Statistical analysis of rare genetic variants in the first exon of the ataxin-2 gene in patients with neurodegenerative diseases

by

Diksha Jethnani

B.Sc. (Hons.), Dayalbagh Educational Institute, 2020

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

© Diksha Jethnani 2024
SIMON FRASER UNIVERSITY
Spring 2024

Declaration of Committee

Name: **Diksha Jethnani**

Degree: **Master of Science**

Thesis title: **Statistical analysis of rare genetic variants in the first exon of the ataxin-2 gene in patients with neurodegenerative diseases**

Committee: **Chair:** Liangliang Wang
Associate Professor, Statistics and Actuarial Science

Jinko Graham
Supervisor
Professor, Statistics and Actuarial Science

Brad McNeney
Committee Member
Associate Professor, Statistics and Actuarial Science

Joanna Lubieniecka
Examiner
Adjunct Professor, Statistics and Actuarial Science

Abstract

The ataxin-2 gene (ATXN2) encodes a ribonucleic acid (RNA) binding protein involved in messenger RNA translation and regulation. Large polyglutamine (CAG) expansions or repeat regions in ATXN2 are causative of the neurodegenerative disease spinocerebellar ataxia type 2 (SCA2) and intermediate expansions are considered to be a risk factor for the neurodegenerative disease amyotrophic lateral sclerosis (ALS). However, most variants in the repeat regions of ATXN2 remain unreported because they are difficult to capture with traditional short-read sequencing. We analyze rare genetic variants found in short-read sequencing of exon 1, a polyglutamine repeat region of the ATXN2 gene. The variants were identified during diagnostic exome sequencing of patients for neurodegenerative disease. After adjusting for potentially confounding variables such as age, biological sex, and the enrichment kit used in the sequencing, we find the variants to be associated with neurodegenerative disease, suggesting their involvement in disease pathology. Our preliminary results with short-read sequencing suggest that re-investigation of the ATXN2 gene with long-read sequencing technologies that allow a better resolution of repeat regions shows promise for new insights into neurodegeneration.

Keywords: ATXN2, neurodegeneration, ALS, SKAT

Dedication

To my parents, Radha and Jitendra Jethnani.

Acknowledgements

I wish to extend my heartfelt gratitude to my supervisor, Dr. Jinko Graham. Thank you for your encouragement, guidance, support, and patience throughout this journey. My sincere appreciation also goes to Dr. Joanna Lubieniecka; this research would not have been possible without your invaluable support. Additionally, I am deeply thankful to committee member Dr. Brad McNeney and examination chair Dr. Liangliang Wang for their insightful contributions.

I seize this moment to express my gratitude to the faculty members in the Department of Statistics and Actuarial Science. My time at SFU has been enriching and enjoyable, and I am forever indebted to the wealth of knowledge and experiences I have gained. I want to extend a special thank you to Dr. Ailene MacPherson for her invaluable guidance and support. To my family and friends, I offer my deepest appreciation for your enduring presence and unwavering support. Your love, kindness, sacrifices, and constant encouragement have been my pillars of strength. To my friends, my extended family here in BC, this would not have been possible without you all. Finally, thank you Divyansh and Piyu for all the motivation you put into me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Genes play a big part in how the cells work in our bodies. One such critical gene is the ATXN2 gene. This gene makes a protein that helps with several important jobs inside our cells. It helps bring things into the cell (endocytosis), control how cells grow and stay healthy, and make sure the cell creates the proteins it needs (ribosomal translation). It also helps keep our cell's mitochondria powerhouses working properly.

Central to the exploration, we have the N-terminal region of ATXN2, which represents the starting point or the beginning section of the protein. This N-terminal region contains a polyglutamine tract (PolyQ) which refers to a portion of the protein consisting of a sequence of several glutamine units stacked together akin to having a row of the same type of building block repeated several times in a row. And in this case, there are usually between 14 to 31 of these, also termed residues. When this tract expands, it triggers a cascade of neurological complications, with various disease manifestations [Lubieniecka et al., 2022]. A lot of times these diseases vary according to the length of the expansion. The scientific community has closely associated intermediate-length expansions of the PolyQ tract (27-34 repeats) with heightened susceptibility to amyotrophic lateral sclerosis (ALS)[Chio et al., 2022]. Even further along the spectrum, when the PolyQ tract expands beyond 34 repeats, it precipitates the onset of spinocerebellar ataxia-2 (SCA2), characterized by a progressive loss of coordination and motor control.[Egorova and Bezprozvanny, 2019].

Various studies have demonstrated that this gene (ATXN2) can modify the toxicity of TDP43, a protein closely linked with ALS pathology in a complex manner. All this research and findings paved the way for an ongoing clinical trial (Clinical Trial NCT04494256) that harnesses ATXN2 antisense oligonucleotides to lower ataxin-2 protein levels. This could be a potential direction towards the treatment of individuals battling with ALS. However, this clinical trial exclusively recruits patients with PolyQ expansions in the ATXN2 gene. In contrast, we question whether other mutations in the ATXN2 gene, especially those residing in exon 1 where the PolyQ region is nestled, influence protein function and disease susceptibility similarly.

To address this question, our clinical-genetics collaborators in the diagnostic genome-sequencing facility at the Ruhr University, Bochum, evaluated rare genetic variants, including single nucleotide variants (SNVs) and insertions/deletions (indels), within the first exon of the ATXN2 gene. They aim to evaluate and understand rare genetic variants in the exon 1 of ATXN2 and their potential contributions to the pathogenesis of neurodegenerative diseases. Their study aims to contribute to understanding the genetic underpinnings of neurological disorders, with potential implications for diagnosis and treatment.

Our collaborators collected data from patients with various diseases. They classified the diseases based on the current knowledge as ND/non-ND along with the spotted variants in the exon 1 of the gene (if any). We received their sampled data in two Excel files for our statistical analysis. In this analysis, we aim to understand the association between the rare variants in exon 1 of the ATXN2 gene and neurological disorder (ND) status. Here, we define “rare” to be frequencies of less than 0.01 in the gnomAD public database.

Chapter 2

Data

This analysis focuses on the ATXN2 gene, specifically, the variants found in exon 1 of the gene. We look at their association with neurodegenerative diseases. In this chapter, we perform some exploratory analysis to understand the data provided to us by our clinical genetics collaborators. We received the data from 358 people, comprising 134 cases with neurodegenerative diseases, 161 cases with non-neurodegenerative diseases, and 63 unclassified cases due to mixed symptoms. All of them had a normal number of polyglutamine repeats in exon 1 of the ATXN2 gene.

The data was delivered in two Excel files for the analysis. These were the <persons> data file and the <variants> data file respectively. The <persons> data file consists of the information about the subjects including the disease they suffer from, its classification (ND or not), the clinical information, the variants in the exon 1 of the ATXN2 gene found in the sample (if any), etc. The sample index column indicates the suspected diagnosis (disease). As specified by our collaborators, the abbreviations used for the sample index are as stated: P-ALS = Amyotrophic lateral sclerosis, P-SPG = Spastic paraplegia, P-AX = Spinocerebellar ataxia paraplegia, P-AMY = Amyloidosis, P-SY = Syndromes/Global developmental delay, P-MH = Malignant hyperthermia, P-MY = Myopathy, P-NP = Neuropathy, P-DIV = Diverse/Rare/Unclassifiable, P-SW = Metabolic disease, P-DYT = Dystonia, P-BGW = Connective tissue diseases, P-TM, = Cancer, P-HL = Hearing loss and EX = 'Healthy' individuals. The second data file details the 19 different variants of the ATXN2 gene under study. Our collaborators refer to the publicly available gnomAD database to find the corresponding allele frequency of the variants. For 3 of the 19 variants, no allele frequencies are reported in the gnomAD database. We call these variants 'Questionable' in our analysis and assess their relationship with disease status separately to check for potential bias from removing them in the analysis.

The exploratory analysis aims to uncover data patterns and insights that may guide further investigations into the relationship between the variants found in the exon 1 of the gene and neurodegeneration (ND status). We use R programming for the exploratory analysis, including the readxl R package for reading Excel data and the *ggplot2* and *dplyr* libraries for visualization and data manipulation.

2.1 The <persons> data

The data comprises 15 variables with 358 observations, representing the total number of samples. To facilitate analysis, the data is pre-processed as follows. The ‘ND’ (Neurodegenerative Disease) column is converted to a factor with levels *yes*, *no*, and *maybe* where *yes* corresponds to 1 implying the disease is ND, *no* corresponds to 0 implying the disease is non-ND and *maybe* corresponds to 3 in the original data file implying the disease couldn’t be classified due to overlapping symptoms. Additionally, a new column, ‘Variant,’ is created to signify the presence of a genetic variant, with entries “yes” or “no” implying presence or absence respectively. Another column, ‘Total_Variants,’ is added to indicate the number of variants each person possesses.

Furthermore, we introduce a new column, ‘age’, providing valuable information for subsequent analyses. Note that, we use their date of birth and the date of the data collection to calculate the age of the subjects. Finally, to check the questionable variants highlighted in yellow in the Excel files, we added another new column, ‘Questionable Variant’ to identify samples containing these variants. After pre-processing the data, we started with univariate summaries of the variables to gain insights into their distribution.

2.1.1 Univariate summaries

The univariate summaries of the dataset include distributions of age, sample index, sex, disease classification (i.e. neurodegenerative disease (ND), non-ND or undetermined), enrichment kits, clinical information, and relationships. Distributions of selected variables are provided in Appendix 1. Briefly, the main findings are as follows. As shown in Figure 2.1, the majority of individuals in the study are aged between 50-60 years, followed by individuals aged between 60-70 years.

The dataset has a balanced gender representation of 177 females and 181 males. ND and non-ND diseases are fairly distributed, with 161 individuals having non-ND conditions, 134 with ND conditions, and 63 undetermined. Amyotrophic lateral sclerosis (ALS) is the most frequent clinical information category with 74 occurrences comprising approximately 55% of the ND patients and approximately 21% of the study sample. Most subjects are unrelated (97 %) except for 12 in six familial clusters: 4 parent-child trios, one father-son pair and one sibling pair.

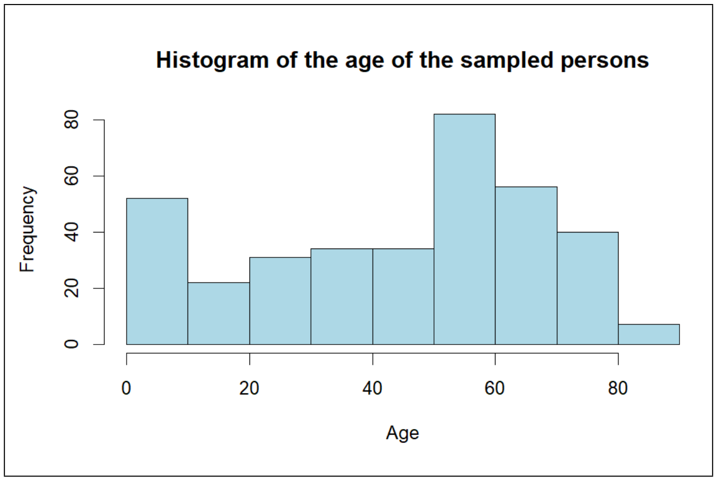

Figure 2.1: Histogram of the age distribution of the sampled individuals.

Amongst the enrichment kits, Twist Comprehensive Exome Refseq vs2 is the most frequently used (67 %), followed by Twist Comprehensive Exome plus Refseq (22 %), Twist Mix (Comprehensive plus 2.0)(9 %), SureSelect All Exon v7(1.2 %) and Twist Comprehensive Exome plus Refseq, Twist Exome 2.0, Twist Mix (Comprehensive plus 2.0) (0.8%). According to our collaborators the enrichment-kit categories of "Twist Comprehensive Exome plus Refseq, Twist Exome 2.0, Twist Mix (Comprehensive plus 2.0)" and "Twist Mix (Comprehensive plus 2.0)" should be merged whereas the "SureSelect All Exon v7" enrichment kit is from a different vendor and needs to be kept separate from the others. These enrichment kits have long names, so we have used abbreviations throughout the analysis to avoid complexity. The abbreviations used are as follows: Twist Comprehensive Exome Refseq vs2 =TCER vs2, SureSelect All Exon v7 = Exon v7 , Twist Comprehensive Exome plus Refseq = TCE (RefSeq), Twist Comprehensive Exome plus Refseq, Twist Exome 2.0, Twist Mix (Comprehensive plus 2.0) = TCE(R, Ex, Mix) and Twist Mix (Comprehensive plus 2.0)= T Mix.

2.1.2 Bivariate summaries

This section explores the associations between pairs of variables in the <persons> data file. These bivariate summaries and association tests provide valuable insights into the relationships between various variables in the dataset, setting the stage for more in-depth multivariate analyses.

Categorical × categorical variables

We constructed contingency tables for all the possible pairs of categorical variables followed by association tests, but report selected results only. Further tables can be found in Appendix 1.

Sample Index	Female	Male
EX	4	4
P-ALS	49	44
P-AMY	0	3
P-AX	13	9
P-BGW	16	4
P-DIV	6	9
P-DYT	3	3
P-HL	1	2
P-MH	3	1
P-MY	19	23
P-NP	22	30
P-SPG	9	10
P-SW	2	1
P-SY	24	34
P-TM	6	4

Table 2.1: Sample Index by Sex.

Table 2.1 summarizes the relationship between sample index and sex ($p = 0.2$ based on an exact permutational test) and suggests gender does not play a role in determining the type of the disease.

Table 2.2 summarizes the association between sample index and enrichment kits ($p = 0.0003$ based on an exact permutational test). The significant association is difficult to explain given that the samples to be sequenced are randomized according to disease type (sample index). However, the association could potentially reflect the different regions enriched and highlighted by changing enrichment kits over time.

Table 2.3 summarizes the association between sample index and ND status ($p = 0.0001$, based on an exact permutational test). A significant association is to be expected because the diseases (in the sample index) have been classified as ND or non-ND based on their type and if they lead to neurodegeneration. Certain diseases that can not be identified because of overlapping symptoms have been classified as unclear or maybe ND.

Table 2.4 summarizes the association between ND status and enrichment kits ($p = 0.01$ based on an exact permutational test). The significant association between ND status and enrichment kits aligns with that between sample index and enrichment kits above, as would be expected because the diseases have been classified as ND or non-ND based on their type and if they lead to neurodegeneration.

Similarly, Tables 2.5, 2.6, and 2.7 summarize, respectively, the associations between ND status and (i) the presence of a variant ($p = 0.02$ based on an exact permutational test), (ii) the presence of a questionable variant ($p = 0.54$ based on an exact permutational test),

	TCER vs2	Exon v7	TCE(RefSeq)	TCE(R,Ex,Mix)	T Mix
EX	8	0	0	0	0
P-ALS	79	0	4	0	10
P-AMY	2	0	1	0	0
P-AX	10	1	6	2	3
P-BGW	14	0	5	0	1
P-DIV	11	0	3	0	1
P-DYT	2	0	3	0	1
P-HL	2	0	1	0	0
P-MH	2	0	2	0	0
P-MY	24	0	14	0	4
P-NP	36	0	9	1	6
P-SPG	10	1	7	0	1
P-SW	2	0	1	0	0
P-SY	35	2	16	0	5
P-TM	3	0	7	0	0

Table 2.2: Sample Index by Enrichment kits.

and (iii) sex ($p = 0.31$ based on an asymptotic chi-squared test). The p-values obtained are listed in Table 2.8.

The results indicate that ND status is significantly associated with sample index (as expected), enrichment kits and the presence of a variant and not significantly associated with sex and the presence of a questionable variant. To explore further the possibility of enrichment kits as a confounding variable in the association between ND status and the presence of a variant, we checked its association with the presence of a variant. Table 2.9 summarizes the association between enrichment kits and the presence of a variant (p=0.0001 based on an exact permutation test). Even though patients are to be randomized to enrichment kits, on testing the association between the enrichment kits and the presence of a variant, we see a strong association conveyed by a very low p-value. Our exploratory analysis tells us that the enrichment kit is a potential confounding variable to adjust for during our formal statistical analysis.

The bivariate summaries and the association tests were done for all the possible combinations of the variables. Selected summary tables and results for the association tests not shown in the main text can be found in Appendix 1.

Age × categorical variables

We use boxplots to visualize the age distribution by different variables like sample index, ND status, sex, etc. The boxplot in Figure 2.2 suggests that people with an ND disease have a higher average age than those with non-ND disease. The colors in the boxplot represent the two sexes. The figure indicates that the average age for males and females is quite similar

	yes	*no*	*maybe*
EX	0	8	0
P-ALS	93	0	0
P-AMY	0	0	3
P-AX	22	0	0
P-BGW	0	20	0
P-DIV	0	13	2
P-DYT	0	0	6
P-HL	0	3	0
P-MH	0	4	0
P-MY	0	42	0
P-NP	0	0	52
P-SPG	19	0	0
P-SW	0	3	0
P-SY	0	58	0
P-TM	0	10	0

Table 2.3: Sample Index by ND status.

	TCER vs2	Exon v7	TCE(RefSeq)	TCE(R,Ex,Mix)	T Mix
yes	99	2	17	2	14
no	99	2	49	0	11
maybe	42	0	13	1	7

Table 2.4: ND status by Enrichment kits.

	no	yes
yes	112	22
no	151	10
maybe	58	5

Table 2.5: ND status by the presence of a variant.

	no	yes
yes	126	8
no	149	12
maybe	61	2

Table 2.6: ND status by the presence of a questionable variant.

for the ND diseases but, for non-ND diseases, males have a slightly lower average age than females.

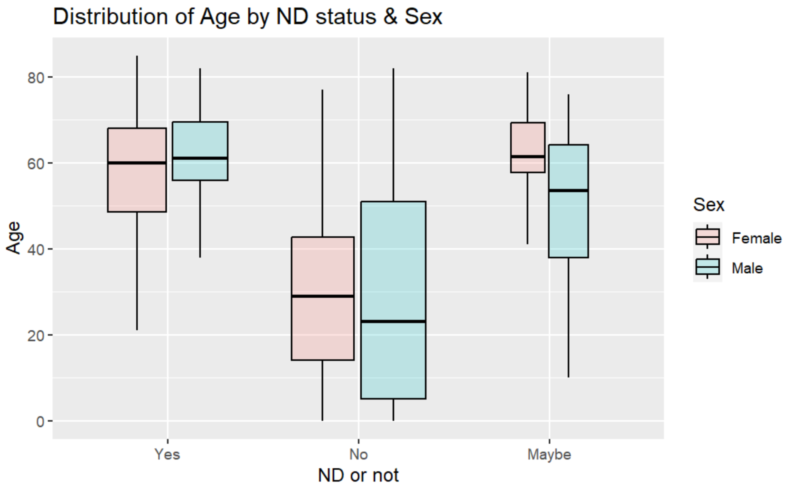

Figure 2.2: Age distribution by ND status and sex

We use an F-test based on a null distribution from parametric bootstrapping (with $B = 1000$ bootstrap replicates) to test the association between the continuous variable, age, and categorical variables such as ND status and clinical information. The bootstrap tests showed significant associations of age with ND status ($p = 0$) and clinical information ($p = 0$) but not with gender ($p = 0.6$), variant presence ($p = 0.29$), enrichment kits ($p = 0.04$), or questionable variants ($p = 0.21$).

To explore whether age was linearly associated with ND status, we apply a generalized additive-logistic model with ND status of *yes* and *maybe* coded as 1 and ND status of *no* coded as 0. We fit a running-average smooth for the predictor 'age' with a span of 1/3 the data for flexibility in capturing potential non-linear relationships. Figure 2.4 shows the resulting fit.

	Female	Male
yes	71	63
no	80	81
maybe	26	37

Table 2.7: ND Status by sex.

Variable 1	Variable 2	p-value
ND status	Sample Index	0.0001
ND status	Enrichment Kits	0.01
ND status	Presence of a Variant	0.02
ND status	Presence of a Questionable Variant	0.54
ND status	Sex	0.30

Table 2.8: Results of tests of association between ND status and different variables.

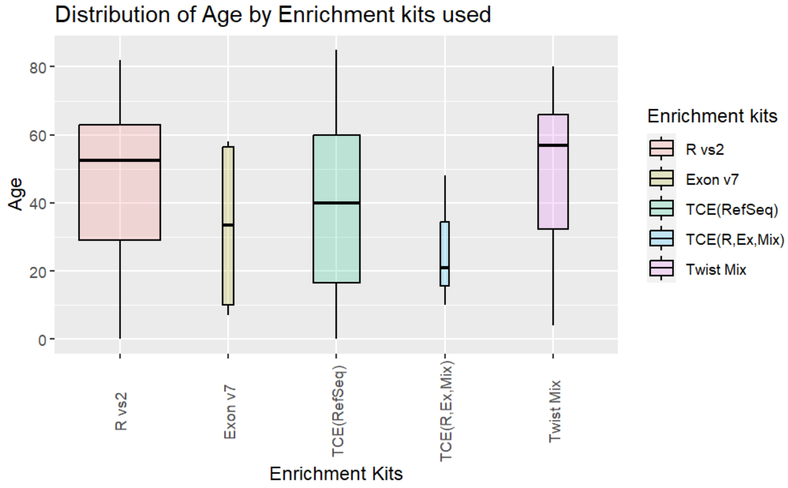

Figure 2.3: Age distribution by the Enrichment Kits used

The linear trend in the plot suggests that a linear term would be sufficient to describe the relationship between age and ND status *yes* or *maybe*.

2.2 The <variants> data

The <variants> data consists of fewer variables than the <persons> data. These variables provide information specific to the different variants such as their genomic position, transcript and protein annotations, functional consequence, relative frequencies in the GnomAD public database of genomic variants, etc. We focus here on just one variable, `func`, describing the predicted functional consequences of the variants.

Table 2.10 summarizes the `func` variable. The *frameshift_variant* category was the most frequently encountered functional consequence followed by *frameshift_truncation*.

	TCER vs2	Exon v7	TCE(RefSeq)	TCE(R,Ex,Mix)	T Mix
no	221	0	67	3	30
yes	19	4	12	0	2

Table 2.9: Presence of a variant by enrichment kits.

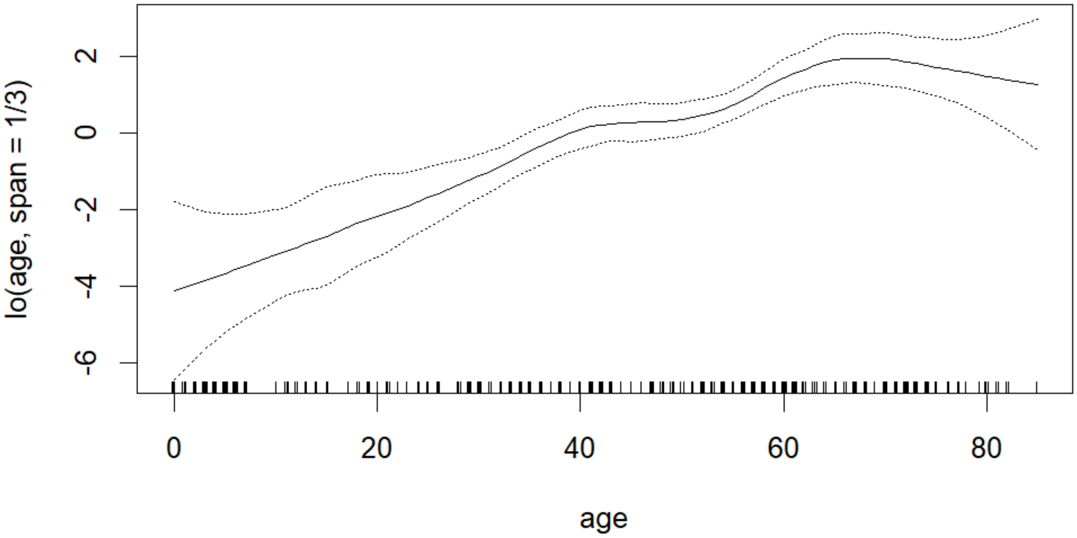

Figure 2.4: Prediction of ND status ‘Yes’ or ‘No’ as a function of age by a generalized additive-logistic model

Type	Count
disruptive_inframe_deletion	5
disruptive_inframe_insertion, direct_tandem_duplication	14
frameshift_elongation	1
frameshift_truncation	15
frameshift_variant	18

Table 2.10: Univariate summary for the func variable

Chapter 3

Analysis

In this chapter, we conduct a formal statistical analysis of the association between neurodegenerative diseases and rare variants in exon 1 of the ATXN2 gene. As discussed in the previous chapter, the data include clinical information from 358 individuals with normal PolyQ repeats. There were 134 cases of neurodegenerative disease, 161 cases of non-neurodegenerative disease, and 63 cases with mixed symptoms where the disease could not be categorized. The dataset containing information on the individuals has 15 variables, including information on whether or not the patient has neurodegenerative disease (ND), the dates of birth and of sampling, the biological sex, the enrichment kit used to sequence the subject's DNA, and information on which of the observed rare variants is carried. The dataset on variants has information about the 19 rare variants observed in the study such as whether or not the variant is thought to be questionable, information on the variant function, the population allele frequency, and the number of allelic copies examined in the gnomAD public database (i.e. the denominator of the population allele frequency).

We use the SNP-set (Sequence) Kernel Association Test (SKAT), which tests the association between a set of SNPS/genes and continuous or dichotomous phenotypes using a kernel regression framework [Wu et al., 2011]. The SKAT package in R ([Lee et al., 2023]) implements this test.

3.1 Preprocessing

After the data exploration, to shape our data for the SKAT package, we perform data preprocessing steps, including:

- Creating covariates matrices (X) to account for demographic and non-genetic variables: Based on investigator input and the results of our exploratory analysis of the data, potential confounding variables for the association between ND status and genetic variants are the variables age, sex, and enrichment kit and so we include these in X. As stated in Chapter 2, we merge certain enrichment kits as suggested by our collaborator. Specifically, two enrichment kits are combined, the 'Twist Comprehensive Exome plus Refseq, Twist Exome 2.0, Twist Mix (Comprehensive plus 2.0)' used for 3 subjects and the 'Twist Mix' used for 32 subjects. In X, they have been grouped into the category coded as 'TCE(R,Ex,Mix)'. We keep the enrichment kit 'SureSelect All Exon v7' coded as 'Exon v7' (4 subjects) separate from the others as it is sourced from a different vendor. We refer to the most common enrichment kit 'Twist Comprehensive Exome Refseq vs2' (240 subjects), coded as 'TCER vs2' in X, as the baseline category for enrichment kits in all our regression analyses.

- Constructing a kinship matrix (K) to account for relatedness among subjects: The data have some subjects that belong to the same families: 4 parent-child trios, 1 father-son duo, and one sibling duo. We can account for kinship in the Gaussian regression framework implemented in the SKAT package, so we create a kinship matrix for this analysis.
- Generating phenotype vector (Y) to indicate the presence of ND diseases: We create the phenotype or the response vector using the ND column from the <persons> data.
- Constructing genotype matrix (Z): Each row of the genotype matrix represents an individual, and each of the 19 columns represents a rare variant in exon 1 of the ATXN2 gene. The matrix was populated with values indicating the copy number of each variant in each individual. Because these are rare variants, the copy number is either 1 or 0.
- Calculation of allele frequencies and imputation of missing values: We use the allele frequencies of the 19 variants reported in the <variants> Excel data file received from our collaborators. The source of these frequencies is the gnomAD database. Two missing allele frequencies were not found in the database which implied zero frequencies were observed in gnomAD-contributed submissions. For our analysis, these 2 frequencies were imputed using the minimum frequency amongst the variants in our study. The missing frequencies were set to be half the minimum frequency.

3.2 SKAT and SKAT-O

Rare-variant association testing plays a crucial role in deciphering the genetic basis of complex traits, particularly with the advent of high-throughput sequencing technologies. Traditional approaches like burden tests, which collapse rare variants into a single genetic variable, suffer from limitations in power, especially in the presence of non-causal variants or when protective and deleterious variants coexist. In response to these challenges, [Wu et al., 2011] introduced the Sequence Kernel Association Test (SKAT) in 2011. The Sequence Kernel Association Test (SKAT) is a statistical method for rare-variant association testing in sequencing data, particularly in genome-wide association studies (GWASs).

SKAT is a regression approach that can assess the association between genetic variants (both common and rare) within a specific genomic region and a continuous or dichotomous trait while adjusting for covariates.

Traditional methods for testing rare variant associations, such as burden tests, often collapse or summarize rare variants within a region into a single value, assuming uniform effects across variants. However, SKAT acknowledges that rare variants may have varying directions and magnitudes of effect on the phenotype, including no effect at all. It uses a multiple regression model to directly regress the phenotype on genetic variants and covariates, allowing for different variant effects.

SKAT employs a kernel association test within a mixed-model framework to assess the regression coefficients of the variants, effectively accounting for rare variants. As it only requires fitting a null model with covariates to calculate p-values using simple analytic formulas, it has good computational efficiency and can also exploit local correlation structures between variants and between subjects, as well as incorporate flexible weights on variants to boost power (e.g., by assigning higher weights to rarer variants). The SKAT methodology implements a score test of the aggregated effects of rare genetic variants in some genomic regions of interest, such as exon 1 of the ATXN2 gene in our analysis.

SKAT-O extends the Sequence Kernel Association Test (SKAT) to optimize power by using the data to adaptively combine the burden test [Madsen and Browning, 2009] and the non-burden variance-component test of SKAT. The test statistic is a linear combination of the burden and SKAT statistics that adapts to the correlation structure of variant effects through a family of kernels and optimally combines the burden and non-burden sequence kernel association tests to maximize power. Burden tests are more powerful when most variants in a region are causal and the effects are in the same direction, whereas SKAT is more powerful when a large fraction of the variants in a region are noncausal or the effects of causal variants are in different directions. The SKAT-O unified test maintains power in both scenarios. For our analysis, we use the default linear kernel in the SKAT package. [Lee et al., 2023] It provides a flexible and efficient approach to identifying associations between rare genetic variants and complex traits, accommodating the inherent variability in variant effects across genomic regions.

3.3 Overview

As already noted, the data has subjects classified as maybe ND because of overlapping symptoms. We therefore work with two sets of data. For the first dataset, we use all the subjects regardless of their diagnosis and combine the subjects who may have a neurodegenerative disease, the *maybe* ND subjects, with subjects who have a neurodegenerative disease, the *yes* ND subjects. The *maybe* and *yes* ND subjects are taken as cases while the *no* ND subjects are taken as controls. For the second dataset, we remove the subjects who may have neurodegenerative disease, the *maybe* ND subjects, and keep only the *yes* ND subjects as cases and the *no* ND subjects as controls.

3.4 Dataset retaining *maybe* ND subjects

As the *maybe* ND subjects are retained in this dataset, the response variable, Y, for the analysis is coded as 1 for ND *yes* and *maybe*, and as 0 for ND *no*.

3.4.1 Null models

We consider two null models to serve as the foundation for subsequent score tests in the SKAT-O framework. Both null models incorporate only the demographic and clinical covariates in X and exclude the genetic variant information Z.

Logistic regression

The first null model that we consider accounts for the binary response variable Y through logistic regression. Importantly, the logistic regression model implemented in the SKAT R package takes only unrelated (i.e. independent) subjects. However, as noted in previous sections, our data contains some related individuals. To ensure independence among the subjects, we remove the children P-SY163, P-SY165, P-SY166 and P-SY169 in the four case-parent trios. We also remove the son from the father-son duo and the younger of the two siblings in the sibling pair. (“P-MY120", “P-MY119"). All these relative clusters involve individuals who do not have a neurodegenerative disease. We keep the older individuals in these clusters so that our non-ND “controls” have ages that better match the older ages of the ND “cases” in our study.

The functions for fitting null models in SKAT do not report the estimated effects of the covariates in X. Since the logistic regression assumes independent (i.e. unrelated) subjects, we can equivalently call the `glm()` function in R to see the effects of age, gender, and enrichment kit. Table 3.1) summarizes the results, which indicate that:

	Estimate	Std. Error	z value	Pr($>|z|$)
(Intercept)	-3.101130	0.425497	-7.288	3.14e-13***
age	0.076831	0.008029	9.1619	<2e-16***
sexMale	0.079097	0.284635	0.278	0.7811
ekit.TCE	-0.846028	0.343175	-2.465	0.0137*
ekit.TwistMix	0.547541	0.521103	1.051	0.2934
ekit.Exon	0.524046	1.469109	0.357	0.7213

Table 3.1: Effect estimates in the logistic regression

- Age is positively associated with neurodegenerative disease($p < 0.001$), indicating that older individuals are more likely to have an ND disease.

- The coefficient for sex (male) is not statistically significant ($p = 0.781$), suggesting no significant difference in the prevalence of ND between genders.
- 'ekit.TCE' is negatively associated with neurodegenerative disease ($p = 0.01$).

As already noted, the 'ekit.TCE' category corresponds to a combination of two enrichment kits in the original <persons> Excel file received from the investigator: 'Twist Comprehensive Exome plus Refseq, Twist Exome 2.0, Twist Mix (Comprehensive plus 2.0)' used for 3 subjects and 'Twist Mix (Comprehensive plus 2.0)' used for 32 subjects. The significant negative association of the response Y with 'ekit.TCE' suggests that patients who don't have neurodegenerative disease tend to be assigned to one of these two enrichment kits. This association is hard to explain as patients are never assigned to any specific enrichment kit, they just end up with a kit that is being used in the lab at that time for all of the patients. As the vendors update the versions of the kits, they are changed in the labs and the most recent version is used. The previous versions of the kits are not used again.

Gaussian regression

The second null model that we fit is with Gaussian regression. This model assumes a continuous Gaussian (rather than binary) response Y but accounts for the relationships among the subjects through the kinship matrix, K.

As the Gaussian regression model implemented in R's `glm()` function incorrectly assumes independence amongst the subjects, the estimates of the effects of the covariates will be biased so we do not bother to call it.

3.4.2 Alternative models

In this section, we add effects for the genetic variants in exon 1 of the ATXN2 gene to our linear models. Therefore, we consider the alternative hypothesis of genetic association against the null hypothesis of no genetic association. We allocate weights to the variants using the allele frequency from the gnomAD database.

Questionable variants

Our collaborator asked us to set the questionable variants c.42del, c.80_85del, and c.39_40del (highlighted in yellow in the original <variants> Excel spreadsheet) to missing in our analyses. We eliminate them from subsequent analyses through a thresholding mechanism based on the sample frequencies. First, we calculate the sample frequencies of these variants which are:

- c.42del: ≈ 0.025
- c.80_85del:≈ 0.0056

- c.39_40del: ≈ 0.020

To exclude these questionable variants from the analysis, we set the missing_cutoff parameter for the variants in the `SKAT()` function to be slightly below the lowest observed frequency, specifically to 0.005. This strategic adjustment ensures the exclusion of questionable variants.

Variant weights

Variant weights are important in the analysis framework. The idea is to upweight rare variants relative to common variants, as population-genetics principles predict they are more likely to be deleterious. We use the population allele frequency of each variant in the publicly available gnomAD database to determine its corresponding weight in the SKAT analysis. Leveraging data from gnomAD, the weights for 19 identified variants are computed. Notably, two variants in our data have no population allele frequencies recorded in gnomAD. We impute the allele frequencies of these variants to be half the minimum gnomAD frequency of the variants with gnomAD frequencies in our dataset.

Logistic regression

Transitioning towards statistical modeling, we apply logistic regression implemented in the SKAT package to the sample of 352 unrelated subjects to assess the association between rare variants in exon 1 of the ATXN2 gene and neurodegenerative diseases. A SKAT-O test based on 100,000 bootstrap replicates under the null hypothesis of no genetic association unveils a significant genetic association ($p = 0.03$) with ND status, following adjustments for potential confounders such as age, sex, and enrichment kit.

Gaussian regression

Expanding the analysis, we apply the SKAT methodology for Gaussian regression to the binary response to account for the familial relationships in the dataset. This time, the analysis encompasses 358 related subjects. Despite the incorrect assumption of a continuous Gaussian response, the results remain consistent with those of the logistic regression, affirming a significant association ($p = 0.02$) between the rare variants in exon 1 of ATXN2 and ND status, after adjusting for age, sex, enrichment kit and familial relatedness.

3.5 Dataset removing *maybe* ND subjects

This section analyses the smaller dataset without subjects of *maybe* ND status. The aim of excluding the *maybe* ND subjects is to remove ambiguity in diagnosis and focus solely on subjects with clear ND status.

We create a new phenotype vector, Y, for this smaller dataset. This vector separates subjects based on their ND status, categorizing them into either *yes* or *no* groups. The *maybe* ND subjects are excluded from this vector, ensuring a more definitive classification. This dataset has 295 subjects with 63 subjects with *maybe* ND status removed. We also set up the covariates matrix (X) and genotype matrix (Z) for this dataset, excluding the persons in the *maybe* ND category.

3.5.1 Null models

We then establish a null model for logistic regression to serve as the foundation for subsequent score tests in the SKAT-O framework. This model incorporates demographic and clinical covariates only. For the smaller dataset excluding the ND subjects, We focus only on logistic regression and do not establish a null model for Gaussian regression.

Logistic regression analysis

On a similar note, as we do for the previous larger dataset, we then use R's `glm()` function to fit a null model with logistic regression to be able to see the effect of the various non-genetic covariates as listed below in Table 3.2.

	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	-2.985171	0.490243	-6.089	1.13e-09***
age	0.052634	0.008225	6.399	1.56e-10***
sexMale	-0.188837	0.259431	-0.728	0.4667
ekit.TCE	0.674749	0.340547	1.981	0.0477*
ekit.TwistMix	0.182780	0.427939	0.427	0.6693
ekit.Exon	-0.163016	1.298819	-0.126	0.9001

Table 3.2: Effect estimates in the logistic regression with unrelated subjects

From the results above, we see that:

- A significant positive association is observed between age and the likelihood of neurological disorder ($p < 0.001$), indicating that older individuals are more likely to have an ND disease.
- The coefficient for sex (male) is not statistically significant ($p = 0.467$), suggesting no significant difference in the prevalence of ND between genders.
- Subjects processed with enrichment kits in the 'ekit.TCE' category shows a statistically significant association with ND ($p = 0.048$). As already mentioned, the 'ekit.TCE' category corresponds to a combination of two enrichment kits in the original <persons> Excel file received from the investigator: 'Twist Comprehensive Exome plus Refseq, Twist Exome 2.0, Twist Mix (Comprehensive plus 2.0)' used for 3 subjects and 'Twist Mix (Comprehensive plus 2.0)' used for 32 subjects.

3.5.2 Alternative models

For the smaller dataset without the *maybe* ND subjects, we apply only the logistic regression analysis under the alternative hypothesis of genetic association. We do not consider Gaussian regression under the alternative hypothesis.

Questionable variants

As described in the previous section, we calculated the sample frequencies of the three questionable variants, c.42del, c.80_85del, and c.39_40del. The sample frequencies in the smaller dataset were,

- $c.42del \approx 0.024$
- $c.80\_85del \approx 0.0067$
- $c.39\_40del \approx 0.02$

To ensure the exclusion of the questionable variants,we set the missing cutoff parameter in the `SKAT()` function slightly below the lowest observed frequency, specifically to 0.005. Note that five variants are removed from the analysis, including two additional variants present only in subjects with ND status *maybe*. These two additional removals are both deletions, 'c.54_58del' and 'c.57_59del'.

Logistic regression analysis

We again apply logistic regression implemented in the SKAT package to the sample of 289 unrelated subjects to assess the association between rare variants in exon 1 of the ATXN2 gene and neurological disorder (ND) status. A SKAT-O test based on 150,000 bootstrap replicates under the null hypothesis of no genetic association unveils genetic association ($p = 0.005 - 0.008$) with ND status, following adjustments for potentially confounding variables such as age, sex, and enrichment kits.

3.6 Summary of results

Our analysis shows that the rare variants in exon 1 of the ATXN2 gene are associated with ND status.

- When subjects who are *maybe* ND are included in the analysis (i.e. *yes* and *maybe* versus *no*), rare variants in exon 1 of the ATXN2 gene are associated with ND status ($p = 0.03$) in a logistic regression analysis of ND status as a binary response. The logistic regression analysis is based on $n = 352$ unrelated subjects and adjusts for age, sex, and enrichment kit as potential confounding variables.

- Again, when subjects who are *maybe* ND are included (i.e. *yes* and *maybe* versus *no*), rare variants in exon 1 of the ATXN2 gene continue to be significantly associated with ND status($p = 0.02$) in a regression analysis of ND status as a Gaussian response. The regression analysis is based on $n = 358$ related subjects and adjusts for age, sex, and enrichment kit as potential confounding variables. The regression analysis of a Gaussian response accommodates the relatedness of four parent-child trios, a father-son duo, and a sibling pair. Still, it incorrectly assumes the binary response is continuous and has a Gaussian distribution.

- When subjects who are *maybe* ND are excluded (i.e. *yes* versus *no*), rare variants in exon 1 of the ATXN2 gene are associated with ND status ($p = 0.005 - 0.008$) in a logistic regression analysis. The logistic regression analysis is based on $n = 289$ unrelated subjects and adjusts for age, sex, and enrichment kit as potential confounding variables.

Chapter 4

Conclusion

We start with an exploratory analysis to understand both the <persons> and <variants> data files that were shared with us. We begin our exploration with univariate summaries, where we examine each variable's unique characteristics and distributions. Through tabular summaries and visual aids such as histograms and box plots, we gain valuable insights into the composition and structure of the data. We identify the key variables: sample index, sex, ND status, enrichment kits, clinical information, age, and presence of a variant, understanding their frequencies, distributions, and potential relationships. We follow this with bivariate summary tables giving us valuable insights into the relationships between various variables in the <persons> data. We use exact association tests between various categorical variables to gain insights about associated variables such as ND status and enrichment kits, enrichment kits and the presence of variant, and ND status and the presence of a variant. We use boxplots to visualize the age distribution with different variables and test the associations using an F-test based on a null distribution from parametric bootstrapping with $B = 1000$ replicates. The bootstrap F-tests indicate a positive association between ND status and age.

The data chapter lays the groundwork for more in-depth multivariate analyses and further research into the interplay between variants and neurodegenerative diseases. The insights gained in this chapter include the identification of the enrichment kit as a potential confounding variable in the association between ND status and the presence of a variant since it is significantly associated with both ND status and the presence of a variant. Another insight is that age is significantly associated with ND status.

Moving on, the analysis chapter explores the association between neurodegenerative disease and rare variants in exon 1 of the ATXN2 gene by performing a formal statistical analysis using the SKAT-O methodology [Lee et al., 2012]. We look at two datasets. The first dataset retains the *maybe* ND subjects and considers them as cases along with the *yes* ND subjects. The second dataset excludes the *maybe* ND subjects from the analysis (it considers the *no* ND subjects as controls and the *yes* ND subjects as cases. Two null models are fit for the first dataset with logistic and Gaussian regression respectively. R's `glm()` function is used to estimate the effects of non-genetic covariates in the logistic regression in the absence of genetic effects. The results of this logistic regression indicate a positive association between age and neurodegenerative disease and a negative association between the 'ekit.TCE' enrichment-kit category and neurodegenerative disease. The alternative models are formulated removing the questionable variants below a threshold for their population frequencies. The variant weights are set using their population allele frequencies in the gnomAD public database. We perform a bootstrap test of genetic association based on a SKAT-O statistic and 100,000 bootstrap replicates under the null hypothesis. The SKAT-O test unveils a significant genetic association ($p = 0.03$) with neurodegenerative disease, following adjustments for potential confounders such as age, sex, and enrichment kit.

Similarly, we analyze our second dataset that excludes the *maybe* ND subjects. We use a score test with 150,000 bootstrap replicates to assess the association between rare variants in exon 1 of the ATXN2 gene and neurodegenerative disease The test under the null hypothesis of no genetic association unveils a significant genetic association ($p = 0.005$-0.008) with ND status, following adjustments for potential confounders such as age, sex, and enrichment kit.

Bibliography

[Chio et al., 2022] Chio, A., Moglia, C., Canosa, A., Manera, U., Grassano, M., Vasta, R., Palumbo, F., Gallone, S., Brunetti, M., Barberis, M., et al. (2022). Exploring the phenotype of italian patients with als with intermediate atxn2 polyq repeats. *Journal of Neurology, Neurosurgery & Psychiatry*, 93(11):1216–1220.

[Egorova and Bezprozvanny, 2019] Egorova, P. A. and Bezprozvanny, I. B. (2019). Molecular mechanisms and therapeutics for spinocerebellar ataxia type 2. *Neurotherapeutics*, 16(4):1050–1073.

[Lee et al., 2012] Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

[Lee et al., 2023] Lee, S., Zhao, Z., with contributions from Larisa Miropolsky, and Wu, M. (2023). *SKAT: SNP-Set (Sequence) Kernel Association Test*. R package version 2.2.5.

[Lubieniecka et al., 2022] Lubieniecka, J., Lubieniecki, K., Bernsen, S., Hippert, Döring, Jovancevic, Fabian, Weydt, and Nguyen (2022). Atxn2 gene variants in neurodegenerative disease. Poster presented at Conference Name.

[Madsen and Browning, 2009] Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384.

[Wu et al., 2011] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Appendix A

Supplementary Tables

A.1 Univariate summary tables

Sample Index	Count
EX	8
P-ALS	93
P-AMY	3
P-AX	22
P-BGW	20
P-DIV	15
P-DYT	6
P-HL	3
P-MH	4
P-MY	42
P-NP	52
P-SPG	19
P-SW	3
P-SY	58
P-TM	10

Sex	Count
Female	177
Male	181

ND status	Count
Yes	134
No	161
Maybe	63

A.2 Bivariate summary tables

	no	yes
EX	8	0
P-ALS	80	13
P-AMY	2	1
P-AX	17	5
P-BGW	19	1
P-DIV	14	1
P-DYT	5	1
P-HL	3	0
P-MH	3	1
P-MY	41	1
P-NP	50	2
P-SPG	15	4
P-SW	1	2
P-SY	53	5
P-TM	10	0

A.3 R Scripts

An RMarkdown file explaining the R functions described in this thesis along with the entire code used can be found on GitHub at https://github.com/SFUStatgen/DJ

	no	yes
EX	8	0
P-ALS	88	5
P-AMY	3	0
P-AX	19	3
P-BGW	19	1
P-DIV	14	1
P-DYT	6	0
P-HL	3	0
P-MH	2	2
P-MY	37	5
P-NP	51	1
P-SPG	19	0
P-SW	3	0
P-SY	54	4
P-TM	10	0

	TCER vs2	Exon v7	TCE(RefSeq)	TCE(R,Ex,Mix)	T Mix
Female	121	2	37	2	15
Male	119	2	42	1	17