured an *APOE* E3/3 genotype. Additional analysis regarding the pathogenicity of this *NOTCH3* variant is needed.

Although the sample size and variability limited our statistical ability to analyze differences in periventricular/deep WMH, lacunes, and enlarged perivascular spaces (PVS), the effect sizes shown in **Table 7.1** suggest *NOTCH3* variant positive participants had greater burden of periventricular (8.4mL vs. 4.1mL, Cohen's d = 0.78) but not deep WMH, and lacunes (485mm3 vs. 82.6mm3, Cohen's d = 1.22) but not PVS. Apart from some limited case reports (Mestre et al., 2014), previous studies have not examined lacunes and PVS burden in PD. Future studies with larger sample sizes are needed to validate these observations in our study sample.

The inverse correlation between WMH and cognition across the entire sample lends some support to previous studies that suggest WMH influence the cognitive presentation of PD; however, in the absence of a control group and long-term follow-up data, our correlations should be interpreted with caution. A recent cross-sectional analysis of the Dutch PROPARK cohort (n=163) demonstrated a relationship between visuospatial functioning and periventricular WMH in PD patients (de Schipper et al., 2019), while a retrospective analysis of a different Dutch cohort (n=204) found no relationship with self-reported cognition (Ten Harmsen et al., 2018). Longitudinally, the PPMI study reported that higher baseline WMH burden was associated with future cognitive decline and cortical thinning over a four year period (Dadar et al., 2018), while the DeNoPa study reported no modifying effects on cognitive function with a two year follow-up period (Hanning et al., 2019). Another smaller study found various relationships with WMH and brain atrophy, but age-adjustment eliminated the correlations, further emphasizing the well-established finding that WMH are an age-related phenomenon (Acharya et al., 2007). The heterogeneity of these reports is further aggravated by the difficulty interpreting findings from studies using different cognitive assessments tools and approaches to measure WMH burden (Dunet et al., 2019; Vesely et al., 2016).

The main strength of our study was the ability to draw data from ONDRI's study design. ONDRI implemented a standardized multi-centre, multi-platform approach which enabled the volumetric analysis of MRI-based neurovascular biomarkers and genomic information using validated quantification methods in their respective fields. As previously mentioned, the main limitations include the limited statistical power, the lack of a normative sample, and the cross-sectional nature of the study.

7.6.   Conclusion

Here, we present the first report of rare *NOTCH3* variants potentially influencing cerebrovascular sub-phenotypes in individuals with idiopathic PD. As ONDRI's longitudinal data becomes available, serial MRI processing and long-term disease tracking, combined with a more comprehensive analysis of the neuropsychological profile of our participants, will be possible. The findings reported here should encourage further studies on the occurrence of *NOTCH3* variants in PD and their clinical, imaging and pathological correlates.

7.7. References

Acharya, H. J., Bouchard, T. P., Emery, D. J., & Camicioli, R. M. (2007). Axial signs and magnetic resonance imaging correlates in Parkinson's disease. *Can J Neurol Sci, 34*(1), 56-61. doi:10.1017/s0317167100005795

Adib-Samii, P., Brice, G., Martin, R. J., & Markus, H. S. (2010). Clinical spectrum of CADASIL and the effect of cardiovascular risk factors on phenotype: study in 200 consecutively recruited individuals. *Stroke, 41*(4), 630-634. doi:10.1161/STROKEAHA.109.568402

Bohnen, N. I., & Albin, R. L. (2011). White matter lesions in Parkinson disease. *Nat Rev Neurol, 7*(4), 229-236. doi:10.1038/nrneurol.2011.21

Dadar, M., Zeighami, Y., Yau, Y., Fereshtehnejad, S. M., Maranzano, J., Postuma, R. B., . . . Collins, D. L. (2018). White matter hyperintensities are linked to future cognitive decline in de novo Parkinson's disease patients. *Neuroimage Clin, 20*, 892-900. doi:10.1016/j.nicl.2018.09.025

de Schipper, L. J., Hafkemeijer, A., Bouts, M., van der Grond, J., Marinus, J., Henselmans, J. M. L., & van Hilten, J. J. (2019). Age- and disease-related cerebral white matter changes in patients with Parkinson's disease. *Neurobiol Aging, 80*, 203-209. doi:10.1016/j.neurobiolaging.2019.05.004

DeCarli, C., Massaro, J., Harvey, D., Hald, J., Tullberg, M., Au, R., . . . Wolf, P. A. (2005). Measures of brain morphology and infarction in the framingham heart study: establishing what is normal. *Neurobiol Aging, 26*(4), 491-510. doi:10.1016/j.neurobiolaging.2004.05.004

Dilliott, A. A., Evans, E. C., Farhan, S. M. K., Ghani, M., Sato, C., Zhang, M., . . . Hegele, R. A. (2019). Genetic Variation in the Ontario Neurodegenerative Disease Research Initiative. *Can J Neurol Sci, 46*(5), 491-498. doi:10.1017/cjn.2019.228

Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2018). Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J Vis Exp*(134). doi:10.3791/57266

Duchesne, S., Chouinard, I., Potvin, O., Fonov, V. S., Khademi, A., Bartha, R., . . . the, C. g. (2019). The Canadian Dementia Imaging Protocol: Harmonizing National Cohorts. *J Magn Reson Imaging, 49*(2), 456-465. doi:10.1002/jmri.26197

Dunet, V., Fartaria, M. J., Deverdun, J., Le Bars, E., Maury, F., Castelnovo, G., . . . de Champfleur, N. M. (2019). Episodic memory decline in Parkinson' s disease: relation with white matter hyperintense lesions and influence of quantification method. *Brain Imaging Behav, 13*(3), 810-818. doi:10.1007/s11682-018-9909-x

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., . . . Strong, M. J. (2017). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci, 44*(2), 196-202. doi:10.1017/cjn.2016.415

Farhan, S. M. K., Dilliott, A. A., Ghani, G., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*(16032), 1-11.

Foo, H., Mak, E., Yong, T. T., Wen, M. C., Chander, R. J., Au, W. L., . . . Kandiah, N. (2016). Progression of small vessel disease correlates with cortical thinning in Parkinson's disease. *Parkinsonism Relat Disord, 31*, 34-40. doi:10.1016/j.parkreldis.2016.06.019

Gesierich, B., Opherk, C., Rosand, J., Gonik, M., Malik, R., Jouvent, E., . . . Ewers, M. (2016). APOE varepsilon2 is associated with white matter hyperintensity volume in CADASIL. *J Cereb Blood Flow Metab, 36*(1), 199-203. doi:10.1038/jcbfm.2015.85

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., . . . Movement Disorder Society, U. R. T. F. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord, 23*(15), 2129-2170. doi:10.1002/mds.22340

Guerreiro, R. J., Lohmann, E., Kinsella, E., Bras, J. M., Luu, N., Gurunlian, N., . . . Singleton, A. (2012). Exome sequencing reveals an unexpected genetic cause of disease: NOTCH3 mutation in a Turkish family with Alzheimer's disease. *Neurobiol Aging, 33*(5), 1008 e1017-1023. doi:10.1016/j.neurobiolaging.2011.10.009

Hanning, U., Teuber, A., Lang, E., Trenkwalder, C., Mollenhauer, B., & Minnerup, H. (2019). White matter hyperintensities are not associated with cognitive decline in early Parkinson's disease - The DeNoPa cohort. *Parkinsonism Relat Disord, 69*, 61-67. doi:10.1016/j.parkreldis.2019.10.016

Hughes, A. J., Daniel, S. E., Kilford, L., & Lees, A. J. (1992). Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J Neurol Neurosurg Psychiatry, 55*(3), 181-184. doi:10.1136/jnnp.55.3.181

Joutel, A., Andreux, F., Gaulis, S., Domenga, V., Cecillon, M., Battail, N., . . . Tournier-Lasserve, E. (2000). The ectodomain of the Notch3 receptor accumulates within the cerebrovasculature of CADASIL patients. *J Clin Invest, 105*(5), 597-605. doi:10.1172/JCI8047

Joutel, A., Corpechot, C., Ducros, A., Vahedi, K., Chabriat, H., Mouton, P., . . . Tournier-Lasserve, E. (1996). Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature, 383*(6602), 707-710. doi:10.1038/383707a0

Joutel, A., Monet-Lepretre, M., Gosele, C., Baron-Menguy, C., Hammes, A., Schmidt, S., . . . Hubner, N. (2010). Cerebrovascular dysfunction and microcirculation rarefaction precede white matter lesions in a mouse genetic model of cerebral ischemic small vessel disease. *J Clin Invest, 120*(2), 433-445. doi:10.1172/JCI39733

Joutel, A., Vahedi, K., Corpechot, C., Troesch, A., Chabriat, H., Vayssiere, C., . . . Tournier-Lasserve, E. (1997). Strong clustering and stereotyped nature of Notch3 mutations in CADASIL patients. *Lancet, 350*(9090), 1511-1515. doi:10.1016/S0140-6736(97)08083-5

Kandiah, N., Mak, E., Ng, A., Huang, S., Au, W. L., Sitoh, Y. Y., & Tan, L. C. (2013). Cerebral white matter hyperintensity in Parkinson's disease: a major risk factor for mild cognitive impairment. *Parkinsonism Relat Disord, 19*(7), 680-683. doi:10.1016/j.parkreldis.2013.03.008

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res, 44*(D1), D862-868. doi:10.1093/nar/gkv1222

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. doi:10.1038/nature19057

Mak, E., Dwyer, M. G., Ramasamy, D. P., Au, W. L., Tan, L. C., Zivadinov, R., & Kandiah, N. (2015). White Matter Hyperintensities and Mild Cognitive Impairment in Parkinson's Disease. *J Neuroimaging, 25*(5), 754-760. doi:10.1111/jon.12230

Mestre, T. A., Armstrong, M. J., Walsh, R., Al Dakheel, A., Moro, E., Stoessl, A. J., & Lang, A. E. (2014). Can Isolated Enlarged Virchow-Robin Spaces Influence the Clinical Manifestations of Parkinson's Disease? *Mov Disord Clin Pract, 1*(1), 67-69. doi:10.1002/mdc3.12009

Muino, E., Gallego-Fabrega, C., Cullell, N., Carrera, C., Torres, N., Krupinski, J., . . . Fernandez-Cadenas, I. (2017). Systematic Review of Cysteine-Sparing NOTCH3 Missense Mutations in Patients with Clinical Suspicion of CADASIL. *Int J Mol Sci, 18*(9). doi:10.3390/ijms18091964

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., . . . Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc, 53*(4), 695-699. doi:10.1111/j.1532-5415.2005.53221.x

Opherk, C., Duering, M., Peters, N., Karpinska, A., Rosner, S., Schneider, E., . . . Dichgans, M. (2009). CADASIL mutations enhance spontaneous multimerization of NOTCH3. *Hum Mol Genet, 18*(15), 2761-2767. doi:10.1093/hmg/ddp211

Pozorski, V., Oh, J. M., Okonkwo, O., Krislov, S., Barzgari, A., Theisen, F., . . . Gallagher, C. L. (2019). Cross-sectional and longitudinal associations between total and regional white matter hyperintensity volume and cognitive and motor function in Parkinson's disease. *Neuroimage Clin, 23*, 101870. doi:10.1016/j.nicl.2019.101870

R Core Team. (2014). R: A language and environment for statistical computing. . Retrieved from http://www.R-project.org/

Ragno, M., Berbellini, A., Cacchio, G., Manca, A., Di Marzio, F., Pianese, L., . . . De Michele, G. (2013). Parkinsonism is a late, not rare, feature of CADASIL: a study on Italian patients carrying the R1006C mutation. *Stroke, 44*(4), 1147-1149. doi:10.1161/STROKEAHA.111.000458

Ramirez, J., Holmes, M. F., Scott, C. J. M., Ozzoude, M., Adamo, S., Szilagyi, G. M., . . . Investigators, O. (2020). Ontario Neurodegenerative Disease Research Initiative (ONDRI): Structural MRI Methods and Outcome Measures. *Front Neurol, 11*, 847. doi:10.3389/fneur.2020.00847

Ramirez, J., McNeely, A. A., Scott, C. J. M., Masellis, M., Black, S. E., & Alzheimer's Disease Neuroimaging, I. (2016). White matter hyperintensity burden in elderly cohort studies: The Sunnybrook Dementia Study, Alzheimer's Disease Neuroimaging Initiative, and Three-City Study. *Alzheimers Dement, 12*(2), 203-210. doi:10.1016/j.jalz.2015.06.1886

Ramirez, J., Scott, C. J., McNeely, A. A., Berezuk, C., Gao, F., Szilagyi, G. M., & Black, S. E. (2014). Lesion Explorer: a video-guided, standardized protocol for accurate and reliable MRI-derived volumetrics in Alzheimer's disease and normal elderly. *J Vis Exp*(86). doi:10.3791/50887

Raz, N., Yang, Y., Dahle, C. L., & Land, S. (2012). Volume of white matter hyperintensities in healthy adults: contribution of age, vascular risk factors, and inflammation-related genetic variants. *Biochim Biophys Acta, 1822*(3), 361-369. doi:10.1016/j.bbadis.2011.08.007

Rutten, J. W., Dauwerse, H. G., Gravesteijn, G., van Belzen, M. J., van der Grond, J., Polke, J. M., . . . Lesnik Oberstein, S. A. (2016). Archetypal NOTCH3 mutations frequent in public exome: implications for CADASIL. *Ann Clin Transl Neurol, 3*(11), 844-853. doi:10.1002/acn3.344

Sakamoto, K., Chao, W. S., Katsube, K., & Yamaguchi, A. (2005). Distinct roles of EGF repeats for the Notch signaling system. *Exp Cell Res, 302*(2), 281-291. doi:10.1016/j.yexcr.2004.09.016

Schmidt, H., Zeginigg, M., Wiltgen, M., Freudenberger, P., Petrovic, K., Cavalieri, M., . . . group, C. c. N. w. (2011). Genetic variants of the NOTCH3 gene in the elderly and magnetic resonance imaging correlates of age-related cerebral small vessel disease. *Brain, 134*(Pt 11), 3384-3397. doi:10.1093/brain/awr252

Slawek, J., Roszmann, A., Robowski, P., Dubaniewicz, M., Sitek, E. J., Honczarenko, K., . . . Bialecka, M. (2013). The impact of MRI white matter hyperintensities on dementia in Parkinson's disease in relation to the homocysteine level and other vascular risk factors. *Neurodegener Dis, 12*(1), 1-12. doi:10.1159/000338610

Stebbins, G. T., Goetz, C. G., Burn, D. J., Jankovic, J., Khoo, T. K., & Tilley, B. C. (2013). How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: comparison with the unified Parkinson's disease rating scale. *Mov Disord, 28*(5), 668-670. doi:10.1002/mds.25383

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., & Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet, 133*(1), 1-9. doi:10.1007/s00439-013-1358-4

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. J Stat Softw. *12*, 1-16. Retrieved from https://www.jstatsoft.org/article/view/v012i03

Sunderland, K. M., Beaton, D., Fraser, J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., . . . Binns, M. A. (2019). The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project. *BMC Med Res Methodol, 19*(1), 102. doi:10.1186/s12874-019-0737-5

Ten Harmsen, B. L., van Rumund, A., Aerts, M. B., Bergkamp, M. I., Esselink, R. A. J., Richard, E., . . . van Wamelen, D. J. (2018). Clinical correlates of cerebral white matter abnormalities in patients with Parkinson's disease. *Parkinsonism Relat Disord, 49*, 28-33. doi:10.1016/j.parkreldis.2017.12.029

Toda, K., Iijima, M., & Kitagawa, K. (2019). Periventricular White Matter Lesions Influence Gait Functions in Parkinson's Disease. *Eur Neurol, 81*(3-4), 120-127. doi:10.1159/000499908

Ungaro, C., Mazzei, R., Conforti, F. L., Sprovieri, T., Servillo, P., Liguori, M., . . . Quattrone, A. (2009). CADASIL: extended polymorphisms and mutational analysis of the NOTCH3 gene. *J Neurosci Res, 87*(5), 1162-1167. doi:10.1002/jnr.21935

Vesely, B., Antonini, A., & Rektor, I. (2016). The contribution of white matter lesions to Parkinson's disease motor and gait symptoms: a critical review of the literature. *J Neural Transm (Vienna), 123*(3), 241-250. doi:10.1007/s00702-015-1470-9

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., . . . nEuroimaging, S. T. f. R. V. c. o. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol, 12*(8), 822-838. doi:10.1016/S1474-4422(13)70124-8

Yoon, C. W., Kim, Y. E., Seo, S. W., Ki, C. S., Choi, S. H., Kim, J. W., & Na, D. L. (2015). NOTCH3 variants in patients with subcortical vascular cognitive impairment: a comparison with typical CADASIL patients. *Neurobiol Aging, 36*(8), 2443 e2441-2447. doi:10.1016/j.neurobiolaging.2015.04.009

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage, 31*(3), 1116-1128. doi:10.1016/j.neuroimage.2006.01.015

Chapter 8 – Discussion

8.1.   Overview

Within this Dissertation, I have comprehensively characterized the genetic contributors to various neurodegenerative diseases, including: 1) Alzheimer's disease (AD); 2) amyotrophic lateral sclerosis (ALS); 3) frontotemporal dementia (FTD); 4) mild cognitive impairment (MCI); and 5) Parkinson's disease (PD), as well as their genetic overlap with cerebrovascular disease (CVD). Throughout this work I have contributed a vast amount of data to the Ontario Neurodegenerative Disease Research Initiative (ONDRI) and leveraged the data I produced using the ONDRISeq next-generation sequencing (NGS) targeted gene panel to elucidate the genetic contributors to the participants' diagnoses. Specifically, I assessed the contribution of common genotypes and haplotypes, rare single nucleotide variants (SNVs), and copy number variants (CNVs) to the various diagnoses and their clinical intermediate phenotypes to gain a greater understanding of the spectrum of genetic variation that can contribute to disease presentation.

8.2.   Summary of research findings

Throughout this Dissertation, I have genetically characterized 519 participants from ONDRI. A summary of the relevant genetic contributors to each ONDRI disease cohort, as determined by the analyses presented herein are summarized in **Figure 8.1**.
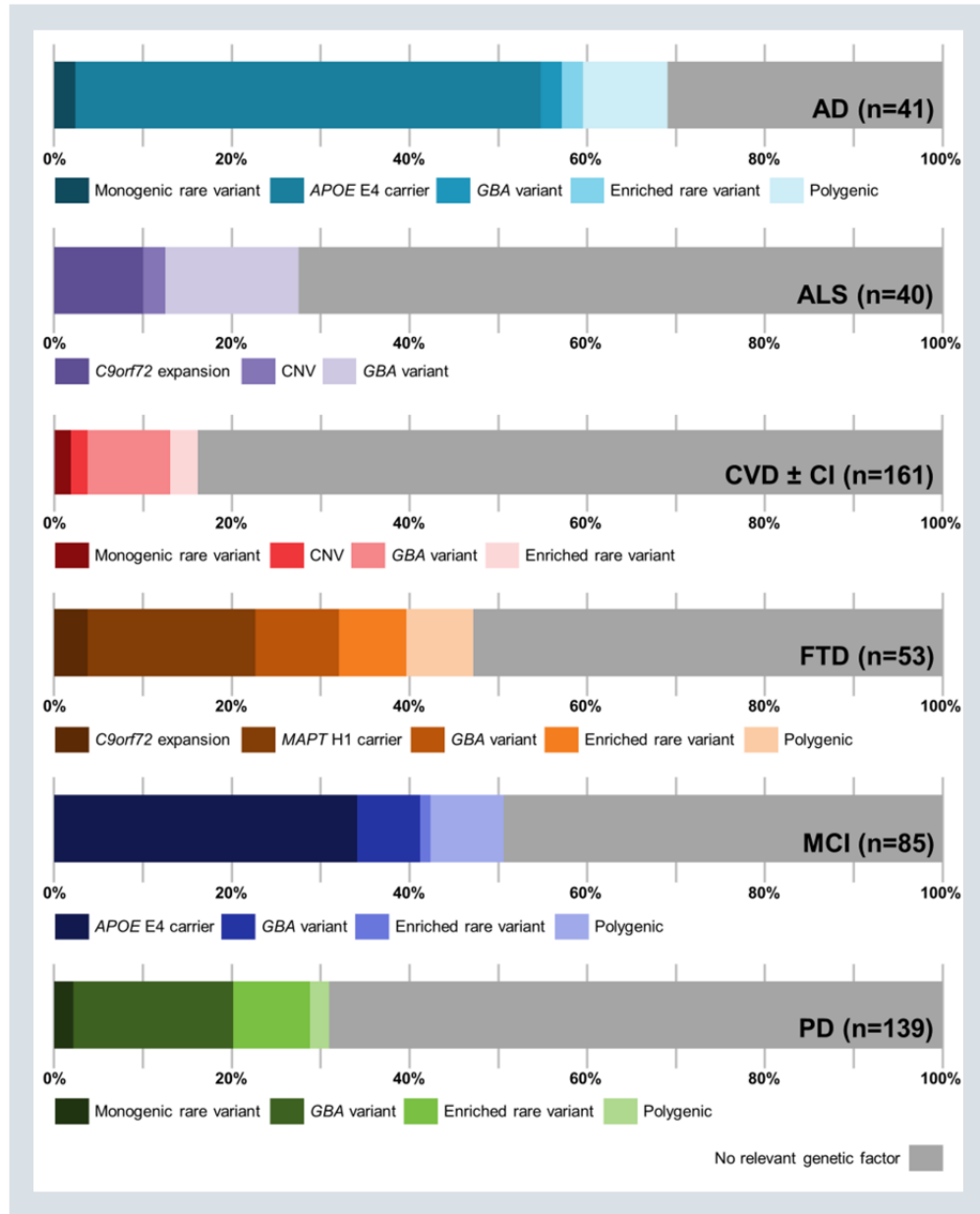
Figure 8.1 Summary of the relevant genetic factors carried by participants of the Ontario Neurodegenerative Disease Research Initiative (ONDRI) that were found to be associated with the disease cohorts.
In total, genetic characterization was performed on 519 participants enrolled in ONDRI, including individuals diagnosed with: 1) Alzheimer's disease (AD); 2) amyotrophic lateral sclerosis (ALS); 3) frontotemporal dementia (FTD); 4) mild cognitive impairment (MCI); 5) Parkinson's disease (PD); and 6) cerebrovascular disease with or without cognitive impairment (CVD ± CI). To identify contributors to disease presentation, all participant DNA samples underwent targeted next-generation sequencing using the ONDRISeq panel, TaqMan genotyping to define *MAPT* haplotypes, repeat-primed PCR to identify *C9orf72* repeat expansions, and Sanger sequencing to identify variants within the gene *GBA*.

"*APOE* E4 carrier" refers to individuals harbouring at least one copy of the *APOE* E4 allele within the AD or MCI cohorts. "*C9orf72* expansion" refers to individuals with ALS or FTD harbouring a pathogenic repeat expansion in *C9orf72*. "CNV" refers to individuals harbouring a copy number variant (CNV) of high confidence in a gene encompassed by the ONDRISeq panel. "Enriched rare variant" refers to individuals harbouring a variant within a gene or gene set found to be association with the cohort of interest using rare variant association analysis. "*GBA* variant" refers to individuals harbouring a likely pathogenic variant in *GBA*; *GBA* variation within all cohorts was assessed based on the relatively high number of likely pathogenic variants identified across ONDRI participants and despite the lack of comparison to a control cohort. "*MAPT* H1 carrier" refers to individuals diagnosed with FTD harbouring at least one copy of the H1 *MAPT* haplotype. "Monogenic rare variants" refers to individuals harbouring a rare single nucleotide variant known to cause the disease of interest. "Polygenic" refers to individuals carrying multiple genetic factors likely contributing to their disease presentation.

8.2.1. Common genetic contributors to neurodegenerative diseases and their intermediate phenotypes

Although it is generally accepted that common genetic variants tend to contribute small phenotypic effects to disease risk, there are a few prominent examples of these variants that have moderate-to-large impact on risk of neurodegenerative disease. The most notable example is the apolipoprotein E (*APOE*) genotype. The results described in Chapters 2 and 3 highlight the contribution of common genetic variants to various neurodegenerative diseases and cerebrovascular disease, as well as to clinical intermediate phenotypes of the diagnoses (Dilliott et al., 2019; Dilliott et al., 2021).

As described in Chapter 1, Section 1.2.2.1.1, *APOE* genotype is the strongest genetic risk factor for the development of AD. There are three possible genotypes — E3, E2, and E4 — corresponding to the wild-type allele, protective allele for AD, and risk allele for AD, respectively (Bertram, McQueen, Mullin, Blacker, & Tanzi, 2007). Due to its high importance in the risk of AD, it was imperative that all ONDRI participants were genotyped for *APOE* and that we replicated the association between AD and E4 in the study cohort as a means to validate the AD diagnoses across the patient cohort. Further, various previous studies have attempted to identify whether *APOE* variation is associated with any other neurodegenerative diseases. ONDRI's concurrent and consistent assessment of multiple diagnoses offered a unique opportunity to pursue this analysis, and specifically to determine whether *APOE* E4 or E2 conferred risk to other neurodegenerative diagnoses. The results of this analysis were included in Chapter 2.

Within the AD cohort, 65.9% of participants carried at least one copy of the E4 allele. Unsurprisingly, the allele accounted for the largest percentage of genetic explanation across any single ONDRI disease cohort (**Figure 8.1**). When allele frequencies were compared to a cognitively normal, elderly control cohort, I identified an expected dose-dependent association between the *APOE* E4 allele and AD presentation, such that the E4 allele itself increased odds of AD presentation by 5.24 ($p < 1.0E-4$) and the E4/4 genotype increased odds of AD presentation by 10.36 ($p < 1.0E-4$). I also identified an association between MCI presentation and the E4 allele, with an increased odds of MCI

presentation by 1.94 (p = 4.9E-3). No other ONDRI cohorts showed any association with the *APOE* E4 allele.

I also analyzed the *APOE* E2 allele, as previous studies had shown a protective effect of the variant for AD presentation. Indeed, I was able to again replicate this result in the ONDRI cohort. The allele was associated with a significantly decreased odds of presenting with either AD or MCI (OR = 0.10, p = 2.7E-2 and OR = 0.26, p = 5.8E-3, respectively), when compared to the cognitively normal, elderly control cohort. Interestingly, no AD or MCI participants in the ONDRI study harboured an E2/2 genotype. While the genotype is considered relatively uncommon in the general population, this observation may also have been due to the incremental decrease in AD or MCI risk that would result from two copies of the allele.

Although the results of the study were not novel and acted as validation of previous work, to our knowledge, it was the first study to concurrently analyze *APOE* genotypes across such a broad spectrum of neurodegenerative diagnoses and cerebrovascular disease phenotypes. Further, our results acted as an important validation of the accurate diagnosis of the ONDRI AD and MCI participants, as the *APOE* allele frequencies we observed were relatively similar to those previously reported in the literature (Farrer et al., 1997; Heffernan, Chidgey, Peng, Masters, & Roberts, 2016). *APOE* genotypes are also an important variable for future analyses within ONDRI, as is it often a critical correction factor due to its associations with AD and intermediate phenotypes of neurodegeneration and cerebrovascular disease. For example, in the study of the influence of rare Notch receptor 3 (*NOTCH3*) genetic variants in PD in Chapter 7, the analysis was corrected for the presence of the *APOE* E2 allele. Pathogenic *NOTCH3* variants are typically associated with a monogenic cerebrovascular disease called Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL), and we studied whether variants in the gene could also modify the presentation of PD patients, specifically by increasing their cerebrovascular disease burden. Importantly, *APOE* E2 has been previously associated with increased white matter hyperintensity (WMH) volume in CADASIL patients; therefore, it was an important correction factor in

our analysis of PD patients, as well. I anticipate that the *APOE* genotype data will be a critical component in ONDRI analyses for years to come.

While we did not observe any novel associations between *APOE* and the other neurodegenerative diseases encompassed by ONDRI's mandate, I also aimed to determine whether variation in the gene may influence the clinical presentation of neurodegenerative disease patients. As previously described, there is considerable overlap among presenting features of various neurodegenerative diseases, such as those with motor impairments (PD and ALS) presenting with cognitive dysfunction (Kalia & Lang, 2015; Karch et al., 2018). And even within single neurodegenerative cohorts, presentations can be highly heterogenous between patients (Beber & Chaves, 2013; Doran et al., 2007; Selvackadunco et al., 2019). The rigorous assessment of ONDRI participants, particularly with a detailed neuropsychology battery, offered a unique opportunity to assess whether *APOE* E2 or E4 variants contributed to differing functioning across multiple cognitive domains in the ONDRI cohorts, as presented in Chapter 3.

Irrespective of disease cohort, I found that E4 carriers had significantly lower performance in verbal memory and visuospatial domains than those with the E3/3 genotype (p = 9.2E-3 and p = 2.2E-2, respectively). Although the results may be suggesting that the participants carrying the E4 allele are in a prodromal stage of AD with increased deposition of early-AD pathology, the influence of *APOE* E4 on cognition appears to be irrespective of AD pathology, rather relative to synaptic plasticity or neurotoxicology. The influence of the E4 allele on verbal memory was also compounded in the AD cohort, with an interaction observed between the diagnosis and the allele, although power limitations likely prevented this result from reaching statistical significance (p = 6.8E-2). In contrast, the E2 allele did not display significant association with cognition across the ONDRI cohorts, although E2 carriers in the FTD cohort displayed significantly worse attention/working memory, executive function, and visuospatial abilities (p = 2.1E-3, p = 4.2E-2, and p = 2.4E-4, respectively). Interestingly, the majority of the E2 carriers in the FTD cohort were diagnosed with the progressive supranuclear palsy (PSP) subtype, which presents with a hallmark of tau pathology, as

well as a relatively high degree of cognitive dysfunction. *APOE* E2 has been previously associated with increased tau pathology and so whether the association we observed between E2 and lower cognition in the FTD cohort was a direct effect or whether it was indicative of an increase in tau pathology — and resultant PSP — remains unclear. Nonetheless, my results demonstrate the complex influence of *APOE* on cognition and highlight that genetic factors may be contributing to the vast heterogeneity observed across the presentation of patients with identical diagnoses.

The results presented in Chapter 2 also display the distribution of microtubule-associated protein tau (*MAPT*) haplotypes across the ONDRI cohort. *MAPT* is located within a large region of linkage disequilibrium (LD), inside of which a 900 kb portion of the genome has been inverted, resulting in two haplotypes of *MAPT*, specifically H1 and H2. Interestingly, I did not identify any associations between the haplotypes and any of the full ONDRI cohorts in comparison to the cognitively normal, elderly control cohort, irrespective of previous associations identified between H1 and AD, PD, and FTD (Ferrari et al., 2017; Myers et al., 2005; Seto-Salvia et al., 2011). However, there was an association between the H1 haplotype and the presentation of PSP. Again, these results were not novel, but replicated a previously observed association with the phenotype (Baker et al., 1999). Within the H1 haplotype, there have also been duplicated regions identified that define *MAPT* sub-haplotypes. Due to the small sample sizes of ONDRI, we chose not to pursue further genotyping to define these sub-haplotypes, as statistical analyses would have had markedly limited power. Therefore, it remains unknown whether the lack of associations between *MAPT* and AD or PD were merely a consequence of not further defining the common variants under study.

Although the results of Chapter 2 did not identify novel associations between *APOE* or *MAPT* and the neurodegenerative diseases or CVD, the lack of associations were an important addition to the literature. The associations that I replicated between *APOE* variation and AD/MCI and *MAPT* and PSP acted as a proof of concept for the strength of the analysis and integrity of the approach and study design; the findings allow us to confidently rule out any further associations of moderate to large phenotypic impact with the genes, which have been controversially proposed previously in the literature (Baum et

al., 2006; Huang, Chen, & Poole, 2004; Myers et al., 2005). Furthermore, Chapter 3 demonstrates the importance in considering known neurodegenerative disease associated genes, such as *APOE*, not only when studying disease risk, but also when endeavoring to account for the large amount of heterogeneity among the clinical features of neurodegeneration.

8.2.2. Rare genetic contributors to neurodegenerative diseases and their intermediate phenotypes

Throughout the literature, it is well documented that rare monogenic variants contribute a relatively large amount of risk to neurodegenerative diseases; however, across the spectrum of diagnoses, a large amount of missing heritability remains. Not only is this missing heritability likely due to rare variants of small to moderate phenotypic effect that are difficult to identify using traditional linkage analysis or genome-wide association analysis (GWAS) approaches, but the presence and influence of structural variants remain relatively understudied in neurodegeneration. Throughout Chapters 3–7, I assessed the contribution of rare genetic factors to the diseases encompassed by the ONDRI mandate, particularly to fill the gap of the potential overlapping impact of rare genetic factors to multiple neurogenerative diagnoses.

I began by assessing the number of ONDRI participants that carried rare, small-scale variants such as SNVs and repeat expansions that were likely causes for Mendelian forms of disease. Unsurprisingly, only ~3% of participants harboured either pathogenic SNVs in a gene covered by the ONDRISeq panel or a pathogenic chromosome 9 open reading frame 72 (*C9orf72*) repeat expansion. The ALS cohort had the highest frequency of these variants, with four participants carrying a pathogenic repeat expansion in *C9orf72*, equating to a frequency of 10% across the cohort (**Figure 8.1**). Although this frequency is relatively low compared to the 40–60% typically observed in familial cohorts of ALS, it is relatively similar to the 5–10% frequency estimates of cohorts with sporadic ALS (Umoh et al., 2016). Within ONDRI, the majority of ALS participants were considered to be sporadic cases, with only two participants self-reporting a family history of the disease, one of whom carried the *C9orf72* repeat expansion. In fact, the low frequency of familial cases was observed across all ONDRI cohorts, and many participants presented

with later-onset, more mild forms of the disease as a result of the strict enrollment criteria (Sunderland et al., 2020). Therefore, it was expected that the frequency of participants carrying monogenic rare variants would remain relatively low.

Nonetheless, sporadic forms of neurodegeneration are also known to have a genetic component, with heritability estimates of up to 80% (Cacace, Sleegers, & Van Broeckhoven, 2016; Keller et al., 2012; Mejzini et al., 2019), although much of the genetic risk is largely unexplained. I hypothesized that some of the missing heritability of these conditions may result from the lack of analysis across neurodegenerative disease cohorts and consideration of pleiotropic genes. To assess this possibility, I performed rare variant association analyses (RVAAs) to identify genes, and sets of genes, covered by the ONDRISeq panel that were enriched for rare variants across the neurodegenerative diseases and CVD.

Using an individual gene-based RVAA approach, I identified a novel association between nonsynonymous rare variants in Parkin RBR E3 ubiquitin protein ligase (*PARK2*) and the FTD cohort (OR = 11.602 [1.395–96.519], p = 2.3E-2). Typically, pathogenic variants in *PARK2* cause autosomal recessive familial PD; however, variants within the gene specifically contribute to lysosomal dysfunction — a pathogenic mechanism known to contribute to both PD and FTD (Wallings, Humble, Ward, & Wade-Martins, 2019). Therefore, the finding may suggest that the variants are indeed contributing to increased risk of FTD. An association was also identified between nonsynonymous rare variants in *NOTCH3* and the PD cohort (OR = 2.986 [1.310–6.806], p = 9.0E-3). Homozygous rare variants in *NOTCH3* cause CADASIL, a Mendelian disease characterized by small vessel damage and dementia; although, there is evidence that some CADASIL patients present with a late-onset, slowly progressive parkinsonism (Ragno et al., 2013; Ragno et al., 2016). Together with the RVAA results, this may suggest that *NOTCH3* rare variants increase risk of PD.

To maximize analytic power, I also created gene sets based on the most well-established previous disease associations of the genes encompassed by the ONDRISeq panel. Interestingly, an association was observed between rare, putative loss-of-function (LOF)

variants in PD-associated genes and the entire ONDRI cohort (OR = 7.322 [1.196–44.826], p = 3.1E-2). The association of putative LOF variants in PD-associated genes was largely driven by variants in the gene melanocortin 1 receptor (*MC1R*). While the gene has been controversially associated with increased risk of PD (Gan-Or et al., 2016; Lorenzo-Betancor, Wszolek, & Ross, 2016; Tell-Marti et al., 2015), it has also been associated with red hair and melanoma (Chen, Feng, Schwarzschild, & Gao, 2017). Although the MC1R protein is neuroprotective within dopaminergic neurons, it remains unclear how the variants may contribute to pleiotropic risk across neurodegenerative diseases and CVD. An association was also observed between rare, putative LOF variants in ALS-associated genes and the ALS and MCI cohorts. The association between the genes and ALS was anticipated, but the association with MCI remained novel. Caution was taken so as to not draw inappropriately large conclusions from the results, since sample sizes of variant positive participants and controls were small. However, it was still recognized that the association between putative LOF variants in ALS-associated genes and MCI may suggest potential for a portion of the participants in the MCI cohort to progress to FTD, rather than the anticipated AD.

The ONDRISeq data was further leveraged to identify large-scale CNVs, such as deletions and duplications, among the ONDRI participants. Heretofore, these structural variants have remained under-studied across neurodegenerative disease and CVD cohorts due to the previous technological challenges involved in their detection. I hypothesized that some of the missing heritability of the diseases may be accounted for by undetected CNVs. Recent bioinformatics advances allowed for the utilization of a depth-of-coverage (DOC) method for CNV detection, which I applied in Chapter 5. Following DOC analysis of the ONDRISeq data across all ONDRI participants, we applied breakpoint analysis to confirm the presence of a heterozygous deletion in optineurin (*OPTN*) harboured by an individual with AD, as well as WES to confirm the presence of duplications in parkinsonism associated deglycase (*PARK7*), ATP binding cassette subfamily C member 6 (*ABCC6*), and SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1 (*SAMHD1*), in individuals diagnosed with ALS, CVD, and MCI, respectively (**Figure 8.1**). I also identified two additional duplications of high confidence in ATP binding cassette subfamily A member 7 (*ABCA7*) and ATPase

cation transporting 13A2 (*ATP13A2*) each in an individual with CVD. Although all CNVs identified were novel, and functional analyses are required to confirm their pathogenicity, the identification of CNVs in genes previously associated with neurodegeneration and cerebrovascular disease across the ONDRI cohorts highlights the need for further analysis of the impact of structural variants to these phenotypes.

The glucosylceramidase beta (*GBA*) gene could not be adequately evaluated with traditional NGS techniques applied using the ONDRISeq panel due to a pseudogene located 16 kb downstream, namely glucosylceramidase beta pseudogene 1 (*GBAP*). However, we were able to capture the full coding sequence of the gene by applying a gold-standard Sanger sequencing approach with carefully designed primers. From the *GBA* sequences of all ONDRI participants, variants were identified and prioritized to identify those most likely to be pathogenic. Unfortunately, controls were unavailable for this analysis, preventing the ability to perform traditional RVAA. Yet the analysis was still able to yield novel results, as displayed in Chapter 6. Interestingly, I identified likely pathogenic, rare *GBA* variants, not only in the PD cohort, but in all other cohorts of ONDRI as well (**Figure 8.1**). In fact, the cohort with the lowest variant frequency — namely the CVD cohort with a frequency of 9.3% of participants carrying a variant — still had a frequency that was more than double that of the of the highest frequencies in a previously reported control cohort (~4%) (Sidransky et al., 2009). Across the ONDRI cohorts, there were also 47 participants identified that harboured distinctive collections of variants, referred to as "complex variants," within exons 5 and 6 of *GBA*, which may have resulted from small-scale rearrangements with *GBAP*. The relatively high frequency of variants across the ONDRI cohorts may suggest that *GBA* is indeed contributing to risk of multiple neurodegenerative diseases. The results may also suggest that *GBA* variant carriers within ONDRI not diagnosed with PD have unrecognized α-synuclein co-pathology, as *GBA* variants are also known to increase the presence of Lewy bodies, the main component of which are aggregates of α-synuclein (Mata et al., 2008; Nalls et al., 2013). In turn, the co-pathology may be influencing the presentation of the individuals, and therefore may account for some of the heterogeneity between patients observed in the presentation of neurodegenerative disease and CVD.

In an effort to determine whether the likely pathogenic *GBA* variants influenced disease presentation, I assessed the difference in age of symptom onset, general cognition, and basic motor phenotype in the variant positive individuals in comparison to those variant negative, but no conclusions could be drawn from the analysis. Nevertheless, with the relatively high frequencies of *GBA* variation observed across the ONDRI cohorts, it will be important to continue assessing the influence the variants may have on clinical presentation of the participants. For example, in a similar manner to the analysis of *APOE*'s influence on cognition presented in Chapter 3, utilizing the rich neuropsychology dataset may allow for a more sensitive analysis of how *GBA* variation influences cognition across multiple domains. Similarly, the gait performance data may allow for a more robust analysis of motor phenotypes across the ONDRI cohorts.

As with the differential possible interpretations of the *GBA* data, the RVAA results from Chapter 4 may also not be suggesting pleiotropic effects of the genes significantly enriched for rare variants. Rather, the enrichment of variants in these genes may actually be a result of rare variants of moderate phenotypic effect that influence overlapping clinical features of disease between the diagnoses. As previously stated, this is an area of neurodegenerative disease and CVD research that remains relatively understudied, yet understanding the factors influencing differential presentation of neurodegeneration and CVD is imperative for a full appreciation of the genetic factors that influence these complex phenotypes.

An excellent example of rare genetic variation contributing to different intermediate phenotypes of neurodegenerative disease was encompassed in the comprehensive analysis of *NOTCH3* in the PD cohort. Upon early analysis of the PD participants' ONDRISeq NGS data, I observed a relatively high frequency of likely pathogenic rare variation within the *NOTCH3* gene, with 13 participants carrying a nonsynonymous, rare variant of interest (~9%). As explained previously, heterozygous pathogenic variants in *NOTCH3* are causative for CADASIL, a monogenic form of vascular dementia (VaD). Therefore, the identification of potentially pathogenic rare variants in the *NOTCH3* gene in PD participants was notable. Further collaboration with the ONDRI imaging platform, outlined in Chapter 7, resulted in an analysis through which we identified a doubling of

WMHs volumes in *NOTCH3* variant positive participants, compared to variant negative participants. Additionally, *NOTCH3* genetic variation was associated with significantly increased periventricular WMHs and increased lacune volumes. Although small sample sizes likely limited the ability to identify differences in generalized cognition or motor phenotypes and the genetic variants, we hypothesize that the increased WMHs observed in the PD participants harbouring *NOTCH3* variants likely contribute to differential disease presentation, warranting further analysis.

The association between *NOTCH3* and cerebrovascular phenotypes in PD patients was also supported by the results of the gene-based RVAAs performed in Chapter 4. As described, *NOTCH3* was found to be enriched for rare variants in the PD cohort, albeit in comparison to the combined AD and MCI cohorts rather than the cognitively normal control cohort. Importantly, while the control cohort was known to be cognitively normal, whether they had any form of CVD remained unknown. In contrast, the AD and MCI cohorts' enrollment criteria excluded any participants with significant evidence of vascular pathology (Sunderland et al., 2020), which resulted in the combined cohort likely acting as a more appropriate control for the detection of enriched variants in a VaD-associated gene. The result seemingly supports the observations in Chapter 7 and suggests that rather than contributing to PD pathogenesis, *NOTCH3* rare variants may be altering the presentation of the PD participants carrying the variants.

Similarly, it cannot be denied that the association between *PARK2* and FTD observed in Chapter 4 may actually be a result of variants of moderate phenotypic effect and/or decreased penetrance influencing intermediate features of disease, rather than directly contributing to FTD risk. As this possibility demonstrates, the novel genetic associations I have identified throughout my Dissertation are hypothesis generating and the comprehensive ONDRI dataset will allow for unique cross-platform assessments to further elucidate how the genetic variants may be contributing to disease risk and differential disease presentation. Overall, my work has provided an important tangible contribution to the elucidation of the entire spectrum of genetic variation contributing to neurodegenerative disease and CVD phenotypes.

8.2.3.   Contributions to the goals of ONDRI

Upon its conception, ONDRI set out with goals to identify markers applicable to early and accurate diagnostic and progression prediction tools for neurodegeneration, as well as to thoroughly analyze the contribution of cerebral small-vessel pathology to neurodegenerative phenotypes. The work I have presented throughout this Dissertation clearly demonstrates progress towards both of these goals. Additionally, my PhD work has encompassed data generation that will be continue to be made available through the ONDRI consortium, allowing for continued analysis and, potentially, further discoveries in the field of neurodegeneration.

The data collected through ONDRI are stored in a secure database, run by the Ontario Brain Institute (OBI), called Brain-CODE (braincode.ca). The centralized portal was designed to house the data produced by more than 160 researchers located at over 20 clinical sites across Ontario. Not only has Brain-CODE allowed for effective collaboration, but all non-identifying data will be made available to the greater scientific community on request through this portal. Therefore, it was imperative that all data being produced through ONDRI go through rigorous data cleaning and standards checks prior to upload to Brain-CODE. Upon completion of my PhD, the work presented herein will have contributed to four genomic platform datasets — collectively encompassing over 675,000 data points — which contain the *APOE* genotypes, *MAPT* haplotypes, *C9orf72* repeat expansions, ONDRISeq panel annotated small-scale variants, ONDRISeq panel CNVs, and *GBA* small-scale variants. These data will continue to be accessed by ONDRI researchers for further genomics analyses, as well as by collaborators from across the globe.

In addition to the data produced through my PhD, this Dissertation also reflects successful steps towards the goals of ONDRI. In particular, ONDRI was designed with its unique cross-disease and cross-platform approach, in hopes of breaking down the silos typically observed in the field of neurodegenerative disease research. While Chapters 2, 4, 5, and 6 demonstrated cross-disease analyses with the study of *APOE*, rare variant enrichment, CNVs, and *GBA* variation, respectively, Chapter 7 demonstrated cross-platform analyses in the study of neuroimaging metrics in reference to *NOTCH3*

variation. Most notably, <u>Chapter 3</u> clearly demonstrates the application of both cross-disease and cross-platform analysis and can act as a proof of concept for the unique study design of ONDRI moving forward.

Finally, ONDRI aimed to gain a better understanding of the contribution of cerebral small vessel disease to neurodegenerative phenotypes. Although without functional analyses the results of the RVAA analysis and *NOTCH3* rare variant analysis in PD cannot make definitive pathologic connection between the *NOTCH3* variants and CVD in PD patients, the results still suggest an important co-pathology for the participants carrying these variants. Furthermore, the novel findings may now be further explored to better understand how the increased small vessel disease influences the presentation of neurodegenerative disease patients with *NOTCH3* variants. Therefore, my results represent the first large step in achieving one of the core goals of ONDRI.

8.3. Dissertation strengths and limitations

Throughout this Dissertation, study-specific strengths and limitations have been presented within each relevant data chapter. Here, I will present strengths and limitations relevant to all data and analyses included in <u>Chapters 2–7</u> in context to the overall results of the Dissertation.

8.3.1. Strengths

A distinct strength of this Dissertation is the unique study design of ONDRI. Upon conceptualization of the consortium, priority was placed on the interdisciplinary research model. It was well known that there were many experts in their respective fields studying neurodegenerative diseases and treating neurodegenerative disease patients, but many have traditionally remained in their respective silos. Further, the large size of the province of Ontario posed a challenge to the coordination of collaborative efforts without detailed protocols in place. ONDRI brought together experts in the field, including both scientists and clinicians, for a concerted effort to tackle the large amount left to be understood regarding neurodegenerative and cerebral small vessel diseases. Priority was placed on recruiting a diverse cohort of participants from throughout the province to effectively represent the general population, and on deeply phenotyping participants to

gain an all-encompassing depiction of neurodegenerative disease presentation. Further, the longitudinal nature of evaluation would allow for investigation into the progression of individuals diagnosed with these incurable diseases.

Throughout my Dissertation, I leveraged ONDRI's study design to perform analyses across the neurodegenerative disease cohorts. Not only was this quite novel in the field of neurogenetics, but it allowed for the unique opportunity to identify overlapping genetic risk factors concurrently for all diseases of interest. Many previous studies have only focused on single disease cohorts, or multiple disease cohorts, but then only applied GWAS-based methodology, thereby limiting utility (Karch et al., 2018). In contrast, the work I reported in Chapters 2–6 encompassed five neurodegenerative diagnoses, as well as cerebrovascular disease, and greatly expanded the range of applied methodologies in order to gain a deeper understanding of the full spectrum of genetic variants contributing to these phenotypes. In Chapters 2, 4, 5, and 6, I sought to identify novel gene-disease associations within known neurodegenerative disease genes. This approach proved successful based on the novel associations found between *PARK2* rare variation and FTD, *NOTCH3* rare variation and PD, and all neurodegenerative diseases and LOF variants in PD-associated genes. I also identified CNVs in genes that were not typically associated with the diagnoses of the individuals carrying the large-scale variants, as well as a large amount of likely pathogenic *GBA* variation occurring across all ONDRI cohorts. Additionally, I took advantage of ONDRI's unique study design by driving strong collaborative projects with other ONDRI platforms, as displayed in Chapters 3 and 7, which ultimately resulted in the identification of novel associations between *APOE* and cognitive impairment (Dilliott et al., 2021), as well as between *NOTCH3* and small vessel disease in PD patients (Dilliott et al., 2020).

Another strength of this Dissertation was the use of the ONDRISeq NGS targeted gene panel. The custom-designed panel allowed for the sequencing of 80 genes previously associated with the neurodegenerative and cerebral small vessel diseases encompassed by ONDRI's mandate. From these data, we accurately extracted the genotyping data necessary to map *APOE* genotypes, identify rare variants of various pathogenicity, and capture large-scale CNVs. Upon validation of the ONDRISeq panel, its strengths were

highlighted, as we observed that variant calls using the NGS method were more reliable than those obtained from the NeuroX microarray, as confirmed using Sanger sequencing (S.M.K. Farhan et al., 2016). Furthermore, although NeuroX captured >250,000 SNVs exome-wide (Nalls et al., 2015), ONDRISeq allowed for the identification of novel variants, which may be contributing to disease presentation, as well as the subsequent gene-based binning and RVAAs. During the conceptualization of ONDRI, whole-genome sequencing (WGS) and whole-exome sequencing (WES) methodologies were also explored; however, the costs remained prohibitive, and the computational power necessary for data processing posed a challenge at the time. In contrast, ONDRISeq offered a targeted approach to further explore genes known to contribute to neurodegenerative and cerebrovascular phenotypes in diseases not typically associated with the genes. Combined with the cross-cohort strengths of the ONDRI design, ONDRISeq allowed for analyses to detect genes that may have pleiotropic effects across diseases or may contribute to overlapping disease features, without the noise and added statistical power reduction that would be introduced by unrelated genetic loci captured by WGS or WES.

8.3.2. Limitations

Expectedly, this Dissertation is not without its limitations that must be considered when interpreting the results and conclusions presented. The most fundamental limitation is the relatively modest sample sizes of the ONDRI cohort. Upon study conception, ONDRI aimed to recruit 600 participants evenly distributed across the six disease cohorts (S. M. K. Farhan et al., 2017). However, neurodegenerative disease studies often experience difficulties in patient recruitment, particularly when a large degree of engagement is expected from the enrolled individuals. Due to the deep phenotyping performed on each participant, and the longitudinal nature of the follow up, enrolling in ONDRI proved to be a substantial commitment on the part of the participants and their essential study partners. Obtaining the intended cohort proved difficult. Further, ONDRI had strict enrollment criteria that each participant was required to meet, including successfully completing an analysis by magnetic resonance imaging (MRI), which proved to be burdensome for many participants, particularly those with cognitive and/or behavioural phenotypes that

were relatively advanced. Even so, attaining the intended 600 participants would still have left a relatively small sample for subsequent genetic analyses. Typically, genetic analyses require larger sample sizes in order to provide enough statistical power to identify novel genetic associations, especially when studying rare variants. To mitigate this limitation, I prioritized the RVAAs to bin variants into larger groups to maximize power. Yet, in some instances, lack of significant results may have occurred due to ONDRI's small sample sizes. Additionally, the available sample sizes limited in the ability to further sub-classify participants in an attempt to gain a clearer understanding of potential neurodegenerative disease sub-phenotypes. A clear example of this is evident in Chapter 3, in which I was unable to further elucidate whether the contribution of *APOE* E2 to cognitive dysfunction in FTD was a result of the PSP sub-phenotype, which represented an exceedingly small subgroup (Dilliott et al., 2021). Nonetheless, the work presented throughout this Dissertation offers a proof of concept for the overall strategy and analytic approaches, which can be applied to larger sample sizes in the future.

Additionally, the demanding recruitment model of ONDRI, and the existing ethnic distribution of elderly Ontarians in the general population, resulted in inherent biases within our total cohort. Most notably, the ONDRI cohort displayed a large bias to individuals of European descent, as displayed in Chapter 4. This resulted in the inability to assess genetic factors likely contributing to the phenotypes of interest in individuals of diverse ancestral backgrounds, which remains a long-standing concern in the field of genomics (Sirugo, Williams, & Tishkoff, 2019). Further, the demands of being involved in our study introduced biases towards individuals with milder disease presentation, as well as towards individuals with attainment of greater education levels and of higher socioeconomic status (Sunderland et al., 2020). Finally, our cohort displayed an apparent male bias, which we have hypothesized may have been a result of the need for a study partner — a role that, according to anecdotal evidence, females in a heterosexual domestic partnership with the participant were most willing to serve.

Select analyses presented throughout this Dissertation were also limited by the available control cohort. Unfortunately, ONDRI did not include an internal control cohort in its study design; however, the genomics platform sought out control samples for the

purposes of our analyses. We obtained 189 control samples from cognitively normal, elderly individuals recruited by the GenADA study (Li et al., 2008). We subsequently sequenced the samples using ONDRISeq, and genotyped the samples using the NeuroX array and TaqMan allelic discrimination assays to validate the *APOE* genotypes and obtain the *MAPT* haplotypes. But due to financial constraints, we were unable to perform Sanger sequencing of *GBA* on the control samples, resulting in no control cohort for the study presented in Chapter 6. Furthermore, the control cohort samples from the GenADA study had limited phenotypic data, only comprised of age, sex, and Montreal Cognitive Assessment (MoCA) score. Therefore, I was unable to use the controls for cross-platform ONDRI analyses, including those presented in Chapters 3 and 7. It also remained unknown whether the control cohort had underlying neurodegenerative pathology, yet remained asymptomatic, as neuropathology can appear many years prior to neurodegenerative symptom onset (Katsuno, Sahashi, Iguchi, & Hashizume, 2018). Finally, the lack of phenotypic data on the controls resulted in uncertainty regarding how appropriate the cohort was for analyses of cerebral small vessel disease and its associated genes. While the controls were considered cognitively normal, it was not known whether any had experienced CVD. An example of how this may have impacted analyses is demonstrated in the RVAA results presented in Chapter 4, in which an association between *NOTCH3* rare variation and PD was observed when compared to the combined AD and MCI cohorts, but not when compared to the control cohort.

Some variables also remained unknown among the ONDRI participants, such as their underlying neuropathology. Until neuropathological analyses can be performed on our participants — either through post-mortem analysis or potentially through emerging methods of plasma biomarker assessment (Shen et al., 2020) — diagnoses remain presumptive. As described in Chapter 1, Section 1.2.4, although clinical diagnoses are possible, the rates of misdiagnoses remain relatively high. ONDRI attempted to mitigate these issues with strategies such as excluding participants from the AD and MCI cohorts who had displayed significant amount of cerebral small vessel disease, but it remains unclear whether those participants indeed had AD or MCI, or whether they were experiencing a form of VaD. During enrollment, there were also instances of patient enrollment that may not have fully complied with inclusion criteria — for example, some

participants were enrolled in the PD cohort with large stroke volumes and others with staging below the inclusionary cut-off. Additionally, some individuals enrolled in the original vascular cognitive impairment (VCI) cohort were subsequently found not to have objective cognitive impairment upon completion of the neuropsychology battery, and this resulted in the cohort being reclassified as CVD ± CI. If ONDRI did contain misdiagnosed participants, it may have introduced noise into the genetic analyses and prevented true associations from being detected, particularly when combined with the reduced statistical power as a consequence of ONDRI's small sample sizes.

It is also important to recognize the limitations of the ONDRISeq panel; although, it was considered a strength of the studies presented herein as described in Section 8.2.1. ONDRISeq was designed by an expert panel of scientists and clinicians in 2012–2013, and, at the time, it was considered to include all relevant genes known or thought to contribute to neurodegenerative disease and VaD phenotypes. However, in the near decade since these design decisions were made, the panel is now recognized to be missing more recently discovered genes of importance, including those identified through GWAS, such as ADAM metallopeptidase domain 10 (*ADAM10*) and angiotensin I converting enzyme (*ACE*) that are associated with AD (Kunkle et al., 2019), and those identified through RVAA, such as DnaJ heat shock protein family (Hsp40) member C7 (*DNAJC7*) that is associated with ALS (S. M. K. Farhan et al., 2019). Additionally, targeted gene panels are by design limited to evaluate only specific regions of the genome, which impairs the ability to discover novel loci that have previously not been associated with the diseases under study. The ONDRISeq panel also did not include probes covering intronic or intergenic SNPs, previously associated with neurodegenerative disease or CVD through GWAS analysis, thereby limiting the ability to assess the contributions of common variation of small to moderate phenotypic effect to the diseases encompassed by ONDRI's mandate. The inclusion of associated SNPs on the ONDRISeq panel may have allowed for the design of novel, or the assessment of previously designed, polygenic risk scores (PRSs) within the ONDRI cohorts.

Finally, two important caveats must be considered when interpreting the results of the studies presented throughout this Dissertation. The first caveat is that the analyses were

only comprised of a discovery cohort, and results were not replicated. Replication cohorts are particularly important in genomics studies to account for the possibility of spurious associations resulting from random chance or not controlling for variables of importance (Kraft, Zeggini, & Ioannidis, 2009). However, as previously described, recruitment of neurodegenerative disease patients for large-scale studies remains a challenge, and by introducing our results into the literature without internal replication, we offer the opportunity for other research groups to replicate our findings in their own study cohorts. The second caveat is that the novel genetic associations presented throughout the Dissertation have not yet been validated with functional analyses to verify the gene-disease associations and identify the pathologic mechanisms involved. Although considered beyond the scope of my PhD work, these follow-up studies will be imperative to validate the results. For example, it will be important to further explore the relationship between *NOTCH3*, CVD, and PD to determine whether the *NOTCH3* variants are only modifying disease presentation, or whether they are directly contributing to the PD phenotype through a vascular mechanism.

8.4. Applications and future directions

The work I have presented contributes to the elucidation of the genetic determinants of neurodegenerative diseases and CVD, including identifying genetic risk factors for disease presentation and investigating how genetic factors may influence the heterogeneous features of neurodegeneration. The knowledge I have generated can be combined with the existing literature to direct further neurogenetic studies and can be applied towards clinical applications to improve the diagnosis, progression prediction, and treatment of patients.

8.4.1. Genomics-based diagnostic tools

As we continue to gain a better understanding of the genetic factors that contribute to neurodegenerative phenotypes, we may be able to begin addressing some pervasive issues in neurodegenerative diagnostics. Unfortunately, many neurodegenerative diagnostic processes remain relatively slow, allowing the often quickly developing disease states to further progress prior to being able to intervene with the few therapeutics

available to slow disease progression (Hulisz, 2018; Katsuno et al., 2018). As therapeutics currently available cannot completely prevent disease progression or reverse damage, it is imperative that treatments are initiated as soon as possible to preserve patient's function. Further, misdiagnosis is relatively common due to the large amount of phenotypic overlap between neurodegenerative diagnoses (Beber & Chaves, 2013; Doran et al., 2007). Definitive diagnoses currently require post-mortem neuropathologic analysis; however, genomics could also offer an avenue of definitive diagnoses as we continue to expand our understanding of genetic risk of disease.

Currently, genetic sequencing is suggested in the diagnostic pathways of some early-onset familial cases of neurodegenerative disease, yet even individuals with seemingly sporadic forms of disease may harbour underlying genetic variation driving their disease presentation (Bennion Callister & Pickering-Brown, 2014; Maraganore et al., 2006; Mejzini et al., 2019; Sidransky et al., 2009). Therefore, it may be beneficial to expand genetic analysis to a greater number of neurodegenerative disease cases. It is also imperative that a wide range of genetic variant types are considered, particularly structural variation, which was identified as being of potential importance in Chapter 5, yet can be overlooked in clinical genetic diagnostics.

8.4.2. Influence of genetic variation on neurodegenerative disease progression

Gaining a greater understanding of the genetic spectrum of neurodegeneration and CVD may also allow for more accurate progression prediction. Currently, it remains unclear how individuals may progress through their disease upon diagnoses, as features of disease can vary widely from patient to patient. However, as the results of Chapter 7 displayed, genetics may provide a lens through which progression could be predicted, although my analyses only found that genetic variation contributes to features of disease presentation at a single point in time. ONDRI includes a longitudinal nature of assessment, following participants for up to three years. It will be important for further analyses to prioritize the use of the longitudinal ONDRI data to study how genetic variants may influence disease progression, including the degree to which individuals may progress and at what rate.

It will also be interesting to investigate whether certain genetic variants increase the risk of relatively atypical presentation, such as the co-occurrence of cognitive and/or behavioral impairments in individuals diagnosed with a form of motor impairment. In this manner, in combination with genetic diagnostic tools, progression predications tools may be developed, such that an individual's entire genetic profile can not only offer an accurate diagnosis, but can predict the clinical features a patient might experience, as well as how quickly those phenotypes may develop. Ultimately, in the study of such complex diseases, this will be an important component for future personalized medicine efforts.

8.4.3.   Identification of novel gene-disease relationships

Although novel gene-disease associations were identified using RVAA in Chapter 4, it is recognized that these associations remain to be replicated, which may be possible through collaborative efforts with other consortia, such as the Parkinson's Progression Marker Initiative (PPMI), and validated using functional studies. My analysis, as well as previous studies, focused on gene-wide approaches to prevent analytic biases in the RVAAs; however, it is important to consider that variant frequencies can vary within genes and variant enrichment may be localized to functional domains in the encoded protein (Richardson et al., 2016). Therefore, it may be beneficial to repeat the RVAA performed herein, but instead specifically prioritize rare variants that are located within the functional domains of proteins. If a gene were to exhibit similar rare variant frequencies between cases and controls, but different distributions of variants, specifying the analysis using a domain-based approach may detect additional novel gene-disease relationships.

Similarly, as described in Section 8.2.2, a limitation of this Dissertation was a result of the use of ONDRISeq and its inability to identify novel neurodegenerative disease and/or CVD associated genes. However, ONDRI is currently pursuing WGS of all ONDRI participants, with the exception of the CVD ± CI cohort, as well as a relatively small (n ≈ 50) elderly control cohort. Ideally, this would allow for exome-wide RVAAs to be performed, although the low statistical power resulting from the small sample sizes of ONDRI and the control cohort will remain a limitation. Yet, the use of the data could offer the opportunity to identify novel gene-disease associations, again applying both

gene-wide and domain-based enrichment analyses. One opportunity to maximize statistical power could be by applying a gene-set approach by binning genes into functional sets, based on the pathways in which they may contribute to neurodegenerative pathology, such as a set of genes encoding mitochondria-associated proteins, based on their potential involvement in ALS or PD (Khalil & Lievens, 2017; Park, Davis, & Sue, 2018). This could mitigate the effects of genes unrelated to the neurodegenerative phenotypes and their potential influence on power reduction. If significant signals were observed in these gene sets, the genes may be further explored to identify whether a single gene was driving the association.

8.4.4. Cerebrovascular changes in neurodegenerative disease

One of the pillars of ONDRI's rationale and design was to gain a greater understanding of the contribution of cerebral small vessel disease to neurodegenerative disease risk and presentation. It is now well recognized that the co-occurrence of cerebrovascular injury with neurodegeneration is relatively common (Kummer et al., 2019; Lendahl, Nilsson, & Betsholtz, 2019; Thal et al., 2015), but the exact mechanisms by which CVD may contribute to pathology remains unclear. Here, I initiated these efforts in the analysis of *NOTCH3* genetic variation, WMH volume, and PD in Chapter 7, which now offers an opportunity for expansion. WMHs are also not uncommon among other neurodegenerative diseases, such as AD and FTD, but their effects on disease features are not fully understood (Brickman, Muraskin, & Zimmerman, 2009; Kandel et al., 2016; Woollacott et al., 2018). Further, multiple genes on the ONDRISeq panel have been previously associated with various features of CVD burden. ONDRI's simultaneous analysis of patients across multiple neurodegenerative phenotypes using consistent assessment offers the unique opportunity to extend the novel *NOTCH3* finding across multiple disease cohorts and genes. Efforts are currently underway to assess the influence of rare, likely pathogenic variants within all CVD-associated genes of the ONDRISeq panel (**Appendix D**) on CVD defining brain volumetrics across all ONDRI cohorts in collaboration with the imaging platform. Functional analyses will remain imperative to confirm that variants within the CVD genes are influencing features of cerebrovascular pathology within the participants, and to determine whether the small vessel damage is

contributing to the diseases' pathogenic mechanisms or to differential disease presentation.

8.4.5. Estimation of gene-environment interactions

Although genetic factors contribute a clear risk to neurodegenerative disease and CVD presentation, it is also important to recognize the large amount of influence from environmental factors to the risk of neurodegeneration. The most important risk factor for the diagnoses is age, potentially due to mechanisms involving DNA damage, epigenetic changes such as increased DNA methylation, mitochondrial dysfunction, cellular senescence, and telomere maintenance (Hou et al., 2019). However, these cellular pathways are also highly influenced by genetic factors (Atzmon et al., 2010; Coutts et al., 2019; Larsen, Hanss, & Kruger, 2018), leading to a potential for interaction between risks associated with both aging and genetic variation. In fact, brains of neurodegenerative disease patients have displayed age-correlated variable gene expression (Cao, Chen-Plotkin, Plotkin, & Wang, 2010), suggesting important gene-environment interactions, and certain pathogenic mutations causative of neurodegeneration have displayed age-dependent effects (Ho et al., 2020; Longo et al., 2017). Similar gene-environment interactions have been observed in relation to pesticide exposures and other environmental toxins (Casarejos et al., 2006; Liu et al., 2017).

This Dissertation did not address the potential for significant interactions between genetic and environmental risk factors, yet in order for the entire spectrum of neurodegenerative and cerebral small vessel disease risk to be assessed, we must obtain a comprehensive understanding of the genetic influences on disease. The novel genetic associations I identified demonstrate that there are still important discoveries to be made to contribute to the greater analysis of gene-environment interactions. Further, the identification of potentially pathogenic variation in individuals with unexpected diagnoses, such as the *GBA* variation observed across the non-PD diagnoses of ONDRI, may suggest that there are other genetic and/or environmental factors contributing to the differential phenotypes of these patients. Therefore, studying the potential influence from both environmental and genetic risk factors may prove itself necessary to truly understand disease risk moving forward.

8.4.6. Development of novel therapeutic targets

A common goal for those studying the genetic determinants of neurodegeneration is addressing the considerable need for appropriate therapeutic targets and the development of treatments that may halt disease development or prevent its onset. Although there have been many attempts at this across neurodegenerative diagnoses, to date there remains few substantial breakthroughs in this quest. It is well-established that 85% of clinical trials fail, and neurodegenerative diseases are no exception to this rate (Olanow, Kieburtz, & Schapira, 2008; Oxford, Stewart, & Rohn, 2020; Petrov, Mansfield, Moussy, & Hermine, 2017; Wong, Siah, & Lo, 2019). By gaining a greater understanding of the breadth of genetic variation that may contribute to disease risk, we may identify novel therapeutic targets specific to the underlying genetic risk factors.

Furthermore, it is important to recognize that many of these efforts are currently underway, such as clinical trials of drugs specific to individuals diagnosed with PD that carry pathogenic *GBA* variation or individuals diagnosed with AD carrying the *APOE* E4 genotype (Schneider & Alcalay, 2020; Yang, Kantor, & Chiba-Falek, 2021). However, the trials have yet to involve patients with different diagnoses, even though a substantial amount of genetic overlap has been observed across the diseases. A clear example of this overlap is displayed in Chapter 6, in which a relatively high frequency of likely pathogenic *GBA* variation is observed across all ONDRI cohorts. By applying these findings, if the clinical trials were to prove successful, it will be absolutely crucial that we gain a full understanding of the genetic overlap between neurodegenerative diseases and potential co-pathologies, as therapies may be useful to patients with seemingly different diagnoses.

8.5. Conclusions

Through the genetic analyses of individuals with neurodegenerative disease and CVD from the ONDRI study, I have contributed to the delineation of the complex genetic architecture of the phenotypes. My work has demonstrated the wide range of genetic variants that contribute to neurodegeneration and cerebral small vessel diseases and the potential overlap of genetic risk that has not previously been well defined. Specifically, I

leveraged the ability to concurrently evaluate patients with different neurodegeneration diagnoses using the ONDRISeq targeted NGS panel to characterize common genotypes and haplotypes, rare SNVs, and CNVs contributing to disease presentation and features. Chapters 2 and 3 demonstrated the important influence of *APOE*, not only to AD and MCI risk, but to cognitive impairment across all disease cohorts. Chapter 4 highlighted potentially novel gene-disease relationships and, along with Chapters 5 and 6, demonstrated that there may be a greater amount genetic overlap between neurodegenerative and cerebrovascular diagnoses than previously appreciated. Finally, the results of Chapter 7 supported the theory that aside from rare variants of high phenotypic effect that can drive diagnoses within a Mendelian model, there may also be rare variants of moderate phenotypic effect that are influencing neurodegenerative presentation and contributing to the large amount of phenotypic heterogeneity observed between patients with the same diagnoses and the phenotypic overlap between those with different diagnoses. Although I have applied robust methodologies to maximize the utility of the limited sample sizes within ONDRI, replication of the novel findings and functional analyses of novel variation are required to fully elucidate their contribution to disease pathogenesis. Nonetheless, to my knowledge, this Dissertation represents the first compendium to date of analyses of genetic associations across this number of neurodegenerative diseases at once with such deep, consistent phenotyping of participants. As the view of neurodegenerative diseases continues to evolve to consider diagnoses as sitting on a spectrum with mixed pathologies and overlapping etiologies, genetic factors will continue to become an ever more important indicator of presentation risk, particularly in regards to features and progression of disease.

8.6.   References

Atzmon, G., Cho, M., Cawthon, R. M., Budagov, T., Katz, M., Yang, X., . . . Suh, Y. (2010). Evolution in health and medicine Sackler colloquium: Genetic variation in human telomerase is associated with telomere length in Ashkenazi centenarians. *Proc Natl Acad Sci U S A, 107 Suppl 1*, 1710-1717. doi:10.1073/pnas.0906191106

Baker, M., Litvan, I., Houlden, H., Adamson, J., Dickson, D., Perez-Tur, J., . . . Hutton, M. (1999). Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum Mol Genet, 8*(4), 711-715. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10072441

Baum, L., Lam, L. C., Kwok, T., Lee, J., Chiu, H. F., Mok, V. C., . . . Ng, H. K. (2006). Apolipoprotein E epsilon4 allele is associated with vascular dementia. *Dement Geriatr Cogn Disord, 22*(4), 301-305. doi:10.1159/000095246

Beber, B. C., & Chaves, M. L. F. (2013). Evaluation of patients with behavioral and cognitive complaints: misdiagnosis in frontotemporal dementia and Alzheimer's disease. *Dement Neuropsychol, 7*(1), 60-65. doi:10.1590/S1980-57642013DN70100010

Bennion Callister, J., & Pickering-Brown, S. M. (2014). Pathogenesis/genetics of frontotemporal dementia and how it relates to ALS. *Exp Neurol, 262 Pt B*, 84-90. doi:10.1016/j.expneurol.2014.06.001

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet, 39*(1), 17-23. doi:10.1038/ng1934

Brickman, A. M., Muraskin, J., & Zimmerman, M. E. (2009). Structural neuroimaging in Altheimer's disease: do white matter hyperintensities matter? *Dialogues Clin Neurosci, 11*(2), 181-190. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19585953

Cacace, R., Sleegers, K., & Van Broeckhoven, C. (2016). Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimers Dement, 12*(6), 733-748. doi:10.1016/j.jalz.2016.01.012

Cao, K., Chen-Plotkin, A. S., Plotkin, J. B., & Wang, L. S. (2010). Age-correlated gene expression in normal and neurodegenerative human brain tissues. *PLoS One, 5*(9). doi:10.1371/journal.pone.0013098

Casarejos, M. J., Menendez, J., Solano, R. M., Rodriguez-Navarro, J. A., Garcia de Yebenes, J., & Mena, M. A. (2006). Susceptibility to rotenone is increased in neurons from parkin null mice and is reduced by minocycline. *J Neurochem, 97*(4), 934-946. doi:10.1111/j.1471-4159.2006.03777.x

Chen, X., Feng, D., Schwarzschild, M. A., & Gao, X. (2017). Red hair, MC1R variants, and risk for Parkinson's disease - a meta-analysis. *Ann Clin Transl Neurol, 4*(3), 212-216. doi:10.1002/acn3.381

Coutts, F., Palmos, A. B., Duarte, R. R. R., de Jong, S., Lewis, C. M., Dima, D., & Powell, T. R. (2019). The polygenic nature of telomere length and the anti-ageing properties of lithium. *Neuropsychopharmacology, 44*(4), 757-765. doi:10.1038/s41386-018-0289-0

Dilliott, A. A., Evans, E. C., Farhan, S. M. K., Ghani, M., Sato, C., Zhang, M., . . . Hegele, R. A. (2019). Genetic Variation in the Ontario Neurodegenerative Disease Research Initiative. *Can J Neurol Sci, 46*(5), 491-498. doi:10.1017/cjn.2019.228

Dilliott, A. A., Ramirez, J., Binns, M. A., Breen, D. P., Evans, E. C., Beaton, D., . . . Ontario Neurodegenerative Disease Research Initiative, I. (2020). Parkinson's Disease, NOTCH3 Genetic Variants, and White Matter Hyperintensities. *Mov Disord, 35*(11), 2090-2095. doi:10.1002/mds.28171

Dilliott, A. A., Sunderland, K. M., McLaughlin, P. M., Roberts, A. C., Evans, E. C., Abrahao, A., . . . Hegele, R. A. (2021). Association of apolipoprotein E variation with cognitive impairment across multiple neurodegenerative diagnoses. *Neurobiol Aging*([In Press]). doi:10.1016/j.neurobiolaging.2021.04.011

Doran, M., du Plessis, D. G., Ghadiali, E. J., Mann, D. M., Pickering-Brown, S., & Larner, A. J. (2007). Familial early-onset dementia with tau intron 10 + 16 mutation with clinical features similar to those of Alzheimer disease. *Arch Neurol, 64*(10), 1535-1539. doi:10.1001/archneur.64.10.1535

Farhan, S. M. K., Bartha, R., Black, S. E., Corbett, D., Finger, E., Freedman, M., . . . Strong, M. J. (2017). The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci, 44*(2), 196-202. doi:10.1017/cjn.2016.415

Farhan, S. M. K., Dilliott, A. A., Ghani, G., Sato, C., Liang, E., Zhang, M., . . . Hegele, R. A. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*(16032), 1-11.

Farhan, S. M. K., Howrigan, D. P., Abbott, L. E., Klim, J. R., Topp, S. D., Byrnes, A. E., . . . Neale, B. M. (2019). Exome sequencing in amyotrophic lateral sclerosis implicates a novel gene, DNAJC7, encoding a heat-shock protein. *Nat Neurosci, 22*(12), 1966-1974. doi:10.1038/s41593-019-0530-0

Farrer, L. A., Cupples, L. A., Haines, J. L., Hyman, B., Kukull, W. A., Mayeux, R., . . . van Duijn, C. M. (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA, 278*(16), 1349-1356. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9343467

Ferrari, R., Wang, Y., Vandrovcova, J., Guelfi, S., Witeolar, A., Karch, C. M., . . . Desikan, R. S. (2017). Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer's and Parkinson's diseases. *J Neurol Neurosurg Psychiatry, 88*(2), 152-164. doi:10.1136/jnnp-2016-314411

Gan-Or, Z., Mohsin, N., Girard, S. L., Montplaisir, J. Y., Ambalavanan, A., Strong, S., . . . Rouleau, G. A. (2016). The role of the melanoma gene MC1R in Parkinson disease and REM sleep behavior disorder. *Neurobiol Aging, 43*, 180 e187-180 e113. doi:10.1016/j.neurobiolaging.2016.03.029

Heffernan, A. L., Chidgey, C., Peng, P., Masters, C. L., & Roberts, B. R. (2016). The Neurobiology and Age-Related Prevalence of the epsilon4 Allele of Apolipoprotein E in Alzheimer's Disease Cohorts. *J Mol Neurosci, 60*(3), 316-324. doi:10.1007/s12031-016-0804-x

Ho, P. W., Leung, C. T., Liu, H., Pang, S. Y., Lam, C. S., Xian, J., . . . Ho, S. L. (2020). Age-dependent accumulation of oligomeric SNCA/alpha-synuclein from impaired degradation in mutant LRRK2 knockin mouse model of Parkinson disease: role for therapeutic activation of chaperone-mediated autophagy (CMA). *Autophagy, 16*(2), 347-370. doi:10.1080/15548627.2019.1603545

Hou, Y., Dan, X., Babbar, M., Wei, Y., Hasselbalch, S. G., Croteau, D. L., & Bohr, V. A. (2019). Ageing as a risk factor for neurodegenerative disease. *Nat Rev Neurol, 15*(10), 565-581. doi:10.1038/s41582-019-0244-7

Huang, X., Chen, P. C., & Poole, C. (2004). APOE-[epsilon]2 allele associated with higher prevalence of sporadic Parkinson disease. *Neurology, 62*(12), 2198-2202. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/15210882

Hulisz, D. (2018). Amyotrophic lateral sclerosis: disease state overview. *Am J Manag Care, 24*(15 Suppl), S320-S326. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30207670

Kalia, L. V., & Lang, A. E. (2015). Parkinson's disease. *Lancet, 386*(9996), 896-912. doi:10.1016/S0140-6736(14)61393-3

Kandel, B. M., Avants, B. B., Gee, J. C., McMillan, C. T., Erus, G., Doshi, J., . . . Wolk, D. A. (2016). White matter hyperintensities are more highly associated with preclinical Alzheimer's disease than imaging and cognitive markers of neurodegeneration. *Alzheimers Dement (Amst), 4*, 18-27. doi:10.1016/j.dadm.2016.03.001

Karch, C. M., Wen, N., Fan, C. C., Yokoyama, J. S., Kouri, N., Ross, O. A., . . . International Parkinson's Disease Genomics, C. (2018). Selective Genetic Overlap Between Amyotrophic Lateral Sclerosis and Diseases of the Frontotemporal Dementia Spectrum. *JAMA Neurol, 75*(7), 860-875. doi:10.1001/jamaneurol.2018.0372

Katsuno, M., Sahashi, K., Iguchi, Y., & Hashizume, A. (2018). Preclinical progression of neurodegenerative diseases. *Nagoya J Med Sci, 80*(3), 289-298. doi:10.18999/nagjms.80.3.289

Keller, M. F., Saad, M., Bras, J., Bettella, F., Nicolaou, N., Simon-Sanchez, J., . . . Wellcome Trust Case Control, C. (2012). Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Hum Mol Genet, 21*(22), 4996-5009. doi:10.1093/hmg/dds335

Khalil, B., & Lievens, J. C. (2017). Mitochondrial quality control in amyotrophic lateral sclerosis: towards a common pathway? *Neural Regen Res, 12*(7), 1052-1061. doi:10.4103/1673-5374.211179

Kraft, P., Zeggini, E., & Ioannidis, J. P. (2009). Replication in genome-wide association studies. *Stat Sci, 24*(4), 561-573. doi:10.1214/09-STS290

Kummer, B. R., Diaz, I., Wu, X., Aaroe, A. E., Chen, M. L., Iadecola, C., . . . Navi, B. B. (2019). Associations between cerebrovascular risk factors and parkinson disease. *Ann Neurol, 86*(4), 572-581. doi:10.1002/ana.25564

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., . . . Environmental Risk for Alzheimer's Disease, C. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet, 51*(3), 414-430. doi:10.1038/s41588-019-0358-2

Larsen, S. B., Hanss, Z., & Kruger, R. (2018). The genetic architecture of mitochondrial dysfunction in Parkinson's disease. *Cell Tissue Res, 373*(1), 21-37. doi:10.1007/s00441-017-2768-8

Lendahl, U., Nilsson, P., & Betsholtz, C. (2019). Emerging links between cerebrovascular and neurodegenerative diseases-a special role for pericytes. *EMBO Rep, 20*(11), e48070. doi:10.15252/embr.201948070

Li, H., Wetten, S., Li, L., St Jean, P. L., Upmanyu, R., Surh, L., . . . Roses, A. D. (2008). Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol, 65*(1), 45-53. doi:10.1001/archneurol.2007.3

Liu, H. F., Ho, P. W., Leung, G. C., Lam, C. S., Pang, S. Y., Li, L., . . . Ho, S. L. (2017). Combined LRRK2 mutation, aging and chronic low dose oral rotenone as a model of Parkinson's disease. *Sci Rep, 7*, 40887. doi:10.1038/srep40887

Longo, F., Mercatelli, D., Novello, S., Arcuri, L., Brugnoli, A., Vincenzi, F., . . . Morari, M. (2017). Age-dependent dopamine transporter dysfunction and Serine129 phospho-alpha-synuclein overload in G2019S LRRK2 mice. *Acta Neuropathol Commun, 5*(1), 22. doi:10.1186/s40478-017-0426-8

Lorenzo-Betancor, O., Wszolek, Z. K., & Ross, O. A. (2016). Rare variants in MC1R/TUBB3 exon 1 are not associated with Parkinson's disease. *Ann Neurol, 79*(2), 331. doi:10.1002/ana.24581

Maraganore, D. M., de Andrade, M., Elbaz, A., Farrer, M. J., Ioannidis, J. P., Kruger, R., . . . Genetic Epidemiology of Parkinson's Disease, C. (2006). Collaborative analysis of alpha-synuclein gene promoter variability and Parkinson disease. *JAMA, 296*(6), 661-670. doi:10.1001/jama.296.6.661

Mata, I. F., Samii, A., Schneer, S. H., Roberts, J. W., Griffith, A., Leis, B. C., . . . Zabetian, C. P. (2008). Glucocerebrosidase gene mutations: a risk factor for Lewy body disorders. *Arch Neurol, 65*(3), 379-382. doi:10.1001/archneurol.2007.68

Mejzini, R., Flynn, L. L., Pitout, I. L., Fletcher, S., Wilton, S. D., & Akkari, P. A. (2019). ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Front Neurosci, 13*, 1310. doi:10.3389/fnins.2019.01310

Myers, A. J., Kaleem, M., Marlowe, L., Pittman, A. M., Lees, A. J., Fung, H. C., . . . Hardy, J. (2005). The H1c haplotype at the MAPT locus is associated with Alzheimer's disease. *Hum Mol Genet, 14*(16), 2399-2404. doi:10.1093/hmg/ddi241

Nalls, M. A., Bras, J., Hernandez, D. G., Keller, M. F., Majounie, E., Renton, A. E., . . . Singleton, A. B. (2015). NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiol Aging, 36*(3), 1605 e1607-1612. doi:10.1016/j.neurobiolaging.2014.07.028

Nalls, M. A., Duran, R., Lopez, G., Kurzawa-Akanbi, M., McKeith, I. G., Chinnery, P. F., . . . Sidransky, E. (2013). A multicenter study of glucocerebrosidase mutations in dementia with Lewy bodies. *JAMA Neurol, 70*(6), 727-735. doi:10.1001/jamaneurol.2013.1925

Olanow, C. W., Kieburtz, K., & Schapira, A. H. (2008). Why have we failed to achieve neuroprotection in Parkinson's disease? *Ann Neurol, 64 Suppl 2*, S101-110. doi:10.1002/ana.21461

Oxford, A. E., Stewart, E. S., & Rohn, T. T. (2020). Clinical Trials in Alzheimer's Disease: A Hurdle in the Path of Remedy. *Int J Alzheimers Dis, 2020*, 5380346. doi:10.1155/2020/5380346

Park, J. S., Davis, R. L., & Sue, C. M. (2018). Mitochondrial Dysfunction in Parkinson's Disease: New Mechanistic Insights and Therapeutic Perspectives. *Curr Neurol Neurosci Rep, 18*(5), 21. doi:10.1007/s11910-018-0829-3

Petrov, D., Mansfield, C., Moussy, A., & Hermine, O. (2017). ALS Clinical Trials Review: 20 Years of Failure. Are We Any Closer to Registering a New Treatment? *Front Aging Neurosci, 9*, 68. doi:10.3389/fnagi.2017.00068

Ragno, M., Berbellini, A., Cacchio, G., Manca, A., Di Marzio, F., Pianese, L., . . . De Michele, G. (2013). Parkinsonism is a late, not rare, feature of CADASIL: a study on Italian patients carrying the R1006C mutation. *Stroke, 44*(4), 1147-1149. doi:10.1161/STROKEAHA.111.000458

Ragno, M., Sanguigni, S., Manca, A., Pianese, L., Paci, C., Berbellini, A., . . . De Michele, G. (2016). Parkinsonism in a pair of monozygotic CADASIL twins sharing the R1006C mutation: a transcranial sonography study. *Neurol Sci, 37*(6), 875-881. doi:10.1007/s10072-016-2497-x

Richardson, T. G., Shihab, H. A., Rivas, M. A., McCarthy, M. I., Campbell, C., Timpson, N. J., & Gaunt, T. R. (2016). A Protein Domain and Family Based Approach to Rare Variant Association Analysis. *PLoS One, 11*(4), e0153803. doi:10.1371/journal.pone.0153803

Schneider, S. A., & Alcalay, R. N. (2020). Precision medicine in Parkinson's disease: emerging treatments for genetic Parkinson's disease. *J Neurol, 267*(3), 860-869. doi:10.1007/s00415-020-09705-7

Selvackadunco, S., Langford, K., Shah, Z., Hurley, S., Bodi, I., King, A., . . . Al-Sarraj, S. (2019). Comparison of clinical and neuropathological diagnoses of neurodegenerative diseases in two centres from the Brains for Dementia Research (BDR) cohort. *J Neural Transm (Vienna), 126*(3), 327-337. doi:10.1007/s00702-018-01967-w

Seto-Salvia, N., Clarimon, J., Pagonabarraga, J., Pascual-Sedano, B., Campolongo, A., Combarros, O., . . . Kulisevsky, J. (2011). Dementia risk in Parkinson disease: disentangling the role of MAPT haplotypes. *Arch Neurol, 68*(3), 359-364. doi:10.1001/archneurol.2011.17

Shen, X. N., Li, J. Q., Wang, H. F., Li, H. Q., Huang, Y. Y., Yang, Y. X., . . . Alzheimer's Disease Neuroimaging, I. (2020). Plasma amyloid, tau, and neurodegeneration biomarker profiles predict Alzheimer's disease pathology and clinical progression in older adults without dementia. *Alzheimers Dement (Amst), 12*(1), e12104. doi:10.1002/dad2.12104

Sidransky, E., Nalls, M. A., Aasly, J. O., Aharon-Peretz, J., Annesi, G., Barbosa, E. R., . . . Ziegler, S. G. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N Engl J Med, 361*(17), 1651-1661. doi:10.1056/NEJMoa0901281

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell, 177*(1), 26-31. doi:10.1016/j.cell.2019.02.048

Sunderland, K. M., Beaton, D., Arnott, S. R., Kleinstiver, P. W., Kwan, D., Lawrence-Dewar, J. M., . . . Binns, M. A. (2020). The Ontario Neurodegenerative Disease Research Initiative. *medRxiv*, 2020.2007.2030.20165456. doi:10.1101/2020.07.30.20165456

Tell-Marti, G., Puig-Butille, J. A., Potrony, M., Badenas, C., Mila, M., Malvehy, J., . . . Puig, S. (2015). The MC1R melanoma risk variant p.R160W is associated with Parkinson disease. *Ann Neurol, 77*(5), 889-894. doi:10.1002/ana.24373

Thal, D. R., von Arnim, C. A., Griffin, W. S., Mrak, R. E., Walker, L., Attems, J., & Arzberger, T. (2015). Frontotemporal lobar degeneration FTLD-tau: preclinical lesions, vascular, and Alzheimer-related co-pathologies. *J Neural Transm (Vienna), 122*(7), 1007-1018. doi:10.1007/s00702-014-1360-6

Umoh, M. E., Fournier, C., Li, Y., Polak, M., Shaw, L., Landers, J. E., . . . Glass, J. D. (2016). Comparative analysis of C9orf72 and sporadic disease in an ALS clinic population. *Neurology, 87*(10), 1024-1030. doi:10.1212/WNL.0000000000003067

Wallings, R. L., Humble, S. W., Ward, M. E., & Wade-Martins, R. (2019). Lysosomal Dysfunction at the Centre of Parkinson's Disease and Frontotemporal Dementia/Amyotrophic Lateral Sclerosis. *Trends Neurosci, 42*(12), 899-912. doi:10.1016/j.tins.2019.10.002

Wong, C. H., Siah, K. W., & Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics, 20*(2), 273-286. doi:10.1093/biostatistics/kxx069

Woollacott, I. O. C., Bocchetta, M., Sudre, C. H., Ridha, B. H., Strand, C., Courtney, R., . . . Rohrer, J. D. (2018). Pathological correlates of white matter hyperintensities in a case of progranulin mutation associated frontotemporal dementia. *Neurocase, 24*(3), 166-174. doi:10.1080/13554794.2018.1506039

Yang, A., Kantor, B., & Chiba-Falek, O. (2021). APOE: The New Frontier in the Development of a Therapeutic Target towards Precision Medicine in Late-Onset Alzheimer's. *Int J Mol Sci, 22*(3). doi:10.3390/ijms22031244

Appendices

Appendix A. Copyright Permissions

Dilliott A.A., Evans E.C., Farhan S.M.K., Ghani M., Sato C., Zhang M., McIntyre A.D., Cao H., Racacho L., Robinson J.F., Strong M.J., Masellis M., Bulman D.E., Rogaeva E., Black S., Finger E., Frank A., Freedman M., Hassan A., Lang A., Shoesmith C., Swartz R., Tang-Wai D., Tartaglia C., Turnbull J., Zinman L., the ONDRI Investigators, and Hegele R.A. (2019). Genetic variation in the Ontario Neurodegenerative Disease Research Initiative. *Can J Neurol Sci,* 46(5):491-498

Dilliott A.A., Sunderland K.M., McLaughlin P.M., Roberts A.C., Evans E.C., Abrahao A., Binns M.A., Black S.E., Borrie M., Casaubon L.K., Dowlatshahi D., Finger E., Fischer C.E., Frank A., Freedman M., Grimes D., Hassan A., Jog M., Kumar S., Kwan D., Lang A.E., Mandzia J., Marras C., Masellis M., McIntyre A.D., Pasternak S., Pollock D.G., Rajji T.K., Robinson J.F., Rogaeva E., Sahlas D.J., Saposnik G., Sato C., Seitz D., Shoesmith C., Steeves T., Strother S.C., Swartz R.H., Tan B., Tang-Wai D., Tartaglia M.C., Troyer A.K., Turnbull J., Zinman L., ONDRI Investigators, and Hegele R.A. (2020). Association of apolipoprotein E variation with cognitive impairment across multiple neurodegenerative diagnoses. *Neurobiol Aging*, [In Press]. doi: https://doi.org/10.1016/j.neurobiolaging.2021.04.011
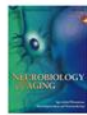
Association of apolipoprotein E variation with cognitive impairment across multiple neurodegenerative diagnoses

Author:
Allison A. Dilliott,Kelly M. Sunderland,Paula M. McLaughlin,Angela C. Roberts,Emily C. Evans,Agessandro Abrahao,Malcolm A. Binns,Sandra E. Black,Michael Borrie,Leanne.K. Casaubon,Dar Dowlatshahi,Elizabeth Finger,Corinne.E. Fischer,Andrew Frank et al.

Publication: Neurobiology of Aging

Publisher: Elsevier

Date: Available online 24 April 2021

Journal Author Rights

BACK   CLOSE WINDOW

Dilliott A.A.*, Ramirez J.*, Binns M., Breen D., Evans E., Beaton D., McLaughlin P.M., Kwan D., Holmes M.F., Ozzoude M., Scott C.J.M., Strother S.C., Symons S., Swartz R.H., Grimes D., Jog M., Masellis M., Black S.E., Joutel A., Marras C., Rogaeva E., Hegele R.A., and Lang A.E., on behalf of the ONDRI Investigators (2020). Parkinson's disease, *NOTCH3* genetic variants, and white matter hyperintensities. *Mov Disord*, 35(11):2090-2095

BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library
http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Farhan, S.M.K., **Dilliott, A.A.**, Ghani, M., Sato, C., Liang, E., Zhang, M., McIntyre, A.D., Cao, H., Racacho, L., Robinson, J.F., ONDRI Investigators, Strong, M.J., Masellis, M., St George-Hyslop, P., Bulman, D.E., Rogaeva, E., Hegele, R. A. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*, 1:16032

Dilliott A.A., Farhan S.M.K., Ghani M., Sato C., Liang E., Zhang M., McIntyre A.D., Cao H., Racacho L., Robinson J.F., Strong M.J., Masellis M., Bulman D.E., Rogaeva E., Lang A., Tartaglia C., Finger E., Zinman L., Turnbull J., Freedman M., Swartz R., Black S., the ONDRI Investigators, and Hegele R.A. (2018). Targeted next-generation sequencing and bioinformatics pipeline to evaluate genetic determinants of constitutional disease. *J Vis Exp*, (134): e57266

jove
JOURNAL OF VISUALIZED EXPERIMENTS

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article: Targeted next-generation sequencing and bioinformatics pipeline to evaluate genetic determinants of constitutional disease.

Author(s): Allison A Dilliott, Sali M K Farhan, Mahdi Ghani, Christine Sato, Eric Liang, Ming Zhang, Adam D McIntyre, Hanbin Cao, Lemuel Racacho, John F Robinson, Michael J Strong, Mario Masellis, Dennis E Bulman, Ekaterina

Item 1 (check one box): The Author elects to have the Materials be made available (as described at

http://www.jove.com/author) via: ☐ Standard Access ■ Open Access

Item 2 (check one box):

■ The Author is NOT a United States government employee.

☐ The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

☐ The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. Defined Terms. As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of *The Journal of Visualized Experiments;* "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. Background. The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. Grant of Rights in Article. In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

1

612542.6

ARTICLE AND VIDEO LICENSE AGREEMENT

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

4.   Retention of Rights in Article.   Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5.   Grant of Rights in Video – Standard Access. This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video.   To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6.   Grant of Rights in Video – Open Access. This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this Section 6 is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7.   Government Employees. If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8.   Likeness, Privacy, Personality. The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

9.   Author Warranties. The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

10.   JoVE Discretion. If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE.   JoVE and its employees, agents and independent contractors shall have

2

jove
JOURNAL OF VISUALIZED EXPERIMENTS

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

11. Indemnification. The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in *JoVE* or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's

expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

12. Fees. To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

13. Transfer, Governing Law. This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement required per submission.

CORRESPONDING AUTHOR:

Name: Robert A Hegele

Department: Schulich School of Medicine and Dentistry

Institution: Robarts Research Institute

Article Title: Targeted next-generation sequencing and bioinformatics pipeline to evaluate genetic determinants of constitutional disease.

Signature: ▆▆▆▆

Date: 13/12/2017

Please submit a signed and dated copy of this license by one of the following three methods:
1) Upload a scanned copy of the document as a pfd on the JoVE submission site;
2) Fax the document to +1.866.381.2236;
3) Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02139

For questions, please email submissions@jove.com or call +1.617.945.9051

612542.6

jove

Hi Allison,

Thank you for publishing your article Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease with JoVE.

You have permission to reuse the following material from it in your thesis or dissertation, pursuant to your Author License Agreement:

Figure(s):
1, 2, 3

Table(s):
1

Please ensure that JoVE is properly cited in the legends as well as the References: "This is adapted from Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., McIntyre, A. D., Cao, H., Racacho, L., Robinson, J. F., Strong, M. J., Masellis, M., Bulman, D. E., Rogaeva, E., Lang, A., Tartaglia, C., Finger, E., Zinman, L., Turnbull, J., Freedman, M., Swartz, R., Black, S. E., Hegele, R. A. Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J. Vis. Exp.* (134), e57266, doi:10.3791/57266 (2018)."

Best regards,

Review
JoVE
617.674.1888
Follow us: Facebook | Twitter | LinkedIn
About JoVE

Appendix B. University of Western Ontario – Ethics Approval

Western Research

Research Ethics

Western University Health Science Research Ethics Board
HSREB Amendment Approval Notice

Principal Investigator: Dr. Michael Strong
Department & Institution: Schulich School of Medicine and Dentistry\Pathology, Western University

Review Type: Full Board
HSREB File Number: 104915
Study Title: The Ontario Neurodegenerative Disease Research Initiative
Sponsor: Ontario Brain Institute

HSREB Amendment Approval Date: November 09, 2016
HSREB Expiry Date: March 04, 2017

Documents Approved and/or Received for Information:

Document Name	Comments	Version Date
Revised Western University Protocol	Amendment 17	2016/10/13
Revised Letter of Information & Consent	Main	2016/10/13
Letter of Information	Brain Autopsy - Version 3	2016/10/13
Wallet Card	Version 3	2016/10/13
Other	After Death Instructions - Version 4	2016/10/13
Other	Informed Verbal Re-Consent - Version 1	2016/09/21
Instruments	ONDRI Newsletter for participants - Fall 2016	2016/08/19
Other	ONDRI Research Protocol - Version 17	2016/10/13
Sponsor Protocol	Protocol Amendment Summary of Changes - Version 17	2016/10/13

The Western University Health Science Research Ethics Board (HSREB) has reviewed and approved the amendment to the above named study, as of the HSREB Initial Approval Date noted above.

HSREB approval for this study remains valid until the HSREB Expiry Date noted above, conditional to timely submission and acceptance of HSREB Continuing Ethics Review.

The Western University HSREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use Guideline for Good Clinical Practice Practices (ICH E6 R1), the Ontario Personal Health Information Protection Act (PHIPA, 2004), Part 4 of the Natural Health Product Regulations, Health Canada Medical Device Regulations and Part C, Division 5, of the Food and Drug Regulations of Health Canada.

Members of the HSREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB.

The HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Ethics Officer: Erika Basile ___ Katelyn Harris ___ Nicole Kaniki ___ Grace Kelly ___ Vikki Tran ___ Karen Gopaul

Western University, Research, Support Services Bldg., Rm. 5150
London, ON, Canada N6G 1G9 t. 519.661.3036 f. 519.850.2466 www.uwo.ca/research/ethics

Appendix C. The Ontario Neurodegenerative Disease Research Initiative (ONDRI) Investigators

ONDRI Investigator	ONDRI Affiliation
Sandra Black	AD/MCI Recruitment (Lead); FTD Recruitment
Morris Freedman	AD/MCI Recruitment (Lead); FTD Recruitment
Michael Borrie	AD/MCI Recruitment
Corinne Fischer	AD/MCI Recruitment
Sanjeev Kumar	AD/MCI Recruitment
Stephen Pasternak	AD/MCI Recruitment
Bruce Pollock	AD/MCI Recruitment
Tarek Rajji	AD/MCI Recruitment
Dallas Seitz	AD/MCI Recruitment
Andrew Frank	AD/MCI Recruitment; FTD Recruitment
David Tang-Wai	AD/MCI Recruitment; FTD Recruitment
Lorne Zinman	ALS Recruitment (Lead)
Agessandro Abrahao	ALS Recruitment
Marvin Chum	ALS Recruitment
Christen Shoesmith	ALS Recruitment
John Turnbull	ALS Recruitment
Donna Kwan	Clinical (Lead)
Brian Tan	Clinical (Co-Lead)
Jane Lawrence Dewar	Clinical
Richard H. Swartz	CVD Recruitment (Lead)
Leanne Casaubon	CVD Recruitment
Dar Dowlatshahi	CVD Recruitment
Ayman Hassan	CVD Recruitment
Jennifer Mandzia	CVD Recruitment
Demetrios Sahlas	CVD Recruitment
Gustavo Saposnik	CVD Recruitment
Susan Bronskill	Executive Member
Doug Munoz	Eye Tracking (Lead)
Don Brien	Eye Tracking
Brian Coe	Eye Tracking
Ying Chen	Eye Tracking Scholar
Elizabeth Finger	FTD Recruitment (Lead)
Carmela Tartaglia	FTD Recruitment (Lead); AD/MCI Recruitment
Bill McIlroy	Gait and Balance (Lead)
Manuel Montero-Odasso	Gait and Balance (Lead)
Karen Van Ooteghem	Gait and Balance
Alanna Black	Gait and Balance
Ben Cornish	Gait and Balance

Yanina Sarquis Adamson	Gait and Balance
Frederico Faria	Gait and Balance Scholar
Julia Fraser	Gait and Balance Scholar
Rob Hegele	Genomics (Lead)
Dennis Bulman	Genomics
Sali Farhan	Genomics
Mahdi Ghani	Genomics
John Robinson	Genomics
Ekaterina Rogaeva	Genomics
Allison Dilliott	Genomics Scholar
Rob Bartha	Neuroimaging (Lead)
Sean Symons	Neuroimaging (Lead)
Sabrina Adamo	Neuroimaging
Courtney Berezuk	Neuroimaging
Melissa Holmes	Neuroimaging
Nuwan Nanayakkara	Neuroimaging
Miracle Ozzoude	Neuroimaging
Christopher Scott	Neuroimaging
Hassan Haddad	Neuroimaging Scholar
Joel Ramirez	Neuroimaging Scholar
Stephen Strother	Neuroinformatics (Lead)
Malcolm Binns	Neuroinformatics (Co-Lead)
Stephen Arnott	Neuroinformatics
Wendy Lou	Neuroinformatics
Kelly Sunderland	Neuroinformatics
Sujeevini Sujanthan	Neuroinformatics
Athena Theyers	Neuroinformatics
Abiramy Uthirakumaran	Neuroinformatics
Mojdeh Zamyadi	Neuroinformatics
Guangyong (GY) Zou	Neuroinformatics
Derek Beaton	Neuroinformatics Scholar
Pradeep Raamana	Neuroinformatics Scholar
Paula McLaughlin	Neuropyschology (Lead)
Brian Levine	Neuropsychology
JB Orange	Neuropsychology
Alicia Peltsch	Neuropsychology
Angela Roberts	Neuropsychology
Angela Troyer	Neuropsychology
Michael Strong	ONDRI 1 Lead
David Munoz	Pathology and Biomarkers (Lead)
Roger A. Dixon	Pathology and Biomarkers

John Woulfe	Pathology and Biomarkers
Tony Lang	PD Recruitment (Lead)
Connie Marras	PD Recruitment (Lead)
David Grimes	PD Recruitment
Mandar Jog	PD Recruitment
Tom Steeves	PD Recruitment
Mario Masellis	PD Recruitment; FTD Recruitment
David Breen	PD Scholar
Peter Kleinstiver	Program Manager ONDRI 1
Natalie Rashkovan	Program Manager ONDRI 1 & 2
Barry Greenberg	SD-OCT (Lead)
Chris Hudson	SD-OCT (Lead)
Sherif Defrawy	SD-OCT
Wendy Hatch	SD-OCT
Efrem Mandelcorn	SD-OCT
Ed Margolin	SD-OCT
Faryan Tayyari	SD-OCT
Elena Leontieva	SD-OCT Scholar

Abbreviations: AD/MCI, Alzheimer's disease/mild cognitive impairment; ALS, amyotrophic lateral sclerosis; CVD, cerebrovascular disease; FTD, frontotemporal dementia; ONDRI, Ontario Neurodegenerative Disease Research Initiative; PD, Parkinson's disease; SD-OCT, spectral domain optical coherence tomography.

Appendix D. Genes included on the ONDRISeq targeted next-generation sequencing panel.

Gene	Chromosomal location	Affected protein	RefSeq Transcript	Associated phenotype	Mode of inheritance
Amyotrophic lateral sclerosis/frontotemporal dementia					
ALS2	2q33.1	Alsin	NM_020919	ALS2	AR (HZ), juvenile onset
ANG	14q11.2	Angiogenin	NM_001145	ALS9	ADm, late onset
ARHGEF28	5q13.2	Rho guanine nucleotide exchange factor 28	NM_001080479	ALS and FTD	AR (HZ) and ADm, late onset
ATXN2	12q24.12	Ataxin 2	NM_002973	ALS13	ADm, late onset
CENPV	17p11.2	Centromere protein V	NM_181716	ALS	Genetic association, late onset
CHMP2B	3p11.2	CHMP family member 2B	NM_014043	ALS17, FTD	ADm, late onset
DAO	12q24.11	D-amino acid oxidase	NM_001917	ALS, schizophrenia	ADm, late onset
DCTN1	2p13.1	Dynactin 1	NM_004082	ALS, HMN7B, Perry syndrome	ADm, late onset
FIG4	6q21	FIG4 homolog, SAC1 lipid phosphatase domain containing	NM_014845	ALS11, CMT disease, YV syndrome	ADm, late onset; AR (HZ and CH), infantile onset; AR (HZ and CH), infantile onset
FUS	16p11.2	Fused in sarcoma	NM_004960	ALS6, FTD, HET4	AR (HZ), ADm, late onset
GRN	17q21.31	Granulin precursor	NM_002087	FTD, NCL	ADm, late onset; AR (HZ), juvenile onset
HNRNPA1	12q13.13	Heterogeneous nuclear ribonucleoprotein A1	NM_002136	ALS20, inclusion body myopathy with early-onset Paget disease with/without FTD 3	ADm, late onset; ADm, early onset

HNRNPA2B1	7p15.2	Heterogeneous nuclear ribonucleoprotein A2/B1	NM_031243	Inclusion body myopathy with early-onset Paget disease with/without FTD 2	ADm, early onset
MAPT/STH	17q21.31	Microtubule-associated protein tau	NM_001123066	ALS, FTD with parkinsonism, PD, AD, Pick disease, supranuclear palsy, tauopathy	ADm, late and early onset
NEFH	22q12.2	Neurofilament protein, heavy polypeptide	NM_021076	ALS1	ADm, late onset
OPTN	10p13	Optineurin	NM_001008211	ALS12, glaucoma	AR (HZ) and AD, early onset
PFN1	17p13.2	Profilin 1	NM_005022	ALS18	ADm, earlier onset
PNPLA6	19p13.2	Patatin-like phospholipase domain-containing protein 6	NM_001166111	Spastic paraplegia, Boucher-Neuhauser syndrome	AR (HZ and CH), early onset
PRPH	12q13.12	Peripherin	NM_006262	ALS1	ADm, late onset
SETX	9q34.13	Senataxin	NM_015046	ALS4, spinocerebellar ataxia 1	ADm and AR, juvenile onset
SIGMAR1	9p13.3	Sigma nonopioid intracellular receptor 1	NM_001282208	ALS16, FTD	AR (HZ); ADm, early onset
SOD1	21q22.11	Superoxide dismutase 1	NM_000454	ALS1	AR (HZ and CH), ADm, age of onset varies from 6–94 years old
SQSTM1	5q35.3	Sequestosome 1	NM_001142298	Paget disease of bone	ADm, late onset
TAF15	17q12	TAF15 RNA polymerase II, TATA box binding protein associated factor	NM_139215	Chondrosarcoma	
TARDBP	1p36.22	Tar DNA-binding protein	NM_007375	ALS10, FTD	ADm, late onset
UBQLN2	Xp11.21	Ubiquilin 2	NM_013444	ALS15, FTD	X-linked, juvenile and late onset

UNC13A	19p13.11	Unc-13 homolog A (*C. elegans*)	NM_001080421	ALS	Genetic association, late onset
VAPB	20q13.33	Vesicle-associated membrane protein (VAMP)-associated protein B and C	NM_004738	ALS, spinal muscular atrophy (Finkel type)	ADm, early and late onset
VCP	9p13.3	Valosin-containing protein	NM_007126	ALS14, FTD, inclusion body myopathy with early-onset Paget disease with/without; FTD1	ADm, early onset
Alzheimer's disease/mild cognitive impairment					
ABCA7	19p13.3	ATP-binding cassette, subfamily a, member 7	NM_019112	AD	Genetic association, late onset
APOE	19q13.32	Apolipoprotein E	NM_001302688	AD2, lipoprotein glomerulopathy, sea-blue histiocyte disease, macular degeneration	ACD, ADm, AR (HZ and CH), late onset
APP	21q21.3	Amyloid-β A4 precursor protein	NM_000484	AD 1, cerebral amyloid angiopathy	ADm and AR (HZ), early and late onset
BIN1	2q14.3	Bridging integrator 1	NM_001320642	AD	Genetic association, late onset
CD2AP	6p12.3	CD2-associated protein	NM_012120	AD	Genetic association, late onset
CD33	19q13.41	CD33 antigen	NM_001772	AD	Genetic association, late onset
CLU	8p21.1	Clusterin	NM_001831	AD	Genetic association, late onset
CR1	1q32.2	Complement component; receptor 1	NM_000651	AD	Genetic association, late onset

CSF1R	5q32	Colony-stimulating factor 1 receptor	NM_001349736	HDLS with dementia	ADm, early and late onset
DNMT1	19p13.2	DNA methyltransferase 1	NM_001130823	HSN1E with dementia	ADm, early onset dementia
ITM2B	13q14.2	Integral membrane protein 2B	NM_021999	Dementia	ADm, early and late onset
MS4A4E	11q12.2	Membrane-spanning 4-domains, subfamily A, member 4E	NM_011545418	AD	Genetic association, late onset
MS4A6A	11q12.2	Membrane-spanning 4-domains, subfamily A, member 6A	NM_152852	AD	Genetic association, late onset
PICALM	11q14.2	Phosphatidylinositol-binding clathrin assembly protein	NM_007166	AD	Genetic association, late onset
PLD3	19q13.2	Phospholipase D family, member 3	NM_012268	AD19	Genetic association, late onset
PSEN1	14q24.2	Presenilin 1	NM_000021	AD3, dilated cardiomyopathy, FTD, Pick disease, acne inversa	ADm, early onset
PRNP	20p13	Prion protein	NM_000311	Dementia	ADm, early onset
PSEN2	1q32.13	Presenilin 2	NM_000447	AD4, dilated cardiomyopathy	ADm, early onset
SORL1	11q24.1	Sortilin-related receptor	NM_003105	AD	ADm, combined gene burden, late onset
TREM2	6p21.1	Triggering receptor expressed on myeloid cells 2	NM_018965	AD Nasu-Hakola disease (dementia and psychotic symptoms)	Genetic association, late onset
TYROBP	19q13.12	Tyro protein tyrosine kinase-binding protein	NM_003332	Nasu-Hakola disease (dementia and psychotic symptoms)	AR (HZ), juvenile onset

Parkinson's disease

ADH1C	4q23	Alcohol dehydrogenase 1C, gamma polypeptide	NM_000669	PD, alcohol dependence protection	Genetic association, late onset
ATP13A2 (PARK9)	1p36.13	ATPase, type 13A2	NM_022089	PD, ceroid lipofuscinosis, dementia	Genetic association, early onset and late onset
DNAJC13	3q22.1	DNAJ/HSP40 homolog, subfamily C, member 13	NM_001329126	PD	ADm, late onset
EIF4G1	3q27.1	Eukaryotic translation initiation factor 4-gamma	NM_182917	PD18	ADm, late onset
FBXO7	22q12.3	F-box only protein 7	NM_012179	PD15	AR (HZ and CH), early onset
GAK	4p16.3	Cyclin G-associated kinase	NM_005255	PD	Genetic association, late onset
GCH1	14q22.2	GTP cyclohydrolase I	NM_000161	PD, dystonia	Genetic association, early onset
GIGYF2	2q37.1	GRB10-interacting GYP protein 2	NM_015575	PD11	Genetic association, early and late onset
HTRA2	2p13.1	HTRA serine peptidase 2	NM_013247	PD13	ADm and genetic association, early and late onset
LRRK2	12q12	Leucine-rich repeat kinase 2	NM_198578	PD8	ADm and genetic association, early and late onset
MC1R	16q24.3	Melanocortin 1 receptor	NM_002386	PD; melanoma, UV induced skin damage	Genetic association, late onset
NR4A2	2q24.1	Nuclear receptor subfamily 4, group A, member 2	NM_006186	PD	Genetic association, late onset
PANK2	20p13	Pantothenate kinase 2	NM_001324191	Neurodegeneration	AR (HZ and CH), early onset

PARK2 (PRKN)	6q26	Parkin	NM_004562	PD2	AR (HZ and CH), juvenile onset; heterozygotes have late onset
PARK7(DJ1)	1p36.23	Oncogene DJ1	NM_007262	PD7	AR (HZ and CH), early onset
PARL	3q27.1	Presenilin-associated rhomboid-like protein	NM_018622	PD (based on biological mechanisms, no linkage confirmed)	NA
PINK1	1p36.12	Pten-induced putative kinase 1	NM_032409	PD6	AR (HZ and CH), ADm, early onset
PLA2G6	22q13.1	Phospholipase A2, group VI	NM_001349867	PD14, NBIA2A, NBIA2B	AR (HZ and CH), early and late onset
PM20D1	1q32	Peptidase M20 domain, containing 1	NM_152491	PD16	Genetic association, late onset
RAB7L1	1q32.1	RAB7-like 1	NM_001135663	PD	Genetic association, late onset
SNCA	4q22.1	α-synuclein	NM_000345	PD1, PD4, LBD	ADm, early onset
UCHL1	4p13	Ubiquitin carboxyl-terminal esterase L1	NM_004181	PD5, neurodegeneration with optic atrophy	ADm, AR (HZ), juvenile-onset
VPS35	16q11.2	Vacuolar protein sorting 35	NM_018206	PD17	ADm, early and late onset
Vascular cognitive impairment					
ABCC6	16p13.11	ATP-binding cassette, subfamily C, member 6	NM_001171	Arterial calcification; pseudoxanthoma elasticum; pseudoxanthoma elasticum forme fruste	AR (HZ), infantile onset; AR; ADm
COL4A1	13q34	Collagen type IV, alpha-1	NM_001845	Angiopathy, brain small vessel disease, porencephaly 1,	ADm, infantile onset

				intracerebral haemorrhage susceptibility	
COL4A2	13q34	Collagen type IV, alpha-2	NM_001846	Porencephaly 2, intracerebral haemorrhage susceptibility	ADm, infantile onset
HTRA1	10q26.13	HTRA serine peptidase 1	NM_002775	CARASIL syndrome, macular degeneration	AR (HZ), early onset
NOTCH3	19p13.12	Notch homology protein 3	NM_000435	Infantile myofibromatosis 2, CADASIL	ADm, early onset
SAMHD1	20q11.23	SAM domain and HD domain 1	NM_015474	Aicardi-Goutieres syndrome 5, Chilblain lupus 2	AR (HZ and CH), AD, infantile onset
TREX1	3p21.31	3-prime repair exonuclease 1	NM_016381	Aicardi-Goutieres syndrome 1, Chilblain lupus, Vasculopathy, retinal, with cerebral leukodystrophy	AD, AR (HZ and CH), juvenile onset

Age of onset was classified as 'late onset' if greater than 65 years of age. Figure adapted from Farhan, S., Dilliott, A., Ghani, M. et al. (2016) The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *npj Genomic Med* 1, 16032. https://doi.org/10.1038/npjgenmed.2016.32

Abbreviations: ACD, autosomal co-dominant; AD, Alzheimer's disease; ADm, autosomal dominant; ALS, amyotrophic lateral sclerosis; AR, autosomal recessive; CARASIL syndrome, cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy; CADASIL, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy; CH, compound heterozygous; CMT disease, Charcot-Marie-Tooth disease; FTD, frontotemporal dementia; HDLS, leukoencephalopathy, diffuse hereditary, with spheroids; HET4, hereditary essential tremor, 4; HMN7B, neuropathy, distal hereditary motor, type VIIB; HSN1E, hereditary sensory neuropathy type 1E; HZ, homozygous; LBD, Lewy body dementia; NCL, neuronal ceroid-lipofuscinoses; NBIA2A, neurodegeneration with brain iron accumulation 2A; NBIA2B, neurodegeneration with brain iron accumulation 2B; NGS, Next-generation sequencing; PD, Parkinson's disease; YV syndrome, Yunis–Varon syndrome.

Appendix E. Regression coefficients (standard error) of the multinomial logistic regressions used for rare variant association analyses in <u>Chapter 4</u>.

Cohort	AD/MCI Associated Genes	ALS/FTD Associated Genes	CVD Associated Genes	PD Associated Genes
Putative loss-of-function variants				
ONDRI	0.423 (0.534)	0.725 (0.576)	-0.194 (0.822)	1.991 (0.924)*
AD	-0.560 (1.601)	0.646 (1.014)	1.007 (1.233)	2.510 (1.107)*
ALS	3.242 (1.805)	3.502 (1.747)*	4.058 (2.272)	4.850 (1.947)*
CVD	0.904 (1.039)	1.156 (1.039)	0.680 (1.499)	2.370 (1.313)
FTD	2.415 (1.591)	2.713 (1.607)	2.790 (2.049)	3.948 (1.833)*
MCI	0.626 (0.575)	1.066 (0.529)*	-1.125 (1.862)	1.836 (0.950)
PD	2.039 (1.555)	2.420 (1.531)	2.196 (1.910)	3.428 (1.832)
Missense variants				
ONDRI	0.047 (0.206)	-0.183 (0.205)	-0.116 (0.217)	-0.051 (0.208)
AD	0.240 (0.384)	-0.239 (0.386)	-0.355 (0.425)	-0.350 (0.384)
ALS	0.291 (1.113)	-0.048 (1.115)	0.441 (1.125)	-0.383 (1.119)
CVD	-0.088 (0.442)	-0.384 (0.440)	-0.248 (0.477)	0.047 (0.443)
FTD	0.065 (0.986)	0.045 (0.987)	0.258 (1.013)	-0.109 (1.001)
MCI	0.107 (0.228)	-0.056 (0.228)	-0.028 (0.237)	-0.065 (0.230)
PD	0.242 (0.873)	-0.339 (0.876)	0.229 (0.909)	-0.083 (0.885)
Possibly deleterious missense variants				
ONDRI	-0.209 (0.235)	-0.056 (0.216)	-0.141 (0.236)	0.123 (0.210)
AD	0.146 (0.420)	0.275 (0.391)	-0.709 (0.512)	-0.526 (0.439)
ALS	1.010 (1.134)	0.345 (1.145)	0.615 (1.158)	0.324 (1.134)
CVD	-0.222 (0.513)	-0.306 (0.480)	-0.189 (0.523)	0.302 (0.444)
FTD	0.478 (1.024)	0.299 (1.009)	0.263 (1.069)	0.670 (1.003)
MCI	-0.109 (0.262)	0.133 (0.237)	-0.137 (0.261)	0.381 (0.230)
PD	0.171 (0.943)	0.043 (0.905)	0.489 (0.934)	0.178 (0.907)

Multinomial logistic regressions adjusted for age, sex, and disease prevalence were performed to analyze enrichment of rare variants identified in the 80 genes encompassed by the ONDRISeq panel. The brglm2 R package was used to fit the regression model and apply a mean bias reduction accounting for the low variant positive counts. *p-value < 0.05.

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CVD, cerebrovascular disease; FTD, frontotemporal dementia; MCI mild cognitive impairment; ONDRI, Ontario Neurodegenerative Disease Research Initiative; PD, Parkinson's disease.

Appendix F. Rare variant association analysis comparing the enrichment of rare missense variants in four disease-associated gene groupings in the ONDRI cohorts compared to cognitively normal controls presented in <u>Chapter 4</u>.
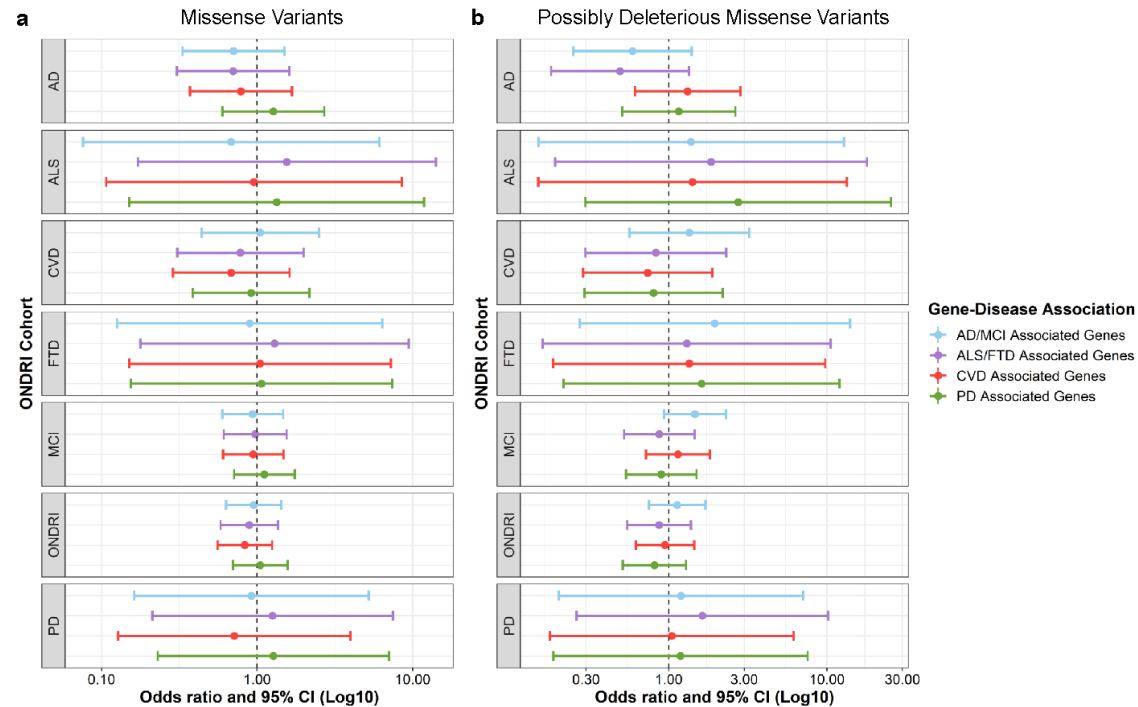
Multinomial logistic regressions adjusted for age, sex, and disease prevalence were performed to analyze enrichment of **(a)** missense variants, and **(b)** possibly deleterious missense variants identified in the 80 genes encompassed by the ONDRISeq panel, which were binned into four disease-associated gene groupings: AD associated genes, ALS/FTD associated genes, CVD associated genes, and PD associated genes, across the ONDRI cohorts compared to the control cohort. Only ancestry matched participants were included in the analyses. The brglm2 R package was used to fit the regression model and apply a mean bias reduction accounting for the low variant positive counts. Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CVD, cerebrovascular disease; FTD, frontotemporal dementia; MCI mild cognitive impairment; ONDRI, Ontario Neurodegenerative Disease Research Initiative; PD, Parkinson's disease.

Appendix G. Copy number variants (CNVs) detected by the VarSeq® v1.4.3 CNV Caller tool in 519 ONDRI participants in 80 neurodegenerative disease genes covered by the ONDRISeq panel in <u>Chapter 5</u>.

Subject	Phenotype	CNV State	Gene	Exon(s)	Span (bp)	Z-score	Ratio	p-value	Validation
1	AD	Het Deletion	*OPTN*	5	2357	-6.851	0.487	1.10E-12	Yes (BA)
2	ALS	Duplicate	*PARK7*	1–5	9810	6.816	1.455	0	Yes (WES)
3	CVD	Duplicate	*ABCC6*	1–31	74407	6.371	1.482	0	Yes (WES)
4	MCI	Duplicate	*SAMHD1*	7–11	8802	5.872	1.474	1.70E-21	Yes (WES)
5	CVD	Duplicate	*ABCA7*	16	276	9.961	3.221	3.14E-13	N/A
6	CVD	Duplicate	*ATP13A2*	16	707	7.175	1.430	2.02E-08	N/A
7	CVD	Duplicate	*ABCA7*	4–5	611	4.156	1.357	3.51E-07	No (WES)
7	CVD	Het Deletion	*LRRK2*	9	170	-3.990	0.636	4.09E-05	N/A[a]
7	CVD	Duplicate	*PRPH*	7–8	453	3.991	1.338	6.34E-05	No (WES)
8	AD	Duplicate	*ABCA7*	15–16	630	4.454	1.845	4.63E-07	N/A
9	CVD	Het Deletion	*TREM2*	4	156	-3.753	0.678	7.30E-07	N/A
10	CVD	Duplicate	*COL4A2*	2	220	4.212	1.382	7.34E-07	N/A
11	PD	Duplicate	*UNC13A*	36	530	6.097	1.356	8.18E-07	N/A
12	CVD	Duplicate	*SQSTM1*	0	532	6.045	1.712	8.62E-07	N/A
13	PD	Duplicate	*ABCA7*	16–17	1627	4.163	1.748	9.65E-07	N/A
14	CVD	Duplicate	*CLU*	8	395	4.731	1.405	1.72E-06	N/A
15	ALS	Duplicate	*DNMT1*	35–36	678	6.623	1.319	2.14E-06	N/A
16	FTD	Het Deletion	*TREM2*	4	156	-3.980	0.653	3.77E-06	N/A
17	CVD	Duplicate	*ATP13A2*	28	700	5.397	1.493	4.39E-06	N/A
18	CVD	Duplicate	*CENPV*	1	972	5.323	1.441	5.14E-06	N/A
18	CVD	Duplicate	*PSEN2*	1	578	4.559	1.350	9.32E-05	N/A
19	CVD	Duplicate	*NOTCH3*	3–4	641	3.901	1.351	5.44E-06	N/A
20	PD	Duplicate	*GCH1*	5	621	6.382	1.322	8.90E-06	N/A[b]
21	PD	Duplicate	*UCHL1*	0–1	150	4.799	2.454	9.01E-06	N/A
22	PD	Duplicate	*ALS2*	34	2127	4.658	1.380	1.07E-05	N/A

23	MCI	Duplicate	TAF15	15	210	4.235	1.513	1.26E-05	N/A
24	CVD	Duplicate	NOTCH3	4–6	944	3.946	1.304	1.33E-05	No (WES)
25	MCI	Het Deletion	GAK	20–23	623	-5.521	0.567	1.35E-05	N/A
26	CVD	Duplicate	ABCC6	3	265	4.652	1.786	1.40E-05	N/A
27	CVD	Duplicate	PNPLA6	1–3	375	3.722	1.596	1.48E-05	N/A
28	CVD	Duplicate	CD2AP	1	960	3.853	1.407	1.85E-05	N/A
29	AD	Duplicate	UNC13A	36	530	5.228	1.304	2.72E-05	N/A
30	ALS	Duplicate	COL4A2	7	3715	3.817	1.314	2.87E-05	N/A
31	CVD	Duplicate	SOD1	1-2	4765	3.541	2.019	3.13E-05	N/A
32	CVD	Duplicate	UNC13A	35	2216	4.415	1.374	3.21E-05	N/A
33	CVD	Het Deletion	NEFH	1	571	-5.899	0.564	3.22E-05	N/A
34	CVD	Duplicate	NOTCH3	21–24	1756	3.884	1.322	3.29E-05	N/A
35	ALS	Duplicate	UCHL1	7–8	2361	3.981	1.301	3.55E-05	N/A
36	CVD	Duplicate	NOTCH3	3–4	641	3.543	1.323	3.63E-05	N/A
37	MCI	Duplicate	TARDBP	1	622	5.506	1.426	4.24E-05	N/A
38	AD	Duplicate	ABCA7	13–16	1925	3.568	1.388	5.11E-05	N/A
38	CVD	Het Deletion	SIGMAR1	1	351	-3.757	0.538	5.76E-05	N/A
40	CVD	Duplicate	ATP12A2	1	700	5.030	1.460	6.29E-05	N/A
41	MCI	Duplicate	PNPLA6	2–5	232	5.132	1.421	7.97E-05	N/A
42	PD	Het Deletion	UNC13A	8	506	-4.214	0.598	8.88E-05	N/A
43	MCI	Het Deletion	ABCC6	3–4	629	-3.744	0.516	9.89E-05	N/A
44	PD	Het Deletion	ABCA7	8	222	-4.021	0.451	9.94E-05	N/A

For multi-exon CNVs, the reported ratio and Z-score values are averaged across each affected region. A response of "No" in respect to validation indicated that the WES did not identify the CNV that had been identified using the ONDRISeq panel. [a]The WES performed did not have probes adequately covering exon 9 of LRRK2. [b]WES exhibited unmappable and incorrectly mapped reads, failing to pass the quality control standards of the CNV Caller tool algorithm. Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; BA, breakpoint analysis; bp, base pairs; CNV, copy number variant; CVD, cerebrovascular disease; FTD, frontotemporal dementia; Het, heterozygous; MCI, mild cognitive impairment; N/A, not applicable; PD, Parkinson's disease; WES, whole exome sequencing.

Appendix H. Polymerase chain reaction (PCR) and Sanger sequencing primers used to capture the sequence of *GBA* in <u>Chapter 6</u>.

Target Exon(s)	PCR Primer Direction	Sequencing Primer Direction	Primer Sequence
1	FWD	FWD	GCCTCACTTCCTGTGTCATG
	-	REV	TTCCCACCCATTTCAACTTC
2	-	FWD	GTGGGCCTTGTCCTAATGAA
	-	REV	CAAAGGACTATGAGGCAGAA
3–4	-	FWD	CAAGGGGTGAGGAATTTTGA
	REV	REV	ACGAAAAGTTTCAATGGCTCT
5–6	FWD	FWD	GCAAGTGATAAGCAGAGTCC
	-	REV	CTAGGTTGAGGGTTGGGACA
7	-	FWD	AGGCTGTTCTCGAACTCCTG
	REV	REV	AGTTTGGGAGCCAGTCATTT
8	FWD	FWD	TGTGTGCAAGGTCCAGGATCAG
	-	REV	CTGGACAGGAAGGGCTTCTG
9	-	FWD	GCAACTCTGGGGAACCA
	-	REV	CATTGGGGTTTTCTGTTGCT
10–11	-	FWD	CAGGAGTTATGGGGTGGGTC
	REV	REV	TTCTAGGGGCCTCCAGCC

The primers listed were used for PCR amplifications, yielding three amplicons that contained: 1) exons 1–4; 2) exons 5–7; and 3) exons 8–11; the most highly specific of previously published primers were used to provide specificity for isolating *GBA* (Mata *et al*. *Mov Disord*, 2016). All 11 exons of GBA were Sanger sequenced, including 100 base pairs at each intron-exon boundary.

Abbreviations: FWD, forward; *GBA*, glucosylceramidase beta gene; PCR, polymerase chain reaction; REV, reverse.

Appendix I. *GBA* **exons unable to be sequenced in ONDRI participants stratified by disease cohort presented in <u>Chapter 6</u>.**

Region unable to be sequenced	Number of ONDRI participants					
	AD (n=41)	ALS (n=39)	CVD ± CI (n=161)	FTD (n=51)	MCI (n=84)	PD (n=139)
Exons 1–4; 9–11	0	0	0	0	0	1
Exon 2	0	1	1	0	0	0
Exons 2–4; 8–11	0	0	0	0	0	1
Exons 3–4	0	0	4	1	2	0
Exon 4	0	0	0	0	1	0

Abbreviations: AD, Alzheimer's disease; ALS, amyotrophic lateral sclerosis; CVD ± CI, cerebrovascular disease with or without cognitive impairment; FTD, frontotemporal dementia; *GBA*, glucosylceramidase beta gene; MCI, mild cognitive impairment; ONDRI, Ontario Neurodegenerative Disease Research Initiative; PD, Parkinson's disease.

Appendix J. Targeted next-generation sequencing and bioinformatics pipeline to evaluate genetic determinants of constitutional disease.

The work presented in Appendix J. has been used with permission from the *Journal of Visualized Experiments*. The video component of this article has accumulated over 20,000 views and can be found at: https://www.jove.com/v/57266/targeted-next-generation-sequencing-bioinformatics-pipeline-to

Dilliott A.A., Farhan S.M.K., Ghani M., Sato C., Liang E., Zhang M., McIntyre A.D., Cao H., Racacho L., Robinson J.F., Strong M.J., Masellis M., Bulman D.E., Rogaeva E., Lang A., Tartaglia C., Finger E., Zinman L., Turnbull J., Freedman M., Swartz R., Black S., the ONDRI Investigators, and Hegele R.A. (2018). Targeted next-generation sequencing and bioinformatics pipeline to evaluate genetic determinants of constitutional disease. J Vis Exp, (134): e57266.

Abstract

Next-generation sequencing (NGS) is quickly revolutionizing how research into the genetic determinants of constitutional disease is performed. The technique is highly efficient with millions of sequencing reads being produced in a short time span and at relatively low cost. Specifically, targeted NGS is able to focus investigations to genomic regions of particular interest based on the disease of study. Not only does this further reduce costs and increase the speed of the process, but it lessens the computational burden that often accompanies NGS. Although targeted NGS is restricted to certain regions of the genome, preventing identification of potential novel loci of interest, it can be an excellent technique when faced with a phenotypically and genetically heterogeneous disease, for which there are previously known genetic associations. Because of the complex nature of the sequencing technique, it is important to closely adhere to protocols and methodologies in order to achieve sequencing reads of high coverage and quality. Further, once sequencing reads are obtained, a sophisticated bioinformatics workflow is utilized to accurately map reads to a reference genome, to call variants, and to ensure the variants pass quality metrics. Variants must also be annotated and curated based on their clinical significance, which can be standardized by applying the American College of Medical Genetics and Genomics Pathogenicity Guidelines. The methods presented herein will display the steps involved in generating and analyzing NGS data from a targeted sequencing panel, using the ONDRISeq neurodegenerative disease panel as a model, to identify variants that may be of clinical significance.

Introduction

As defining the genetic determinants of various conditions takes on a higher priority in research and in the clinic, next-generation sequencing (NGS) is proving to be a high-throughput and cost-effective tool to achieve these goals[1,2,3]. For almost 40 years, Sanger sequencing had been the gold standard for identifying genetic variants[4]; however, for diseases with genetic heterogeneity or unknown genetic etiology, many possible candidate genes must be evaluated, often concurrently. In this context, Sanger sequencing becomes expensive and time-consuming. However, NGS involves massive parallel sequencing of millions of DNA fragments, allowing for a cost and time efficient technique to simultaneously detect a wide range of genetic variation across various regions of the genome.

There are three types of NGS for sequencing DNA: 1) whole-genome sequencing (WGS), 2) whole-exome sequencing (WES), and 3) targeted sequencing[5]. WGS evaluates the entire genomic content of an individual, while WES involves sequencing only the protein-coding regions of the genome[6]. Targeted sequencing, in contrast, focuses on specific regions of the genome based on relatively few specific genes linked by common pathological mechanisms or known clinical phenotype. Either the exons or introns, or any intergenic regions of a gene or specific group of genes can be specified using this approach. Therefore, targeted sequencing can be an excellent approach when there is already a foundation of candidate genes known to be associated with the disease of interest. Targeting specific regions of the genome allows for elimination of superfluous and irrelevant genetic variation that can cloud or distract from clinical interpretation. While WGS and WES both produce a large amount of high-quality data, the amount of data can be overwhelming. Not only does this large amount of data require computationally intensive bioinformatics analysis, but data storage can frequently present problems[7]. This challenge of data storage also adds additional costs to both WGS and WES, which is often not initially considered when calculating the expense of sequencing. Further, although it is decreasing, the cost of WGS and WES remain relatively high. Targeted sequencing can be a more cost-efficient option, particularly when sequencing of a large number of individuals is required.

The Ontario Neurodegenerative Disease Research Initiative (ONDRI) is a multi-platform, provincial-wide, observational cohort study characterizing five neurodegenerative diseases, including: 1) Alzheimer's disease and mild cognitive impairment, 2) amyotrophic lateral sclerosis, 3) frontotemporal dementia, 4) Parkinson's disease, and 5) vascular cognitive impairment[8]. The ONDRI genomics subgroup is aiming to elucidate as part of the baseline characterization of this cohort the often discounted, yet extremely important genetic landscape of these phenotypically and genetically heterogeneous diseases. Neurodegenerative diseases are thus appropriate candidates for NGS methodologies and for targeted sequencing in particular.

We have custom-designed a targeted NGS panel, ONDRISeq, to sequence 528 participants involved in ONDRI for the protein-coding regions of 80 genes that have been previously associated with the five diseases of interest. With this methodology, we are able to harness the high-quality NGS data in a focused and efficient manner. The design and validation of the ONDRISeq panel with multiple concordance studies has been previously described, for which the ONDRISeq panel was able to identify novel, rare variants of possible clinical significance in 72.2% of 216 cases used for panel validation[9]. Although NGS technology has advanced rapidly and remarkably in recent years, many researchers face a challenge when processing the raw data into a list of usable, annotated variants[10]. Further, interpretation of the variants can be complex, especially when faced with many that are rare or novel[11].

Here, we describe in a step-by-step manner, the methodology of targeted NGS and the associated bioinformatics workflow required for resequencing, variant calling, and variant annotation using the ONDRISeq study as an example. After the generation of NGS data, raw sequencing files must be aligned to the human reference genome in order to accurately call variants. Variants must then be annotated in order to perform subsequent variant curation. We will also explain our implementation of the American College of Medical Genetics' Standards and Guidelines to accurately classify variant pathogenicity.

Protocol

For the purposes of ONDRI, ethics protocols and informed consent were obtained based on the Research Ethic Boards at Baycrest Centre for Geriatric Care (Toronto, Ontario, Canada); Centre for Addiction and Mental Health (Toronto, Ontario, Canada); Elizabeth Bruyère Hospital (Ottawa, Ontario, Canada); Hamilton General Hospital (Hamilton, Ontario, Canada); London Health Sciences Centre (London, Ontario, Canada); McMaster (Hamilton, Ontario, Canada); The Ottawa Hospital (Ottawa, Ontario, Canada); Parkwood Hospital (London, Ontario, Canada); St Michael's Hospital (Toronto, Ontario, Canada); Sunnybrook Health Sciences Centre (Toronto, Ontario, Canada); and University Health Network-Toronto Western Hospital (Toronto, Ontario, Canada).

1. DNA Isolation from Human Blood Samples

    1.1. **Collect samples from sequencing participants in accordance with appropriate ethics protocols and informed consent.**

    1.1.1. To obtain DNA of high quality, draw blood samples for the purposes of extraction.

    NOTE: DNA can also be extracted from saliva or buccal cells, ensuring that an appropriate DNA extraction kit is used.

    1.1.2. If extracting from blood, to obtain a high yield of DNA, collect the sample in three 4 mL EDTA K2 tubes, providing a sample of total volume ~12 ml.

    1.1.3. Centrifuge blood samples for 20 min at 750 x g to fraction into an upper phase of plasma, thin, middle phase of leukocytes, and a bottom phase of erythrocytes.

    1.2. Remove the plasma from the blood sample by pipetting it off the top of the sample with a disposable transfer pipette. Appropriately discard the plasma or dispense into multiple 500 µL aliquots for storage at -80 °C for future biochemical analyses. Ensure that a new, sterile pipette is used for each sample.

    1.3. Extract DNA from the blood sample with a blood extraction kit[12] (**Table of Materials**) according to manufacturer's instructions.

    NOTE: If a sample of the volume described above is obtained, ~3 mL of leukocytes will be obtained to use in the DNA extraction.

1.4. Measure initial DNA concentration in ng/µL using a full-spectrum spectrophotometer[13] (**Table of Materials**), according to manufacturer's instructions.

1.5. Proceed directly to step 2. Alternatively, store DNA at 4 °C.

2. Sequencing Library Preparation

2.1. **Perform serial dilutions on the DNA samples over the course of three days to obtain a final concentration of 5.0 ± 1.0 ng/µL.**

2.1.1. Dilute 1 M Tris buffer pH 8.5 to 10 µM with deionized water.
NOTE: The volume diluted will depend on the number of DNA samples that will need to be diluted in the subsequent steps.

2.1.2. If performing the DNA dilution directly after step 1.4, proceed to the following step. If not on the same day, measure the DNA concentration as was done in step 1.4.

2.1.3. Based on the concentration measured, dilute 40 µL of the DNA to ~10 ng/µL using 10 µM Tris buffer pH 8.5 and allow the sample to sit overnight at 4 °C.

2.1.4. Measure DNA concentration with a fluorometer[14] appropriate for the quantification of DNA (**Table of Materials**), according to manufacturer's instructions.
NOTE: The concentration of the sample should be >10 ng/µL because of the lower sensitivity of the spectrophotometer used previously.

2.1.5. Based on the concentration measured, dilute 20 µL of the DNA to 10 ng/µL using 10 µM Tris buffer pH 8.5 and allow the sample to sit overnight at 4 °C.

2.1.6. Measure DNA concentration with the fluorometer[14], according to manufacturer's instructions.

2.1.7. Based on the concentration measured, dilute 10 µL of the DNA to 5 ng/µL using 10 µM Tris-HCl pH 8.5 and allow the sample to sit overnight at 4 °C.

2.2. **Prepare sequencing library according to manufacturer's instructions with the targeted NGS panel's appropriate target enrichment kit[15] (Table of**

Materials). Ensure that the enrichment kit is appropriate for the NGS platform being used.

    2.2.1. Follow manufacturer's instructions[16] regarding the plexity and pooling of libraries.

    NOTE: For ONDRISeq, libraries are composed of 12 DNA samples, pooled in sets of two, and run on the NGS desktop instrument (**Table of Materials**). The number of samples that can be run in a single reaction will depend on the sequencing kit and platform used.

    2.2.2. To achieve higher quality sequencing data, perform the optional step to validate the DNA library quality following tagmentation, described in manufacturer's instruction of the target enrichment kit[15].

        2.2.2.1.   Analyze each library in triplicate to ensure the quality of the library yield.

    2.2.3. If pooling libraries, measure DNA concentration with the fluorometer[14], according to manufacturer's instructions. Use this concentration to determine the volume of each DNA library to pool to obtain the equimolar ratios recommended by the target enrichment kit being used.

3. Next-generation Sequencing

   3.1. **Sequence the library according to the NGS desktop instrument's reagent kit manufacturer's instructions[17,18] (Table of Materials).**

    3.1.1. Prepare a sample sheet according to manufacturer's instructions[18] using the appropriate NGS technology software (**Table of Materials**), which will be imported into the NGS desktop instrument's workflow.

    NOTE: For the purposes of ONDRISeq, the application option chosen is 'other', with only the FASTQ files requested (**Figure J.1**). Subsequent steps will process these FASTQ files, to allow for full customization of alignment and quality parameters. However, if targeted sequencing is chosen, some NGS instruments are able to process the sequencing data into VCF files themselves. The manufacturer's instructions[18] may be consulted for a full selection of options.

3.1.2. If using a cloud-based computing environment[19] (**Table of Materials**), log in when setting up the sequencing run. Do this after clicking "Sequencing" on the NGS desktop instrument home page.

3.1.3. Following library denaturation[18] according to the manufacturer's instructions, measure DNA library concentration with the fluorometer[14].

3.1.4. Validate the DNA library quality using an appropriate automated electrophoresis system and DNA quality analysis kit[20] (**Table of Materials**), as per manufacturer's instructions.

3.1.5. To convert the DNA concentration from ng/µL to nM, use the following formula[16]

$$\frac{(concentration\ in\ ng/\mu L)}{(660\ g/mol * average\ library\ size)} * 10^6 = concentration\ in\ nM$$

NOTE: Average library size will be specific to target enrichment kit being used, and can be obtained from the electrophoresis trace observed in step 3.1.4.

3.1.6. Dilute the sequencing library to a final concentration of 6–20 pM, as appropriate, and volume of 600 μL, according to manufacturer's instructions[21].
NOTE: The exact concentration needed is dependent on the sequencing kit used. Consult the enrichment kit manufacturer to determine the proper loading concentration.

3.1.7. Dilute, denature, and include a positive control sequencing library[21], according to the manufacturer's instructions.

3.1.8. Keep a log of every sequencing run, which includes the DNA library concentration loaded (pM), the percentage of positive control added, reagent cartridge barcode, application chosen in step 3.1.1, number of index reads, enrichment kit used, read length(s), and the sample sheet name.
NOTE: The run time of the NGS desktop instrument will depend on the

instrument, enrichment kit, and read lengths chosen (4–56 h for the sequencer used in this experiment[22]).

3.2. Upon completion of the sequencing run, access the "Run Folder", which includes all outputs, by navigating to the NGS desktop instrument home page and clicking "Manage Files". Move the files to a local drive for later access. For a separate option, on a computer, find the files within the cloud-based computing environment[19] by selecting "Runs" on the navigation panel. Select the appropriate sequencing run to navigate to the Run Summary page. Select "Download" to obtain data from the cloud. From the dialog box that appears, select the FASTQ files as the file type to download and click "Download".

3.3. **From the Run Summary page of the cloud-based computing environment[19,23], navigate to "Charts" to analyze the quality of the sequencing run with the various figures produced by the computing environment. Refer to the manufacturer's instructions[23] for details regarding each figure produced.**

   3.3.1. From the Run Charts page, find the figure labeled "Data by Cycle". Under chart, select "Intensity" and under channel select "All Channels". Ensure that this signal intensity plot produced is similar to that produced by sequencing runs performed in the past with the same enrichment kit and NGS desktop instrument.

   NOTE: This reflects the percentage of intensity shown by each base across all 150 cycles. The figure can vary widely depending on the enrichment kit used, which is why it must be compared to past sequencing runs of the same panel.

   3.3.2. Select the "Indexing QC" tab within the run navigation panel to find the indexing quality control (QC) histogram, which is on the right-hand side of the page. Ensure that a relatively uniform distribution of % Reads Identified (PF) is observed across all samples.

   NOTE: If any samples have a much lower % Reads Identified (PF) than the rest of the samples, note that the quality of the sequencing data may be affected.

3.4. **From the Run Summary page of the cloud-based computing environment, navigate to the quality metrics by clicking "Metrics" within the run navigation panel.**

NOTE: Metrics cut-offs will depend on the sequencing platform and enrichment kit being used. There are many metrics that can be utilized based on manufacturer's instructions[23], with the following steps highlighting three that are highly recommended for quality control.

3.4.1. Under "DENSITY (K/MM^2)" ensure the cluster density is within the range recommended by the enrichment kit being used (in this case 1,200–1,400 K/mm^2).

3.4.2. Under the total "%≥Q30" ensure that the value is ≥85%, reflecting the quality of the sequencing reads.

NOTE: If lower than this threshold of 85%, note that the quality of the sequencing may be compromised.

3.4.3. Under "ALIGNED (%)" ensure that the value is similar to the % of positive control that was included in the sequencing run.

NOTE: This acts as a measure of positive control, such that only this percentage of total reads were found to align to the positive control genome. If 1% positive control was used it would be expected that the Aligned (%) would be ~1–5%.
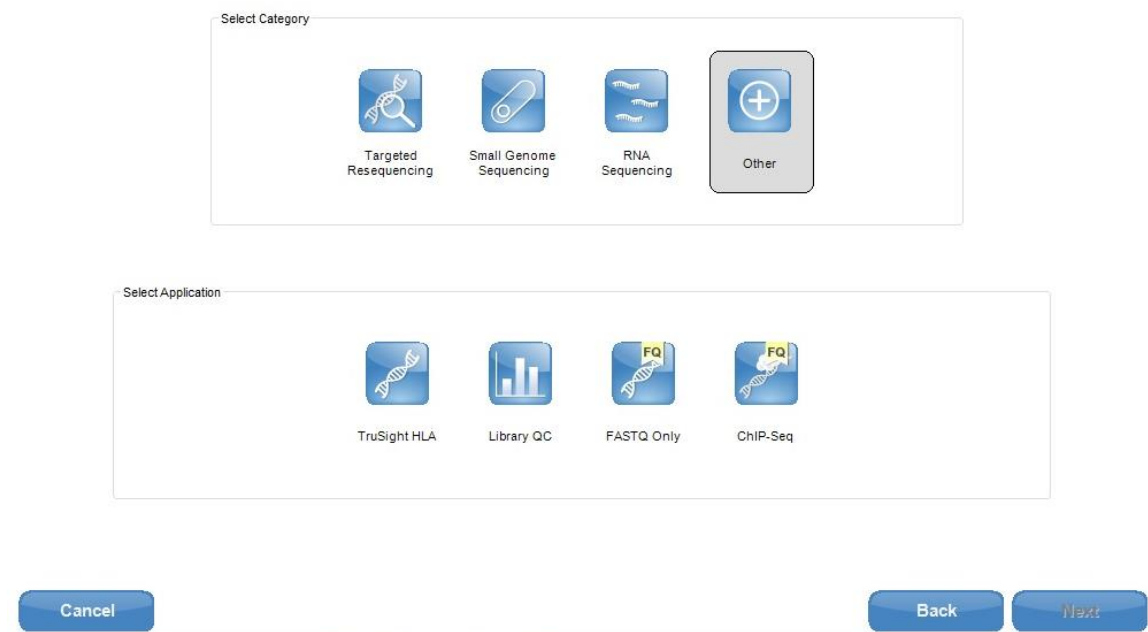
Figure J.1: Screenshot of the NGS technology software's (Table of Materials) sample sheet creator application options.
For the purposes of ONDRISeq, the FASTQ only application is used. However, if the user would like other files produced, such as VCF files, it is recommended that an application within the targeted resequencing category is used.

4. Resequencing and Variant Calling

    4.1. For data pre-processing, select appropriate software to align the raw FASTQ files to the human reference genome and to call variants (**Table of Materials**).

    4.2. **Import FASTQ sequencing reads into the data pre-processing software.**

        NOTE: For the purposes of ONDRISeq, the 48 FASTQ files produced from a single sequencing run of 24 samples are imported and processed through the software. The number of samples processed at once can vary depending on the needs of the researcher and size of the NGS panel.

        4.2.1. Within the "Navigation Area", right click and select "New Folder". Name the folder such that there is clarity as to the sequencing run that was performed.

        4.2.2. From the toolbar at the top, select "Import". From the dropdown list of sequencing platforms shown chose the platform with which the sequencing was performed.

            NOTE: For the purposes of ONDRISeq, "Illumina" is chosen. However, if using a different sequencing platform consult the manufacturer's instructions for the remainder of the FASTQ importing steps[24].

        4.2.3. In the dialog box, navigate to and select the FASTQ files from the sequencing run that is being processed. Ensure that the files being imported are stored in and imported from the local drive, if using a computer with multiple servers.

        4.2.4. From the "General options" of the dialog box, click the box beside "Paired reads" if sequencing used paired end chemistries.

            NOTE: In this case, there should also be two FASTQ samples imported for each sample - one forward and one reverse.

        4.2.5. From the Paired read information of the dialog box, select "Paired-end (forward-reverse)" if the forward read FASTQ file appears before the reverse read in the file list. If the files appear in the opposite order, select "Mate-pair (reverse-forward)". Set the paired read minimum distance to 1 and maximum distance to 1000, to allow for the detection of small scale structural rearrangements within the sample sequences.

4.2.6. From the "Illumina options" of the dialog box, select "Remove failed reads", to remove the reads that failed sequencing. If the NGS desktop instrument de-multiplexed the data before exporting the FASTQ files do not select the "MiSeq de-multiplexing" box.

4.2.7. From the "Quality score" dropdown list, select the NGS Pipeline that was utilized for sequencing. Select "Next" at the bottom of the dialog box. NOTE: The pipeline used will affect the format of the FASTQ file quality scores. For more information about which pipeline to select, consult the manufacturer's instructions[24].

4.2.8. From the new dialog box, select "Save" and "Create subfolders per bath unit to put each sample's FASTQ files into their own individual folder. Select "Next" at the bottom of the dialog box.

4.2.9. From the new dialog box, choose the folder that was created in step 4.2.1. This is where the FASTQ files will be imported. Select "Finish" at the bottom of the dialog box and wait until the FASTQ files are imported. Click the "Processes" tab to see the status of the file import.

4.3. **Design a workflow within the software to perform resequencing and variant calling, according to manufacturer's instructions.**

NOTE: This workflow can vary based on the needs of the researcher, but the following steps encompass what is included for the purposes of ONDRISeq (**Figure J.2**). The steps in this workflow can be applied to other NGS resequencing and variant calling software as appropriate. All bioinformatics processing for the purposes of ONDRI is performed in reference to human reference genome GRCH37/hg19, for consistency of data processing and analysis.

4.3.1. Map the sequencing reads to the reference genome.

4.3.1.1. When configuring, choose the reference genome as appropriate, ensuring that it is the same reference genome that is used for all bioinformatics steps.

4.3.1.2. From the masking mode drop-down list select "No masking" so that no regions of the reference sequence are masked.

4.3.1.3.   Use the default mapping options assigned by the software. Review the manufacturer's instructions[24] to verify that this is acceptable based on the purposes of the research.

4.3.2. Include in the workflow local realignment to the human reference genome to resolve any read mapping errors, particularly surrounding insertion-deletion variants.

4.3.2.1.   Use the default local realignment options assigned by the software. Review the manufacturer's instructions[24] to verify that this is acceptable based on the purposes of the research.

4.3.3. Remove duplicated mapped reads produced by PCR within the NGS protocol to reduce the effect of the PCR amplification bias, which may produce false positives[25].

4.3.3.1.   Set the "Maximum representation of minority sequence (%)", based on the needs of the research.

NOTE: A lenient setting, as used for the purposes of ONDRISeq, is 5%; however, the software's default setting is more stringent 20%. When two reads are very similar, this setting determines if the sequence with fewer read counts should be considered a sequencing error from the PCR amplification bias. Therefore, by setting 5%, the minority read count must be ≤ 5% of the majority read count to be corrected to be identical to the majority read.

4.3.4. Export statistics for the target regions in the form of a coverage summary text file from the read tracks generated in step 4.3.3. Ignore non-specific matches and broken pairs in the settings. Choose a destination on the local drive for these files.

4.3.5. Export a binary sequence alignment map (BAM) file for each sample from the read tracks generated in step 4.3.3. This contains sequence alignment data, if needed in future analyses. Choose a destination on the local drive for these files.

4.3.6. Choose a method of variant detection to call variants within the sequence.

NOTE: When assumptions can be made about the ploidy of the samples, it is

recommended that a fixed ploidy variant detection algorithm be used, as is used for the purposes of ONDRISeq. If this assumption cannot be made, refer to the manufacturer's instructions[24] to determine the best algorithm for the purposes of the research.

4.3.6.1.   When configuring, from the fixed ploidy variant parameters options set the ploidy as appropriate for the sample organism. Set the "required variant probability", or the probability that a variant has been correctly called in order for it to be retained, at 90.0%.

4.3.6.2.   Use the following recommended settings for the general filters: "Minimum coverage" of 10x, "Minimum count" of 2, "Minimum read frequency" of 20%, "Ignore broken pairs", ignore nonspecific matches based on "Reads", and "Minimum read length" of 20.
NOTE: These parameters are based on the purposes of ONDRISeq. Refer to the manufacturer's instructions[24] to ensure they are appropriate for the research being done.

4.3.6.3.   Use the following recommended settings for the noise filters: "Base quality filters" with a "Neighbourhood radius" mapping quality score of 5, "Minimum central quality" mapping score of 20, and "Minimum neighbourhood quality" mapping score of 15; a "Read direction filter" of 5.0%; and "Relative read direction filter" of 1.0% significance.
NOTE: These parameters are based on the purposes of ONDRISeq. Refer to the manufacturer's instructions[24] to ensure they are appropriate for the research being done.

4.3.7.   Filter the variants that have been called based on their overlap with the targeted panel's target regions as specified by the Browser Extensible Data (BED) file, allowing only variants occurring within the genomic regions selected for the targeted NGS panel to be retained.
NOTE: The BED file will be unique to the targeted NGS panel that is being utilized, based on the regions of the genome that the panel is able to cover.

4.3.8. Export a variant report in a variant calling format (VCF) file from the variant track produced in step 4.3.7. Choose a destination on the local drive for these files.

4.3.9. Save and install the workflow according to manufacturer's instructions[24], to make it available in the software's "Toolbox". Ensure the workflow is named such that it is clear in the future what NGS panel it is appropriate for.

4.3.9.1. In the dialog box with the "Exporting reference data" options during installation, set all options to "Bundle".

4.3.9.2. In the dialog box with the "Install location" options during installation, click "Install the workflow on your local computer".

4.4. **Run imported FASTQ sequencing read files through the customized bioinformatics workflow designed in step 4.3, according to manufacturer's instructions[24].**

4.4.1. Identify the workflow designed in step 4.3 in the software's "Toolbox" and double-click it.

4.4.2. Within the dialog box that appears, locate the folders of FASTQ files that were imported in step 4.2 within the "Navigation Area". Highlight all folders by selecting them within the "Navigation Area" and then click the box beside "Batch". Use the right-facing arrow to move the files to "Selected elements". Click "Next" at the bottom of the dialog box.

4.4.3. Within the dialog box, review the "Batch overview" to ensure the correct FASTQ files were selected and then click "Next".

4.4.4. Review the following steps of the workflow within the dialog box to ensure the correct files and export locations were selected when designing the workflow in step 4.3: "Map Reads to Reference"; Remove Duplicate Mapped Reads"; "Create Statistics for Target Regions"; "Export BAM"; "Export Tab delimited text"; "Filter Based on Overlap"; and "Export VCF"

4.4.5. Within the final step in the dialog box -"Result handling"- select the option "Save in input folder". Click "Finish" at the bottom of the dialog box. NOTE: This means that the files produced for each sample will be placed

into the same folder that stores the FASTQ file within the data pre-processing software.
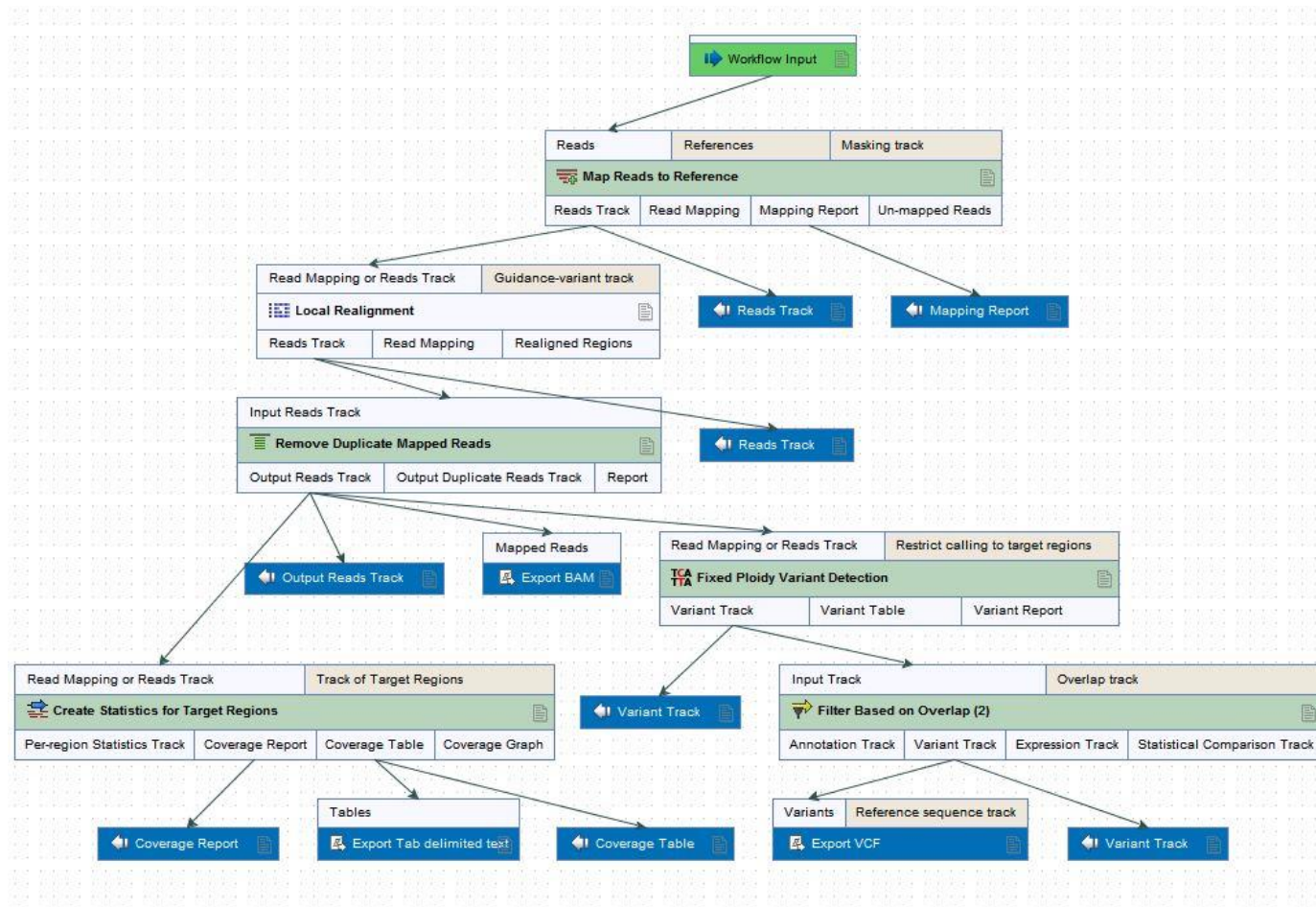
Figure J.2: Workflow for the resequencing and variant calling of FASTQ files within the data pre-processing software (Table of Materials) customized for the purposes of ONDRISeq.
The steps in the workflow can be applied to other NGS resequencing and variant calling software based on the needs of the researcher.

5. Variant Annotation

5.1. **Download and customize the Annotate Variation (ANNOVAR)[26] script to perform variant annotation upon the VCF file of each sample.**

    5.1.1. Download the following databases from ANNOVAR to be included as annotations: 1) RefSeq[27] (August 2015 update); 2) dbSNP138[28] (September 2014 update); 3) the Exome Aggregation Consortium[29] (ExAC, version 0.3 November 2015 update); 4) the National Heart, Lung, and Blood Institute Exome Sequencing Project European Cohort[30] (ESP, March 2015 update); 5) the 1000 Genomes Project European Cohort[31] (1KGP, August 2015 update); 6) ClinVar[32] (March 2016 update); and 7) Combined Annotation Dependent Depletion[33] (CADD), Sorting Intolerant from Tolerant[34] (SIFT), and PolyPhen-2[35].

    NOTE: Genome coordinates and all databases referenced by ANNOVAR referred to human genome build GRCh37/hg19. Additionally, the database versions listed are those used for the purposes of ONDRISeq, when downloading the databases use the most up to date versions available.

    5.1.2. If desired, customize ANNOVAR to output the complete list of annotated variants, as well as a reduced compilation of annotated variants using the --filter operation[26].

    NOTE: The reduced list can be customized based on the needs of the researcher. For the purposes of ONDRISeq, the reduced list of annotated variants does not include variants that occur further than 15 bases from the nearest exon or any variants with a minor allele frequency (MAF) >3% in any of the three databases: 1) ExAC; 2) ESP; and 3) 1KGP. This step is highly recommended.

    5.1.3. If desired, customize ANNOVAR to single out specific allele calls based on the needs of the researcher[26].

    NOTE: For the purposes of ONDRISeq, ANNOVAR assesses the sequencing calls made for the *APOE* risk alleles rs429358(C>T):p.C130R and rs7412(C>T):p.R176C in order to output the overall *APOE* genotype, of which there are six possible combinations, including: 1) E2/E2; 2) E3/E2; 3)

E4/E2; 4) E3/E3; 5) E4/E3; 6) E4/E4. Of these six possible *APOE* genotypes, E4/E4 is the most commonly accepted genetic risk factor for developing late-onset Alzheimer's disease[36].

5.2. **Query disease mutation databases (Table of Materials) to determine if variants have been previously associated with disease, with reasonable evidence. Consider any variants that have not been previously reported as a novel variant.**

   5.2.1. Assess the ANNOVAR annotations from ClinVar, such that the disease-associated variants include any classified as likely pathogenic or pathogenic.

5.3. Process splicing variants through the *in silico* prediction tools Splicing-based Analysis of Variants[37] (SPANR) and Human Splicing Finder[38] (HSF, version 3.0).

5.4. If processing a large number of samples, compare the variant calls within each sample to determine which variants are shared by various samples. Do this manually or with a custom-designed script, allowing for the detection of possible sequencing artifacts and contamination events.

   NOTE: For the purposes of ONDRI, a custom script is used to annotate the ANNOVAR output files by comparing them to one another. The script incorporates an annotation, per variant, with the subject ID of any other samples harboring the same variant, otherwise termed the variant's history in the study cohort.

5.5. Classify variants based on the American College of Medical Genetics (ACMG) Pathogenicity Guidelines[39], assigning each variant a classification as one of the following: 1) pathogenic; 2) likely pathogenic; 3) variant of uncertain significance; 4) likely benign; or 5) benign.

   NOTE: For the purposes of ONDRI, an in-house designed Python script is used to perform ACMG classification on a semi-automated basis. Although not used for this study, InterVar[40] is a similarly designed tool that can be utilized in an analogous manner.

5.6. Sanger sequence any variants with a sequencing coverage of <30x and/or variants that have been identified in > 10% of the study cohort to validate that they are not sequencing artifacts[41].

Representative Results

The methodologies described herein were applied to 528 participant DNA samples from individuals that have been enrolled in ONDRI. Samples were run on the ONDRISeq panel in 22 runs of 24 samples per run. Overall, sequencing data were determined to be of high quality with a mean sample coverage of $78 \pm 13x$ and all individual runs expressed a mean sample coverage >30x. Further, on average, 94% of all target regions were covered at least 20x (**Table J.1**).

A mean 95.6% of reads were mapped to the reference sequence and all ONDRISeq runs had >90% of reads mapped (**Table J.1**). Of the mapped reads, 92.0% had a Phred score ≥Q30, with only one run having <80% of mapped reads meeting this quality metric. However, this run still displayed a mean coverage of 79x and 93% of target regions were covered at least 20x.

Table J.1: Sequencing quality metrics for 22 runs on ONDRISeq.

Parameter	Mean (±sd)	Best performance	Poorest performance
Cluster Density (x10^3/mm^2)	1424 (±269)	1347	1835
Total Reads (10^6)	43.1 (±6.0)	48.7	47.4
Mapped Reads (10^6)	40.1 (±6.0)	47.1	25.7
Mapped Reads (%)	95.6 (±1.3)	96.8	92.6
Phred Quality Score ≥Q30 (%)	92.0 (±6.0)	92	68.3
Sample Coverage (x)	78 (±13)	99	51

Case Study: Identification of rare variants in a PD patient.

To demonstrate the utility of our targeted NGS workflow, we present the example of a 68 year-old, male, Parkinson's disease patient. The DNA sample was run on the NGS desktop instrument (**Table of Materials**) using the ONDRISeq panel alongside 23 other ONDRI samples. The run displayed a cluster density of $1,555 \times 10^3/mm^2$. The patient's particular sample displayed a mean coverage of 76x, with 93.9% of the target regions covered at least 20x.

After performing variant calling and annotation with the custom bioinformatics workflow, the patient was found to harbor 1351 variants within the exons and surrounding 250 bp of the 80 genes included on the ONDRISeq panel. However, the ANNOVAR pipeline was able to reduce the number of variants by considering variant sequence ontology and MAF, as described above. This produced a list of seven variants that underwent manual curation (**Figure J.3**). From these seven variants, two were identified as having possible clinical significance. This process is specific to the needs of ONDRI and was done by identifying those that are relatively rare in the general population and are nonsynonymous in ontology thereby causing a change in the protein. Whether the variant had been previously associated with disease, the *in silico* predictions of deleteriousness to the protein and the ACMG pathogenicity classification of the variants were also utilized in this process.

The first identified from the reduced list was a heterozygous variant, namely *LRRK2*:c.T3939A, resulting in the nonsense variant p.C1313*. *LRRK2* encodes the protein Leucine-Rich Repeat Kinase 2, which possesses both GTPase and kinase activity[42]. Further, mutations within this gene are known to be among the leading causes of familial Parkinson's disease[43]. This variant introduces a premature stop codon within *LRRK2*, thereby losing amino-acid residues 1,314–2,527. This prevents the translation of the protein's Ras of complex proteins (Roc), C-terminal of Roc (COR), and protein kinase domains, which are involved in functioning as an atypical Rho GTPase, GTP binding protein, and protein kinase, respectively, and was predicted to be damaging by the *in silico* analysis generated by CADD (CADD Phred = 36). This variant is also

rare with a MAF of 0.004% and 0.01% in ExAC and ESP, respectively, and is absent from the 1000G database. Additionally, this is the only patient out of all 528 sequenced who carries this variant, which is novel since it has not been previously described in disease mutation databases (**Table of Materials**). The confidence of the variant call was confirmed by its deep coverage of 109x. Finally, the variant was assessed with the AMCG Standards and Guidelines for pathogenicity and was classified as being pathogenic.

The patient also carried a second heterozygous variant, *NR4A2*:c.C755A, resulting in the missense change p.P252Q. The protein encoded by *NR4A2*, Nuclear Receptor Subfamily 4 Group A Member 2, is a transcription factor involved in the generation of dopaminergic neurons[44] and mutations within this gene have been previously associated with Parkinson's disease[45]. The substitution of the non-polar proline to the polar glutamine was predicted to be damaging by the *in silico* prediction analysis generated by CADD (CADD Phred = 21.1), but not by the analysis generated by SIFT or PolyPhen-2. The variant is rare, with a MAF of 0.004% in ExAC and absence from both ESP and 1000G. The variant was also identified in an ONDRI participant diagnosed with vascular cognitive impairment, but has not been previously described in disease mutation databases. This variant had coverage of only 18x, however, Sanger sequencing will be performed in order to ensure its validity within the sequence. Finally, the variant was determined to be of uncertain significance when assessed with the ACMG Standards and Guidelines for pathogenicity.

The ONDRISeq panel and bioinformatics pipeline is also able to determine the *APOE* genotype of each sample. This patient was determined to have the *APOE* genotype E3/E3.

Chromosome	Position	Reference	Alternate	RefSeq	dbsnp138	ExAC MAF	ESP MAF	1KGP MAF	CADD Phred	SIFT	PolyPhen-2	ClinVar
2	157185944	G	T	NR4A2:NM_006186:exon3:c.C755A:p.P252Q	rs148826679	3.82E-05	0	0	21.1	0.164	0.35	
4	100263940	C	A	ADH1C:NM_000669:exon6:c.828+12G>T	rs55982726	0.0048	0.0059	0.008	N/A	N/A	N/A	
12	40699748	T	A	LRRK2:NM_198578:exon28:c.T3939A:p.C1313X	rs183902574	4.43E-05	0.0001	0	36	.	N/A	
12	49689009	G	A	PRPH:NM_006262:exon1:c.G26A:p.R9Q	rs57451017	0.017	0.0115	0.0169	24.1	0.006	0.556	Previously identified, no information provided
20	3870023	G	A	PANK2:NM_153638:exon1:c.G276A:p.R92R	rs142832849	0.0049	0.0046	0.003	N/A	N/A	N/A	
22	29885567	A	C	NEFH:NM_021076:exon4:c.A1938C:p.A646A	rs75808076	3.36E-05	0	0	N/A	N/A	N/A	
22	29885594	A	T	NEFH:NM_021076:exon4:c.A1965T:p.P655P	rs79235463	9.0E-4	0	0	N/A	N/A	N/A	

Figure J.3: Example of a reduced output from ANNOVAR displaying manually curated, annotated variants.
The reduced ANNOVAR output from the case study of a 68-year-old, male, patient with Parkinson's disease. Annotated variants are curated to identify those that are most likely to be of clinical significance, as denoted by the red boxes.

Discussion

In the path from DNA sample extraction to identifying variants that may be of interest when considering a patient's diagnosis, disease progression, and possible treatment options, it is important to recognize the multifarious nature of the methodology required for both sequencing and proper data processing. The protocol described herein is an example of the utilization of targeted NGS and subsequent bioinformatic analysis essential to identify rare variants of potential clinical significance. Specifically, we present the approach taken by the ONDRI genomics subgroup when using the ONDRISeq custom-designed NGS panel.

It is recognized that these methods were developed based on a specific NGS platform and that there are other sequencing platforms and target enrichment kits that may be used. However, the NGS platform and desktop instrument (**Table of Materials**) was chosen based on its early US Food and Drug Administration (FDA) approval[46]. This authorization reflects the high-quality sequencing that can be performed with the NGS protocols of choice and the reliability that can be placed on the sequencing reads.

Although obtaining accurate sequencing reads with the depth of coverage is very important, the bioinformatics processing required for final rare variant analysis is vital and can be computationally intensive. Due to the many sources of errors that may occur within the sequencing process, a robust bioinformatics pipeline must correct for the various inaccuracies that can be introduced. They may arise from misalignments in the mapping process, amplification bias introduced by PCR amplification in the library preparation, and the technology producing sequencing artifacts[47]. No matter the software used to perform read mapping and variant calling, there are common ways to reduce these errors including local realignment, removal of duplicate mapped reads, and setting proper parameters for quality control when calling variants. Additionally, the parameters chosen during variant calling may vary based on what is most appropriate for the study at hand[11]. The minimum coverage and quality score of a variant and the surrounding nucleotides that were applied herein were chosen as to create a balance between

appropriate specificity and sensitivity. These parameters have been validated for the ONDRISeq panel based on variant calling concordance with three separate genetic techniques, as previously described, including: 1) chip-based genotyping; 2) allelic discrimination assay; and 3) Sanger sequencing[9].

Following accurate variant calling, in order to determine those of potential clinical significance, annotation and curation are essential. Due to its open access platform, ANNOVAR is an excellent tool for both annotation and preliminary variant screening or elimination. Beyond being easily accessible, ANNOVAR can be applied to any VCF file, no matter what sequencing platform is used, and is customizable based on the needs of the research[26].

After annotation, variants must be interpreted to determine if they should be considered to be of clinical significance. Not only does this process become complex, but it is often prone to subjectivity and human error. For this reason, the ACMG has set guidelines to assess the evidence for pathogenicity of any variant. We apply a non-synonymous, rare variant-based manual curation approach, which is constructed based on these guidelines and safeguarded by individually assessing each variant that is able to pass through the pipeline with a custom-designed Python script that classifies the variants based on the guidelines. In this way, each variant is assigned a ranking of pathogenic, likely pathogenic, uncertain significance, likely benign, or benign, and we are able to add standardization and transparency to the variant curation process. It is important to recognize that the specifics of variant curation, beyond the bioinformatics pipeline, will be individualized based on the needs of the research, and was therefore beyond the scope of the methodologies presented.

Although the methods presented here are specific to ONDRI, the steps described can be translated when considering a large number of constitutional diseases of interest. As the number of gene associations increase for many phenotypes, targeted NGS allows for a hypothesis driven approach that can capitalize on the previous research that has been done in the field. Yet, there are limitations to targeted NGS and the methodology presented. By only focusing on specific regions of the genome, the areas of discovery are

limited to novel alleles of interest. Therefore, novel genes or other genomic loci beyond those covered by the sequencing targets, which could be revealed with WGS or WES approaches, will not be identified. There are also regions within the genome that can be difficult to accurately sequence with NGS approaches, including those with a high degree of repeated sequences[48] or those that are rich in GC content[49]. Fortunately, when utilizing targeted NGS, there is a priori a high degree of familiarity with the genomic regions being sequenced, and whether these might pose technical challenges. Finally, detection of copy number variants from NGS data at present is not standardized[50]. However, bioinformatics solutions to these concerns may be on the horizon; new computational tools may help to analyze these additional forms of variation in ONDRI patients.

Despite its limitations, targeted NGS is able to obtain high-quality data, within a hypothesis-driven approach, while remaining less expensive than its WGS and WES counterparts. Not only is this methodology appropriate for efficient and directed research, the clinical implementation of targeted NGS is growing exponentially. This technology is being used to answer many different questions regarding the molecular pathways of various diseases. It is also being developed into an accurate diagnostic tool at relatively low cost when opposed to WES and WGS. Even when compared to the gold-standard Sanger sequencing, targeted NGS can outcompete in its time- and cost-efficiency. For these reasons, it is important for a scientist or clinician who receives and uses NGS data, for instance, delivered as text in a laboratory or clinical report, to understand the complex "black box" that underlies the results. The methods presented herein should help users understand the process underlying the generation and interpretation of NGS data.

Table of Materials

Name	Company	Catalog Number	Comments
4 mL EDTA K2 tubes	Fisher Scientific	02-689-4	
1 M Tris Buffer	Bio Basic Canada Inc.	SD8141	
Gentra Puregene Blood Kit	Qiagen	158389	1,000 mL Kit. This is the blood extraction kit, referred to in step 1.3.
NanoDrop-1000 Spectrophotometer	Thermo Fisher Scientific	ND-2000	Replaced by the NanoDrop-2000 Spectrophotometer. This is the full-spectrum spectrophotometer, referred to in steps 1.4 and 2.1.2.
Qubit 2.0 fluorometer	Invitrogen	Q32866	This is a fluorometer appropriate for the quantification of DNA, referred to in steps 2.1.4, 2.1.6, 2.2.3, and 3.1.3.
Nextera Rapid Custom Capture Enrichment Kit	Illumina, Inc.	FC-140-1009	Specifically designed for the ONDRISeq panel, sequencing the exons of 80 genes, resulting in 971,388 base pairs of sequence in paired-end reads of 150 bases in length; 288 samples per kit. This is the target enrichment kit, referred to in steps 2.2, 2.2.2, 2.2.3, 3.1.5, 3.1.6, 3.4.1, and the Discussion.
2100 BioAnalyzer	Agilent Technologies	G2939BA	This is an automated electrophoresis system, referred to in step 3.1.4.
High Sensitivity DNA Reagent Kit	Agilent Technologies	5067-4626	110 Samples per kit; This is a DNA quality analysis kit, referred to in step 3.1.4.
MiSeq Reagent Kit v3	Illumina, Inc.	MS-102-3003	600 Cycle Kit; This is the NGS desktop instrument reagent kit, referred to in step 3.1.

MiSeq Personal Genome Sequencer	Illumina, Inc.	SY-410-1003	This is a NGS desktop instrument, referred to in steps 2.2.1, 3.1, 3.1.1, 3.1.2, 3.1.8, 3.2, 4.2.6, the Representative Results, and the Discussion.
Experiment Manager	Illumina, Inc.		This is NGS technology software, referred to in step 3.1.1 and **Figure J.1**. https://support.illumina.com/sequencing/sequencing_software/experiment_manager/downloads.html
BaseSpace	Illumina, Inc.	SW-410-1000	This is a cloud-based computing environment, referred to in steps 3.1.2, 3.2, 3.3, 3.3.1, 3.3.2, 3.4, 3.4.1, 3.4.2 and 3.4.3. https://basespace.illumina.com/
CLC Genomics Workbench 10.1.1	Qiagen	832000	Open source options for data pre-processing are also available that can model the workflow used in this protocol. This is the software used for data pre-processing, referred to throughout step 4 and in **Figure J.2**.
Annotate Variation			http://annovar.openbioinformatics.org/en/latest/user-guide/download/
RefSeq	National Center for Biotechnology Information		https://www.ncbi.nlm.nih.gov/refseq/
dbSNP138	National Center for Biotechnology Information		https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?view+summary=view+summary&build_id=138
Exome Aggregation Consortium	Broad Institute		http://exac.broadinstitute.org/

National Heart, Lung, and Blood Institute Exome Sequencing Project European Cohort	University of Washington and the Broad Institute		http://evs.gs.washington.edu/EVS/
ClinVar	National Center for Biotechnology Information		https://www.ncbi.nlm.nih.gov/clinvar/
Combined Annotation Dependent Depletion	University of Washington and Hudson-Alpha Institute for Biotechnology		http://cadd.gs.washington.edu/
Sorting Intolerant from Tolerant	J. Craig Venter Instutite		http://sift.jcvi.org/
PolyPhen-2	Brigham and Women's Hospital, Harvard Medical School		http://genetics.bwh.harvard.edu/pph2/
Human Gene Mutation Database	Qiagen	834050	This is a disease mutation database, referred to in step 5.2 and the Representative Results. https://portal.biobase-international.com/cgi-bin/portal/login.cgi?redirect_url=/hgmd/pro/start.php
Splicing-based Analysis of Variants	Frey lab, University of Toronto		http://tools.genes.toronto.edu/
Human Splicing Finder	Aix Marseille Université		http://www.umd.be/HSF3/HSF.shtml

Other materials

Centrifuge

Disposable transfer
pipets

References

1. Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet*. **11**, (1), 31-46 (2010).

2. Mardis, E. R. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. **9**, 387-402 (2008).

3. Shendure, J., Ji, H. Next-generation DNA sequencing. *Nat Biotechnol*. **26**, (10), 1135-1145 (2008).

4. Sanger, F., Nicklen, S., Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. **74**, (12), 5463-5467 (1977).

5. Farhan, S. M. K., Hegele, R. A. Exome Sequencing: New Insights into Lipoprotein Disorders. *Current Cardiology Reports*. **16**, (7), (2014).

6. Choi, M., *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*. **106**, (45), 19096-19101 (2009).

7. Mardis, E. R. DNA sequencing technologies: 2006-2016. *Nat Protoc*. **12**, (2), 213-218 (2017).

8. Farhan, S. M., *et al.* The Ontario Neurodegenerative Disease Research Initiative (ONDRI). *Can J Neurol Sci*. **44**, (2), 196-202 (2017).

9. Farhan, S. M. K., *et al.* The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*. (16032), 1-11 (2016).

10. El-Metwally, S., Hamza, T., Zakaria, M., Helmy, M. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput Biol*. **9**, (12), e1003345 (2013).

11. Yohe, S., Thyagarajan, B. Review of Clinical Next-Generation Sequencing. *Arch Pathol Lab Med*. (2017).

12. Qiagen. *Gentra Puregene Handbook*. 4th edn, (2014).

13. NanoDrop Technologies, Inc. *Spectrophotometer V3.5 User's Manual*. (2007).

14. Invitrogen by Life Technologies. *Qubit 2.0 Fluorometer User Manual. Vol. Q32866*. (2010).

15. Illumina, Inc. *Nextera Rapid Capture Enrichment Guide*. Vol. 15037436 v01 (2016).

16. Illumina, Inc. *Nextera Rapid Capture Enrichment Reference Guide*. Vol. 15037436 v01 (2016).

17. Rev. B. Illumina, Inc. *MiSeq Reagent Kit v3 Reagent Preparation Guide*. Vol. 15044932 Rev. B (2013).

18. Illumina, Inc. *MiSeq System Guide*. Vol. 15027617 v01 (2015).

19. *BaseSpace Sequence Hub*. https://basespace.illumina.com/dashboard (2017).

20. Rev. B. Agilent Technologies. *Agilent High Sensitivity DNA Kit Guide*. Vol. G2938-90321 (2013).

21. Illumina, Inc. *MiSeq System Denature and Dilute Libraries Guide*. Vol. 15039740 v01 (2016).

22. Illumina, Inc. *System Specification Sheet: MiSeq System*. (2016).

23. *BaseSpace Sequence Hub Help Center*. Available from: https://help.basespace.illumina.com/ (2017).

24. Qiagen. *Genomics Workbench 10.1.1 User Manual*. (2017).

25. Ebbert, M. T., *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*. **17**, Suppl 7. 239 (2016).

26. Wang, K., Li, M., Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. **38**, (16), e164 (2010).

27. Leary, N. A., *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. **44**, (D1), D733-D745 (2016).

28. Kitts, A., Phan, L., Ward, M., Bradley Holmes, J. *The Database of Short Genetic Variation (dbSNP)*. National Center for Biotechnology Information. Bethesda, MD. (2013).

29. Lek, M., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**, (7616), 285-291 (2016).

30. *Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP)*. http://evs.gs.washington.edu/EVS/ (2017).

31. Auton, A., *et al.* A global reference for human genetic variation. *Nature*. **526**, (7571), 68-74 (2015).

32. Landrum, M. J., *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. **44**, (D1), D862-D868 (2016).

33. Kircher, M., *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. **46**, (3), 310-315 (2014).

34. Kumar, P., Henikoff, S., Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. **4**, (7), 1073-1081 (2009).

35. Adzhubei, I. A., *et al.* A method and server for predicting damaging missense mutations. *Nat Methods*. **7**, (4), 248-249 (2010).

36. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet*. **39**, (1), 17-23 (2007).

37. Xiong, H. Y., *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science*. **347**, (6218), (2015).

38. Desmet, F. O., *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*. **37**, (9), e67 (2009).

39. Richards, S., *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. **17**, (5), 405-424 (2015).

40. Li, Q., Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet*. **100**, (2), 267-280 (2017).

41. Yang, Z. L., Sun, G. L. High-frequency, low-coverage "false positives" mutations may be true in GS Junior sequencing studies. *Scientific Reports*. **7**, (2017).

42. Gandhi, P. N., Wang, X., Zhu, X., Chen, S. G., Wilson-Delfosse, A. L. The Roc domain of leucine-rich repeat kinase 2 is sufficient for interaction with microtubules. *J Neurosci Res*. **86**, (8), 1711-1720 (2008).

43. Goldwurm, S., *et al.* The G6055A (G2019S) mutation in LRRK2 is frequent in both early and late onset Parkinson's disease and originates from a common ancestor. *J Med Genet*. **42**, (11), e65 (2005).

44. Caiazzo, M., *et al.* Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature*. **476**, (7359), 224-227 (2011).

45. Grimes, D. A., *et al.* Translated mutation in the Nurr1 gene as a cause for Parkinson's disease. *Mov Disord*. **21**, (7), 906-909 (2006).

46. Collins, F. S., Hamburg, M. A. First FDA authorization for next-generation sequencer. *N Engl J Med*. **369**, (25), 2369-2371 (2013).

47. Van der Auwera, G. A., *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. **43**, 11-33 (2013).

48. Treangen, T. J., Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. **13**, (1), 36-46 (2011).

49. Shin, S., Park, J. Characterization of sequence-specific errors in various next-generation sequencing systems. *Mol Biosyst*. **12**, (3), 914-922 (2016).

50. Povysil, G., *et al.* panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum Mutat*. **38**, (7), 889-897 (2017).

Curriculum Vitae

Allison A. Dilliott

H-index: 6 | Citations: 230

EDUCATION

Western University, London, ON

Doctor of Philosophy (PhD)   **September 2016 – August 2021**

Biochemistry
- Thesis title: "Identifying the genetic determinants of neurodegenerative disease."
- Supervisor: Robert A. Hegele, MD

Bachelor of Medical Sciences (BMSc)   **September 2012 – April 2016**

Honours Specialization in Biochemistry and Pathology of Human Disease
- Thesis Title: "Identifying the genetic basis of neurodegeneration in 75 patients using a targeted resequencing panel."

LABORATORY AND RESEARCH EXPERIENCE

Postdoctoral Researcher   **September 2021 – TBD**

Departments of Neurology and Neurosurgery, and Human Genetics, Montréal Neurological Institute and Hospital, McGill University, Montréal, QC
- Supervisors: Guy A. Rouleau, OC, OQ, MD, PhD & Sali M.K. Farhan, PhD

PhD Candidate   **September 2016 – August 2021**

Blackburn Cardiovascular Genetics Lab, Robarts Research Institute, Western University, London, ON
- Supervisor: Robert A. Hegele, MD

Undergraduate Research Assistant   **May 2016 – August 2016**
                                                  May 2015 – August 2015

Blackburn Cardiovascular Genetics Lab, Robarts Research Institute, Western University, London, ON
- Supervisor: Robert A. Hegele, MD

Undergraduate Honours Research Thesis Student   **September 2015 – April 2016**

Blackburn Cardiovascular Genetics Lab, Robarts Research Institute, Western University, London, ON
- Supervisor: Robert A. Hegele, MD

OTHER EXPERIENCE AND PROFESSIONAL AFFILIATIONS

Research Scholar   **September 2016 – TBD**

The Ontario Neurodegenerative Disease Research Initiative (ONDRI)

Biocurator   **March 2021 - TBD**

Amyotrophic Lateral Sclerosis Spectrum Disorders Gene Curation Expert Panel, ClinGen

HONOURS AND ACHIEVEMENTS

Doctoral Award: Frederick Banting and Charles Best Canada Graduate Scholarship
- Award duration: May 2020 – April 2023
- From: Canadian Institute of Health Research (CIHR)

Top Platform Presentation, Neuroscience Research Day
- Awarded: February 2021
- From: Neuroscience Research Day Planning Committee

Queen Elizabeth II Graduate Scholarship in Science and Technology
- Awarded: May 2020 (DECLINED)
- From: School of Graduate and Postdoctoral Studies, Western University

Queen Elizabeth II Graduate Scholarship in Science and Technology
- Award duration: May 2019 – April 2020
- From: School of Graduate and Postdoctoral Studies, Western University

Outstanding Research Contribution Scholarship
- Awarded: September 2018
- From: Public Service Alliance of Canada (PSAC) Local 610

Alzheimer Society of London and Middlesex Doctoral Scholarship
- Award duration: May 2018 – April 2020
- From: Alzheimer Society of London and Middlesex

Ontario Graduate Scholarship
- Award duration: May 2018 – April 2019
- From: School of Graduate and Postdoctoral Studies, Western University

Graduate Student Teaching Assistant Award Nominee
- Nominated: May 2018; May 2017
- From: Graduate Student Teaching Award Committee, Western University

CIHR 2018 International Symposium on Atherosclerosis Travel Award
- Awarded: June 2018
- From: CIHR – Institute of Circulatory and Respiratory Health

Alzheimer Society of London and Middlesex Masters Scholarship
- Award duration: May 2017 – April 2018
- From: Alzheimer Society of London and Middlesex

PUBLICATIONS

1. **Dilliott A.A.**, Abdelhady A., Sunderland K.M., Farhan S.M.K., Abrahao A., Binns M.A., Black S.E., Borrie M., Casaubon L.K., Dowlatshahi D., Finger E., Fischer C.E., Frank A., Freedman M., Grimes D., Hassan A., Jog M., Kumar S., Kwan D., Lang A.E., Mandzia J., Masellis M., McIntyre A.D., Pasternak S., Pollock D.G., Rajji T.K., Rogaeva E., Sahlas D.J., Saposnik G., Sato C., Seitz D., Shoesmith C., Steeves T., Swartz R.H., Tan B., Tang-Wai D., Tartaglia M.C., Turnbull J., Zinman L., ONDRI Investigators, and Hegele R.A. (2021). Contribution of rare variant associations to neurodegenerative disease presentation in the Ontario Neurodegenerative Disease Research Initiative. *NPJ Genomic Med*, [Accepted]. Ms. No. NPJGENMED-00951

2. **Dilliott A.A.**, Sunderland K.M., McLaughlin P.M., Roberts A.C., Evans E.C., Abrahao A., Binns M.A., Black S.E., Borrie M., Casaubon L.K., Dowlatshahi D., Finger E., Fischer C.E., Frank A., Freedman M., Grimes D., Hassan A., Jog M., Kumar S., Kwan D., Lang A.E., Mandzia J., Marras C., Masellis M., McIntyre A.D., Pasternak S., Pollock D.G., Rajji T.K., Robinson J.F., Rogaeva E., Sahlas D.J., Saposnik G., Sato C., Seitz D., Shoesmith C., Steeves T., Strother S.C., Swartz R.H., Tan B., Tang-Wai D., Tartaglia M.C., Troyer A.K., Turnbull J., Zinman L., ONDRI Investigators, and Hegele R.A. (2021). Association of

apolipoprotein E variation with cognitive impairment across multiple neurodegenerative diagnoses. *Neurobiol Aging*, 378.e1-378.e9

3. Gill P.K., Dron J.S., **Dilliott A.A.**, McIntyre A.D., Cao H., Wang J., Movsesyan I.G., Malloy M.J., Pullinger C.R., Kane J.P., and Hegele R.A. (2020). Ancestry-specific profiles of genetic determinants of severe hypertriglyceridemia. *J Clin Lipidol,* S1933-2874(20)30333-0

4. **Dilliott A.A.**, Wang J., Brown E., Singh G., Shkrum M.J., Clin M., Rupar A., Hegele R.A., and Siu V.M. (2020). A novel homozygous variant in *REN* in a family presenting with classic features of disorders involving the renin-angiotensin pathway, without renal tubular dysgenesis. *Am J Med Genet A*, 182(10):2284-2290

5. Dron J., **Dilliott A.A.**, Lawson A., McIntyre A., David B., Wang J., Cao H., Movsesyan I., Malloy M., Pullinger C., Kane J., and Hegele R.A. (2020). Loss-of-function *CREB3L3* variants in patients with severe hypertriglyceridemia. *Arterioscler Thromb Vasc Biol,* 40(8):1935-1941

6. **Dilliott A.A.***, Ramirez J.*, Binns M., Breen D., Evans E., Beaton D., McLaughlin P.M., Kwan D., Holmes M.F., Ozzoude M., Scott C.J.M., Strother S.C., Symons S., Swartz R.H., Grimes D., Jog M., Masellis M., Black S.E., Joutel A., Marras C., Rogaeva E., Hegele R.A., and Lang A.E., on behalf of the ONDRI Investigators (2020). Parkinson's disease, *NOTCH3* genetic variants, and white matter hyperintensities. *Mov Disord*, 35(11):2090-2095

7. **Dilliott A.A.*,** Zhang M.*, Khallaf R., Robinson J.F., Hegele R.A., Comishen M., Sato C., Tosto G., Reitz C., Mayeux R., St George-Hyslop P., Freedman M., and Rogaeva E. (2019). Genetic and epigenetic study of a family with monozygotic triplets discordant for Alzheimer's Disease. *Brain*, 142(11):3375-3381
         November 2019 Editor's Choice Award

8. **Dilliott A.A.**, Evans E.C., Farhan S.M.K., Ghani M., Sato C., Zhang M., McIntyre A.D., Cao H., Racacho L., Robinson J.F., Strong M.J., Masellis M., Bulman D.E., Rogaeva E., Black S., Finger E., Frank A., Freedman M., Hassan A., Lang A., Shoesmith C., Swartz R., Tang-Wai D., Tartaglia C., Turnbull J., Zinman L., the ONDRI Investigators, and Hegele R.A. (2019). Genetic variation in the Ontario Neurodegenerative Disease Research Initiative. *Can J Neurol Sci,* 46(5):491-498
         Editor's Choice Award

9. Wang L., Radonjic A., **Dilliott A.A.**, McIntyre A.D., and Hegele R.A. (2018). A de novo *POLD1* mutation associated with mandibular hypoplasia, deafness, progeroid features and lipodystrophy (MDPL) syndrome in a family with Werner syndrome. *J Investig Med High Impact Case Rep,* 6:1-6

10. **Dilliott A.A.**, Farhan S.M.K., McIntyre A.D., Robinson J.F., Kwan D., Swartz R., Hassan A., Dowlatshahi D., Sahlas S., Saposnik G., Mandzia J., Casaubon L., Strong M.J., Masellis M., Bulman D.E., Rogaeva E., Hegele R.A., and the ONDRI Investigators (2018). Identifying the Genetic Basis of Vascular Cognitive Impairment Using a Custom Designed Next-generation Sequencing-based Gene Panel [Abstract]. *Atheroscler Suppl,* 80:80-81

11. **Dilliott A.A.**, Farhan S.M.K., Ghani M., Sato C., Liang E., Zhang M., McIntyre A.D., Cao H., Racacho L., Robinson J.F., Strong M.J., Masellis M., Bulman D.E., Rogaeva E., Lang A., Tartaglia C., Finger E., Zinman L., Turnbull J., Freedman M., Swartz R., Black S., the ONDRI Investigators, and Hegele R.A. (2018). Targeted next-generation sequencing and bioinformatics pipeline to evaluate genetic determinants of constitutional disease. *J Vis Exp,* (134):e57266

12. Badalato L., Farhan S.M.K., **Dilliott A. A.**, Care4Rare Canada, Bulman D. E., Hegele R.A., Goobie S. (2017). *KMT2D* p.Gln3575His segregating in a family with autosomal dominant choanal atresia strengthens the Kabuki/CHARGE connection. *Am J Med Genet A,* 173(1):183-189

13. Wang, J., Ban, M., Dron, J., Robinson, J., McIntyre, A., Alazzam, M., **Dilliott, A.**, Cao, H., Huff, M., Rhainds, D., Dube, M-P., Lettre, G., Tardiff, J-C., Hegele, R. (2016). Polygenic versus monogenic causes of hypercholesterolemia ascertained clinically. *Arterioscler Thromb Vasc Biol*, 36(12):2439-2445

14. Farhan, S.M.K., **Dilliott, A.A.**, Ghani, M., Sato, C., Liang, E., Zhang, M., McIntyre, A.D., Cao, H., Racacho, L., Robinson, J.F., ONDRI Investigators, Strong, M.J., Masellis, M., St George-Hyslop, P., Bulman, D.E., Rogaeva, E., Hegele, R. A. (2016). The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. *NPJ Genom Med*, 1:16032

*Co-first authors

INVITED PRESENTATIONS AND SEMINARS

1. **Neurodegeneration Genetics Seminar**, Farhan Lab, McGill University, Montréal, QC (Visiting Speaker)
   "Rare variant enrichment across neurodegenerative and cerebrovascular diseases." | May 27, 2021

2. **ONDRI Friends and Family Day**, London, ON (Provincial, <u>Audience:</u> General Public)
   "Investigating genetics' role in neurodegenerative disease." | September 17, 2019

3. **Alzheimer's Society of London and Middlesex Scholarship Research Showcase**, Alzheimer's Society of London and Middlesex, London, ON (Provincial, <u>Audience:</u> General Public)
   "Investigating genetics' role in neurodegenerative disease." | May 29, 2019

4. **ONDRI Friends and Family Day**, Toronto, ON (Provincial, <u>Audience:</u> General Public)
   "Investigating genetics' role in neurodegenerative disease." | February 20, 2019

5. **London Regional Genomics Centre and Illumina's NextSeq Launch**, Robarts Research Institute, London, ON (Local)
   "Identifying rare disease-causing mutations using whole-exome sequencing." | September 26, 2016

PLATFORM PRESENTATIONS

1. **Neuroscience Research Day**, Western University, London, ON (Provincial; Competitive) | February 19, 2021
   *Award-winning

2. **Robarts Research Retreat**, Western University, London, ON (Institutional; Competitive) | June 19, 2020

3. **Leena Peltonen School of Human Genomics**, Les Diablerets, Switzerland (International; Competitive) | August 20, 2019

4. **8th Annual Canadian Human and Statistical Genetics Meeting**, Montebello, QC (National; Competitive) | June 17, 2021

5. **ONDRI Scientific Retreat**, Toronto, ON (Provincial) | May 27, 2019

6. **Biochemistry Spring Menten Symposium**, Western University, London, ON (Institutional) | May 3, 2019

7. **Biochemistry Graduate Student Spring Symposium**, Western University, London, ON (Institutional) | June 13, 2018
8. **London Health Research Day**, London, ON (Local; Competitive) | May 10, 2018
9. **Molecular Medicine Data Club**, Robarts Research Institute, Western University, London, ON (Institutional) | May 2, 2018
10. **ONDRI Workshop**, Toronto, ON (Provincial) | October 3, 2017
11. **Robarts Research Retreat**, Western University, London, ON (Institutional; Competitive) | June 20, 2017
12. **Molecular Medicine Data Club Three Minute Theses**, Robarts Research Institute, Western University, London, ON (Institutional) | May 31, 2017
13. **ONDRI Workshop**, Mississauga, ON (Provincial) | April 20, 2017

POSTER PRESENTATIONS

1. **Biochemistry Fall Symposium**, Western University, London, ON (Institutional) | December 10, 2020
2. **American Society of Human Genetics Meeting**, Virtual (International) | October 27–30, 2020
3. **American Society of Human Genetics Meeting**, Houston, TX (International) | October 17, 2019
4. **8th Annual Canadian Human and Statistical Genetics Meeting**, Montebello, QC (National) | June 18, 2019
5. **Department of Medicine Resident Research Day**, Western University, London, ON (Institutional) | May 10, 2019
6. **London Health Research Day**, London, ON (Local) | April 30, 2019
7. **American Society of Human Genetics Meeting**, San Diego, CA (International) | October 19, 2018
8. **Schulich Clinician Scientist Trainee Symposium**, Western University, London, ON (Institutional) | August 14, 2018
9. **International Symposium on Atherosclerosis**, Toronto, ON (International) | June 10, 2018
10. **Robarts Research Retreat**, Western University, London, ON (Institutional) | June 1, 2018
11. **Department of Medicine Resident Research Day**, Western University, London, ON (Institutional) | May 11, 2018
12. **American Society of Human Genetics Meeting**, Orlando, FL (International) | October 17, 2017
13. **Robarts Research Retreat**, Western University, London, ON (Institutional) | June 20, 2017
14. **8th Annual Canadian Human and Statistical Genetics Meeting**, Quebec City, QC (National) | April 23, 2017
15. **London Health Research Day**, London, ON (Local) | March 28, 2017
16. **Harold Stewart Research Showcase**, Western University, London, ON (Institutional) | January 20, 2017
17. **ONDRI Annual Workshop**, London, ON (Provincial) | October 17, 2016
18. **Robarts Research Retreat**, Western University, London, ON (Institutional) | June 13, 2016
19. **Department of Medicine Resident Research Day**, Western University, London, ON (Institutional) | May 26, 2016

20. **Harold Stewart Research Showcase**, Western University, London, ON
    (Institutional) | January 22, 2016

TEACHING EXPERIENCE

Graduate Teaching Assistantships – Western University, London, ON

MEDSCI 4931G: Selected Topics in Medical Sciences	**January 2021 – April 2021**
MEDSCI 3900G: Medical Science Laboratory	**January 2020 – April 2020**
MEDSCI 4900F/G: Advanced Medical Science Laboratory	**September 2019 – April 2020**
	September 2018 – April 2019
BIOCHEM 3380G: Biochemistry Laboratory	**January 2018 – April 2018**
	January 2017 – April 2017
BIOCHEM 3382A: Biochemical Regulation	**September 2017 – December 2017**
	September 2016 – December 2016

MENTORING AND LEADERSHIP EXPERIENCE

- **Summer Research Assistant Mentor (6)**   **May 2017 – August 2021**
  Western University, London, ON
- **Undergraduate Thesis Student Mentor (8)**   **September 2016 – April 2021**
  Western University, London, ON
- **Schulich Graduate Students' Council Chair**   **March 2019 – May 2021**
  Schulich School of Medicine and Dentistry, Western University, London, ON

INTERVIEWS AND MEDIA RELATIONS

- **Impact Stories**, The Ontario Neurodegenerative Disease Research Initiative
  "ONDRI Scholars Promoting Novel Scientific Discoveries." | October 2020
- **The Collaborator**, Schulich School of Medicine and Dentistry
  "Solving a genetic puzzle." | November 2019
- **The Collaborator**, Schulich School of Medicine and Dentistry
  "On a mission to increase diversity and inclusion in the world of science." |
  April 2019
- **Trainee Q&A**, Robarts Research Institute Discovery Newsletter
  "The research gene." | October 2017