Title
The prevalence and effect of expanded repeat alleles in neurological disorders

Author
Chubick, Alex Lee

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The prevalence and effect of expanded repeat alleles in neurological disorders

A dissertation submitted in partial satisfaction

of the requirement for the degree

Doctor of Philosophy in Human Genetics

by

Alex Lee Chubick

2022

ABSTRACT OF THE DISSERTATION

The prevalence and effect of expanded repeat alleles in neurological disorders

by

Alex Lee Chubick

Doctor of Philosophy in Human Genetics
University of California, Los Angeles, 2022
Professor Roel A. Ophoff, Chair

The gap between heritability estimates from genotype data and the heritability estimates from familial studies is known as missing heritability. For example, the heritability of autism spectrum disorder (ASD) is estimated to be 50-90% in twin studies but the disease genes and susceptibility loci identified in studies on the genetic variation of ASD only explain 11% of heritability. A part of the missing heritability of ASD may reside in repetitive regions of the human genome. Which are currently understudied in whole genome-association studies (GWAS) and large-scale sequencing efforts. A subclass of these repetitive regions consist of short tandem repeats (STRs), which are repetitive DNA sequence units composed of 2-6 nucleotides. There are some 40 monogenic disorders for which expansion of STRs are reported to be the cause of disease. With the increasing availability of large whole genome sequence (WGS) data sets, there have also been advances with the detection of computational tools to detect expanded alleles at STR loci. One of these methods is called ExpansionHunter, which aims to accurately detect and measure repeat expansions longer than the short sequencing read length itself. We used ExpansionHunter to identify FMR1 and C9orf72 repeat expansions in the WGS data of 20,576 samples from the largest available dataset of families affected by ASD, the Simons Simplex

Collection. We also used WGS data from 1,359 healthy control subjects in the Australian Medical Reference Genome Bank to establish the repeat expansion prevalences for FMR1 and C9orf72 independently from ASD. We observed 180 samples with repeats expanded into the premutation range of FMR1 and 50 samples with repeats in the pathogenic range of C9orf72. We also submitted 7 pedigrees with at least one family member with an expanded repeat in the FMR1 or C9orf72 gene and validated the repeat lengths by PCR. The prevalence of the FMR1 premutation was observed to be significantly higher in ASD, however, no significant difference was observed in the prevalence of the C9orf72 pathogenic repeat. Furthermore, neither the FMR1 or C9orf72 repeat expansion was observed to have an effect on ASD susceptibility. We also used the familial structure of the ASD cohort to establish inheritance patterns of the alleles at the two repeat loci. For FMR1, we observed instances of both transmission of premutation alleles and the occurrence of *de novo* events leading to expanded alleles in the offspring. For the c9orf72 hexanucleotide repeat locus, we only observed parental transmission for expanded alleles of the C9orf72 repeats. In an analysis specific to the C9orf72 repeat, we identified risk haplotypes previously reported to be associated with the expanded C9orf72 repeat. We observed no evidence that expanded repeats contribute to ASD disease susceptibility. However, the repeat expansion prevalences that we observe in this dataset are higher than reports in the general population, which suggest repeat expansions may be more common than previously thought. Overall, our analyses provide further insight on the prevalence of pathogenic repeat expansions in large population samples and their contribution to neurological disorders.

The dissertation of Alex Lee Chubick is approved.

Esteban C. Dell'Angelica

Paivi Elisabeth Pajukanta

Martina Hildegard Wiedau

Noah A. Zaitlen

Roel A. Ophoff, Committee Chair

University of California, Los Angeles

2022

This work is dedicated to my Father, Sister, and the loving memory of my mother.

Table of contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

There are many people that I want to thank and acknowledge for their contributions through the course of my PhD. None of this would have been possible without my father, my sister, and her family who have always supported me and my career. I also want to thank my mother, who passed too soon, but always pushed me to better myself and would have been proud of this accomplishment. Through my work in the Ophoff lab I have also been fortunate to meet so many people including Merel Bot, Toni Boltz, Marcelo Francia, Lingyu Zhan, Lianne Reus, Carolinne Alvarado, Kevin Wojta, Loes Olde Loohuis, Anil Ori, and Catherine Krebs who helped develop a welcoming environment to learn in and were a positive force in my PhD. I am also grateful to James Boocock, Arun Durvasala, and Colin Farrell for their friendship and encouragement that helped get me to this accomplishment. I want to recognize all the people that I have met in the Department of Human Genetics, Biological Sciences Council, Graduate Student Association, and the UAW who I have worked with in my PhD to make graduate school a fair and equal place for all. I also want to thank Evan Wang, Michael Thompson, Zhongan Yang, Cora Au, and Wayne Grody who made the validations and other analyses in my thesis possible. I also want to thank the members of my committee, Esteban Dell'Angelica, Paivi Pajukanta, Martina Wiedau, and Noah Zaitlen whose advice was instrumental in my development as a researcher. My final thanks goes to my PhD advisor Roel Ophoff, who welcomed me into his lab and provided experiences that not only bettered myself as a scientist but also as a person.

VITA

EDUCATION

BA Biochemistry and Molecular Biology, Gustavus Adolphus College, US 2013

PRESENTATIONS & POSTERS

American Society of Human Genetics (ASHG) Annual Meeting *Oct 2020*

"Enrichment of FMR1 premutation alleles in female probands with autism spectrum disorder detected through whole genome sequence analysis" (Poster)

World Congress of Psychiatric Genetics (WCPG) *Oct 2020*

"Enrichment of FMR1 premutation alleles in female probands with autism spectrum disorder detected through whole genome sequence analysis" (Poster)

American Society of Human Genetics (ASHG) Annual Meeting *Oct 2018*

"Analysis of rare coding variants in ALS C9orf72 HRE Carriers" (Poster)

International Plant and Animal Genome Conference XXIV *Jan 2016*

"Genotyping by sequencing through Ion AmpliSeq technology: a tool for genetic trait selection" (Talk)

Introduction

Over the past two decades genomics research has seen an exponential increase in the number of human genomes that have been sequenced[1–3]. This collection of genetic data has enabled studies to determine how the genetic variation between people and populations may affect susceptibility to disease. For example, Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) that are associated with many complex diseases and/or traits[4]. While these studies have led to gene discovery for human diseases that have a considerable genetic component, many of the genetic variants that have been identified only explain a small proportion of heritability, which is the proportion of the disease and/or trait that is explained by genetic variation[5,6]. A recurrent issue of GWAS, however, is that the explained SNP heritability is substantially lower than the heritabilities estimated in familial studies[7,8]. For example, familial studies of schizophrenia provide evidence of an estimated heritability of 80% but the SNPs identified by GWAS are only able to explain approximately 20% of the genetic variation[9]. To further characterize this missing heritability, studies have begun to analyze other complex variants such as short tandem repeats (STRs)[10,11].

STRs are DNA sequences composed of 2-6 nucleotides that are repeated directly adjacent to each other[12]. Approximately 3-6% of the human genome consists of discrete STR loci that are prone to frequent mutations and can be highly polymorphic[13,14]. While the variability of these loci have allowed previous studies to utilize STRs for genetic mapping, expansion of these repeats at some loci have been found to cause over 40 monogenic disorders[15]. The majority of diseases caused by repeat expansions are observed to affect the nervous system such as the CGG trinucleotide repeat expansion in the FMR1 gene that causes Fragile X Syndrome and the GGGGCC hexanucleotide repeat expansion in the C9orf72 gene that causes amyotrophic lateral

sclerosis/frontotemporal dementia[16,17]. Discovery of these disease mechanisms led to hypotheses that repeat expansions may also be causal in autism spectrum disorder (ASD) and other severe mental illnesses. The initial studies exploring this hypothesis found no evidence for repeat expansions being a major risk factor for ASD and psychiatric disorders[18–21]. However, technical limitations for repeat detection as well as small sample sizes of these studies have hampered these efforts.

Historically, Southern blotting and/or PCR-based assays have been used as gold standards for detecting repeat expansions but both methods have their limitations[22]. For example, Southern blotting requires very large amounts of DNA (~10 μg) per sample, while PCR-based assays such as repeat primed PCR may be impacted by indels and variants present in the DNA sequence of primer binding sites, which can lead to false negative findings[23,24]. Additionally, the actual repeat lengths measured by repeat primed PCR are oftentimes imprecise[23]. The methods used in Southern blotting and repeat primed PCR are also very labor intensive, which makes these approaches less viable for large, population-based studies[23,24]. The increase in available whole genome sequencing (WGS) data has also led to the development of standard variant calling pipelines, which are capable of identifying STRs. However, most variant calling pipelines are only capable of identifying STRs up to 50 base pairs (bp) long and are limited to the short read length used in sequencing (~150 bp). This makes it impossible to identify disease-causing repeat expansions which are observed to have repeats in the length of hundreds to thousands of repeat units[25–31].

To address the issues of previous repeat expansion detection methods and the increasing rates of available WGS data, computational methods have been developed to accurately detect and measure repeat expansions longer than the sequencing read length itself [32–34]. Utilization of these tools have allowed studies to reassess the effect that repeat expansions have on neurological

disease and disorders. Recent studies have been able to use these computational repeat expansion detecting methods to identify complex pathogenic repeat expansions that may increase disease susceptibility in several neurological and neuropsychiatric diseases[35–38] . In our own study, we utilized one of the computational repeat expansion detecting methods, ExpansionHunter, to further examine whether specific repeat expansions play a role in autism spectrum disorder [32,33].

ASD is a neurodevelopmental disorder characterized by deficits in social communication and repetitive patterns of behavior [39]. The heritability of ASD has been estimated to be between 50-90%, which means up to 90% of ASD disease variation can be explained by genetic variation[40]. While hundreds of disease genes and susceptibility loci have been identified in studies on the genetic variation of ASD, they explain approximately 11% of heritability [41,42]. To determine if the part of the missing heritability in ASD could be attributed to complex variants such as STRs and their expansions, we used ExpansionHunter to identify repeat expansions at several loci in the WGS data of 20,576 samples from the Simons Simplex Collection, the largest publicly available dataset of families affected by ASD [42,43] .

In our analysis, we validated the computational method by processing 118 subjects with known repeat expansion status through ExpansionHunter and confirmed that we could correctly determine the status of the samples [32]. After confirmation that we could accurately detect expanded repeats within samples we processed the ASD probands and their family members through ExpansionHunter. We identified a subset of pedigrees with at least one family member with an expanded repeat in the FMR1 (4 pedigrees) or C9orf72 (3 pedigrees) gene and validated the repeat lengths by PCR. Based on our findings we determined the prevalence and effect of expanded FMR1 and C9orf72 repeats in the ASD cohort. WGS data from 1,359 independent healthy subjects in the Australian Medical Reference Genome Bank (MRGB) were used to compare the repeat

expansion prevalences we observed in the parents of the cohort[44]. We also utilized the familial structure of the dataset to establish patterns of paternal and maternal inheritance as well as to identify putative *de novo* repeat expansion events. In a final analysis specific to C9orf72, we phased haplotypes associated with expanded C9orf72 repeats.

While ExpanionHunter calculated repeat length genotypes for multiple loci in the Simons Simplex Collection samples, we observed that most samples possessed non-expanded repeats at these loci. However, we did observe 180 samples with repeats expanded into the premutation range of the FMR1 trinucleotide repeat. Alleles with more than 200 repeat units are considered pathogenic and are observed in cases of Fragile X Syndrome (FXS)[45]. FXS symptoms such as avoidance of eye contact, social withdrawal, communication difficulties, and repetitive behaviors are observed to overlap with ASD, which should come as no surprise as FMR1 pathogenic repeats are considered the most common monogenic cause of ASD[46–48]. A premutation range exists between 55 and 200 repeats and males and females who possess these repeats have been observed to develop Fragile X-associated tremor/ataxia syndrome and females have also been observed to develop Fragile X-associated primary ovarian insufficiency[49,50]. While premutation range repeats are associated with increased instability in the gametes of a parent, which may result in further *de novo* allelic expansions transmitted to the offspring, the effect that FMR1 premutation repeats have on ASD remains unclear[46,49,51–53]. In Chapter 1, we explain in greater detail the contribution of FMR1 premutation repeats on ASD .

We also observed 50 samples in the Simons Simplex Collection with repeats in the pathogenic range of the C9orf72 hexanucleotide repeat. Expansion of the C9orf72 repeat has been reported to be the most common genetic cause of amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD)[54,55]. Patients are reported to carry hundreds to thousands of repeat

units, however, the lower limit has been reported to be at 30 repeats[56,57]. Both ALS and FTD are neurodegenerative diseases, however, ALS results in degeneration of motor neurons that affect the movement of the body and FTD is characterized by neuronal degeneration in the frontal and temporal brain lobes that affects cognitive function, creating a potential link to psychiatric disorders[58–60]. Furthermore, behavioral variant FTD, the most common form of FTD, is also reported to have clinical similarities to ASD such as alterations in social interaction, obsessive-compulsive traits, mental inflexibility, and stereotypy of speech[61]. While symptomatic overlap is observed between C9orf72-mediated ALS/FTD and psychiatric disorders, it remains unclear if other genetic variants contribute to this overlap[38]. In Chapter 2, we perform an in depth analysis to determine the effect of C9orf72 repeat expansions on ASD and identify other genetic variants that may explain shared genetic risk between C9orf72-mediated ALS/FTD and ASD.

Furthermore, we provide a framework on how tools such as ExpansionHunter can be used to detect repeat expansions in WGS data to shed light on the prevalence of pathogenic repeat expansions in large population samples and determine their effect on complex diseases and traits. In Chapter 3, we evaluate our use of ExpansionHunter in these analyses and suggest future directions.

Chapter 1: Large-scale whole genome sequence analysis of >22,000 subjects provides no evidence of FMR1 premutation allele involvement in autism spectrum disorder

Alex Chubick,[1] Evan Wang,[4] Cora Au,[2] Wayne W. Grody,[1,2,3] Roel A. Ophoff[1,4,5]

1) Department of Human Genetics, University of California Los Angeles, Los Angeles, CA USA

2) Department of Pathology & Laboratory Medicine, University of California Los Angeles, Los Angeles, CA USA

3) Department of Pediatrics, University of California Los Angeles, Los Angeles, CA, USA

4) Center of Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, CA USA

5) Department of Psychiatry, Erasmus University Medical Center, Rotterdam, The Netherlands.

**The contents of Chapter 1 have been submitted to the American Journal of Human Genetics and is currently under review. To ensure credit is given to those involved in this work, the author list is included in Chapter 1.*

Introduction

An expansion of a CGG trinucleotide repeat in the 5' UTR of the FMR1 gene on the X chromosome is the cause of Fragile X Syndrome (FXS)[62]. Unaffected individuals in the general population carry FMR1 repeat alleles ranging from 5 to 40 repeats, while repeat lengths of 200 or more are pathogenic, resulting in FX[45]. Two intermediate allelic ranges of 41-54 and 55-200 repeat units defined as the gray zone and premutation range, respectively, have also been identified and are associated with increased instability in the gametes of a parent, that may result in further *de novo* allelic expansions transmitted to the offspring[49,50]. While carriers of FMR1 premutation alleles do not develop FXS, 40% of male and 16-20% of female premutation carriers suffer from fragile X-associated tremor/ataxia syndrome (FXTAS) and 20% of female carriers are reported to develop fragile X-associated primary ovarian insufficiency (FXPOI)[63]. The reported frequency of

FMR1 premutation range alleles in the general population is estimated to be roughly 1 in 300 for females and 1 in 900 for males[64–68]. A number of relatively small clinical and community-based studies of FMR1 premutation carriers have suggested that the FMR1 premutation may increase the risk for developing symptoms related to social anxiety, depression, and attention deficit disorder (ADHD)[46,52].

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by deficits in social communication and repetitive patterns of behavior[39]. Through twin studies the heritability of ASD has been estimated to be between 50-90% and to date more than 100 disease genes and susceptibility loci have been identified[40,69]. ASD patients may present additional medical complications such as intellectual disability or epilepsy, characteristics that are also core observations in patients with FXS[70]. It is therefore no surprise that the pathogenic FMR1 trinucleotide repeat expansion that underlies FXS is also reported to be the most common monogenic cause of ASD[47]. However, diagnostic studies of FXS testing in males with ASD have shown pathogenic FMR1 repeats may be less common in cases of developmental delay, intellectual disability, and ASD[71]. To address this more systematically, we performed an analysis of the FMR1 trinucleotide repeat, in the largest available whole-genome sequence (WGS) datasets of ASD families with a total of 20,578 subjects, collected through the Simons Foundation Autism Research Initiative (SFARI) and the Simons Foundation Powering Autism Research (SPARK) initiative, resources of the Simons Simplex Collection[72]. These families (n=5,467) consist of an ASD proband with both biological parents, and in many instances also an unaffected sibling. We used an independent WGS control sample of 1,359 independent healthy subjects of the Australian Medical Reference Genome Bank (MRGB) to examine the prevalence of FMR1 premutation

alleles as well as a sample of 118 subjects with known FMR1 repeat expansion status to confirm/establish the computational pipeline.

We performed large-scale WGS-based analyses using the computational tool ExpansionHunter[32,33], to examine the FMR1 CGG trinucleotide repeat in 22,053 subjects (Figure 1.1). While we did not observe a significant increased burden of premutation alleles in either male or female ASD probands versus unaffected siblings, we did observe an increased overall prevalence of FMR1 premutation alleles in the SFARI and SPARK female subjects compared to independent control subjects and previous literature findings. While positive controls (i.e. FXS patients) were correctly identified through analysis of WGS data, molecular validation in ASD families suggests a degree of overestimation of computationally predicted FMR1 repeat allele sizes. Overall, we observed stable transmission of FMR1 alleles and our findings suggest that FMR1 expanded alleles in the premutation range do not play an important role in ASD susceptibility.

Materials and Methods

Samples with WGS data

To confirm that our protocol would correctly detect FMR1 repeat expansions through analysis of whole genome sequence (WGS) data using ExpansionHunter, we obtained PCR-free HiSeqX WGS data on 118 samples with triplet repeat expansions (premutation and full expansions) via the European Genome-phenome Archive (EGA, accession number EGAS00001002462). These samples were previously used for the development of ExpansionHunter to demonstrate the ability to call large repeats from high throughput, PCR-free WGS data[32]. The DNA samples were obtained from Coriell Repository representing subjects with

validated repeat expansion including 32 patients with Fragile X Syndrome (FXS). As was previously described, samples were acquired from the Coriell Institute and sequenced at an average 45x coverage[32].

We also obtained access to WGS data of samples from the Simons Foundation Autism Research Initiative (SFARI) of the Simons Simplex Collection, the largest ASD collection of trio (proband with both biological parents) and quad (proband with unaffected sibling and both biological parents) families as described before[72]. Briefly, families admitted into SFARI were required to meet metrics related to age such as the proband being between the age of 4 years and 17 years and 11 months when the data was collected and also meeting several diagnostic criteria based on Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS) scores. Probands with known FXS diagnosis, a known genetic risk factor for ASD, were excluded from the repository[72]. Sequence alignment files (CRAM format) from 9,031 samples representing a combination of 2,380 trio and quad families containing WGS reads aligned to the hg38 reference genome were obtained from SFARI base (https://base.sfari.org). WGS was performed at 30x coverage on genomic DNA extracted from whole blood as described before[72].

Our analysis also included samples from the Simons Foundation Powering Autism Research (SPARK) initiative, an additional dataset from the Simons Simplex Collection aimed to create the largest recontactable research cohort of families affected with ASD in the United States for longitudinal phenotypic and genomic characterization research studies[42]. Briefly, families admitted into SPARK were required to meet three criteria: 1) have at least one family member with an ASD diagnosis, 2) currently live in the United States and 3) able to read and speak English[42]. Samples recruited into the dataset were enriched for affected individuals whose parents were also available to participate. Participants registered for SPARK online

(www.SPARKforAutism.org) or at a clinical site by completing questionnaires on medical history and social communication, meaning case status in SPARK is based on patient/parent report[42]. Sequence alignment files (CRAM format) from 11,545 samples representing a combination of 3,087 trio and quad families containing WGS reads aligned to the hg38 reference genome were also obtained from SFARI base (https://base.sfari.org). WGS was performed at 30x coverage on genomic DNA extracted from whole blood as described before [73].

For independent reference samples, we utilized the WGS data of 1,359 subjects from the Australian Medical Reference Genome Bank (MRGB) available via the European Genome-Phenome Archive (accession code EGAD00001005095). Sequence alignment files (CRAM format) were downloaded to determine the frequency of the FMR1 trinucleotide repeat allele within the MRGB. Data from the MRGB were generated from individuals of European descent, aged 60-95 years, and confirmed to be healthy with no reported history of cancer, cardiovascular disease, or dementia[44]. The MRGB samples were sequenced at an average of >38x coverage. The WGS data was aligned to the hs37d5 reference genome with decoys, with no further processing applied [44].

Genotype assessment of the CGG trinucleotide repeat at FMR1.

The length of the CGG trinucleotide repeat in the FMR1 gene was determined by processing WGS alignment files (formatted in CRAM or BAM format) from the European Genome-Phenome Archive and SFARI base through ExpansionHunter v5.0.0 (https://github.com/Illumina/ExpansionHunter)[32,33]. Following the established protocol, files processed by ExpansionHunter were run with the reference genome assembly matching the reference of the aligned reads and the variant catalog associated with that reference[32,33].

Additionally, as FMR1 is located on the X chromosome, ExpansionHunter was run with the --sex option as it specifies the sex of the sample and affects the estimated genotype on sex chromosomes. An output JSON and VCF file containing the genotype of FMR1 were generated for each of the processed files . Additionally, ExpansionHunter estimates a confidence interval for each allele of the genotype using a parametric bootstrap method and the information is included in the output files[32,74].

We also applied Hardy-Weinberg equilibrium testing in females for quality control. Moreover, we used the available pedigree information to examine compatibility with paternal and maternal inheritance patterns. We also utilized a Z-test to determine if there were any differences between the age of parents with offspring who were observed to be *de novo* for FMR1 premutation alleles and the age of parents with offspring who inherited premutation range repeat alleles.

PCR validation of FMR1 premutation carriers and their pedigrees

PCR validation was performed by our collaborators Cora Au and Dr. Wayne Grody. CGG repeat lengths in the FMR1 genes of the subjects were determined in genomic DNA extracted from whole blood by standard methods[72]. They used the AmplideX FMR1 PCR reagents (Asuragen, Austin, TX) for amplification and capillary electrophoresis on the ABI 3730 analyzer (ThermoFisher Scientific, Waltham, MA), according to manufacturer specifications. As explained by our collaborators, the technique uses repeat-primed PCR (32 cycles) to robustly amplify and accurately size FMR1 CGG repeats ranging from the normal alleles through premutation alleles to full expansion mutations. They performed the analysis of the capillary electrophoresis profiles using GeneMapper software, Microsatellite subroutine (ThermoFisher Scientific, Waltham, MA),

with final results being expressed in base-pair length, trinucleotide repeat number, and normal/premutation/full mutation range based on the accepted cut-offs discussed above.

Repeat burden statistical testing

The burden of repeats in the premutation range of FMR1 in probands compared to unaffected siblings was determined using Fisher's exact test. Since FMR1 is located on the X chromosome and because the prevalence of FMR1 premutation range alleles has been reported to be different in male and females, we performed sex-specific analyses in which we compared male probands to unaffected male siblings and female probands to unaffected female siblings.

Results

Validation of FMR1 Repeat Alleles Calculated by ExpansionHunter

ExpansionHunter calculated genotypes for all 118 CRAM files downloaded from the European Genome-phenome archive (EGA). We used the FMR1 repeat lengths calculated by ExpansionHunter to identify the 34 samples that were listed in the Coriell database to have repeats exceeding the normal range of FMR1 (Figure 1.2), of which 17 were of pathogenic repeat length. We observed that ExpansionHunter correctly identifies expanded repeat carriers but underestimates the length of repeats in the pathogenic range (Figure 1.3). In the 84 samples observed in the normal range/gray zone by both PCR and ExpansionHunter, we observed a discrepancy in a single sample observed with an allele size in the gray zone by PCR but in the premutation range by ExpansionHunter. Overall, we observe that ExpansionHunter underestimates the repeats of alleles in the pathogenic range and in rare instances overestimates repeat lengths of shorter alleles.

We selected four families (11372, 11676, 13390, and 14489) from the SFARI dataset with at least one member with a premutation range allele, for PCR validation (Table 1.1). Among the four families, seven samples were predicted to be premutation carriers by ExpansionHunter. Based on the repeat lengths determined by PCR, three samples were observed to have repeats in the gray zone and the remaining samples were observed to have normal range repeats. Our comparison of repeat lengths determined by ExpansionHunter and PCR of SFARI samples show that ExpansionHunter may overestimate repeat lengths at the FMR1 locus.

Off-target reads were observed to contribute to these overestimated read lengths and we reprocessed the PCR tested samples through ExpansionHunter only using the FMR1 reference region. The results showed that the newly calculated length of the FMR1 repeat alleles lengths were much closer to the repeat lengths determined by PCR. We also reprocessed the 110 samples from the EGA through ExpansionHunter only using the FMR1 reference region and observed that the calculated repeat lengths were based on reads only from the X chromosome. To reduce the likelihood of overestimating the repeat length of samples, we reprocessed samples originally determined to have FMR1 premutation repeats and calculated their repeat length only using the FMR1 reference region.

Prevalence of Individuals with FMR1 Repeat Alleles

ExpansionHunter calculated genotypes for 9,016 of 9,031 (99.8%) CRAM files that were downloaded from SFARI base. ExpansionHunter does not calculate genotypes for samples with less than 10x coverage or do not possess reads that span the regions in the variant catalog for the FMR1 locus. We did not observe any samples with repeats that extended into the pathogenic range, which is not surprising as families that were identified to have family members with FXS were

excluded from the dataset. We also identified 102 samples with repeats in the premutation range of FMR1 (Figure 2). Father and male offspring that were carriers of the FMR1 premutation were observed at a prevalence of 0.93% and 0.71%, while mother and female offspring premutation carriers were observed at a prevalence of 1.78% and 1.28% (Table 1.2). The prevalences of FMR1 premutation carriers in the parents of the Simons Simplex Collection compared to the prevalences observed in samples from the MRGB were significantly higher for mothers (p-value = 4.057e-05) and nominally significant for fathers (p-value = 0.010). We found no evidence that the observed genotype distribution of the FMR1 premutation in females, diploid for the X chromosome, deviated from what is expected based on the Hardy-Weinberg equilibrium (p-value = 0.89, χ^2= 0.24). Furthermore, we observed no evidence for bias of FMR1 premutation allele frequencies in the different sequencing batches in the Simons Simplex Collection.

We used ExpansionHunter to calculate an additional 11,520 genotypes from 11,545 (99.8%) CRAM files from the SPARK dataset. We observed no samples with repeats in the pathogenic range of FMR1. We also identified 78 samples with repeats in the premutation range of FMR1 (Figure 3). Father and male offspring that were carriers of the FMR1 premutation were observed at a prevalence of 0.62% and 0.55%, while mother and female offspring premutation carriers were observed at a prevalence of 0.71% and 0.97% (Table 1.2). We also observed higher prevalences of FMR1 premutation carriers in the mothers of the SPARK dataset compared to the prevalences observed in samples from the MRGB, albeit at nominal significant levels (p-value = 0.02) while the prevalence in fathers was observed to be nonsignificant (p-value = 0.25). Again, we observed no deviation from what is expected from the Hardy-Weinberg equilibrium (p-value = 0.96, χ^2= 0.08) in female carriers and no evidence for bias of FMR1 premutation allele frequencies in the different sequencing batches.

Burden of FMR1 premutation range alleles on ASD in males and females

In the comparison of male probands and unaffected male siblings, we did not observe a significant difference in allele frequencies (p-value = 0.48, OR = 0.71). We also observed no significant difference in the allele frequencies (p-value=0.27, OR=1.73) between female probands and unaffected female siblings. In the SPARK dataset, no significant difference was observed between the allele frequencies of male proband and unaffected male siblings (p-value = 0.34, OR = 1.81) or female probands and unaffected female siblings (p-value = 0.62, OR = 1.30). We also performed a burden test combining the allele frequencies observed in SFARI and SPARK and no significant differences were observed between the allele frequencies of male probands and unaffected male siblings (p-value = 0.99, OR = 1.08) or female probands and unaffected female siblings (p-value = 0.35, OR = 1.43).

Parental transmission of FMR1 premutation range alleles

We utilized the FMR1 repeat length genotypes from the trio and quad families to establish patterns of paternal and maternal inheritance as well as for the identification of putative *de novo* repeat expansion events. There were 79 families (including 62 quads and 17 trios) with 141 offspring of which 32 (22.7%) inherited the premutation allele from one of their parents (Table 1.4). Transmissions from mother to male offspring were observed two times more often than transmission from mother to female offspring, which corresponds to the ratio of male and female offspring in the SFARI dataset (Table 1.2). Additionally, we observed 5 offspring (3.5%) with premutation range alleles that are best explained by *de novo* expansion of the FMR1 repeat (Table 1.4). While we observed that offspring with the premutation more commonly inherit the allele

from one of their parents instead of *de novo* expansions occurring, we also observed 42 offspring 29.8%) who did not inherit a premutation allele from one of their parents who was identified to be a carrier (Table 1.4). No father to male offspring transmission was observed as expected by X-linked inheritance (Table 1.4).

In the SPARK dataset, we observed within 58 families (including 21 trios and 37 families with 4 or more members) that 34 offspring out of 99 (34.3%) inherited the premutation allele from one of their parents (Table 1.5). Similar to the Simons Simplex Collection, male probands constitute a majority of the offspring in the SPARK dataset (Table 1.3). In correspondence, we observed the transmission from mother to male proband to be the most common form of transmission from parent to offspring (Table 1.5). We also observed 3 offspring (3.03%) where expansion of the FMR1 repeat allele occurred *de novo* (Table 1.5). In contrast, we observed 26 offspring with premutation allele sizes (26.3%) who showed no inheritance of premutation range alleles from their parents (Table 1.5). The patterns of inheritance that we observed in the SPARK dataset were further validated by the fact that no father to male offspring transmissions were observed (Table 1.5).

In a combined analysis based on the transmission patterns in the Simons Simplex Collection and SPARK we compared the age of parents who transmitted premutation alleles to their offspring to the age of parents with offspring that possessed premutation alleles best explained by *de novo* expansion. We observed no significant differences in the age of the parents (p-value = 0.98) between the two groups of offspring.

Discussion

In a large-scale and systematic study of whole genome sequencing (WGS) data of nuclear families with autism spectrum disorder (ASD) affected probands and unaffected siblings, we found

no evidence of increased burden of FMR1 premutation range alleles in male or female ASD patients compared to unaffected siblings. At the same time, we observed an increased number of FMR1 premutation carriers compared to previous studies, which reported a population prevalence of FMR1 premutation carriers at 0.12% (with 95% confidence interval of 0.06-0.19%) in males and 0.34% (with 95% confidence interval of 0.06-0.83%) in females[75]. We observed a six-fold higher prevalence of 0.72% in fathers and a four-fold higher prevalence of 1.47% in mothers (Table 1.2 and 1.3). The prevalence of female FMR1 premutation carriers in the Simons Simplex Collection was also observed to be significantly higher than the prevalence observed in the Medical Reference Genome Bank (MRGB), which served as an independent control sample of healthy subjects of advanced age. Furthermore, the prevalence of FMR1 premutation carriers that we report in the Simons Simplex Collection is similar to the prevalence reported in gnomAD (males = 0.70% , females = 0.98%)[76]. Our results should be further validated in additional independent datasets to determine if the prevalence and effect of the FMR1 premutation is larger than what has been reported in the general population.

We also used PCR to validate the FMR1 premutation repeats identified in a subset of samples (i.e. four quad families) from the SFARI dataset. None of the tested samples were observed to have alleles in the premutation range of FMR1. Furthermore, in the samples that were identified to have premutation range alleles, only three were detected to have an allele in the gray zone of FMR1. There may be a number of reasons as to why we observe a discrepancy in the repeat lengths calculated by ExpansionHunter. ExpansionHunter has been reported to overestimate the length of some repeats[76]. We also observed that the confidence intervals estimated by ExpansionHunter broaden with increased repeat unit length, which could explain the overestimated repeat lengths calculated in the Simons Simplex Collection. Moreover, mosaicism,

higher sequencing error rates, and GC biases are factors that can contribute to this discrepancy at this locus[32,33]. Studies have also shown that the use of off-target regions in the variant catalog used in the ExpansionHunter algorithm can lead to the overestimation of some genotypes[76]. We observed that off-target reads from chromosomes 2, 6, and 7 lead to an overestimation of the genotypes calculated by ExpansionHunter in the PCR tested samples (Table 1.1), which suggests similar repeat sequences exist in these chromosomes and should be further explored to determine if they have an effect on ASD or other traits. Evidence suggests the number of individuals carrying premutation range alleles identified by ExpansionHunter is overestimated. However, further analyses are needed to identify the factors that lead to these discrepancies and better determine the frequency of premutation alleles within the population.

The stability and consistency we observed in the length of alleles transmitted from parent to offspring in the Simons Simplex Collection suggests that ExpansionHunter does accurately calculate the repeat length from WGS data. Utilizing the family structure within the Simons Simplex Collection, we established whether FMR1 premutation range alleles in the offspring were inherited from the father or the mother, or arose *de novo*. We observed that offspring with premutation alleles more commonly inherited the alleles from one of their parents versus the expansion occurring *de novo* (Table 1.4 and 1.5). Additionally, we observed that the age of the parents had no effect on whether a premutation allele was transmitted from the parent or occurred *de novo* in an offspring. Among the transmissions that we observed mother to male proband were the most common type of transmission. The commonality of this transmission pattern is best explained by the preponderance of male probands in the Simons Simplex Collection and the X-linked inheritance pattern of FMR1. We also observed instances where the offspring in families with parents who carried an FMR1 premutation allele did not inherit the allele from their parents

(Table 1.4 and 1.5). The transmission of normal range alleles over premutation alleles or *de novo* expansion of the allele could provide insight on the selective pressures that affect expanded alleles and should be explored in other family cohorts.

In the Simons Simplex Collection, we also note that the observed FMR1 premutation genotype distribution in females (with diploid genotypes) did not violate the expected distribution according to the Hardy-Weinberg equilibrium, which suggests that there is no inherent technical deviation in the detection of premutation range alleles in females using ExpansionHunter. Furthermore, the full mutation of the pathogenic FMR1 repeat expansion causing FXS, is the most common monogenic cause of ASD accounting for up to 6% of cases[77]. While probands with FXS were excluded from the Simons Simplex Collection, ascertainment in the ASD pedigree cohort with parents and offspring may have led to enrichment of FMR1 premutation alleles in the dataset. Lastly, our conservative approach for association testing by using unaffected siblings instead of unrelated subjects as controls is considered more robust against false positive results[78]. The ambiguity of true and false positive cases of FMR1 premutation carriers identified in our analyses demonstrates the need for further validation in additional independent datasets. It will therefore be worthwhile to examine the allelic distribution of FMR1 repeats in other ASD and non-ASD cohorts with PCR-free WGS data using computational methods.

Overall, our study demonstrates the utility of large-scale repeat expansion screening in available WGS datasets as an initial diagnostic test. While we observe no evidence the FMR1 premutation is associated with ASD disease risk, we do observe the Simons Simplex Collection potentially possesses an increased prevalence of individuals with alleles in the premutation range of FMR1 compared to previous studies. However, more information on the FMR1 locus and loci with similar sequences are required to provide more accurate genotypes.

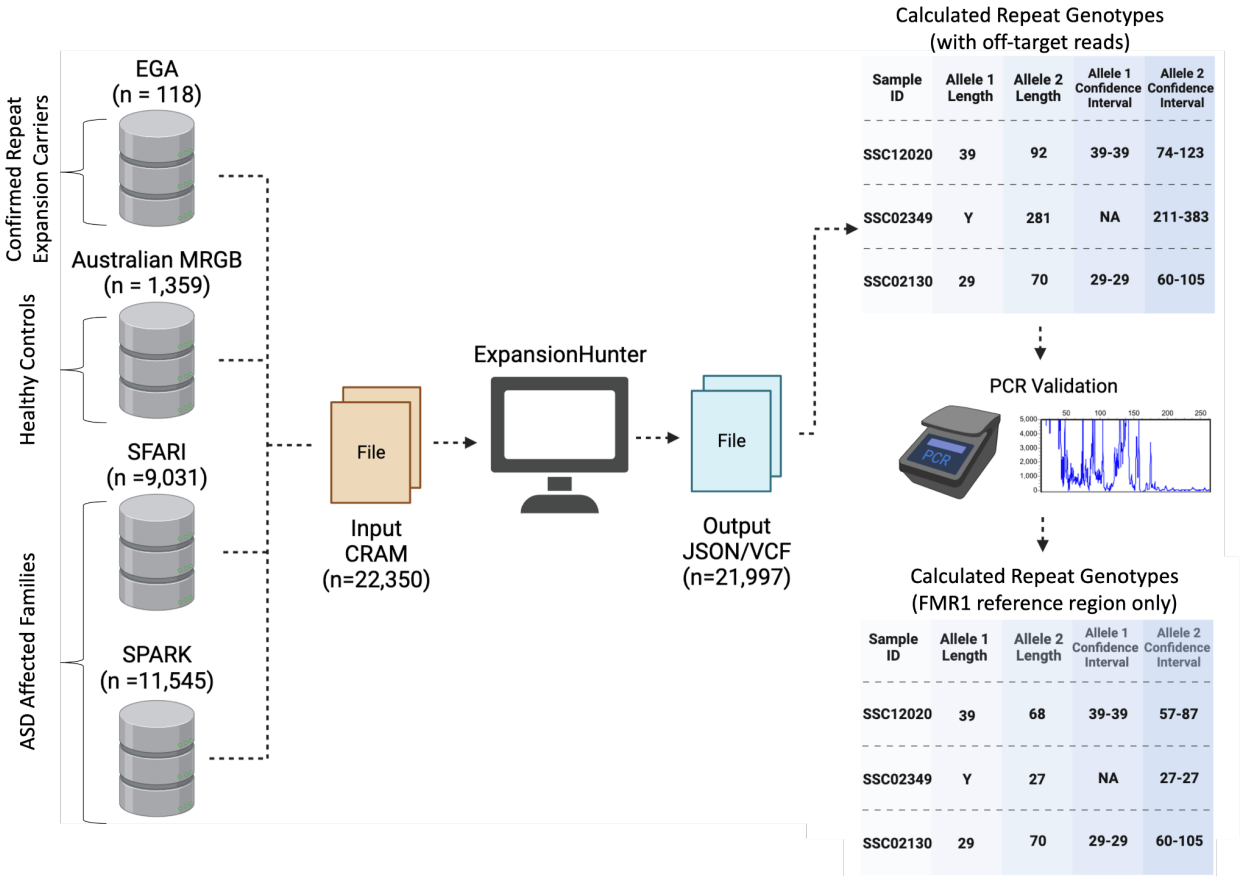

Figure 1.1 Workflow of calculating the FMR1 repeat length of 22,350 samples using ExpansionHunter: 22,350 sequence alignment files (CRAM format) were downloaded from either the 1) European Genome Archive (EGA), 2) Australian Medical Reference Genome Bank (MRGB), 3) Simons Foundation Autism Research Initiative (SFARI) ,or 4) the Simons Foundation Powering Autism Research (SPARK) initiative. ExpansionHunter was able to calculate repeat length genotypes for 21,977 samples (98.4%). Four families that were identified to have at least one member with an expanded repeat were validated by PCR. PCR showed overestimation in the repeat lengths calculated by ExpansionHunter, which was shown to be caused by the use of off-target reads in repeat length calculation. To reduce overestimating expanded repeat lengths, samples originally identified to have repeats exceeding the normal range were calculated using only the FMR1 reference region.

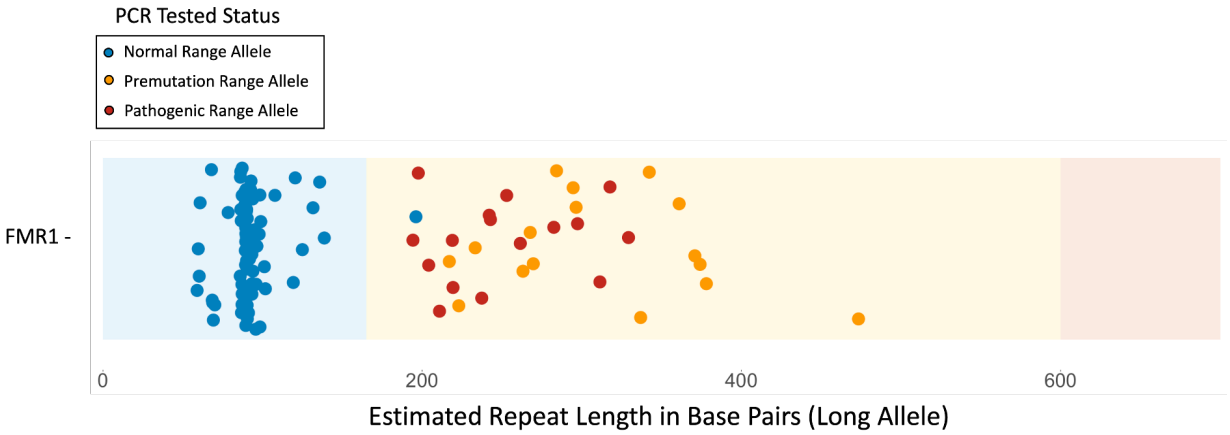

Figure 1.2 Validating repeat lengths calculated by ExpansionHunter: 118 samples identified to have expanded repeats at loci associated with eight different diseases were processed through ExpansionHunter. ExpansionHunter correctly identified all samples with expanded FMR1 repeat lengths, however, discrepancies were observed in the exact length of the repeats as the samples known to be carriers of pathogenic repeats were classified as being in the premutation range by ExpansionHunter.

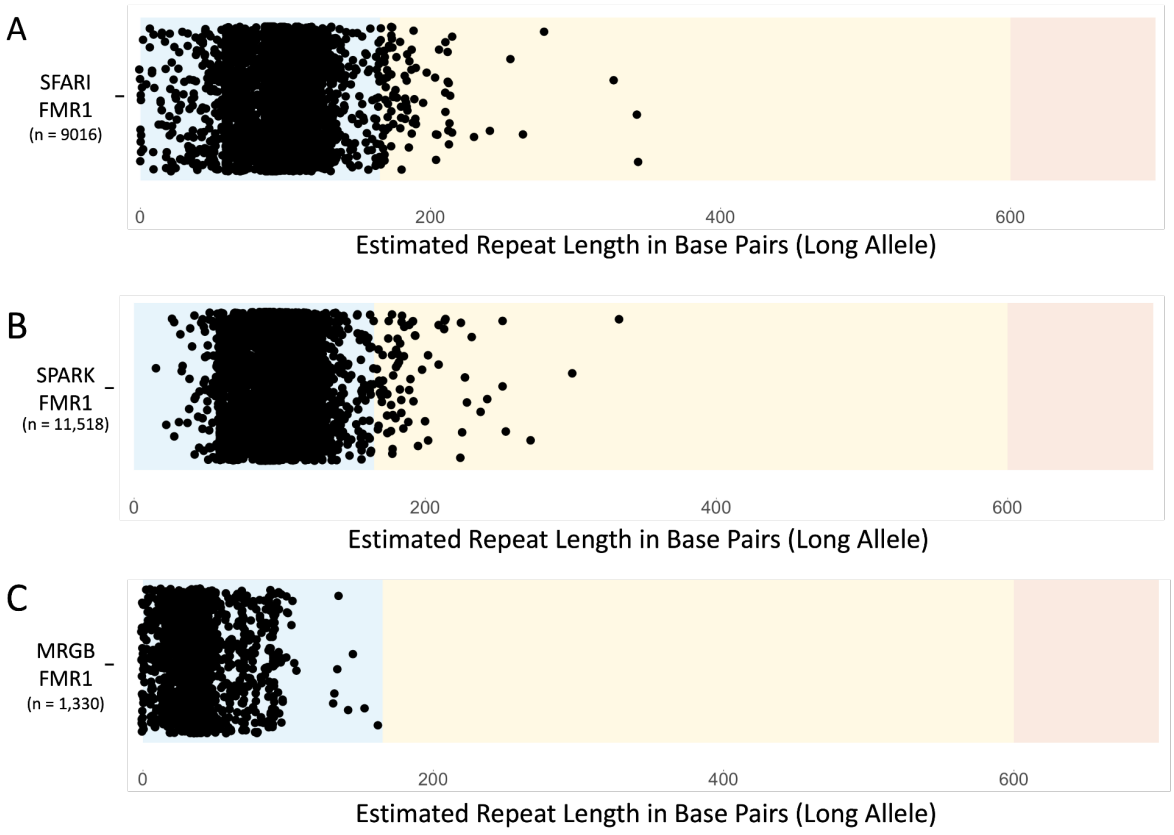

Figure 1.3 Estimated repeat length of the FMR1 trinucleotide repeat in SFARI: For each subject the longer allele for the FMR1 genotype is displayed in A) SFARI, B) SPARK, and C) Australia MRGB. Alleles have repeat lengths within the normal (blue), premutation (yellow), or pathogenic (red) range. Repeat length estimates were observed to fall within distributions within the normal range (30 bp - 120 bp) or around the start of the premutation range.

Sample ID	Family ID	Family Member	PCR Determined Repeat Length Genotype	ExpansionHunter Determined Repeat Length Genotype (with off-target reads)	ExpansionHunter Determined Repeat Length Genotype (FMR1 Reference Region)
SSC02135	11372	Father	Y/23	Y/23	Y/23
SSC02130	11372	Mother	29/48	29/70	29/70
SSC02128	11372	Male Proband	Y/48	Y/46	Y/46
SSC02136	11372	Male Sibling	Y/48	Y/57	Y/51
SSC02349	11676	Father	Y/27	Y/281	Y/27
SSC02338	11676	Mother	29/30	29/30	29/30
SSC02330	11676	Female Proband	27/30	27/30	27/30
SSC02350	11676	Female Sibling	27/29	27/29	27/29
SSC07297	13390	Father	Y/30	Y/30	Y/30
SSC07289	13390	Mother	30/33	30/33	30/33
SSC07282	13390	Male Proband	Y/33	Y/157	Y/71
SSC07298	13390	Female Sibling	30/30	30/30	30/30
SSC12025	14489	Father	Y/30	Y/30	Y/30
SSC12020	14489	Mother	39/46	39/92	38/68
SSC12016	14489	Male Proband	Y/39	Y/128	Y/71
SSC12026	14489	Male Sibling	Y/39	Y/168	Y/

Table 1.1 Validation of FMR1 Repeat Alleles Calculated by ExpansionHunter. The FMR1 repeat allele estimated by ExpansionHunter was validated by comparing the allele to repeat length measured by PCR. PCR detected no alleles in the premutation range of FMR1. Only three of the premutation carriers identified by ExpansionHunter were observed to have longer alleles in the gray zone based on PCR.

Family Member	Processed Samples	Samples with FMR1 Genotype	FMR1 Premutation Carriers	FMR1 Premutation Non-carriers	Percentage of FMR1 Premutation Carriers
Father	2,364	2,362	22	2,340	0.93%
Mother	2,369	2,365	42	2,323	1.78%
Male Proband	2,060	2,055	13	2,042	0.63%
Male Sibling	906	905	8	897	0.88%
Female Proband	321	321	6	315	1.87%
Female Sibling	1,011	1008	11	997	1.09%
Total	9,031	9,016	102	8,914	-

Table 1.2 Simons Simplex Collection Samples Processed and Genotyped for FMR1 CGG Trinucleotide Repeat by ExpansionHunter. We genotyped the CGG trinucleotide repeat of the FMR1 locus in 9,031 samples from 2,380 families using ExpansionHunter. Of the samples that were genotyped, we identified 139 individuals with an FMR1 premutation allele. In comparison to previous reports, we observe an increased baseline prevalence of premutation carriers in SFARI.

Family Member	Processed Samples	Samples with FMR1 Genotype	FMR1 Premutation Carriers	FMR1 Premutation Non-carriers	Percentage of FMR1 Premutation Carriers
Father	3,075	3,067	19	3,048	0.62%
Mother	3,078	3,077	22	3,055	0.71%
Male Proband	2,509	2,497	16	2,481	0.64%
Male Sibling	1,133	1,127	4	1,123	0.35%
Female Proband	614	614	7	607	1.14%
Female Sibling	1,136	1,136	10	1,126	0.88%
Total	11,545	11,518	78	11,446	-

Table 1.3 Simons Foundation Powering Autism Research (SPARK) Initiative Samples Processed and Genotyped for FMR1 CGG Trinucleotide Repeat by ExpansionHunter. We genotyped the CGG trinucleotide repeat of the FMR1 locus in 11,545 samples from 3,087 families using ExpansionHunter. Of the samples that were genotyped, we identified 92 individuals with an FMR1 premutation allele. In comparison to previous reports, we observe an increased baseline prevalence of premutation carriers in the SPARK.

Parent to Offspring Transmission/ de novo expansion in Offspring	Allele Transmitted	Count	Percentage of Offspring in FMR1 Premutation Families
Mother to Male Proband	Premutation Range Allele	10	7.09%
Mother to Male Sibling	Premutation Range Allele	7	4.96%
Mother to Female Proband	Premutation Range Allele	3	2.12%
Mother to Female Sibling	Premutation Range Allele	6	4.26%
Father to Male Proband	Premutation Range Allele	0	0.00%
Father to Male Sibling	Premutation Range Allele	0	0.00%
Father to Female Proband	Premutation Range Allele	2	1.42%
Father to Female Sibling	Premutation Range Allele	4	2.83%
Mother to Male Proband	Normal Range Allele	19	13.48%
Mother to Male Sibling	Normal Range Allele	14	9.92%
Mother to Female Proband	Normal Range Allele	2	1.42%
Mother to Female Sibling	Normal Range Allele	7	4.96%
Father to Male Proband	Normal Range Allele	0	0.00%
Father to Male Sibling	Normal Range Allele	0	0.00%
Father to Female Proband	Normal Range Allele	0	0.00%
Father to Female Sibling	Normal Range Allele	0	0.00%
Father to Female Sibling	Normal Range Allele	0	0.00%
Male Proband	de novo Expansion	3	2.12%
Male Sibling	de novo Expansion	1	0.71%
Female Proband	de novo Expansion	0	0.00%
Female Sibling	de novo Expansion	1	0.71%

Table 1.4 Parental Transmission of the FMR1 Premutation Allele in SPARK. Analysis of the calculated repeat length of the FMR1 trinucleotide repeat in the families within SPARK revealed 35 cases where an offspring was observed to inherit the FMR1 premutation from one of the parents. The transmission from mother to male proband were most commonly observed and no father to male offspring transmission was observed as expected for X-linked inheritance. *Unclassified indicates an offspring with an ungenotyped parent and transmission could not be determined.

Parent to Offspring Transmission/ de novo expansion in Offspring	Allele Transmitted	Count	Percentage of Offspring in FMR1 Premutation Families
Mother to Male Proband	Premutation Range Allele	15	15.15%
Mother to Male Sibling	Premutation Range Allele	4	4.04%
Mother to Female Proband	Premutation Range Allele	3	3.03%
Mother to Female Sibling	Premutation Range Allele	4	4.04%
Father to Male Proband	Premutation Range Allele	0	0.00%
Father to Male Sibling	Premutation Range Allele	0	0.00%
Father to Female Proband	Premutation Range Allele	3	3.03%
Father to Female Sibling	Premutation Range Allele	5	5.05%
Mother to Male Proband	Normal Range Allele	11	11.11%
Mother to Male Sibling	Normal Range Allele	6	6.06%
Mother to Female Proband	Normal Range Allele	2	2.02%
Mother to Female Sibling	Normal Range Allele	7	7.07%
Father to Male Proband	Normal Range Allele	0	0.00%
Father to Male Sibling	Normal Range Allele	0	0.00%
Father to Female Proband	Normal Range Allele	0	0.00%
Father to Female Sibling	Normal Range Allele	0	0.00%
Male Proband	de novo Expansion	1	0.01%
Male Sibling	de novo Expansion	0	0.00%
Female Proband	de novo Expansion	1	0.00%
Female Sibling	de novo Expansion	1	0.01%

Table 1.5 Parental Transmission of the FMR1 Premutation Allele in SFARI. Analysis of the calculated repeat length of the FMR1 trinucleotide repeat in the families within SFARI revealed 40 cases where an offspring was observed to inherit the FMR1 premutation from one of the parents. The transmission from mother to male proband were most commonly observed and no father to male offspring transmission was observed as expected for X-linked inheritance. *Unclassified indicates an offspring with an ungenotyped parent and transmission could not be determined.

Chapter 2: Large-scale whole genome sequence analysis of the C9orf72 hexanucleotide repeat in >22,000 subjects

Introduction

An expansion of a GGGGCC hexanucleotide repeat in the *C9orf72* gene on chromosome 9 is the most common genetic cause of amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD)[54,55]. Unaffected individuals in the general population usually carry 2-8 repeat units, while repeat lengths of hundreds to thousands of units are observed in cases of ALS and FTD[54,55,79,80]. The threshold of 30 or more repeat units at the C9orf72 repeat locus is considered to be pathogenic[56,57]. Based on the prevalence of ALS and FTD and the frequency of C9orf72 repeat expansions in ALS and FTD cohorts, the prevalence of the C9orf72 repeat expansion in the general population is estimated to be approximately 1:10,000[81,82]. C9orf72 repeat expansions have also been observed in patients with other neurodegenerative disorders, however, the mechanisms of neuronal dysfunction in the presence of C9orf72 repeat expansions remain unknown[83–85].

Faimiles with carriers of the C9orf72 pathogenic repeat are reported to have higher rates of schizophrenia, psychosis, suicide, and ASD [86]. While expansions of the C9orf72 repeat have not been reported to cause ASD, symptoms such as social isolation, obsessive compulsion, inflexibility, and stereotypy of speech are observed in both behavioral variant FTD and ASD[87]. To determine the effect that C9orf72 pathogenic repeats have on ASD susceptibility, we compared ASD probands and unaffected siblings who were carriers of the C9orf72 pathogenic repeats in the largest available datasets of ASD families, the Simons Simplex Collection[42,72].

Similar to our analysis of the FMR1 trinucleotide repeat, we performed a large-scale WGS-based analysis using ExpansionHunter[32,33], to examine the C9orf72 repeat in 22,981 subjects. We did not observe a significant increased burden of pathogenic alleles in the probands versus

unaffected siblings. We also did not observe an increased overall prevalence of C9orf72 pathogenic repeats in the parents of the ASD cohort compared to the 1,359 independent healthy subjects of the Australian Medical Reference Genome Bank (MRGB)[3,44]. Through PCR validation we also determined that ExpansionHunter can accurately calculate repeats at the C9orf72 locus. We further observe stable transmission of C9orf72 repeat alleles on specific haplotypes, without evidence for *de novo* expansion events in the ~10,000 meioses in the data set. The higher than expected prevalence of C9orf72 repeat alleles and lack of evidence of *de novo* repeat expansion may have clinical relevance for ALS and FTD.

Materials and Methods

Samples with WGS Data

In our analysis we utilized multiple data sets of ASD families and control subjects. First, we included data of ASD families from the Simons Simplex Collection (SSC) , a resource of the Simons Foundation Autism Research Initiative (SFARI)[72]. As previously mentioned, families admitted into SSC were required to meet metrics related to age such as the ASD proband being between the age of 4 years and 17 years and 11 months when the data was collected and also meeting several diagnostic criteria based on Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS) scores. Additionally, probands with known FXS diagnosis were excluded from the repository [72]. Sequence alignment files (CRAM format) from 9,031 samples representing a combination of 2,380 trio and quad families containing WGS reads aligned to the hg38 reference genome were obtained from SFARI base (https://base.sfari.org). WGS was performed at 30x coverage on genomic DNA extracted from whole blood as described before[72].

Our analysis also included samples from the SPARK (Simons Foundation Powering Autism Research) initiative, another resource of SFARI[42]. SPARK is an autism research initiative that aims to recruit, engage and retain a community of 50,000 individuals with autism and their family members living in the U.S. Families admitted into SPARK were required to meet three criteria: 1) have at least one family member with an ASD diagnosis, 2) currently live in the United States and 3) able to read and speak English[42]. Samples recruited into the dataset were enriched for affected individuals whose parents were also available to participate. Participants registered for SPARK online (www.SPARKforAutism.org) or at a clinical site by completing questionnaires on medical history and social communication[42]. Sequence alignment files (CRAM format) from 11,545 samples representing a combination of 3,087 trio and quad families containing WGS reads aligned to the hg38 reference genome were also obtained from SFARI base (https://base.sfari.org). WGS was performed at 30x coverage on genomic DNA extracted from whole blood as described before[73].

We used the WGS data of the 1,359 healthy subjects from the Australian Medical Reference Genome Bank (MRGB) available via the European Genome-Phenome Archive (accession code EGAD00001005095) as an independent reference sample. Data from the MRGB were generated from individuals of European descent, aged 60-95 years, and confirmed to be healthy with no reported history of cancer, cardiovascular disease, or dementia[44]. The MRGB samples were sequenced at an average of >38x coverage. The WGS data was aligned to the hs37d5 reference genome with decoys, with no further processing applied [44].

Genotype assessment of the GGGGCC hexanucleotide repeat at C9orf72.

The length of the GGGGCC hexanucleotide repeat in the C9orf72 gene was determined by processing WGS alignment files (formatted in CRAM or BAM format) from SFARI base and the Australian Medical Reference Genome Bank through ExpansionHunter v5.0.0 (https://github.com/Illumina/ExpansionHunter)[32,33]. Following the established protocol, files processed by ExpansionHunter were run with the reference genome assembly matching the reference of the aligned reads and the variant catalog associated with that reference[32,33]. An output JSON and VCF file containing the genotype of the C9orf72 repeat were generated for each of the processed files. We applied Hardy-Weinberg equilibrium testing for quality control and used the available pedigree information to examine inheritance patterns.

PCR validation of C9orf72 pathogenic carriers and their pedigrees

The presence of a pathological hexanucleotide repeat expansion in C9orf72 was detected by our collaborator Dr. Zhongan Yang, a Staff Research Associate at UCLA, who used both fluorescent and repeat-primed polymerase chain reaction (PCR), as described previously[54]. They performed the fragment length analysis on an ABI 3730 genetic analyzer, and they analyzed the data using the Peak Scanner Software, including a positive control sample for reference.

Repeat burden statistical testing

We classified individuals to be carriers of pathogenic repeats if they possessed at least one allele with a repeat length that was either equal to or greater than 30 repeat units. We performed a

burden analysis using Fisher's Exact, which compared the frequency of pathogenic alleles in the proands versus the unaffected siblings.

Analysis of expanded C9orf72 repeat haplotypes

We downloaded the chromosome 9 VCF file for samples in the SFARI dataset from SFARI base (https://base.sfari.org) to allow our collaborator Dr. Michael Thompson to perform haplotype phasing using Beagle v5.4 (https://faculty.washington.edu/browning/beagle/beagle.html)[88,89]. To perform the analysis he first filtered the VCF files to 1 Mb from the start and end of the C9orf72 locus and combined them into a single file. Following the established protocol, he processed the merged VCF file through Beagle [88,89]. For haplotype identification he used a 170Kb block at the chromosomal location of the C9orf72 repeat. Based on previous literature and methods, Dr. Thompson phased haplotypes using a cut-off of 23 repeats, which also broadened the range of samples included in the analysis[57,90–92].

Results

Prevalence of C9orf72 Pathogenic Carriers

ExpansionHunter calculated genotypes for 9,019 of 9,031 (99.9%) CRAM files that were downloaded from SFARI base. ExpansionHunter does not calculate genotypes for samples with less than 10x coverage or do not possess reads that span the regions in the variant catalog for the C9orf72 locus. We identified 24 samples with repeats that extended into the pathogenic range of C9orf72. Among the parents, representing an independent sample of unrelated subjects, the pathogenic C9orf72 repeat alleles were observed at a prevalence of 0.25% (Table 2.1). The prevalence of C9orf72 pathogenic carries in the parents of the SFARI dataset compared to the prevalence observed in samples from the MRGB were not found to be significantly different from

each other (p-value = 0.08). We found no evidence that the observed genotype distribution of the pathogenic C9orf72 repeat in SFARI deviated from what is expected based on the Hardy-Weinberg equilibrium (p-value = 0.99, χ^2= 0.02). Furthermore we observed no evidence for bias of C9orf72 pathogenic allele frequencies in the different sequencing batches in SFARI.

We selected three families (11599, 12323, and 12524) from the SFARI dataset with at least one family member with a pathogenic allele, for PCR validation (Table 2.2). Among the three families, seven samples were predicted to be pathogenic carriers by ExpansionHunter. All samples were found to match the repeat lengths determined by PCR. Our comparison of repeat lengths determined by ExpansionHunter and PCR demonstrate that ExpansionHunter accurately estimates repeat lengths of the alleles at the C9orf72 locus.

We used ExpansionHunter to calculate an additional 11,542 genotypes from 11,545 (99.9%) CRAM files from the SPARK dataset. We identified 26 samples with repeats in the pathogenic range of C9orf72 (Figure 3). Parents that were carriers of the C9orf72 pathogenic repeat were observed at a prevalence of 0.20%; we observed no instances of C9orf72 *de novo* events in the offspring (Table 2.3). Similar to the SFARI dataset, no significant difference was observed between the prevalence of C9orf72 pathogenic repeat carriers in the parents of the SPARK dataset compared to the prevalences observed in samples from the MRGB (p-value = 0.14). Again, we observed no deviation from what is expected from the Hardy-Weinberg equilibrium (p-value = 0.99, χ^2= 0.01) in SPARK and no evidence for bias of C9orf72 pathogenic allele frequencies in the different sequencing batches.

Burden of C9orf72 pathogenic range alleles on ASD in affected probands

In a comparison of probands and unaffected siblings in the SFARI dataset, we did not observe a significant difference in the frequency of pathogenic alleles (p-value = 0.39, OR = 0.57). Furthermore, in the SPARK dataset, no significant difference was observed between the frequency of pathogenic alleles in probands and unaffected siblings (p-value = 0.99, OR = 0.97). We also performed a burden test combining the allele frequencies observed in SFARI and SPARK and no significant differences were observed between the allele frequencies of probands and unaffected siblings (p-value = 0.55, OR = 0.75).

Parental transmission of C9orf72 pathogenic range alleles

We used the C9orf72 repeat length genotypes from the trio and quad families to establish patterns of paternal and maternal inheritance and to examine whether *de novo* repeat expansion events were detected in our data set. There were 12 families (including 10 quads and 2 trios) with 23 offspring of which 12 offspring (52.2%) were observed to inherit the pathogenic allele from one of their parents and a majority of the transmission (75.0%) were observed to be paternal (Table 2.4). Notably, we observed no offspring with pathogenic range alleles that were best explained by *de novo* expansion of the C9orf72 repeat (Table 2.4).

In the SPARK dataset, we observed 16 families (including 2 trios and 14 quads) with parental alleles in the pathogenic range showing an expected transmission rate of 50% in the offspring without evidence of paternal or maternal bias, nor any evidence for *de novo* repeat expansions (Table 2.5).

Haplotypes associated with C9orf72 Expansions

In the 12 families from the SFARI dataset where transmission was observed, we also identified that carriers with pathogenic alleles in 10 of the 12 families (83.3%) possessed the rs3849942A allele, which has previously been as marker for the risk (R) haplotype with the pathogenic hexanucleotide repeat expansions among cases of ALS and FTD[93]. We broadened the analysis of the haplotype distribution among 32 subjects with repeat lengths of n>23. We used the difference in MAF between high repeats and non-high repeats to identify marker alleles in the haplotypes and the greatest difference was found in SNPs rs3849945, rs245355, and rs2453565, which were also characterized in the R haplotype[91] . We also filtered SNPs from haplotypes previously identified in individuals with C9orf72 pathogenic repeats within the Swedish population and determined that 12 of the 32 haplotypes (37.5%) were not observed in the Swedish population and were unique to the SFARI dataset.

Discussion

In a secondary large-scale and systematic study of the WGS data from 5,467 families affected by ASD, we observed a prevalence of the pathogenic C9orf72 repeat alleles at 0.22% among the unrelated parents, which was much higher than previously reported in the general population (0.01%)[82]. In the parents of the Simons Simplex Collection, we observed a twenty-two-fold higher prevalence of 0.22% compared to the prevalence of the C9orf72 pathogenic repeat in the general population (Table 1.2 and 1.3). The prevalence of C9orf72 pathogenic carriers we observed in the Simons Simplex Collection is not significantly different to a recently performed analysis performed in gnomAD (n_{total} =19,239; 0.15%) (p-value = 0.75) or what we observed in our independent control sample (n_{total}=1,359) of the Medical Reference Genome Bank[76]. We do note that the number of carriers with C9orf72 pathogenic alleles in the Simons Simplex Collection

is still relatively small and a power analysis suggests the sample sizes would need to be doubled to observe a significant difference in the prevalence of repeat expansion carriers between different datasets. Our results should be further validated in additional independent datasets to determine if the prevalence of the C9orf72 pathogenic repeat that we observe is actually larger than the estimates in the general population or is specific to these datasets.

We used PCR to validate the C9orf72 pathogenic repeats identified in a subset of samples (i.e. three quad families) from the SFARI dataset. All samples predicted to be carriers of the pathogenic repeat by ExpansionHunter were confirmed by PCR. Our comparison of the C9orf72 repeat length calculated by ExpansionHunter to PCR results demonstrates that ExpansionHunter provides accurate estimates of the length of the GGGGCC hexanucleotide repeat at the C9orf72 locus. Furthermore, our results are consistent with other studies that have used ExpansionHunter to identify repeat expansions in known carriers, including C9orf72[32,76,94].

In our burden analysis we found no evidence of increased burden of C9orf72 pathogenic range alleles in ASD probands compared to unaffected siblings. While pathogenic C9orf72 alleles have not been reported to have a direct effect on ASD, longer or intermediate range alleles might cause specific symptoms such as those observed in ASD[95,96]. Further characterization of the C9orf72 repeat in additional datasets will be vital to further delineate the frequency of intermediate and expanded C9orf72 repeats and elucidate their potential influence on phenotypes such as ASD.

Utilizing the family structure within the Simons Simplex Collection, we also established whether C9orf72 pathogenic range alleles in the offspring were inherited paternally, maternally, or arose *de novo*. We observed that all offspring with pathogenic range alleles inherited them from one of their parents (Table 2.4 and 2.5). We also observed that approximately half of the offspring did not inherit the pathogenic allele (Table 2.4 and 2.5), which is expected based on the rules of

probability that any given gamete in the parent has a 50 percent chance of having the normal range allele and a 50 percent chance of having the pathogenic range allele. The higher prevalence of C9orf72 pathogenic alleles and the lack of evidence of de novo expansion events at this locus may also shed light on ALS and FTD caused by expansion of the C9orf72 repeat. Previous studies have reported that 10% of ALS cases and 20% of FTD cases are sporadic (i.e., without family history of the disease) and carry the C9orf72 pathogenic repeat expansion. Based on the fact that we observe no evidence for *de novo* expansion of the C9orf72 repeat expansion in our analysis, we hypothesize that patients diagnosed with sporadic C9orf72 mediated ALS or FTD are likely to have a family member who carries the C9orf72 pathogenic expansion also[82,97]. Furthermore, our results strengthen the reports of previous studies that show C9orf72 repeat expansions are more commonly observed in familial cases of ALS and FTD[98,99].

We further utilized the family structure of the Simons Simplex Collection to identify haplotypes associated with expanded C9orf72 repeats. We first identified that a majority (83.3%) of the families in the SFARI dataset with C9orf72 pathogenic repeat carriers also possessed the rs3849942A alleles, which has previously been identified as the hallmark for the risk (R) haplotype commonly observed in cases of ALS and FTD caused by C9orf72 expansions[93]. The identification of the R haplotype in these families further validates that the C9orf72 repeats that they carry are in the pathogenic range. Based on previous methods, we also phased 32 haplotypes associated with high-repeats (>23 repeats), including the R haplotype[57,90–92,100]. Furthermore, we utilized SNPs from haplotypes observed in individuals with C9orf72 pathogenic repeats within the Swedish population to identify 12 haplotypes unique to the SFARI dataset[101].

While we observe no evidence of C9orf72 pathogenic repeat allele involvement with ASD disease risk, we do observe the Simons Simplex Collection possesses an increased prevalence of

individuals with alleles in the pathogenic range of C9orf72 compared to the prevalence estimated in the general population. Furthermore, the C9orf72 repeat expansions we identified were validated by PCR, allele transmission, and haplotype phasing. It is interesting that a repeat expansion reported to cause neurodegenerative disorders is observed at a higher prevalence and on the same haplotype in a cohort focused on collecting families with a member affected by a neurodevelopmental disorder. Our observations may suggest that the C9orf72 repeat expansion's contribution to ALS and FTD have been overestimated and C9orf72 repeat expansion may confer less risk to ALS and FTD overall. Our observation may suggest that the C9orf72 pathogenic repeat expansion is less penetrant and that additional genetic variants with the expansion are needed for ALS and/or FTD to develop. Additional analyses will be required to determine if expanded C9orf72 repeats are more common in the general population and to determine their overall contribution to phenotypic variation.

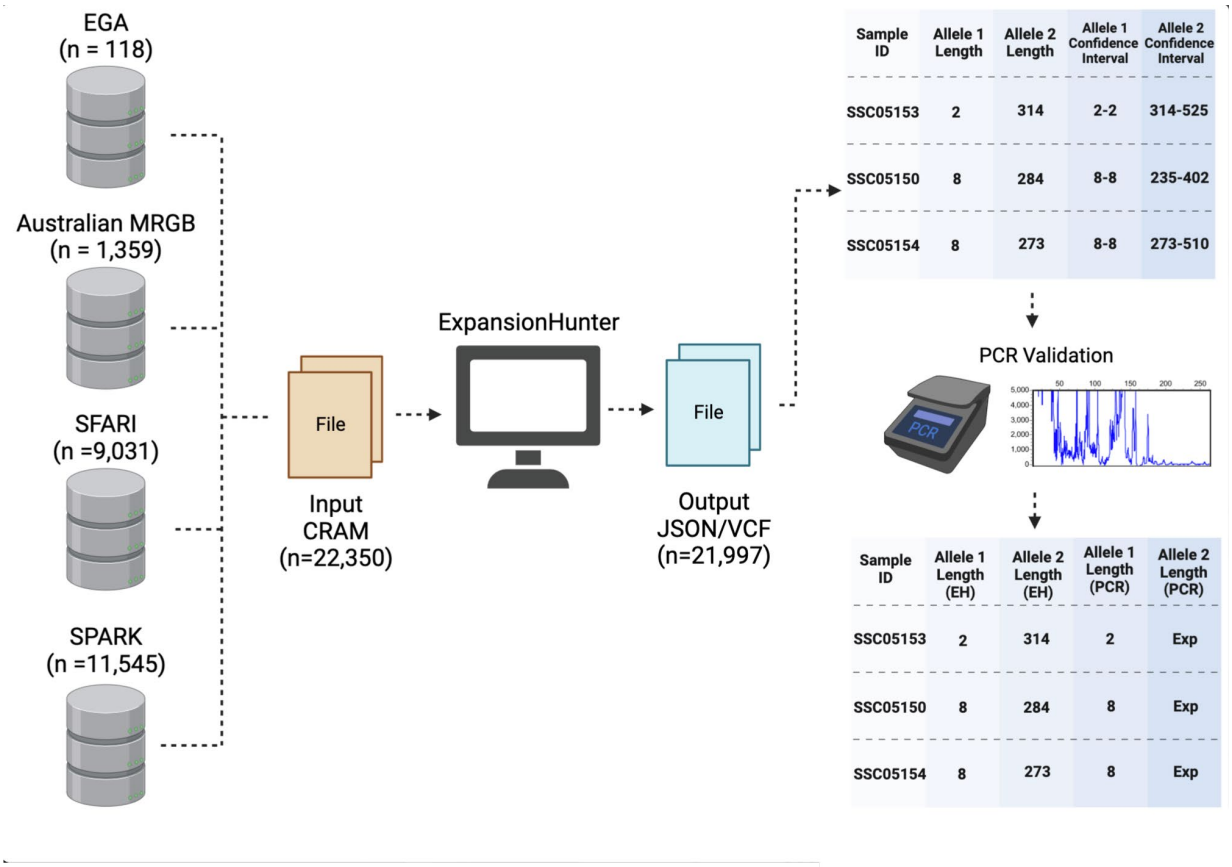

Figure 2.1 Workflow of calculating the C9orf72 repeat length of 22,350 samples using ExpansionHunter: 22,350 sequence alignment files (CRAM format) were downloaded from either the 1) European Genome Archive (EGA), 2) Australian Medical Reference Genome Bank (MRGB), 3) Simons Foundation Autism Research Initiative (SFARI) ,or 4) the Simons Foundation Powering Autism Research (SPARK) initiative. ExpansionHunter was able to calculate repeat length genotypes for 21,977 samples (98.4%). Three families that were identified to have at least one member with an expanded repeat were validated by PCR.

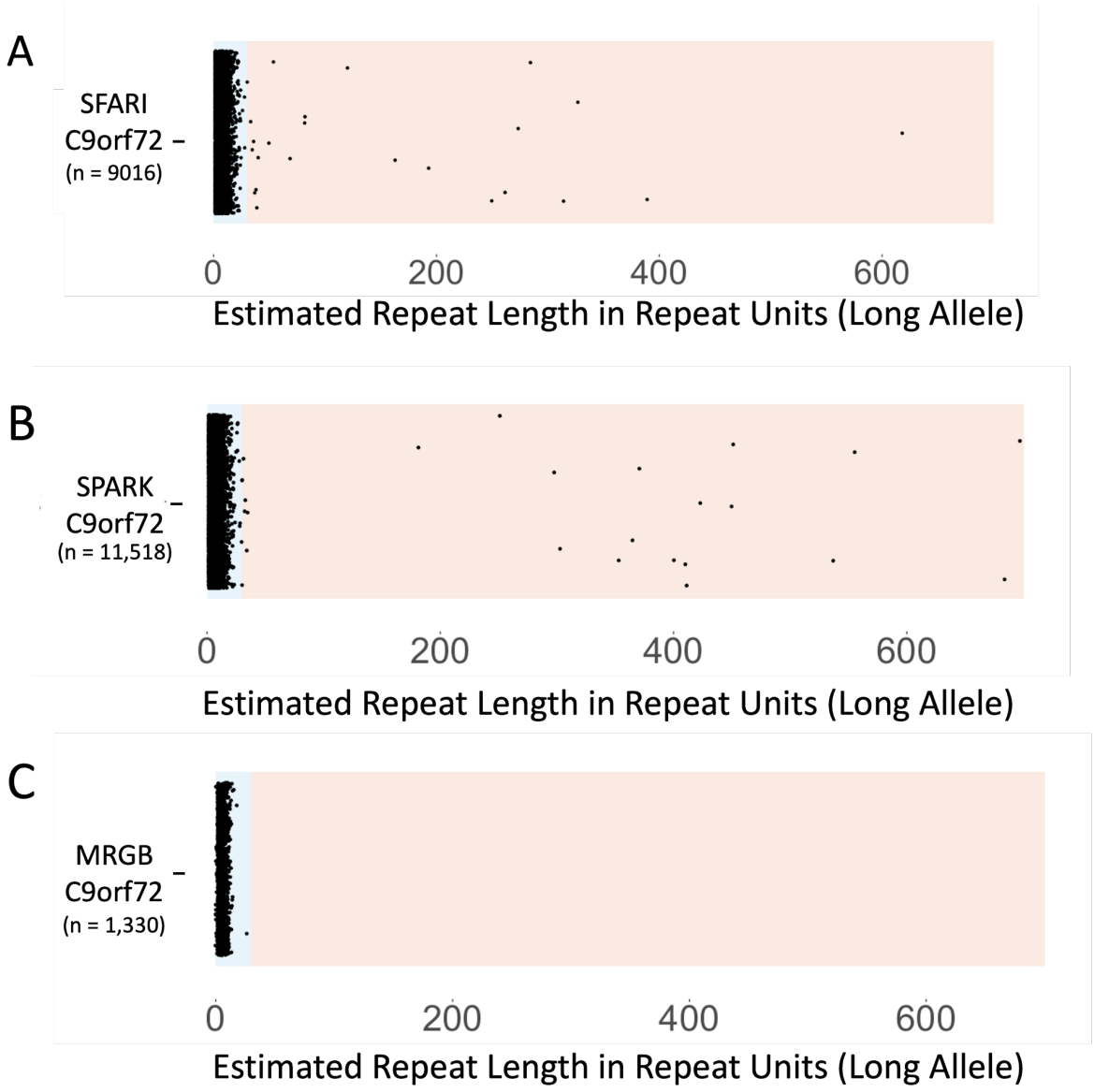

Figure 2.2 Estimated repeat length of the FMR1 trinucleotide repeat in SFARI: For each subject the longer allele for the FMR1 genotype is displayed in A) SFARI, B) SPARK, and C) Australia MRGB. Alleles have repeat lengths within the normal (blue), premutation (yellow), or pathogenic (red) range. Repeat length estimates were observed to fall within distributions within the normal range (30 bp - 120 bp) or around the start of the premutation range.

Family Member	Processed Samples	Samples with C9orf72 Genotype	C9ORF72 Pathogenic Carriers	C9ORF72 Pathogenic Non-carriers	Percentage of C9ORF72 Pathogenic Carriers
Father	2,364	2,362	9	2,353	0.38%
Mother	2,369	2,366	3	2,363	0.13%
Male Offspring	2,966	2,962	5	2,957	0.17%
Female Offsping	1,332	1,329	7	1,322	0.53%
Total	9,031	9,019	24	8,995	-

Table 2.1 Simons Simplex Collection Samples Processed and Genotyped for C9orf72 GGGGCCC hexanucleotide repeats by ExpansionHunter. We genotyped the GGGGCC hexanucleotide repeat of the C9orf72 locus in 9,031 samples from 2,380 families using ExpansionHunter. Of the samples that were genotyped, we identified 24 individuals with a C9orf72 pathogenic allele. In comparison to previous reports, we observe an increased baseline prevalence of pathogenic carriers in SFARI.

Sample ID	Family ID	Family Member	PCR Determined Repeat Length Genotype	ExpansionHunter Determined Repeat Length Genotype (FMR1 Reference Region)
SSC02760	11599	Father	2/5	2/5
SSC02742	11599	Mother	2/618	2/Exp
SSC02730	11599	Male Proband	2/2	2/2
SSC02761	11599	Male Sibling	5/193	5/Exp
SSC06410	12323	Father	2/30	2/~29
SSC06407	12323	Mother	2/5	2/5
SSC06404	12323	Male Proband	5/36	5/~41
SSC06411	12323	Female Sibling	2/82	2/Exp
SSC05153	12524	Father	2/314	2/Exp
SSC05152	12524	Mother	2/8	2/8
SSC05150	12524	Female Proband	8/284	8/Exp
SSC05154	12524	Female Sibling	8/273	8/Exp

Table 2.2 Validation of C9orf72 Repeat Alleles Calculated by ExpansionHunter. The C9orf72 repeat allele estimated by ExpansionHunter was validated by comparing the allele to repeat length measured by PCR. With the exception of a single allele, all C9orf72 repeats measured by ExpansionHunter were found to match the repeats lengths measured by PCR.

Family Member	Processed Samples	Samples with FMR1 Genotype	FMR1 Premutation Carriers	FMR1 Premutation Non-carriers	Percentage of FMR1 Premutation Carriers
Father	3,075	3,074	6	3,068	0.19%
Mother	3,078	3,077	6	3,071	0.19%
Male Offspring	3,642	3,641	6	3,635	0.16%
Female Offspring	1,750	1,750	4	1,756	0.23%
Total	11,545	11,542	26	11,516	-

Table 2.3 Simons Foundation Powering Autism Research (SPARK) Initiative Samples Processed and Genotyped for C9orf72 GGGGCCC hexanucleotide repeats by ExpansionHunter. We genotyped the GGGGCC hexanucleotide repeat of the C9orf72 locus in 11,545 samples from 3,087 families using ExpansionHunter. Of the samples that were genotyped, we identified 26 individuals with a C9orf72 pathogenic allele. In comparison to previous reports, we observe an increased baseline prevalence of premutation carriers in the SPARK.

Parent to Offspring Transmission	Allele Transmitted	Count	Percentage of Offspring in C9orf72 Pathogenic Families
Mother to Male Offspring	Pathogenic Range Allele	2	14.3%
Mother to Female Offspring	Pathogenic Range Allele	1	11.1%
Father to Male Offspring	Pathogenic Range Allele	3	21.4%
Father to Female Offspring	Pathogenic Range Allele	6	66.7%
Mother to Male Offspring	Normal Range Allele	3	21.4%
Mother to Female Offspring	Normal Range Allele	0	0.00%
Father to Male Offspring	Normal Range Allele	6	42.9%
Father to Female Offspring	Normal Range Allele	2	22.2%

Table 2.4 Parental Transmission of the C9orf72 Pathogenic Allele in SFARI. Analysis of the calculated repeat length of the C9orf72 hexanucleotide repeat in the families within SPARK revealed 12 cases where an offspring was observed to inherit the C9orf72 pathogenic repeat from one of the parents. The transmission from father to offspring was the most commonly observed.

Parent to Offspring Transmission	Allele Transmitted	Count	Percentage of Offspring in C9orf72 Pathogenic Families
Mother to Male Offspring	Pathogenic Range Allele	4	19.0%
Mother to Female Offspring	Pathogenic Range Allele	3	33.3%
Father to Male Offspring	Pathogenic Range Allele	6	28.6%
Father to Female Offspring	Pathogenic Range Allele	2	22.2%
Mother to Male Offspring	Normal Range Allele	3	14.3%
Mother to Female Offspring	Normal Range Allele	1	11.1%
Father to Male Offspring	Normal Range Allele	8	38.1%
Father to Female Offspring	Normal Range Allele	3	33.3%

Table 2.5 Parental Transmission of the C9orf72 Pathogenic Allele in SPARK. Analysis of the calculated repeat length of the C9orf72 hexanucleotide repeat in the families within SFARI revealed 15 cases where an offspring was observed to inherit the C9orf72 pathogenic repeat from one of the parents. Transmissions from father to offspring and mother to offspring were observed to occur evenly in SPARK.

Chapter 3: Conclusion

Over the past few years computational tools have been developed to detect and measure repeat expansion from WGS data[11,22,32,33]. The development of these tools have allowed studies to reassess the effect that expanded repeats have on complex diseases and disorders, many of which are neurological. Recent reports in literature have shown that both known and novel repeat expansions contribute to the disease susceptibility of several neurological and neuropsychiatric disorders including spinocerebellar ataxia type 31, myoclonic epilepsy, autism spectrum disorder (ASD), and schizophrenia[35–38]. In our studies, we examined the distribution/presence of repeat expansions at two loci known to cause disease and determined their effect on ASD.

Our first analysis focused on repeats in the premutation range of FMR1. Repeats of 200 or more are classified as being pathogenic and are observed in cases of Fragile X Syndrome (FXS)[45]. Furthermore, the FMR1 pathogenic repeat is reported to be the most common monogenic cause of ASD[46–48]. However, the effect that FMR1 premutation repeats (55-200 repeats) have on ASD remains unclear [46,52]. We observed no evidence that the FMR1 premutation affects ASD susceptibility and none of the samples in the ASD cohort were observed to have repeats in the pathogenic range of FMR1. However, our results do help support previous literature that states expanded FMR1 repeats contribute solely to FXS[71]. In our second analysis we focused on the C9orf72 repeat expansion which is the most common genetic cause of amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD)[54,55]. Both ALS and FTD are neurodegenerative diseases, however, symptoms involved in cognitive function and behavior have been observed in these diseases, which may suggest a possible link with ASD and other neuropsychiatric disorders. We observed no evidence that the C9orf72 pathogenic repeat affects ASD susceptibility but we did identify a risk haplotype previously reported to be associated with the expanded C9orf72 repeat

and was also present in carriers in our study. The fact that we observed the same haplotype in a cohort that does not have any association with a neurodegenerative disorder and observed the prevalence of C9orf72 repeat expansion carriers to be higher than reports in previous literature suggest the C9orf72 repeat expansion may confer less risk to ALS and FTD overall and that other genetics variants may be required to be present with the C9orf72 repeat expansion for the disease to develop. While we observed no evidence that the repeat expansions observed in our study affect ASD, our results still provide information that is valuable to both the medical and research communities.

Our study was focused on the utilization of ExpansionHunter, which is limited by the software's user defined variant catalog that specifies reference coordinates and the structure of loci known to have disease causing expansions[32,33]. However, other tools such as ExpansionHunter Denovo have been developed to detect repeat expansion throughout the entire genome[34]. ExpansionHunter Denovo was recently utilized in a similar study that processed data from the Simons Simplex Collection to perform a genome-wide search for expanded repeats that contribute to ASD[37]. The genome-wide study reported 2,588 tandem repeat-containing regions that were significantly more prevalent among individuals with ASD and that these expansions occurred in or near genes involved in nervous system development, the cardiovascular system and muscle tissues[37]. The study also suggests that tandem repeat expansions make a collective contribution of 2.6% to the explained phenotypic variance of ASD[37]. Another study also used ExpansionHunter Denovo to examine the role of expanded repeats in schizophrenia [38]. The sample set in this study was much smaller in size, consisting of 257 unrelated adults with SCZ, 225 congenital heart disease patients, and 2,504 individuals from the 1000 Genomes Project. The authors identified 436 distinct regions containing expanded repeats. After determining where these regions were in

the genome, the authors determined that intronic repeat expansions and repeat expansions close to exons were major contributors to the etiology of schizophrenia[38]. These studies demonstrate that ExpansionHunter Denovo can be used to detect expanded repeats throughout the entire genome and to examine the role of these loci in disease susceptibility. We made a first effort to examine genome-wide repeat expansions in severe mental illness (i.e. schizophrenia and bipolar disorder). We processed WGS data from schizophrenia cases, bipolar disorder cases, and controls for a total of 7,960 subjects of African-American ancestry from the Whole Genome in Psychiatric Disorders (WGSPD) Consortium [102]. We were able to identify 45,233 repeat sequences in 36,717 distinct regions of the human genome, which is similar to the ExpansionHunter Denovo analysis of ASD in the Simons Simplex Collection that identified 37,865 tandem repeat sequences in 31,793 distinct regions of the human genome[37]. However, further characterization of the repeat sequences identified in our initial analysis will be necessary to determine if they impact schizophrenia and/or bipolar disorder. We also note that the samples from the WGSPD Consortium are African American and our analyses could provide insight on how ancestry may affect the prevalence and effect of repeat expansions in neuropsychiatric disorders. Additional analyses similar to this will need to be conducted to fully determine the genetic variation explained by repeat expansions and validate the results of currently published literature.

While the use of ExpansionHunter and other computational tools have allowed our and other studies to identify STRs and their expansion in WGS data, there are still other structural variants that are not well detected by existing methods. For example, there are variable number tandem repeats (VNTRs) which are similar to STRs except they consist of longer tandem repeat sequences and they vary greatly in repeat length between individuals[103]. Together STRs and VNTRs represent one of the largest sources of polymorphisms in humans, however, the complex

variability of VNTRs between individuals have led them to being ignored by currently available genomic pipelines[104,105]. To ensure we are able to identify all of the genetic variation that explains a disease or trait we will need methods to detect and measure these complex structural variants from WGS data or other data types. For example, a computational tool called adVNTR was recently developed to detect VNTRs from WGS data [106]. More importantly, the development of long-read sequencing will play a crucial role in the detection of structural variants. Long-read sequencing is capable of generating reads in excess of 10 kb that span repetitive regions and provides unique anchors that can be utilized in calling structural variants on specific haplotype[107]. Future studies on expanded repeats or structural variants must incorporate high-throughput long-read sequencing and development of computational tools for analysis of this type of sequence data into their analyses to ensure that all expanded repeats that contribute to a disease or trait can be identified.

Overall, our analysis demonstrates the potential use of WGS data for the detection of pathogenic repeats through computational analysis to determine their effect on complex diseases and traits. In the framework that we developed we did identify six factors that should be considered in future studies. 1) While we solely utilized ExpansionHunter in our analysis, we do recognize that there are other computational software capable of detecting repeat expansions in WGS data, such as exSTRa, STRetch, TREDPARSE, and gangSTR[22,32,33,108–110]. However, we were unable to use tools such as STRetch and TREDPARSE as they only accept sequencing alignment files in BAM format and many of our files were in CRAM format as they are more suitable for the storage of large WGS datasets. We were also unable to use exSTRa and STRetch as their software was not compatible with the computational node we processed our data on. We also did not process our samples through gangSTR because the software took 24 hours to process while

ExpasnionHunter only took an hour. Overall, we experienced ExpansionHunter to be the most efficient and user friendly tool. 2) Also, in our study we processed > 22,000 sequence alignment files which required more than 100TB of storage. Studies of large WGS datasets (most likely in CRAM format) will require a significant amount of data storage for their analyses. 3) Furthermore, when we validated the 118 samples with known repeat expansion length we noted that one sample with a repeat measured in the normal range of FMR1 by PCR was predicted to be in the premutation range by ExpansionHunter. The false positive that we observe suggests that we can expect to detect approximately 1% of samples in the Simons Simplex Collection to be false positive for permutation repeats. Our PCR validation of samples predicted to be FMR1 premutation carriers by ExpansionHunter also showed a 100% mismatch between the two methods. While ExpansionHunter is capable of calculating repeat genotypes for the FMR1 locus, we recommend PCR validation for expanded repeats that are identified using this computational method. However, we also note that ExpansionHunter was highly accurate in detecting C9orf72 repeat expansion based of our PCR validation and that we were able to identify the risk (R) haplotype previously associated with the expanded repeat, which suggest studies could use ExpansionHunter as a gold standard to detect expanded C9orf72 repeats. 5) In our study we also recognize that the use of ExpansionHunter's off-target analysis can lead to the method overestimating the read length of repeats such as the FMR1 trinucleotide repeat. Similar to another analysis we recommend that the reference region for the repeat be used in ExpansionHunter's calculation of the repeat length [76]. 6) Further validation can also be done by using tools such as REViewer to observe the reads that ExpansionHunter used to calculate the read length genotype to ensure the genotype is not being overestimated or underestimated [76,94]. The framework

developed in this study can be utilized to develop best practices and a pipeline for the analysis of tandem repeat sequences in larger datasets.

REFERENCES

1. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol.* **13**, e1002195 (2015).
2. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
3. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
4. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).
5. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
6. Young, A. I. Solving the missing heritability problem. *PLoS Genet.* **15**, e1008222 (2019).
7. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1193–1198 (2012).
8. Jelenkovic, A. *et al.* Genetic and environmental influences on adult human height across birth cohorts from 1886 to 1994. *Elife* **5**, (2016).
9. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
10. Hannan, A. J. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet.* **26**, 59–65 (2010).
11. Bahlo, M. *et al.* Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Res.* **7**, (2018).

12. Fan, H. & Chu, J.-Y. A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics* **5**, 7–14 (2007).
13. Nakamura, Y. *et al.* Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**, 1616–1622 (1987).
14. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
15. López Castel, A., Cleary, J. D. & Pearson, C. E. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat. Rev. Mol. Cell Biol.* **11**, 165–170 (2010).
16. Paulson, H. Repeat expansion diseases. *Handb. Clin. Neurol.* **147**, 105–123 (2018).
17. Ramakrishnan, S. & Gupta, V. Trinucleotide Repeat Disorders. in *StatPearls* (StatPearls Publishing, 2022).
18. Lindblad, K. *et al.* Detection of expanded CAG repeats in bipolar affective disorder using the repeat expansion detection (RED) method. *Neurobiol. Dis.* **2**, 55–62 (1995).
19. Klauck, S. M. *et al.* Molecular genetic analysis of the FMR-1 gene in a large collection of autistic patients. *Hum. Genet.* **100**, 224–229 (1997).
20. Oruc, L. *et al.* CAG repeat expansions in bipolar and unipolar disorders. *Am. J. Hum. Genet.* **60**, 730–732 (1997).
21. Margolis, R. L., McInnis, M. G., Rosenblatt, A. & Ross, C. A. Trinucleotide repeat expansion and neuropsychiatric disease. *Arch. Gen. Psychiatry* **56**, 1019–1031 (1999).
22. Tankard, R. M. *et al.* Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *Am. J. Hum. Genet.* **103**, 858–873 (2018).
23. Buchman, V. L. *et al.* Simultaneous and independent detection of C9ORF72 alleles with low and high number of GGGGCC repeats using an optimised protocol of Southern blot

hybridisation. *Molecular Neurodegeneration* vol. 8 12 Preprint at https://doi.org/10.1186/1750-1326-8-12 (2013).

24. Akimoto, C. *et al.* A blinded international study on the reliability of genetic testing for GGGGCC-repeat expansions in C9orf72 reveals marked differences in results among 14 laboratories. *J. Med. Genet.* **51**, 419–424 (2014).
25. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
26. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
27. Raczy, C. *et al.* Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041–2043 (2013).
28. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
29. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
30. Roller, E., Ivakhno, S., Lee, S., Royce, T. & Tanner, S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* **32**, 2375–2377 (2016).
31. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
32. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
33. Dolzhenko, E. *et al.* ExpansionHunter: a sequence-graph-based tool to analyze variation in

short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).

34. Dolzhenko, E. *et al.* ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.* **21**, 102 (2020).
35. Sato, N. *et al.* Spinocerebellar Ataxia Type 31 Is Associated with 'Inserted' Penta-Nucleotide Repeats Containing (TGGAA)n. *The American Journal of Human Genetics* vol. 85 544–557 Preprint at https://doi.org/10.1016/j.ajhg.2009.09.019 (2009).
36. Ishiura, H. *et al.* Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.* **50**, 581–590 (2018).
37. Trost, B. *et al.* Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**, 80–86 (2020).
38. Mojarad, B. A. *et al.* Genome-wide tandem repeat expansions contribute to schizophrenia risk. *Mol. Psychiatry* **27**, 3692–3698 (2022).
39. Geschwind, D. H. Genetics of autism spectrum disorders. *Trends Cogn. Sci.* **15**, 409–416 (2011).
40. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).
41. Hansen, S. N., Overgaard, M., Andersen, P. K. & Parner, E. T. Estimating a population cumulative incidence under calendar time trends. *BMC Med. Res. Methodol.* **17**, 7 (2017).
42. Matoba, N. *et al.* Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. *Transl. Psychiatry* **10**, 265 (2020).
43. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
44. Pinese, M. *et al.* The Medical Genome Reference Bank contains whole genome and

phenotype data of 2570 healthy elderly. *Nat. Commun.* **11**, 435 (2020).

45. Garber, K. B., Visootsak, J. & Warren, S. T. Fragile X syndrome. *Eur. J. Hum. Genet.* **16**, 666–672 (2008).
46. Abbeduto, L., McDuffie, A. & Thurman, A. J. The fragile X syndromeâ€“autism comorbidity: what do we really know? *Front. Genet.* **5**, 355 (2014).
47. Fernandez, B. A. & Scherer, S. W. Syndromic autism spectrum disorders: moving from a clinically defined to a molecularly defined approach. *Dialogues Clin. Neurosci.* **19**, 353–371 (2017).
48. Abbeduto, L. *et al.* ASD Comorbidity in Fragile X Syndrome: Symptom Profile and Predictors of Symptom Severity in Adolescent and Young Adult Males. *J. Autism Dev. Disord.* **49**, 960–977 (2019).
49. Hall, D. A. In the Gray Zone in the Fragile X Gene: What are the Key Unanswered Clinical and Biological Questions? *Tremor Other Hyperkinet. Mov.* **4**, 208 (2014).
50. Nolin, S. L. *et al.* Expansions and contractions of the FMR1 CGG repeat in 5,508 transmissions of normal, intermediate, and premutation alleles. *Am. J. Med. Genet. A* **179**, 1148–1156 (2019).
51. Zeesman, S. *et al.* Paternal transmission of fragile X syndrome. *Am. J. Med. Genet. A* **129A**, 184–189 (2004).
52. Hagerman, R., Au, J. & Hagerman, P. FMR1 premutation and full mutation molecular mechanisms related to autism. *J. Neurodev. Disord.* **3**, 211–224 (2011).
53. Alvarez-Mora, M. I. *et al.* Paternal transmission of a FMR1 full mutation allele. *Am. J. Med. Genet. A* **173**, 2795–2797 (2017).
54. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding

region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).

55. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
56. Byrne, S., Heverin, M., Elamin, M., Walsh, C. & Hardiman, O. Intermediate repeat expansion length in C9orf72 may be pathological in amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Frontotemporal Degener.* **15**, 148–150 (2014).
57. Iacoangeli, A. *et al.* C9orf72 intermediate expansions of 24–30 repeats are associated with ALS. *Acta Neuropathologica Communications* **7**, 1–7 (2019).
58. Graff-Radford, N. R. & Woodruff, B. K. Frontotemporal dementia. *Semin. Neurol.* **27**, 48–57 (2007).
59. Devenney, E. M. *et al.* Psychiatric disorders in C9orf72 kindreds: Study of 1,414 family members. *Neurology* **91**, e1498–e1507 (2018).
60. Masrori, P. & Van Damme, P. Amyotrophic lateral sclerosis: a clinical review. *Eur. J. Neurol.* **27**, 1918–1929 (2020).
61. Rhodus, E. K. *et al.* Behaviors Characteristic of Autism Spectrum Disorder in a Geriatric Cohort With Mild Cognitive Impairment or Early Dementia. *Alzheimer Dis. Assoc. Disord.* **34**, 66–71 (2020).
62. Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
63. Hunter, J. E., Berry-Kravis, E., Hipp, H. & Todd, P. K. FMR1 disorders. *GeneReviews®[Internet]* (2019).

64. Rousseau, F., Rouillard, P., Morel, M. L., Khandjian, E. W. & Morgan, K. Prevalence of carriers of premutation-size alleles of the FMRI gene--and implications for the population genetics of the fragile X syndrome. *Am. J. Hum. Genet.* **57**, 1006–1018 (1995).

65. Hantash, F. M. *et al.* FMR1 premutation carrier frequency in patients undergoing routine population-based carrier screening: insights into the prevalence of fragile X syndrome, fragile X-associated tremor/ataxia syndrome, and fragile X-associated primary ovarian insufficiency in the United States. *Genet. Med.* **13**, 39–45 (2011).

66. Seltzer, M. M. *et al.* Prevalence of CGG expansions of the FMR1 gene in a US population-based sample. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **159B**, 589–597 (2012).

67. Tassone, F. *et al.* FMR1 CGG allele size and prevalence ascertained through newborn screening in the United States. *Genome Med.* **4**, 100 (2012).

68. Maenner, M. J. *et al.* FMR1 CGG expansions: prevalence and sex ratios. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **162B**, 466–473 (2013).

69. Sandin, S. *et al.* The Heritability of Autism Spectrum Disorder. *JAMA* **318**, 1182–1184 (2017).

70. Niu, M. *et al.* Autism Symptoms in Fragile X Syndrome. *J. Child Neurol.* **32**, 903–909 (2017).

71. Mullegama, S. V. *et al.* Is it time to retire fragile X testing as a first-tier test for developmental delay, intellectual disability, and autism spectrum disorder? *Genetics in medicine: official journal of the American College of Medical Genetics* vol. 19 (2017).

72. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. *Neuron* vol. 68 192–195 Preprint at https://doi.org/10.1016/j.neuron.2010.10.006 (2010).

73. Wilfert, A. B. *et al.* Recent ultra-rare inherited variants implicate new autism candidate risk genes. *Nat. Genet.* **53**, 1125–1134 (2021).

74. Rice, J. A. *Mathematical Statistics and Data Analysis*. (Cengage Learning, 2006).

75. Sherman, S. L. & Hunter, J. E. Epidemiology of Fragile X Syndrome. *Fragile X Syndrome* 57–76 Preprint at https://doi.org/10.1016/b978-0-12-804461-2.00004-4 (2017).

76. Weisburd, B., VanNoy, G. & Watts, N. The addition of short tandem repeat calls to gnomAD. https://gnomad.broadinstitute.org/news/2022-01-the-addition-of-short-tandem-repeat-calls-to-gnomad/.

77. Kaufmann, W. E. *et al.* Autism Spectrum Disorder in Fragile X Syndrome: Cooccurring Conditions and Current Treatment. *Pediatrics* **139**, S194–S206 (2017).

78. Curtis, D. Use of siblings as controls in case-control association studies. *Annals of Human Genetics* vol. 61 319–333 Preprint at https://doi.org/10.1017/s000348009700626x (1997).

79. Gami, P. *et al.* A 30-unit hexanucleotide repeat expansion in C9orf72 induces pathological lesions with dipeptide-repeat proteins and RNA foci, but not TDP-43 inclusions and clinical disease. *Acta Neuropathol.* **130**, 599–601 (2015).

80. Van Mossevelde, S., van der Zee, J., Cruts, M. & Van Broeckhoven, C. Relationship between C9orf72 repeat size and clinical phenotype. *Curr. Opin. Genet. Dev.* **44**, 117–124 (2017).

81. van Blitterswijk, M., DeJesus-Hernandez, M. & Rademakers, R. How do C9ORF72 repeat expansions cause amyotrophic lateral sclerosis and frontotemporal dementia: can we learn from other noncoding repeat expansion disorders? *Curr. Opin. Neurol.* **25**, 689–700 (2012).

82. Gossye, H., Engelborghs, S., Van Broeckhoven, C. & van der Zee, J. C9orf72 Frontotemporal Dementia and/or Amyotrophic Lateral Sclerosis. (2020).

83. Hensman Moss, D. J. *et al.* C9orf72 expansions are the most common genetic cause of Huntington disease phenocopies. *Neurology* **82**, 292–299 (2014).

84. Schottlaender, L. V. *et al.* The analysis of C9orf72 repeat expansions in a large series of clinically and pathologically diagnosed cases with atypical parkinsonism. *Neurobiol. Aging* **36**, 1221.e1–1221.e6 (2015).

85. Shu, L. *et al.* The Association between C9orf72 Repeats and Risk of Alzheimer's Disease and Amyotrophic Lateral Sclerosis: A Meta-Analysis. *Parkinsons Dis.* **2016**, 5731734 (2016).

86. O'Brien, M. *et al.* Clustering of Neuropsychiatric Disease in First-Degree and Second-Degree Relatives of Patients With Amyotrophic Lateral Sclerosis. *JAMA Neurol.* **74**, 1425–1430 (2017).

87. Breevoort, S., Gibson, S., Figueroa, K., Bromberg, M. & Pulst, S. Expanding Clinical Spectrum of C9ORF72-Related Disorders and Promising Therapeutic Strategies: A Review. *Neurol Genet* **8**, e670 (2022).

88. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).

89. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).

90. Fahey, C. *et al.* Analysis of the hexanucleotide repeat expansion and founder haplotype at C9ORF72 in an Irish psychosis case-control sample. *Neurobiol. Aging* **35**, 1510.e1–5 (2014).

91. Ben-Dor, I., Pacut, C., Nevo, Y., Feldman, E. L. & Reubinoff, B. E. Characterization of C9orf72 haplotypes to evaluate the effects of normal and pathological variations on its

expression and splicing. *PLoS Genet.* **17**, e1009445 (2021).

92. Reus, L. M. *et al.* Genome-wide association study of frontotemporal dementia identifies a C9ORF72 haplotype with a median of 12-G4C2 repeats that predisposes to pathological repeat expansions. *Transl. Psychiatry* **11**, 451 (2021).
93. Laaksovirta, H. *et al.* Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. *Lancet Neurol.* **9**, 978–985 (2010).
94. Halman, A., Dolzhenko, E. & Oshlack, A. STRipy: A graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. *Hum. Mutat.* **43**, 859–868 (2022).
95. Hagerman, P. J., Greco, C. M. & Hagerman, R. J. A cerebellar tremor/ataxia syndrome among fragile X premutation carriers. *Cytogenet. Genome Res.* **100**, 206–212 (2003).
96. Terracciano, A. *et al.* Expansion to full mutation of a FMR1 intermediate allele over two generations. *Eur. J. Hum. Genet.* **12**, 333–336 (2004).
97. Manzoni, C. & Ferrari, R. Mendelian and Sporadic FTD: Disease Risk and Avenues from Genetics to Disease Pathways Through In Silico Modelling. in *Frontotemporal Dementias : Emerging Milestones of the 21st Century* (eds. Ghetti, B., Buratti, E., Boeve, B. & Rademakers, R.) 283–296 (Springer International Publishing, 2021).
98. Majounie, E. *et al.* Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *Lancet Neurol.* **11**, 323–330 (2012).
99. Marogianni, C. *et al.* The role of C9orf72 in neurodegenerative disorders: a systematic review, an updated meta-analysis, and the creation of an online database. *Neurobiol. Aging* **84**, 238.e25–238.e34 (2019).

100. Mok, K. *et al.* The chromosome 9 ALS and FTD locus is probably derived from a single founder. *Neurobiol. Aging* **33**, 209.e3–209.e8 (2012).

101. Chiang, H.-H. *et al.* No common founder for C9orf72 expansion mutation in Sweden. *J. Hum. Genet.* **62**, 321–324 (2017).

102. Sanders, S. J. *et al.* Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nat. Neurosci.* **20**, 1661–1668 (2017).

103. Bakhtiari, M. *et al.* Variable number tandem repeats mediate the expression of proximal genes. *Nat. Commun.* **12**, 2075 (2021).

104. Willems, T. F. *et al.* The Landscape of Human STR Variation. Preprint at https://doi.org/10.1101/004671.

105. Gymrek, M. A genomic view of short tandem repeats. *Curr. Opin. Genet. Dev.* **44**, 9–16 (2017).

106. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* **28**, 1709–1719 (2018).

107. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).

108. Tang, H. *et al.* Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am. J. Hum. Genet.* **101**, 700–715 (2017).

109. Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 121 (2018).

110. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).