Discovery to clinical utility of genome sequencing in autism

A thesis submitted for the degree of Doctor of Philosophy (Ph.D.)

by

Fiana Ní Ghrálaigh, B.A (Mod)

Supervisors: Dr Lorna Lopez (Maynooth University) & Prof. Louise Gallagher (Trinity College Dublin)

Head of Department: Prof. Paul Moynagh

Department of Biology

Maynooth University

June 2022

Summary

Autism is a highly heritable complex trait, heterogenous in genotype and phenotype. Rare genetic variants, both inherited and *de novo*, typically have larger effect sizes and are more penetrant than common variants in the population. Next-generation sequencing technologies facilitate simultaneous investigation of variation across the allele frequency spectrum.

This thesis aims to investigate rare variation and its contribution to the genetic basis of autism. This study applies genome sequencing to an autism cohort of affected individuals in a family-based study design. (WES n=42, WGS n=35). Variants emerging from these analyses contribute to the existing evidence supporting association of relevant genes with autism. Additionally, this thesis investigates the clinical utility of genome sequencing in autism. Genetic diagnosis in autism is limited by the ability to robustly determine the relevance of putatively pathogenic genetic variation. Through application of an evidence-based gene curation framework and through investigation of the diagnostic yield of commercial gene panels available for use in autism, this thesis informs on current strategies to translate genomics findings into the clinic.

Insights into the biological mechanisms underlying autism arising from this research, will lead to a greater understanding of the condition and potentially benefit clinical intervention and treatment plans in the future.

Contributions to this Work

All studies were designed with my supervisors Dr Lorna Lopez (MU) and Professor Louise Gallagher (TCD). I was responsible for the data analysis and interpretation of the results described in this thesis, under the supervision of Dr Lopez and Prof. Gallagher.

The cohorts analysed in Chapter 3 and Chapter 4 were recruited under the long-standing MolGen study of the Autism and Neurodevelopmental Research Group at Trinity College Dublin, led by Professor Louise Gallagher. Dr Nadia Bolshakova (TCD) and past members of the research group were involved in management, ascertainment, enrolment, sample collection and clinical phenotyping of this study. The cohort analysed in Chapter 5 was newly recruited under a study led by Dr Lorna Lopez, with Prof. Louise Gallagher. Mr Richard O'Conaill (TCD), Mr Matthew O'Sullivan (TCD), Dr Nadia Bolshakova, Ms Aoife Coghlan (MU) and Ms Aoife Brennan (MU) were involved in management, ascertainment, enrolment, sample collection and clinical phenotyping of this study.

Dr Elaine Kenny (TCD), TrinSeq and ELDA Biotech, contributed to this work by providing technical guidance, sample preparation and sequencing of the cohorts analysed in Chapter 3 and Chapter 4. Genuity Science contributed to this work by providing technical guidance, sample preparation and sequencing of the cohort analysed in Chapter 5.

Dr Cathal Ormond (TCD) and Dr Niamh Ryan (TCD) in the Neuropsychiatric Genetics Research Group at Trinity College Dublin advised on next-generation sequencing analyses. Ms Ellen McCarthy (MU) and Dr Daniel Murphy (MU) contributed to analyses outlined in Chapter 6, respectively performing literature searching and advising on gene panel inclusion.

The thesis has emanated from research conducted with the financial support of Science Foundation Ireland under grant number 15/SIRG/3324.

Finally, the families included within these sequencing cohorts have contributed their time and their biological samples to participate in the research studies of the Autism and Neurodevelopmental Research Group at Trinity College Dublin and the Family Genomics Research Group at Maynooth University.

Acknowledgements

I would like to express my gratitude to the many people I have been supported by during my PhD.

Lorna, thank you for your mentorship over the years. I have learned so much from you and I am so honoured to be your first PhD student. I know that I will be the first of many more to come after me. Thank you for giving me an inspiring and supportive environment to develop as a researcher- no matter where we were. You have shown me the importance of always putting people first through your research, your leadership, and your engagement. I hope that as I move through my career, I can share with others what I have learned from you.

Louise, I am so grateful for the opportunity to have worked with you and to have been part of your amazing research team at TCD. Thank you for your guidance and encouragement during my PhD, from day one to now. I will continue to follow from afar as you begin your next chapter in Toronto.

Thank you to everyone, past and present, that I have been lucky to work with in the Family Genomics Research Group at MU and the Autism and Neurodevelopmental Research Group at TCD. Tom and Sarah-Marie it has been so fun to go through this PhD journey with you both. I am excited to see what is next as it comes to an end for all of us. Cathal, Niamh, and Ciara, I cannot thank you enough for the advice and motivation- especially at the times that you probably didn't even realise you were giving it. Thank you for letting me laugh with you in your office for too many hours! Thank you to Nadia for always keeping me on track. To Aoife, Aoife, Cathy, and Shane- it has been so energising to work with you all at the beginning of our new research team. Thank you for all your help!

Thank you to the Neuropsychiatric Genetics Research Group at TCD. Elaine, thank you for the opportunities that you extended to me at the beginning of my PhD. The training and insights you shared with me have carried me through to this point. To Office 1.17 and all of TTMI, thank you for the fun and for lovely place to work for over 2 years. Thank you to the Biology Department at Maynooth University who were so welcoming to me when I joined. Thank you, Dr Adeline Cooney for facilitating Writing Retreat sessions without which this thesis would not be written. Thank you to my PhD Progress Committee for your advice.

Thank you, Mam, and to all of my friends and family- I could not have done it without you. Finally, thank you to Joseph for your love and encouragement always.

Related Publications and Presentations

Journal Articles

Ní Ghrálaigh, F., Gallagher, L. and Lopez, L. M. (2020) 'Autism spectrum disorder genomics: The progress and potential of genomic technologies', Genomics, 112(6). doi: 10.1016/j.ygeno.2020.09.022 (Appendix IV-I).

Ní Ghrálaigh, F. *et al.* (2022) 'Brief Report: Evaluating the Diagnostic Yield of Commercial Gene Panels in Autism', Journal of Autism and Developmental Disorders 2022. Springer, pp. 1–5. doi: 10.1007/S10803-021-05417-7 (Appendix IV-II).

Conference Presentations

"Genomic syndromes in autism: Using whole genome sequencing to investigate multiplex families with autism and associated neurodevelopmental conditions." Fiana Ní Ghrálaigh, Aoife Coghlan, Louise Gallagher & Lorna M. Lopez
Poster presented at Genomics of Rare Diseases (Wellcome Connecting Science), April 2022 (Appendix III-I).

"Determining the clinical utility of gene panels in autism; a study of diagnostic yield, relevance, and penetrance." Fiana Ní Ghrálaigh, Thomas Dinneen, Ellen McCarthy, Daniel N. Murphy, Louise Gallagher & Lorna M. Lopez
Poster presented at the World Congress of Psychiatric Genetics, October 2021 (Appendix III-II).

"Evaluating the diagnostic yield of commercial gene panels in autism." Fiana Ní Ghrálaigh, Ellen McCarthy, Daniel N. Murphy, Louise Gallagher & Lorna M. Lopez
Poster presented at the Irish Society for Human Genetics, September 2021 (Appendix III-III).

"Application of an evidence-based curation framework to aid gene discovery: a pilot investigation in an autism family cohort." Fiana Ní Ghrálaigh, Louise Gallagher & Lorna M. Lopez
Poster presented at the World Congress of Psychiatric Genetics, October 2020 (Appendix III-IV).

“Rare genetic variation in autism; an exome sequencing study.” Fiana Ní Ghrálaigh, Cathal Ormond, Elaine Kenny, Louise Gallagher & Lorna M. Lopez
Poster presented at the Irish Society for Human Genetics, September 2020 (Appendix III-V).

“Analysis Pipeline of Whole Genome Sequencing Data in Neurodevelopmental Disorders.” Fiana Ní Ghrálaigh, Niamh M. Ryan, Louise Gallagher, Lorna M. Lopez
Poster presented at the British Neuroscience Association Festival of Neuroscience, April 2019 (Appendix III-VI).

“A Search for Rare Variants in a Family-Based Study of ASD.” Fiana Ní Ghrálaigh, Jessica E. Smith, Elaine Kenny Louise Gallagher & Lorna M. Lopez
Poster presented at the World Congress of Psychiatric Genetics, October 2018, and the Irish Society for Human Genetics, September 2018 (Appendix III-VII).

List of Abbreviations

(g)VCF	(genome) Variant Call Format
ACMG	American College of Medical Genetics and Genomics
ADHD	Attention Deficit Hyperactivity Disorder
alt	Alternative Allele
ASD	Autism Spectrum Disorder
BAM	Binary (Sequence) Alignment Map
BQSR	Base Quality Score Recalibration
BWA	Burrow-Wheeler Aligner
CADD	Combined Annotation-Dependent Depletion
ClinGen	Clinical Genome Resource
CMA	Chromosomal Microarray
CNV	Copy Number Variant
dbNSFP	Database of Non-Synonymous Functional Prediction
DDD	Deciphering Developmental Disorders
DSM-V	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
DTC	Direct-To-Consumer
DZ	Dizygotic
ExAC	Exome Aggregation Consortium
FPE	Female Protective Effect
GATK	Genome Analysis Tool-Kit
GWAS	Genome-Wide Association Study
HGMD	Human Gene Mutation Database
HGNC	HUGO Gene Nomenclature Committee
HUGO	Human Genome Organisation
HWE	Hardy-Weinburg Equilibrium
IBD	Identity By Descent
IBS	Identity By State
ID	Intellectual Disability
Indel	Insertion Deletion
ISHG	Irish Society for Human Genetics
ISPG	International Society for Psychiatric Genetics
KB	Kilobase
LoF	Loss-Of-Function
MAF	Minor Allele Frequency
MZ	Monozygotic

NGS	Next-Generation Sequencing
OMIM	Online Mendelian Inheritance in Man
PCA	Principal Component Analysis
PGS	Polygenic Score(ing)
QC	Quality Control
ref	Reference allele
SFARI	Simon's Foundation Powering Autism Research Initiative
SNV/SNP	Single Nucleotide Variant/Polymorphism
SPARK	Simons Power Autism Research Knowledge
SV	Structural Variant
TCD	Trinity College Dublin
Ts/Tv	Transition/Transversion
UCSC	University California Santa Cruz
VEP	Variant Effect Predictor
VQSR	Variant Quality Score Recalibration
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Table Of Contents

List of Tables

List of Figures

Chapter 1. Introduction

The contents of this chapter have been adapted from the following article:
Ní Ghrálaigh, F., Gallagher, L. and Lopez, L. M. (2020) 'Autism spectrum disorder genomics: The progress and potential of genomic technologies', Genomics, 112(6). doi: 10.1016/j.ygeno.2020.09.022 (Appendix IV-I).

1.1.1 An overview of the autism phenotype

Autism Spectrum Disorder (ASD), hereafter referred to as autism, is a prevalent neurodevelopmental condition occurring in around 1% of individuals in a population (Baird *et al.*, 2006). Autism is characterised by social communication difficulties and restricted repetitive behaviours (American Psychiatric Association, 2013). Gene discovery is complicated by the complexity of phenotypic heterogeneity in autism (Figure 1-1). Autism ranges in severity and manifestation both between affected individuals and in the same individuals across their lifespan and across behavioural, cognitive and language domains (Anderson *et al.*, 2007). Improving genetic understanding of autism and autism-relevant phenotypes will help in defining endophenotypes within autism and developing a more targeted approach to clinical management (Jeste and Geschwind, 2014).

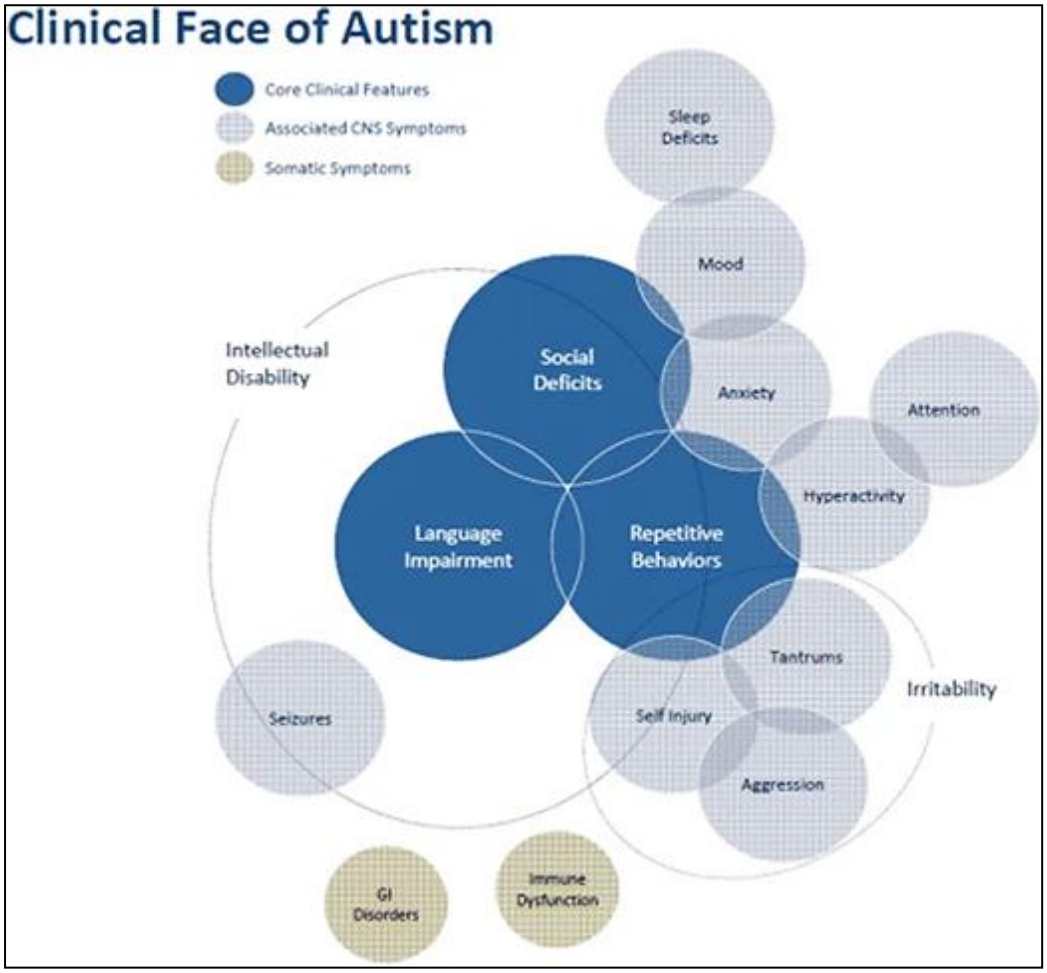


Figure 1-1 Phenotypic heterogeneity in autism.

Taken from (Kas *et al.*, 2014)*: "A schematic representation of core clinical features of ASD, associated central nervous system symptoms and somatic symptoms that are often observed in ASD patients".*

1.2 The genetic basis of autism

Autism is a highly heritable complex trait. The heritability of autism measures the genomic variation contributing to the phenotype and in autism has been estimated at ~ 80-90% (Tick *et al.*, 2016; Sandin *et al.*, 2017). The genetic risk of autism is contributed to by both rare and common genetic variants, and as yet the majority of the genetic risk remains unexplained (Gaugler *et al.*, 2014; De La Torre-Ubieta *et al.*, 2016). Rare variants refer to those occurring at less than 5% of the population and very rare variants occur at a minor allele frequency (MAF) of less than 1%. Common genetics variants typically refer to genetic variants with a MAF of greater than 5%. Rare variants, particularly those occurring *de novo*, have the potential to occur at higher effect sizes than common variants. The larger effect size of rare variants is in line with the hypothesis that variants of a higher effect sizes have a more detrimental effect on brain development resulting in the early-life manifestation of the autistic phenotype, when compared to neuropsychiatric conditions most commonly arising later in life, such as schizophrenia and psychosis.

This introduction aims to inform on state-of-the-art autism genomics research. The focus is on the application of genome sequencing technologies to search for these genetic variants in extensive sample collections that have transformed our understanding of autism genomics. This introduction reviews cutting-edge research that uses genome sequencing methods, bioinformatic processing and clinical implementation for improved diagnosis and medical decision-making in autism and other neurodevelopmental conditions. It explains the value of genome sequencing technologies and highlights what they can achieve for neurodevelopmental and neuropsychiatric conditions.

1.3 The heritability of autism

There is clear evidence that autism has a genetic basis. The heritability (denoted as h^2) of autism measures the genomic variation contributing to the phenotype. Heritability measures the proportion of genetic variance in a phenotype in the population. Measures range from 0 to 1, with a measure towards 0 indicating high environmental contribution and a measure towards 1 indicating a strong genetic contribution to a phenotype. Recent meta-analysis investigating heritability estimate h^2 at 80-90% (Tick *et al.*, 2016; Sandin *et al.*, 2017; Bai *et al.*, 2019).

Traditionally measures of heritability arise from twin studies, on the basis that monozygotic twins will possess nearly identical DNA sequences while dizygotic twins should by chance share approximately 50% of their genetic sequence. Higher concordance, i.e., affectation in both twins, in monozygotic pairs (MZ) when compared to dizygotic pairs (DZ) indicates a

strong genetic component to condition manifestation. With relatively constant environment between twins, the heritability estimate of 0.8 in autism indicates that 80% of the variability of the condition in the population is due to genetic differences between individuals. While twin studies highlight the expected contribution of genetic variation to autism risk, it is important to note that heritability estimates based on clinical genetics studies are limited, but the evidence is still strongly in favour of a heritable component (Wray & Visscher, 2008). Although the accuracy of these heritability estimates for autism are not exact, there is evidence for a major genetic component to the occurrence of the condition.

Recurrence risk represents a further key measure of genetic effects on autism. It refers to the probability of parents of a child with autism giving birth to another affected child. Recurrence risk measures the level of aggregation of the condition in a family, in turn giving insight into the contribution of shared genetics to the phenotype. The recurrence risk of autism is estimated to fall between 3% and 10% (Chakrabarti and Fombonne, 2001; Lauritsen, Pedersen and Mortensen, 2005). This estimate is true for the first affected sibling in a family and increases with additional affected siblings, moderated by sex of the affected proband (Werling and Geschwind, 2015). A major limitation is that the sibling recurrence risk may be underestimated because of genetic stoppage. A study controlling for this factor reported a higher recurrence estimate of 18.7% of infants with at least one affected sibling developing autism (Ozonoff *et al.*, 2011), indicating genetic stoppage occurring in families affected by autism. Taken together these lines of evidence from twin and family studies in autism clearly show a substantial genetic contribution to susceptibility of the condition.

International collaborative efforts accelerated by advances in sequencing technologies aim to discover genetic variation associated with autism (Table 1-2). Genetic variation can come in the form of highly penetrant rare genetic variation, or variants that are common in the population and typically having a lower effect on genetic risk. Rare genetic variant discovery is particularly successful when using a family-based study design, while common genetic variants are identified through population-based studies (Yuen *et al.*, 2015; Feliciano *et al.*, 2019; Grove *et al.*, 2019).

Many genes associated with autism affect synapses or gene regulation and some more broadly affect gene regulation (Satterstrom et al., 2020). Yuen *et al.*, along with other large studies, compiled a series of functionally annotated gene lists against which rare variants may be searched, such as axon guidance pathways, synapse pathway or neuron projection (Yuen *et al.*, 2015). Further to these autism-associated processes, genes associated with

schizophrenia and ID may be informative to consider in analyses (Iossifov *et al.*, 2014). This is due to the shared global gene expression pathways identified among some psychiatric conditions (Gandal *et al.*, 2018).

1.4 Sequencing technologies have advanced the identification of rare variants

Genome sequencing, specifically whole exome sequencing (WES) and whole genome sequencing (WGS), has transformed variant discovery. These technologies give the opportunity for more widespread and in-depth genomic analysis than older techniques, such as microarray studies and candidate gene studies, have allowed. Table 1-1 lists the next-generation sequencing (NGS) technologies that can identify single nucleotide variants (SNVs) and insertion-deletion variants (indels), as well as larger genomic hits, including structural variants (SVs) or copy number variants (CNVs), across the allele frequency spectrum. In the past decade, sequencing technologies have stretched from covering select points across to genome to cover up to 100%, when sequenced at high coverage with *de novo* assembly (Table 1-1) (Miga *et al.*, 2020). Higher coverage WGS results in more precise variant calls across the coding and non-coding regions of the genome.

These advances in genomic technologies and decreasing costs have enabled large sequencing cohorts (Table 1-2), allowing key strides to be made in the field of autism genomics. Large-scale analyses of these cohorts have identified hundreds of autism-associated genetic variants across the genome. For example, discovery of rare variants, particularly rare CNVs, affecting *SHANK3* and *NRXN1* among other genes, implicated synaptic transmission and plasticity in autism neurobiology (De Rubeis *et al.*, 2014). Extending beyond variant discovery, combining rare variant analysis with single-cell investigation in the developing human cortex showed enriched expression of particular autism-associated genes in maturing and mature excitatory and inhibitory neurons from mid-fetal development, and helped to validate the role of these genes in neuronal communication and regulation of gene expression (Satterstrom *et al.*, 2020). Impactful findings such as these, suggest great potential for advancing our understanding of autism neurobiology through rare variant discovery. Key findings arising from these data and the impact of the variation detected is detailed in section 1.7.

	Exome Sequencing		**Whole Genome Sequencing**	
	Clinical Exome Sequencing	**Whole Exome Sequencing**	**Short-Read**	**Long-Read**
% Genome covered	~0.5%	~1%	~90%	Potential for up to 100%
Types of variant detected	SNVs Indels CNVs (limited)	SNVs Indels CNVs (limited) SVs (limited) Mitochondrial	SNVs Indels CNVs SVs Mitochondrial Repeat expansions (including tandem repeats (Mousavi *et al.*, 2019; Mitra *et al.*, 2021)	SNVs Indels CNVs SVs Mitochondrial Repeat expansions Complex SVs Haplotype phased variants Methylation
Diagnostic yield in autism	Limited application	31% (Srivastava *et al.*, 2019)	42.4% (Yuen *et al.*, 2015)	Not yet available
Cost estimate	€37.19[a]	€79.33[b]	€1,239.50[c]	€918[d]

Table 1-1 Genomic technologies compared.

Outlined are four key sequencing technologies with potential for use to identify rare autism genetic variants. Note that these costs are estimates and do not include library preparation costs, barcodes, access fees, labour, VAT, service provider, data processing and data storage and other associated sequencing costs. a) SOPHiA GENETICS Clinical Exome Solution (12Mb covering ~4500 genes (2.5Gb/sample/800 samples/flowcell)) b) Illumina Nextera Rapid Capture Exome (37Mb (8Gb/sample/375 samples/flowcell)) c) WGS (120Gb/sample/24 sample/flowcell). Estimates a), b) and c), are based on sequencing with Illumina NovaSeq S4 flowcell (2x150) up to 3000Gb/flowcell. d) Oxford Nanopore Technologies (60X; 1 sample/flow cell/180GB) Sequencing metrics: https://nanoporetech.com/accuracy Acronyms; SNV single nucleotide variant, Indel insertion deletion, CNV copy number variant, SV structural variant.

1.5 Common genetic variants have been challenging to associate with autism

The search for common genetic variants has been less successful than that in more typically adult-onset neuropsychiatric conditions, in particular schizophrenia (~7% of variance on the liability scale) (Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.*, 2014) and bipolar disorder (~2.5% of variance on the liability scale) (Psychiatric GWAS Consortium Bipolar Disorder Working Group *et al.*, 2011; Creese *et al.*, 2019; Stahl *et al.*, 2019). The largest study to date investigating common genetic variants in autism, using genome-wide genotyping, provides evidence for statistically significant association of the first common risk variants with autism. A Genome Wide Association Study (GWAS) was carried out on 18,381 autism cases and 27,969 controls. While this sample size is large in terms of autism, it is smaller than that of other traits such as schizophrenia with 36,989 cases or bipolar disorder with 20,352 cases (Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.*, 2014; Stahl *et al.*, 2019). Five loci showed significant association with autism alone and seven further loci were identified upon analysis of schizophrenia, depression and educational attainment together (Grove *et al.*, 2019). Polygenic burden, measured by a polygenic score (PGS), is the combined impact of common variants on the probability of a phenotype. In autism this explains just 2.5% of the observed variance in risk (Grove *et al.*, 2019). The lower yield of common variant loci in autism may be because of a greater relative contribution of rare genetic variants than common variants in the genetic architecture of autism (Vorstman *et al.*, 2013). However, the current smaller sample sizes in GWAS of autism fail to validate this hypothesis.

Cohort	Size of Cohort	Study Design	Dataset	Reference
Australian National Autism Consortium	48 cases and 80 parent controls	Simplex & Multiplex	WES	(An *et al.*, 2014)
Autism Genetic Resource Exchange (AGRE)	> 1,700 families	Simplex & Multiplex	Genome-Wide Genotyping	https://www.autismspeaks.org/agre (Leppa *et al.*, 2016)
Autism Sequencing Consortium (ASC)	12,772 individuals	Case Control, Simplex	WES	(Buxbaum *et al.*, 2012; Satterstrom *et al.*, 2020)
Deciphering Developmental Disorders (DDD)	12,000 individuals and their parents	Simplex	Genotyping & WES	https://www.ddduk.org/ (Wright *et al.*, 2015; Gardner *et al.*, 2019)
iHART	2,308 individuals (from 493 AGRE families)	Quad & Multiplex	WGS	(D. Kashef-Haghighi et al., 2016; Ruzzo *et al.*, 2019)
iPsych Danish Cohort	16,146 cases (Genotyping) and 4,811 cases (WES)	Case Control	Genome-Wide Genotyping (PsychChip array from Illumina) & WES	(Pedersen *et al.*, 2018; Satterstrom *et al.*, 2018, 2020)
MSSNG	11,312 individuals (4,258 families)	Simplex & Multiplex	WGS	https://research.mss.ng/ (Yuen *et al.*, 2015, 2016, 2017; Woodbury-Smith *et al.*, 2017; Brandler *et al.*, 2018)
Simons Foundation Powering Autism Research for Knowledge	27,615 individuals (Genotyping & WES) and 400 quad families (WGS)	Simplex & Multiplex	Genome-Wide Genotyping (Illumina InfiniumCoreExome-24), WES & WGS	https://sparkforautism.org/ (Feliciano *et al.*, 2019)

Simons Simplex Collection (SSC)	8,975 individuals (WGS) 2,517 families (WES) and 10,220 individuals (Genotyping)	Quartet (Phase 1-3) & Trio/Incomplete family data (Phase 4)	Genome-Wide Genotyping, WES & WGS	https://www.sfari.org/resource/simons-simplex-collection (Levy *et al.*, 2011; Sanders *et al.*, 2011; An *et al.*, 2018; Brandler *et al.*, 2018; Werling *et al.*, 2018; Zhou *et al.*, 2019; Satterstrom *et al.*, 2020)
The Autism Genome Project *a consortium including TASC and AGRE samples	7,917 individuals (1,492 families)	Simplex & Multiplex	Genome-Wide Genotyping (10K SNP array and 400 microsatellite marker panel)	(The Autism Genome Project Consortium *et al.*, 2007; Pinto *et al.*, 2010)
The Autism Simplex Collection (TASC)	5,444 individuals (1,719 families)	Simplex	Genome-Wide Genotyping (Illumina 1M SNP) & WES	(Buxbaum *et al.*, 2014; Sanders *et al.*, 2015)
The Psychiatric Genetics Consortium	18,381 cases	Case Control	Genome-Wide Genotyping	(Grove *et al.*, 2019)

Table 1-2 Key autism genomics cohorts.

Featured in the table are large-scale autism cohorts used in genomic studies to date. Note that there is significant overlap of samples between these cohorts, for example, the MSSNG cohort includes samples from both AGRE and TASC. These details are subject to frequent update. Reference refers to the original research article/website linked to the cohort and research studies cited in this review that analyse these cohorts.

1.6 Heterogeneity in the genetic architecture of autism

Autism displays a high level of heterogeneity across a phenotypic spectrum, both between individuals and within the same individual throughout the lifespan. It is estimated that around 10% of individuals affected with autism have a syndromal form of the condition, for which each single autism risk gene accounts for at most 1% of overall cases on average (Abrahams and Geschwind, 2008). Rare disorders often manifest with an underlying autistic phenotype (MENDELIAN.CO, 2019). These syndromes are frequently caused by highly penetrant variants in single genes, such as Fragile X syndrome, MIM:30024 (*FMR1*), and Tuberous Sclerosis Complex, MIM: 613254 (*TSC2*) (reviewed in Betancur, 2011). These syndromic forms of autism are frequently associated with intellectual disability (ID) and developmental delay, suggesting that autism may only form part of the overall behavioural phenotype of the syndrome.

Autism cases that do not fall into clinically defined syndromes appear to have more complex genetic architecture and various models of risk have been suggested to encompass this. The polygenic model, strongly supported in schizophrenia (Tansey *et al.*, 2016), proposes that multiple loci, each contributing a small effect, accumulate to surpass a threshold of disease liability. In contrast, Boyle *et al.* proposed the omnigenic model (Boyle, Li and Pritchard, 2017; Liu, Li and Pritchard, 2019). This model suggests that all genes expressed in disease-relevant cells can influence pathogenesis, through their interference with the expression of "core genes." In that, it may be hypothesised that most of the heritability of autism could be explained by the effect of variation on genes outside of the core autism pathways.

Understanding gene regulation is critical to parsing out the relative contribution of common and rare variants to autism heritability. Whichever model is most appropriate in describing its architecture, rare genetic variants are crucial to understanding autism. Further to heterogeneity in the genetic architecture among autism cases, there is heterogeneity, both genetically and clinically, between males and females. Males are more frequently affected with autism than females (Fombonne, 2003). Although factors such as hormonal sex differences, sex-specific epigenetic factors and genetic factors related to sex chromosomes have been hypothesised to play a role in this bias, the biological basis remains unclear. A large-scale family study interrogating *de novo* variants in autism reinforces the importance of evaluation of the X chromosome, identifying 5 of 7 genes replicated in the study are located on the X chromosome (Turner

et al., 2019). Together with the evidence of sex biases of autosomal genes, this study highlights the potential for genomic studies to elucidate this phenomenon.

1.7 Rare variants disrupt gene function, dosage, and regulation in autism

Current WGS and WES technologies enable investigation of most genomic variant classes (Table 1-1). The consequences of such variants in the genome occur to varying effects with different degrees of penetrance, as outlined below.

1.7.1 Gene disruption

Gene disruption refers to the disturbance of gene expression and the impact of variation on overall gene function. The consequence of a genetic variant can be detrimental to gene function or can have little effect depending on the variant in question and the overall genome environment. Genes disrupted in autism often include those related to brain development, post-synaptic density, nerve impulse and neuron projection (Abrahams *et al.*, 2013). Much focus lies on the importance of LoF variants and damaging missense variants in the evaluation of genetic variation on autism. In particular, variants impacting evolutionarily conserved genes to the detriment of crucial cellular processes.

Another mechanism of gene disruption is gene rearrangement, encompassing translocations, inversions and large-scale insertions and deletions. Although varying between studies, the estimated rate of large variants in autism is approximately 5-10% (Veenstra-VanderWeele, Christian and Cook, Jr., 2004). A recent study implicates rare retro-transposition derived disruption in neurodevelopmental conditions through trio-based exome sequencing analysis from the Deciphering Developmental Disorders (DDD) cohort. This mechanism of disruption is an avenue for pathogenesis which has been largely unexplored in neurodevelopmental conditions to date (Gardner *et al.*, 2019).

1.7.2 Gene dosage

Gene dosage refers to the number of copies of a given gene that are present in the genome of an individual. Dosage has been found to play a substantial role in autism pathogenesis, as demonstrated through CNV analysis, i.e. analysis of duplication or deletion variants of >1Kb (Sanders *et al.*, 2015). In 2004, two groups independently identified that large scale CNVs were often overlapping with genic regions (Iafrate *et al.*,

2004; Sebat *et al.*, 2004). The influence of these CNVs means either an increase or depletion in activity of the contained genes with potential for damaging functional consequences. A comprehensive analysis identified clinically relevant CNVs in 10.5% of neurodevelopmental condition cases investigated, with 11.4% in autism cases. Importantly many of the CNVs identified were found to occur across multiple neurodevelopmental conditions (Zarrei *et al.*, 2019).

1.7.3 Gene regulation

As a complex trait, non-coding variants, particularly variants affecting gene regulation are likely to influence autism (Botstein and Risch, 2003). Advances in WGS and bioinformatic tools are enabling studies of non-coding regions of the genome. Yuen *et al.* estimated that non-coding and genic non-coding *de novo* variants account for 15.6% and 22.5% respectively, of predicted damaging *de novo* variants in autism cases. Non-coding elements, e.g. untranslated regions, regulatory sequences involved in exon skipping and DNAse hypersensitivity regions were most enriched for *de novo* variants (Yuen *et al.*, 2016). The first study significantly associating genome-wide non-coding variants with autism shows convergence in the pathways and processes disrupted by both coding and non-coding variants in autism, specifically in synaptic transmission and neuronal development (Zhou *et al.*, 2019). Ruzzo *et al.* also provided evidence that non-coding variants impact neurobiology in autism, reporting a recurrent 2.5KB deletion within the promoter of *DLG2*, a gene associated with cognition and learning in mice and human (Ruzzo *et al.*, 2019).

Preferential transmission of structural non-coding variants has been reported in autism, specifically the transmission of cis-regulatory elements from father to affected rather than to unaffected offspring (Brandler *et al.*, 2018). These findings are suggestive that not only are rare inherited non-coding variants increasing risk to autism, but also indicate a parent-of-origin effect from this non-coding variant class, highlighting a key benefit to the use of a family-based study design in studies of autism.

1.8 Family-based studies are key to rare variant analysis in autism

Family-based studies, previously the foundation of disease gene discovery, are re-emerging as an effective tool to identify potentially pathogenic variants in neuropsychiatric conditions, including autism (Glahn *et al.*, 2019). Family-based designs facilitate the analysis of parent to offspring variant transmission. These study designs take the form of i) simplex families (trios); parents and their affected child, ii) multiplex

families; parents with more than one affected child, and iii) more complex extended pedigrees with multiple affected individuals. By design, trio studies such as those investigating the MSSNG cohort (Table 1-2), have been particularly key to uncovering the enrichment of *de novo* variants in cases by comparing rates of *de novo* variants in affected offspring with their unaffected respective siblings (Yuen *et al.*, 2016).

Family-based study designs also enable analyses of parent-of-origin effects that are not possible in case-control design. Furthermore, the presence of matched unaffected siblings in these studies, gives a background level of genetic variation that can be used to distinguish between disease relevant variants and those that are unrelated, such as population-specific background variation or biases introduced in sequencing. A number of large-scale genomic investigations of autism apply a family-based approach, including the Simons Simplex Collection (Simplex), Autism Genetic Research Exchange (Simplex and Multiplex) and The Autism Genome Project (Simplex and Multiplex) (Table 1-2).

1.9 Multiplex and simplex cases of autism show different genetic architectures

Family structure plays a major role in the types of putative variants expected to be causative of a given autism proband. Earlier CNV studies in autism provided some evidence of differences in genetic architecture between simplex and multiplex families (Sebat *et al.*, 2007). These differences are centred on the contribution of *de novo* and inherited variants to autism susceptibility.

1.9.1 *De novo* variants

A lower rate of *de novo* variation is seen in multiplex families compared to simplex families, as expected by study design. Sebat *et al.* reported *de novo* CNVs in 10% of simplex cases and 3% of cases from multiplex families in their cohort (Sebat *et al.*, 2007). Similarly Ruzzo *et al.* give evidence for depletion of rare *de novo* autism risk variants in multiplex families (Ruzzo *et al.*, 2019). While, this is observed across multiple studies, the difference between multiplex and simplex family structures is not consistently evident. In their CNV analyses, Pinto *et al.* did not report such differences (Pinto *et al.*, 2010). A limitation to these analyses, such as analyses involving the Autism Genome Project cohort (Table 1-2), arises from challenges in reporting of simplex/multiplex status, i.e., identifying a family as a true simplex, or as a family for which just one offspring was investigated.

1.9.2 Inherited variants

Consistent with the enrichment of *de novo* variants in simplex cases of autism, there is a depletion of inherited variants associated with autism in these spontaneous cases (Sebat *et al.*, 2007; Ronemus *et al.*, 2014). Klei *et al.* estimate narrow sense heritability to exceed 60% for autism cases in multiplex families but estimate just 40% of narrow sense heritability for simplex families (Klei *et al.*, 2012). This means that 60% of phenotypic variance may be attributed to additive genetic variance in individuals of multiplex families. As in comparison of *de novo* variant enrichment of simplex and multiplex families, this effect is not reported consistently across analyses.

Interestingly, the same putative variant may not be found in all affected individuals within a multiplex family as highlighted recently (Feliciano *et al.*, 2019). This study reports a maternally inherited 15q11.2 deletion in an affected male child and no paternally inherited putative variants from an affected father. Other studies have identified non-sharing of CNVs (Leppa *et al.*, 2016) and SNVs in members of multiply affected families. In the latter study the two affected siblings did not harbour the same rare risk variant in more than half of the multiplex families studied (Yuen *et al.*, 2015). Similarly, pathogenically significant CNVs have been identified that are transmitted to an autism proband from an unaffected parent, and shared with a unaffected sibling (Woodbury-Smith *et al.*, 2017), adding to evidence for asymptomatic carriers of neurodevelopmental condition CNVs.

Family studies in epidemiological cohorts from isolated populations have also confirmed that both rare and common genetic variants contribute to the susceptibility to autism. A study on the Faro Island genetic isolate, affirms the importance of both common and rare variants in autism susceptibility (Leblond *et al.*, 2019). This study identifies in a subset of individuals in the cohort carrying rare deleterious variants in genes known already associated with autism and in this same cohort, common genetic variants were also associated. Given these two mechanisms of genetic variation, *de novo* and inherited in autism, genome sequencing studies in families with multiple affected individuals offers greater opportunity to understand the relative contribution of inherited and *de novo* variation in the genetic architecture of autism.

1.10 Establishing putative autism variants faces many challenges

Heterogeneity in autism diagnoses is a major challenge facing genome sequencing studies in autism. In particular, diagnosis of autism in the presence of intellectual

disability. Diagnostic procedures are found to differ between that used in a clinical and research setting. For a comprehensive discussion on these challenges refer to Schaaf *et al.* 2020 (Schaaf *et al.*, 2020). The greatest challenge in analysis of large-scale genomic data is in the establishment of pipelines for data interpretation. Interpretation of putative variants is complicated by a wide variety of technical factors, such as sequence coverage, variant validation, consistency in sequencing platforms and variant calling and filtering techniques. Robust clinical diagnoses and rich phenotyping increase confidence in variant association (Callaghan *et al.*, 2019). A variant that has been associated with autism and has substantial evidence supporting its validity will be interrogated for its biological role.

Variants associated with autism disrupt a wide variety of pathways and biological processes (De Rubeis *et al.*, 2014). Identifying pathways and processes showing an increased mutational burden enables the isolation of cellular processes and pathways disrupted in autism. Gene-lists are often compiled listing genes involved in a given process (Yuen *et al.*, 2015). These lists are useful in establishing the process which a putative variant may be disrupting, and such gene lists are often consulted for membership when investigating the impact of a variant (Feliciano *et al.*, 2019).

The establishment and maintenance of collective databases, such as SFARI Gene (Abrahams *et al.*, 2013), DDD gene2phenotype (Wright *et al.*, 2015) and ClinVar (Landrum *et al.*, 2014), that are openly shared among researchers give hope for the development of variant specific disease models which will expectedly lead to a greater understanding of autism pathology. Consistent re-analysis of pathogenicity is key to gaining maximum insight from available genomic data, as proven fruitful in the re-annotation of developmental and epileptic encephalopathies genes (Steward *et al.*, 2019) (Figure 7-1). A key stride in the development of an autism gene list comes from Schaaf *et al.* in their proposal to adapt the Clinical Genome Resource (ClinGen) curation framework to autism (Schaaf *et al.*, 2020). Development of a high-confidence gene list for autism would have great use in genomic investigation, specifically in the development of targeted gene panels and a ‘clinical exome’. Without a consensus gene list in autism, attempts to develop such genome analysis strategies have limited application (Table 1-1).

Advances in long-read sequencing technologies hold the potential for sequencing of “dark gene regions,” genomic regions inaccessible through NGS. With high coverage

and *de novo* assembly, Nanopore technologies have potential to sequence up to 100% of the genome (Table 1-1), with the greatest level of 'recovered' genes when compared with other genomic technologies, including the recovery of genes associated with autism (Ebbert *et al.*, 2019). This technology, to our knowledge, has yet to be applied to autism cohorts, aside from use in variant validation (Brandler *et al.*, 2018). Long-read sequencing will enable discovery of genetic variants which have thus far been largely under-explored in autism, such as repeat expansions, haplotype phased variants and methylation changes. Repeat expansion variants have already been associated with autism, most notably the *FMR1* repeat expansion associated with Fragile X syndrome (MIM: 30024). As shown in an early haplotype mapping study, identification of haplotypes can succeed in identifying loci involved in autism susceptibility (Casey *et al.*, 2012). Even more relevant perhaps, long-read sequencing enables the detection of CNVs and rearrangement events without the need for bioinformatic re-assembly and alignment of short reads.

1.11 Putting autism in the context of other neuropsychiatric conditions

WGS has potential to investigate some of the major questions remaining unanswered in autism genomics, including investigation of the overlap of autism with other neurodevelopmental and neuropsychiatric conditions, both clinically and genetically. As highlighted in a review from Lord *et al.*, elucidation of the genetic overlap of autism with other neuropsychiatric conditions is needed (Lord *et al.*, 2020). Clinically, autism frequently co-occurs with other neuropsychiatric conditions, in particular attention-deficit hyperactivity disorder (ADHD) (28%), anxiety disorders (13%) and mood disorders (11%) (Lai *et al.*, 2019).

At the systems-level there is substantial evidence of genetic overlap between autism and neurodevelopmental and neuropsychiatric conditions (An and Claudianos, 2016). There is overlap in the genes associated with autism and those associated with other neuropsychiatric conditions, such as schizophrenia and bipolar disorder (Carroll and Owen, 2009) (Geschwind and Flint, 2015; Lee *et al.*, 2019). This has been demonstrated strongly in a large-scale meta-analysis of eight European psychiatric cohorts identifying 109 pleiotropic loci (Lee *et al.*, 2013). The genetic overlap of autism with other conditions is also evident at the variant-level with *de novo* variation in autism shared with intellectual disabilities (Satterstrom *et al.*, 2020) and shared with epilepsy (Heyne *et al.*, 2018).

1.12 Next-generation sequencing technologies improve diagnostic yield

There is a demand for clinical genetic testing in autism (Barton *et al.*, 2018). Clinical CNV detection has already been translated widely, advancing the clinical genetics understanding of the condition. This translation crystallised some of the issues that will emerge with widespread translation of genomic technologies, namely clinical interpretation, relative contribution of inherited variants and particularly variant specificity to autism. Currently no gene, which when disrupted by a pathogenic variant, has been found to confer risk to autism without conferring risk to ID or other neurodevelopmental conditions. In the absence of appropriate study design and explicit, robust diagnoses, there is insufficient evidence to assign meaningful specificity of gene involvement in autism (Myers, Challman, Bernier, *et al.*, 2020).

Genomic technologies, given the greater proportion of the genome covered, have the potential to transform the clinical genetic understanding of the condition. This is illustrated by the increase in diagnostic yield with genomic technologies. Diagnostic yield refers to the number of cases where a putative genetic variant associated with the condition is identified in a cohort. This can be interpreted as a measure of the utility of the technique and analysis strategy for the condition.

A recent meta-analysis scoping review states that exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental conditions, defined in this study as developmental delay, ID and/or autism (Srivastava *et al.*, 2019). The diagnostic yield for WES overall from these meta-analyses is 36%, surpassing the estimated 15-20% diagnostic yield of candidate gene arrays.

Using WES technologies, Feliciano *et al.* in the SPARK pilot, report a returnable genetic result in 10.4% of their cohorts affected offspring (Feliciano *et al.*, 2019). Importantly, in individuals with more complex phenotypes, such as autism with seizures or co-occurring ID, they report a higher diagnostic yield than overall (27% and 20% respectively). This finding is consistent with other studies (Tammimies *et al.*, 2015; Srivastava *et al.*, 2019). The SPARK study also reports a higher diagnostic yield in cases from multiplex families (15.2%) than simplex families (10.1%) (Feliciano *et al.*, 2019).

Yuen *et al.* find a diagnostic yield of autism-relevant variants using WGS to be 42.4% in their cohort of 85 multiplex families of autism. This mirrors the diagnostic yield estimated in ID using the same sequencing platform (Gilissen *et al.*, 2014; Yuen *et al.*, 2015). The

increased diagnostic yield using WGS highlights the great potential for use of the technology in families with autism. This estimate can be expected to increase further with developments in variant interpretation strategies and increases in sample sizes, giving more power to investigations of common variants and variants in the non-coding regions of the genome.

The clinical utility of WGS holds great promise; however, this sequencing approach also faces major challenges. These include the need for large-cohort analyses and the failure to replicate genomic findings. One example is the report of the enrichment of *de novo* and private disruptive mutations within fetal CNS DNase I hypersensitive sites within 50kb of genes that have been previously associated with autism risk (Turner *et al.*, 2016) that later did not replicate despite a larger sample size (Turner *et al.*, 2017). Furthermore, we face limitations to the current capacity to interpret variants in the non-coding genome, as discussed by Lee & Gleeson (2020) (Lee and Gleeson, 2020). Notwithstanding these challenges, the decrease in sequencing costs (Table 1-1) and the increase in sample sizes under investigation, together with the greater understanding of family inheritance will continue to give a more precise estimate of the diagnostic yield in autism. The return of genetic results, alongside current behavioural diagnoses, may be used to improve therapeutic avenues in the future. Genetic diagnoses may also be used to inform family planning on a family-by-family basis as illustrated by a recent family study showing the CNV findings, which would have been pre-symptomatically predictive of autism or atypical development in 7% (11 of 157) of families analysed (D'Abate *et al.*, 2019).

1.13 Conclusion

WGS is the most effective technology to improve our biological understanding of neurodevelopmental conditions. With near full coverage of the human genome, coupled with the increase in sample sizes, detailed phenotyping, and the development of cutting-edge analytical methods, we now have the potential to identify more variants across the genome, in particular more rare pathogenic genetic variants. The detection of rare variants by genomic technologies will improve our understanding of the genetic architecture of autism and other neurodevelopmental and neuropsychiatric conditions. With advances in biological interpretation enabling delivery of genetic discovery into clinical translation, genomic technologies will become an achievable step towards personalised family medicine, ultimately aiding autism diagnosis and informing medical decision-making.

1.14 Aims

This thesis aims to investigate rare variation and its contribution to the genetic basis of autism. The work outlined in this thesis aims to apply an analysis strategy for isolation of rare exonic pathogenic SNVs from NGS data, specifically genome sequencing of an autism cohort of affected individuals in a family-based study design (WES n=42, WGS n=35). The aim of this work is to identify rare putatively pathogenic SNVs in genes with evidence supporting their role in autism, using a family-based study design to evaluate variant transmission. Variants emerging from these analyses contribute to the existing evidence supporting association of relevant genes with autism.

Additionally, this thesis investigates the clinical utility of genome sequencing in autism. Genetic diagnosis in autism is limited by the ability to robustly determine the relevance of putatively pathogenic genetic variation. Through application of an evidence-based gene curation framework to dissect gene-phenotype relationships and through investigation of the diagnostic yield of commercial gene panels available for use in autism, this thesis aims to inform on current strategies to translate genomics findings into the clinic.

Chapter 2. Materials and Methods

2.1 Dataset description

2.1.1 Dataset description of Cohort 1

2.1.1.1 Ethics and ascertainment

Cohort 1 was selected from the existing Autism and Neurodevelopmental Disorders Research Group TCD DNA biobank (n=808) under ethics approval “Irish Molecular Genetics Study in Autism REC: 2020-01 List 1 (17)” (Appendix I-I).

Candidate sample selection for inclusion was carried out as presented in Figure 2-1. This sample collection has previously been included in TASC and UK10K sequencing studies and samples that were sequenced through these projects were excluded as candidates for this sequencing study to avoid duplicate sequencing (Buxbaum *et al.*, 2014; The UK10K Consortium, 2015). Inclusion and exclusion criteria for TASC is outlined in Buxbaum *et al.,* 2014 and inclusion and exclusion criteria for UK10K are outlined in The UK10K Consortium, 2015. Where DNA samples did not reach criteria for sequencing, samples were excluded as candidates for this cohort (>30ng/ul and 260/280>1.8). Complete phenotypic records were needed for all candidates for this cohort to determine ASD diagnosis status. ASD diagnoses were confirmed from clinical expertise.

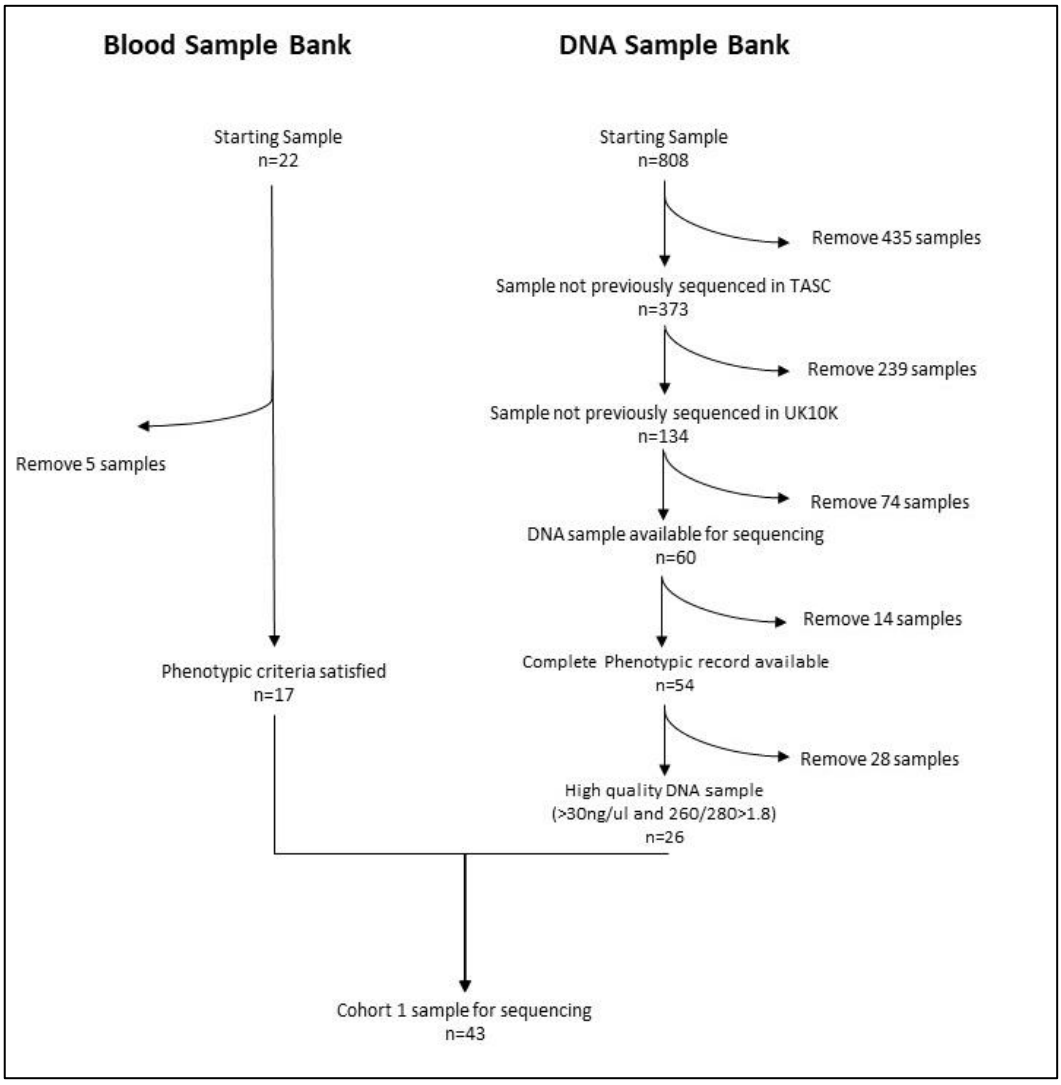


Figure 2-1 Cohort 1 selection criteria.

Presented in the figure is the flow of candidate samples for inclusion within Cohort 1. TASC; is The Autism Simplex Collection.

2.1.1.2 Cohort structure and phenotype

This cohort includes 23 individuals affected with autism (19 male and 4 female). All probands (n=23) have a diagnosis of autism or ASD according to Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) criteria, with n=10 having a co-occurring neurodevelopmental condition diagnosis (Table 2-1).

Cohort Overview	N = 42
Number of families	2 quad families 6 trio families 2 parent-child pairs 1 set of DZ twins 11 affected singletons
Male probands	N=19
Female probands	N= 4

Table 2-1 Overview of Cohort 1.

Outlined are the family structures included in the cohort and proband counts by sex.

2.1.2 Dataset description of Cohort 2

2.1.2.1 Ethics and ascertainment

A subset of Cohort 1 was selected for further analysis using WGS based on phenotype severity, under ethics approval “Irish Molecular Genetics Study in Autism REC: 2020-01 List 1 (17)” (Appendix I-I). Samples were prioritised for WGS from Cohort 1 on the basis of their hypothesised rare variant burden. For this reason, two female probands were selected and one male proband with a complex syndromal phenotype, with hypothesised rare penetrant SNVs being causative.

2.1.2.2 Cohort structure and phenotypes

All probands (n=3) have a diagnosis of autism according to DSM-V criteria (Table 2-2 and Table 2-3).

Cohort Overview	N = 6
Number of families	1 quad family 2 affected singletons
Male probands	N=1
Female probands	N= 2

Table 2-2 Overview of Cohort 2.

Outlined are the family structures included in the cohort and proband counts by sex.

FID	IID	Sex	Phenotype
AS315	AS315C	F	Autism
AS322	AS322C1	F	ASD, ADHD
AS420	AS420C1	M	Autism, moderate ID, self-injurious behaviour, catatonia, dysmorphology
AS420	AS420C2	M	Unknown
AS420	AS420F	M	Unknown
AS420	AS420M	F	Unknown

Table 2-3 Cohort 2 phenotype and sex.

Outlined are reported sex and clinically validated phenotype for individuals analysed within Cohort 2. Unknown is given as the phenotype where no neurodevelopmental or neuropsychiatric phenotype has been reported, however parent phenotyping was not performed.

2.1.3 Dataset description of Cohort 3

2.1.3.1 Ethics and ascertainment

Cohort 3 was ascertained under ethics approval "Genomics of Neurodevelopmental Disorders (Reference number BSRESC-2021-2402328)" (Appendix I-II). Inclusion criteria in the recruitment of this cohort is families with two or more family members affected with autism and a third family member affected with another neurodevelopmental or neuropsychiatric condition, including autism, neuropsychiatric conditions (e.g., schizophrenia or depression), ADHD, learning disability, developmental delay, Tourette's Syndrome, or epilepsy.

2.1.3.2 Cohort structure and phenotypes

Clinical diagnoses are made according to DSM-V criteria. Diagnoses have been confirmed through clinical reports with review and verification by Professor Louise Gallagher, Chair of Child and Adolescent Psychiatry at TCD.

Cohort Overview	N = 29
Number of families	4 multiplex families
Males	N=14
Females	N= 15

Table 2-4 Overview of Cohort 3.

Outlined are the family structures included in the cohort and proband counts by sex.

Cohort 3 is comprised of 4 multiplex pedigrees, as presented in Figure 2-2. Each family has multiple affected individuals, affected by autism and other neurodevelopmental and neuropsychiatric conditions. AS326 and AS328 are multigenerational families, with maternal grandmother samples available.

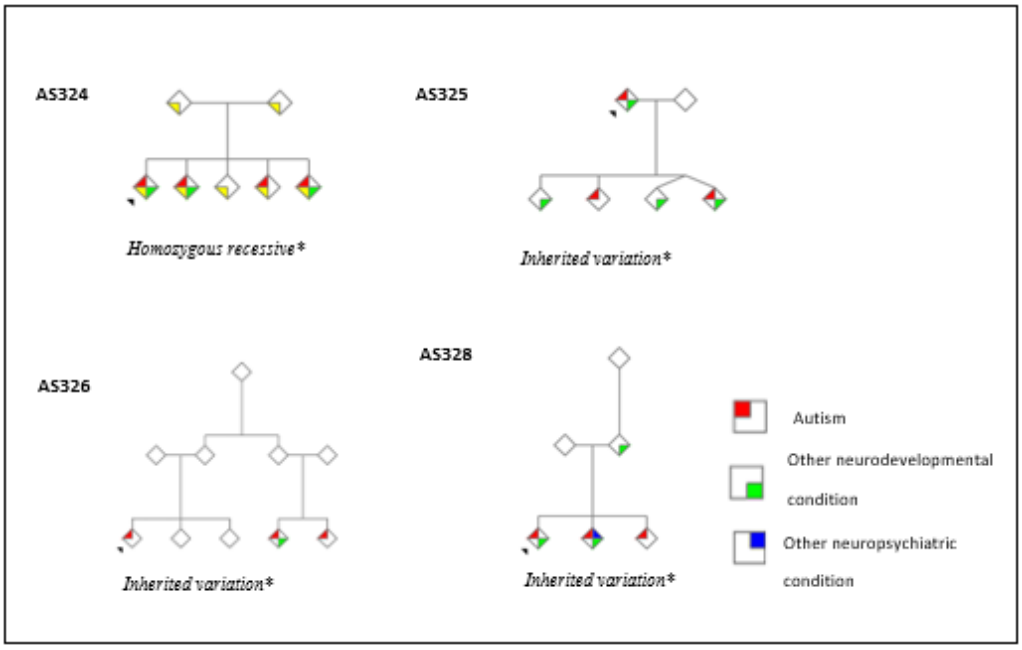


Figure 2-2 Cohort 3 in summary.

*Proposed mode of transmission for variant interpretation.} Presented are the pedigrees of the 4 families sequenced in this rare cohort. The key associated with affection and sequencing status is presented alongside. * Denotes the mode of variant transmission hypothesised to be relevant within each family.*

2.2 Sequencing

Next-generation sequencing technologies were applied to the cohorts under investigation. These technologies give the opportunity for more widespread and in-depth genomic analysis than older techniques, such as microarray studies and candidate gene studies, have allowed. Table 1-1 lists the next-generation sequencing (NGS) technologies that can identify single nucleotide variants (SNVs) and insertion-deletion variants (indels), as well as larger genomic hits, including structural variants (SVs) or copy number variants (CNVs), across the allele frequency spectrum. Here, whole exome and whole genome sequencing have been applied as described as follows, with an aim to detect exonic SNVs.

2.2.1 Sequencing of Cohort 1

WES was carried out on a total of 42 samples. DNA samples (n=17) were extracted from whole blood using Perkin Elmer Prepito DNA cyto kit (CMG-2034). Biobanked DNA samples (n=25) were extracted as previously published (Buxbaum *et al.*, 2014). Biobanked DNA samples were confirmed to have a concentration of >30ng/ul as measured by Qubit and an optical density 260/280>1.8 as measured by Nanodrop prior to library preparation. Samples were whole exome sequenced using the Nextera Rapid Capture Exome (v1.2) on Illumina NovaSeq6000. FASTQ, BAM and VCF files were returned for bioinformatic analyses.

2.2.2 Sequencing of Cohort 2

WGS was carried out on a total of 6 samples. DNA samples (n=6) were extracted from whole blood using Perkin Elmer Prepito DNA cyto kit (CMG-2034). Samples were whole genome sequenced on Illumina NovaSeq6000. FASTQ, BAM and VCF files were returned for bioinformatic analyses.

2.2.3 Sequencing of Cohort 3

Sequencing of this cohort was performed on DNA samples extracted from blood (n=22) and saliva (n=7). As reported by service provider:

“Whole genome library preparation was performed using the Illumina TruSeq PCR Free Library Prep protocol (20015963) with an input amount of 1µg. Library preparation was automated and processed using a Hamilton NGS Star. Library quality was assessed using the Roche KAPA Library Quantification Kit (7960298001). Libraries were pooled and sequenced on an Illumina NovaSeq 6000 instrument using NovaSeq 6000 S4 Reagent Kit (20012866) targeting a mean coverage of 30X.

Genotyping was performed using the Illumina Global Screening Array version 3 (20030772)”. FASTQ, BAM and VCF files were returned for bioinformatic analyses.

2.3 Read alignment

Short sequence reads are generated by NGS and are reported in FASTQ format by the genomic sequencing described in 2.2. These short reads require alignment to a human reference genome and the resulting aligned reads are output in BAM format.

2.3.1 Read alignment of Cohort 1

FASTQ files were generated through Illumina BaseSpace using the function FASTQ Generation (Version: 1.0.0). Sequence data was aligned to the reference genome “Homo sapiens (UCSC hg19)” using bwa-mem through Illumina BaseSpace BWA Enrichment (Version: 2.1.0.0) targeting the regions covered by Nextera Rapid Capture Exome (v1.2) (Li, 2013).

2.3.2 Read alignment of Cohort 2

FASTQ files were generated through Illumina BaseSpace using the function FASTQ Generation (Version: 1.0.0). Sequence data was aligned to the reference genome “Homo sapiens (UCSC hg19)” using bwa-mem through Illumina BaseSpace function BWA Aligner - DEPRECATED (Version: 1.1.1) (Li, 2013).

2.3.3 Read alignment of Cohort 3

As reported by service provider:

“Genuity Science Pipeline Service (GSPS) is a feature which allows importation of raw data from the NovaSeq 6000 (BCL files) or raw sequencing data (in FASTQ format) for analysis with AWS S3 delivery capabilities. The standard workflow for GSPS is to use in-house generated raw data from the NovaSeq 6000 (BCL files), which is demultiplexed in the pipeline to individual samples FASTQ pairs, which then get analysed for variants and have their sequence quality assessed. FastQ generation was performed using BCL2FastQ, adapter trimming using Skewer and assessment of QC using FASTQC.”

2.4 Base read pre-processing

2.4.1 Base read pre-processing of Cohort 1

Software and corresponding versions used during quality control (QC are listed in (Table 2-5). Input files for analysis were sourced from the Genome Analysis Tool Kit (GATK) Resource Bundle for reference genome hg19, available at: ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/ (Table 2-6).

Software	Version
Picard (Broad Institute, 2019)	2.20.2-SNAPSHOT
Genome Analysis ToolKit (GATK) (Van der Auwera *et al.*, 2013)	3.8-0-ge9d806836
Java	openjdk version "1.8.0_212"
R	3.6.0 “Planting of a Tree”

Table 2-5 Software used at QC.

Input	Version
Reference Genome	UCSC hg19
Verified Indel sites	1000G Phase 1 & Mills Gold Standard Indels
Verified SNP sites	dbsnp 138 (hg19)
Exome Target Intervals	NexteraRapidCapture_Exome_TargetedRegions_v1.2

Table 2-6 Input public datasets used at QC.

2.4.1.1 ReorderSam (Picard)

BAM file reads were reordered to match contig ordering in the input reference genome.

2.4.1.2 SortSam (Picard)

BAM file reads were sorted by reference sequence coordinate.

2.4.1.3 MarkDuplicates (Picard)

Duplicate reads arising from technical errors and biases were located and tagged.

2.4.1.4 BuildBamIndex (Picard)

An index file complementary to BAM file was generated to allow fast look-up of data.

2.4.1.5 IndelRealigner (GATK)

Local realignment of insertion and deletions was carried out to minimise the number of mismatching base pairs.

```
### Local Realignment Around Indels ###

java <java_arguments> -jar <GATK_jar_file> \
-T RealignerTargetCreator \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-o <output_file_to_report_intervals> \
--known <1000G_phase1.indels.hg19.sites.vcf \
--known <Mills_and_1000G_gold_standard.indels.hg19.sites.vcf>

java <java_arguments> -jar <GATK_jar_file> \
-T IndelRealigner \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-o <BAM_following_manipulation> \
-known <1000G_phase1.indels.hg19.sites.vcf \
-known <Mills_and_1000G_gold_standard.indels.hg19.sites.vcf> \
-targetIntervals <output_file_to_report_intervals> \
--consensusDeterminationModel USE_READS
```

2.4.1.6 Base Quality Score Recalibration

Systematic technical errors in quality score estimates of each base call were detected and scores were adjusted using *BaseRecalibrator* (GATK). The recalibration was visualised with *AnalyseCovariates* (GATK).

```
### Base Quality Score Recalibration ###

java <java_arguments> -jar <GATK_jar_file> \
-T BaseRecalibrator \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-o <output_BQSR_before> \
-knownSites <dbsnp_138.hg19.vcf> \
-knownSites <1000G_phase1.indels.hg19.sites.vcf> \
-knownSites <Mills_and_1000G_gold_standard.indels.hg19.sites.vcf> \
--sort_by_all_columns

java <java_arguments> -jar <GATK_jar_file> \
-T BaseRecalibrator \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-BQSR <output_BQSR_before> \
-o <output_BQSR_before_after> \
-knownSites <dbsnp_138.hg19.vcf> \
-knownSites <1000G_phase1.indels.hg19.sites.vcf> \
-knownSites <Mills_and_1000G_gold_standard.indels.hg19.sites.vcf> \
--sort_by_all_columns

java <java_arguments> -jar <GATK_jar_file> \
-T PrintReads \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-BQSR <output_BQSR_before> \
-o <new_BAM_post-bqsr>

java <java_arguments> -jar <GATK_jar_file> \
-T AnalyzeCovariates \
-R <reference_genome_fasta> \
-before <output_BQSR_before> \
-after <output_BQSR_before_after> \
-plots <plots_file_generated> \
-csv <csv_file_generated>
```

2.4.1.7 ValidateSamFile (Picard)

To ensure compliance of BAM file format specifications, *ValidateSamFile* was applied.

```
### Validate Bam File ###

#In Summary Mode

java <java_arguments> -jar <picard_jar_file> ValidateSamFile \
I= <BAM_to_investigate> \
O= <document_to_output_errors_to> \
MODE=SUMMARY \
MAX_OUTPUT=null

#In Verbose Mode

java <java_arguments> -jar <picard_jar_file> ValidateSamFile \
I= <BAM_to_investigate> \
O= <document_to_output_errors_to> \
IGNORE_WARNINGS=true \
MODE=VERBOSE
```

2.4.2 Base read pre-processing of Cohort 2

Software and corresponding versions used during QC are listed in (Table 2-7). Input files for analysis were sourced from the GATK Resource Bundle for reference genome hg19, available at: ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/ (Table 2-8).

Software	**Version**
Picard	2.20.2-SNAPSHOT
Genome Analysis ToolKit (GATK)	3.8-0-ge9d806836
Java	openjdk version "1.8.0_212"
R	3.6.0 “Planting of a Tree”

Table 2-7 Software used at QC.

Input	**Version**
Reference Genome	UCSC hg19
Verified Indel sites	1000G Phase 1 & Mills Gold Standard Indels
Verified SNP sites	dbsnp 138 (hg19)

Table 2-8 Input public datasets used at QC.

2.4.2.1 AddOrReplaceReadGroups (Picard)

Read group information missing from the BAM file generated was added to BAM files. Generic read group information was added in the cases of mandatory fields, except for RGSM which corresponds to sample ID.

```
### Edit Read Group Information ###

java <java_arguments> -jar <picard_jar_file> AddOrReplaceReadGroups \
INPUT= <BAM_to_reformat> \
OUTPUT= <new_BAM_reformatted> \
RGID= <default_value_eg.1> \
RGLB= <default_value_eg.library1> \
RGPL= <default_value_eg.illumina> \
RGPU= <default_value_eg.1> \
RGSM= <unique_sample_ID> \
SORT_ORDER=coordinate \
CREATE_INDEX=true
```

Parameters SORT_ORDER=coordinate and CREATE_INDEX=true were applied to simultaneously sort and index the reformatted BAM file.

2.4.2.2 IndelRealigner (GATK)

Local realignment of insertion and deletions was carried out to minimise the number of mismatching base pairs.

```
### Local Realignment Around Indels ###

java <java_arguments> -jar <GATK_jar_file> \
-T RealignerTargetCreator \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-o <output_file_to_report_intervals> \
--known <1000G_phase1.indels.hg19.sites.vcf \
--known <Mills_and_1000G_gold_standard.indels.hg19.sites.vcf>

java <java_arguments> -jar <GATK_jar_file> \
-T IndelRealigner \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-o <BAM_following_manipulation> \
-known <1000G_phase1.indels.hg19.sites.vcf \
-known <Mills_and_1000G_gold_standard.indels.hg19.sites.vcf> \
-targetIntervals <output_file_to_report_intervals> \
--consensusDeterminationModel USE_READS
```

2.4.2.3 Base Quality Score Recalibration

Systematic technical errors in quality score estimates of each base call were detected and scores were adjusted using *BaseRecalibrator* (GATK). The recalibration was visualised with *AnalyseCovariates* (GATK).

```
### Base Quality Score Recalibration ###

java <java_arguments> -jar <GATK_jar_file> \
-T BaseRecalibrator \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-o <output_BQSR_before> \
-knownSites <dbsnp_138.hg19.vcf> \
-knownSites <1000G_phase1.indels.hg19.sites.vcf> \
-knownSites <Mills_and_1000G_gold_standard.indels.hg19.sites.vcf> \
--sort_by_all_columns

java <java_arguments> -jar <GATK_jar_file> \
-T BaseRecalibrator \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-BQSR <output_BQSR_before> \
-o <output_BQSR_before_after> \
-knownSites <dbsnp_138.hg19.vcf> \
-knownSites <1000G_phase1.indels.hg19.sites.vcf> \
-knownSites <Mills_and_1000G_gold_standard.indels.hg19.sites.vcf> \
--sort_by_all_columns

java <java_arguments> -jar <GATK_jar_file> \
-T PrintReads \
-R <reference_genome_fasta> \
-I <BAM_to_manipulate> \
-BQSR <output_BQSR_before> \
-o <new_BAM_post-bqsr>

java <java_arguments> -jar <GATK_jar_file> \
-T AnalyzeCovariates \
-R <reference_genome_fasta> \
-before <output_BQSR_before> \
-after <output_BQSR_before_after> \
-plots <plots_file_generated> \
-csv <csv_file_generated>
```

2.4.2.4 ValidateSamFile (Picard)

To ensure compliance of BAM file format specifications, *ValidateSamFile* was applied in both summary and verbose modes.

```
### Validate Bam File ###

#In Summary Mode

java <java_arguments> -jar <picard_jar_file> ValidateSamFile \
I= <BAM_to_investigate> \
O= <document_to_output_errors_to> \
MODE=SUMMARY \
MAX_OUTPUT=null

#In Verbose Mode

java <java_arguments> -jar <picard_jar_file> ValidateSamFile \
I= <BAM_to_investigate> \
O= <document_to_output_errors_to> \
IGNORE_WARNINGS=true \
MODE=VERBOSE
```

2.4.3 Base read pre-processing of Cohort 3

As reported by service provider:

"FASTQ data is processed by a Sentieon based pipeline and resulting files are uploaded to the user specified S3 buckets. The main components of the pipeline act to align raw sequence reads to a reference, carry out QC and call variants in the genome. The pipeline is based on an AWS infrastructure that is automated."

2.5 Variant calling and genotyping

Variants were called from aligned reads (BAM) format and were output to a readily interpretable called variant file (VCF files). VCF files can then be manipulated and interrogated for biological relevance. The GATK Best Practices provide guidelines for effective use of the tool set, enabling manipulation of parameters to suit the data set under investigation, for example specification of target intervals or specification of reference genome (Li, 2013; Broad Institute, 2019). Here GATK Best Practices were applied to call SNVs reaching the standard minimum threshold for calling using HaplotypeCaller. Genotypes were assigned by joint genotyping. Joint genotyping requires the use of HaplotypeCaller to generate input genotype assignments. The gVCF file format details all variant sites in the genome whether reference (ref) or alternative (alt), as opposed to the traditional VCF file listing alternative variant sites only. Joint genotyping is a more time and computationally intensive approach to genotyping, however, it improves the detection of rare variants in the genome making it beneficial for use in a family-based study design where accurate and sensitive rare variant discovery is required.

2.5.1 Variant calling and genotyping of Cohort 1

Software and corresponding versions used throughout gVCF to cohort VCF analysis are listed in Table 2-9. Input files for analysis were sourced from the GATK Resource Bundle for reference genome hg19, available at: ftp://gsapubftp- anonymous@ftp.broadinstitute.org/bundle/hg19/ (Table 2-10).

Software	**Version**
Genome Analysis ToolKit (GATK)	3.8-0-ge9d806836
Java	openjdk version "1.8.0_212"

Table 2-9 Software used at variant calling and genotyping.

Input	**Version**
Reference Genome	UCSC hg19
Verified SNP sites	dbsnp 138 (hg19)
Exome Target Intervals	NexteraRapidCapture_Exome_TargetedRegions_v1.2

Table 2-10 Input datasets used at variant calling and genotyping.

2.5.1.1 HaplotypeCaller in gVCF Mode (GATK)

SNVs and indels were called and output to gVCF format, recording all site information whether reference or alternative.

```
### Variant Calling ###

java <java_arguments> -jar <GATK_jar_file> \
-T HaplotypeCaller \
-R <reference_genome_fasta> \
-I <post_processing_bam> \
-o <newly_called_gvcf> \
-ERC GVCF \
--dbsnp <dbsnp_138.hg19.vcf> \
--annotation MappingQualityZero \
--annotation VariantType \
--annotation AlleleBalance \
--annotation AlleleBalanceBySample \
--excludeAnnotation ChromosomeCounts \
--excludeAnnotation FisherStrand \
--excludeAnnotation StrandOddsRatio \
--excludeAnnotation QualByDepth \
--GVCFGQBands 10 \
--GVCFGQBands 20 \
--GVCFGQBands 30 \
--GVCFGQBands 40 \
--GVCFGQBands 60 \
--GVCFGQBands 80 \
--standard_min_confidence_threshold_for_calling
```

2.5.1.2 Cohort Joint Genotyping

Joint genotyping was carried out using *GenotypeGVCFs* (GATK) to enable detection of variants to a higher degree of sensitivity and genotype accuracy by leveraging information cohort-wide.

```
### Joint Genotyping ###

java <java_arguments> -jar <GATK_jar_file> \
-T GenotypeGVCFs \
-R <reference_genome_fasta> \
-V <newly_called_gvcf1> \
-V <newly_called_gvcf2> \
-V <newly_called_gvcf3> \
-V <newly_called_gvcf4> \
-V <newly_called_gvcf5> \
-V <newly_called_gvcf6> \
--annotation InbreedingCoeff \
--annotation FisherStrand \
--annotation QualByDepth \
--annotation ChromosomeCounts \
--annotation StrandOddsRatio \
--dbsnp <dbsnp_138.hg19.vcf> \
-o <newly_genotyped_cohort_vcf> \
--standard_min_confidence_threshold_for_calling 10.0 \
--downsample_to_coverage 1000 \
--downsampling_type BY_SAMPLE
```

2.5.2 Variant calling and genotyping of Cohort 2

Software and corresponding versions used throughout gVCF to cohort VCF analysis are listed in Table 2-11. Input files for analysis were sourced from the GATK Resource Bundle for reference genome hg19, available at: ftp://gsapubftp- anonymous@ftp.broadinstitute.org/bundle/hg19/ (Table 2-12).

Software	**Version**
Genome Analysis ToolKit (GATK)	3.8-0-ge9d806836
Java	openjdk version "1.8.0_212"

Table 2-11 Software used at variant calling and genotyping.

Input	**Version**
Reference Genome	UCSC hg19
Verified SNP sites	dbsnp 138 (hg19)

Table 2-12 Input datasets used at variant calling and genotyping.

2.5.2.1 HaplotypeCaller in gVCF Mode (GATK)

SNVs and indels were called and output to gVCF format, recording all site information whether reference or alternative.

```
### Variant Calling ###

java <java_arguments> -jar <GATK_jar_file> \
-T HaplotypeCaller \
-R <reference_genome_fasta> \
-I <post_processing_bam> \
-o <newly_called_gvcf> \
-ERC GVCF \
--dbsnp <dbsnp_138.hg19.vcf> \
--annotation MappingQualityZero \
--annotation VariantType \
--annotation AlleleBalance \
--annotation AlleleBalanceBySample \
--excludeAnnotation ChromosomeCounts \
--excludeAnnotation FisherStrand \
--excludeAnnotation StrandOddsRatio \
--excludeAnnotation QualByDepth \
--GVCFGQBands 10 \
--GVCFGQBands 20 \
--GVCFGQBands 30 \
--GVCFGQBands 40 \
--GVCFGQBands 60 \
--GVCFGQBands 80 \
--standard_min_confidence_threshold_for_calling
```

2.5.2.2 Cohort Joint Genotyping

Joint genotyping was carried out using *GenotypeGVCFs* (GATK) to enable detection of variants to a higher degree of sensitivity and genotype accuracy by leveraging information cohort-wide.

```
### Joint Genotyping ###

java <java_arguments> -jar <GATK_jar_file> \
-T GenotypeGVCFs \
-R <reference_genome_fasta> \
-V <newly_called_gvcf1> \
-V <newly_called_gvcf2> \
-V <newly_called_gvcf3> \
-V <newly_called_gvcf4> \
-V <newly_called_gvcf5> \
-V <newly_called_gvcf6> \
--annotation InbreedingCoeff \
--annotation FisherStrand \
--annotation QualByDepth \
--annotation ChromosomeCounts \
--annotation StrandOddsRatio \
--dbsnp <dbsnp_138.hg19.vcf> \
-o <newly_genotyped_cohort_vcf> \
--standard_min_confidence_threshold_for_calling 10.0 \
--downsample_to_coverage 1000 \
--downsampling_type BY_SAMPLE
```

2.5.3 Variant calling and genotyping of Cohort 3

As reported by service provider:

"The secondary pipeline begins with a Sention backbone which included bwa mem, markduplication, WgsMetricsAlgo, Realigner and QualCal. The resulting VCF files were GORized with Genuity Science proprietary tools and loaded into CSA platform for downstream analysis."

2.6 Variant filtration

2.6.1 Variant filtration of Cohort 1

Software and corresponding versions used throughout variant filtration are listed in Table 2-13. Input files for analysis were sourced from the GATK Resource Bundle for reference genome hg19, available at ftp://gsapubftp- anonymous@ftp.broadinstitute.org/bundle/hg19/ (Table 2-14).

Software	**Version**
Genome Analysis ToolKit (GATK)	3.8-0-ge9d806836
VCFtools (Danecek *et al.*, 2011)	0.1.14-gcc-8.2.0-srywzy

Table 2-13 Software used at variant filtration.

Input	Version
Reference Genome	UCSC hg19
VQSR Model Training Set	dbsnp 138 (hg19) hapmap 3.3 (hg19) omni 2.5 (hg19) 1000G Phase 1 High-Confidence SNPs (hg19) Mills Gold Standard Indels

Table 2-14 Input datasets used at variant filtration.

2.6.1.1 Split Cohort VCF by variant type

The cohort VCF file was split by variant type using *SelectVariants* (GATK) and output into distinct cohort SNV, indel and mixed VCF files.

```
### Split Variants by Variant Type ###

java <java_arguments> -jar <GATK_jar_file> \
-T SelectVariants \
-R <reference_genome_fasta> \
--variant <newly_genotyped_cohort_vcf> \
-L chr1 -L chr2 -L chr3 -L chr4 -L chr5 \
-L chr6 -L chr7 -L chr8 -L chr9 -L chr10 \
-L chr11 -L chr12 -L chr13 -L chr14 -L chr15 \
-L chr16 -L chr17 -L chr18 -L chr19 -L chr20 \
-L chr21 -L chr22 -L chrX -L chrY \
-o <cohort_snp_vcf> \
-selectType SNP

java <java_arguments> -jar <GATK_jar_file> \
-T SelectVariants \
-R <reference_genome_fasta> \
--variant <newly_genotyped_cohort_vcf> \
-L chr1 -L chr2 -L chr3 -L chr4 -L chr5 \
-L chr6 -L chr7 -L chr8 -L chr9 -L chr10 \
-L chr11 -L chr12 -L chr13 -L chr14 -L chr15 \
-L chr16 -L chr17 -L chr18 -L chr19 -L chr20 \
-L chr21 -L chr22 -L chrX -L chrY \
-o <cohort_indel_vcf> \
-selectType INDEL

java <java_arguments> -jar <GATK_jar_file> \
-T SelectVariants \
-R <reference_genome_fasta> \
--variant <newly_genotyped_cohort_vcf> \
-L chr1 -L chr2 -L chr3 -L chr4 -L chr5 \
-L chr6 -L chr7 -L chr8 -L chr9 -L chr10 \
-L chr11 -L chr12 -L chr13 -L chr14 -L chr15 \
-L chr16 -L chr17 -L chr18 -L chr19 -L chr20 \
-L chr21 -L chr22 -L chrX -L chrY \
-o <cohort_other_vcf> \
-xlSelectType SNP \
-xlSelectType INDEL
```

2.6.1.2 Evaluating Variant Confidence Pre-Filtering

The Transition/Transversion (Ts/Tv) ratio presented in Table 2-15 as 2.594. is low when compared with the guide Ts/Tv of 3.0 in human exome sequencing (Bainbridge *et al.*, 2011). This highlights the need for low confidence variant call removal from the call-set to achieve a high confidence variant set for downstream analysis.

Model	**Count**
AC	8,231
AG	42,065
AT	4,546
CG	11,607
CT	41,748
GT	7,921
Ts	83,813
Tv	32,305
Ts/Tv Ratio	**2.594**

Table 2-15 Ts/Tv ratio of SNV call-set.

Presented in the table are genotype variant transitions (Ts) and Transversions (Tv) across variant sites within the cohort. Transitions are defined as a change of purine bases or pyrimidine bases, i.e. A with G or C with T. Transversion are defined as changes between purine and pyrimidine bases, i.e. A with C/T, C with G or G with T.

2.6.1.3 Variant Quality Score Recalibration (VQSR) and Hard Filtering

VariantRecalibration (GATK) and *ApplyRecalibration* (GATK) were run independently on SNV and indel raw variants. Mixed variants not falling into these variant types were hard filtered using GATK Best Practices for indel filtration, as too few variants (fewer equivalent variant sites than one genome or 30 exomes) fell into this category to train the Gaussian mixture model of VQSR (Table 2-16).

Filter	**Description**	**Threshold**
QD	Variant quality / depth	< 2.0
FS	Phred-score Fisher's test p-value for strand bias	> 200.0
ReadPosRankSum	Distance of alternative allele from the end of the reads	< -20.0

Table 2-16 GATK recommended variant quality filters for mixed variant sites.

```
### VQSR ###

java <java_arguments> -jar <GATK_jar_file> \
-T VariantRecalibrator \
-R <reference_genome_fasta> \
-input <cohort_snp_vcf> \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
<hapmap_3.3.hg19.sites.vcf> \
-resource:omni,known=false,training=true,truth=true,prior=12.0
<1000G_omni2.5.hg19.sites.vcf> \
-resource:1000G,known=false,training=true,truth=false,prior=10.0
<1000G_phase1.snps.high_confidence.hg19.sites.vcf> \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
<dbsnp_138.hg19.vcf> \
-an DP -an QD -an FS -an SOR -an MQ -an MQRankSum -an ReadPosRankSum \
-mode SNP \
-tranche 100.0 -tranche 99.99 -tranche 99.98 -tranche 99.97 -tranche 99.96 \
-tranche 99.95 -tranche 99.94 -tranche 99.93 -tranche 99.92 -tranche 99.91 \
-tranche 99.90 -tranche 99.80 -tranche 99.70 -tranche 99.60 -tranche 99.50 \
-tranche 99.00 -tranche 98.00 -tranche 97.00 -tranche 96.00 -tranche 95.00 \
-tranche 94.00 -tranche 93.00 -tranche 92.00 -tranche 91.00 -tranche 90.00 \
-recalFile <snp.recalfile> \
-tranchesFile <snp.tranchesfile> \
-rscriptFile <snp.plotsfile> \
--maxGaussians 4

java <java_arguments> -jar <GATK_jar_file> \
-T ApplyRecalibration \
-R <reference_genome_fasta> \
-input <cohort_snp_vcf> \
-mode SNP \
-recalFile <snp.recalfile> \
-tranchesFile <snp.tranchesfile> \
--ts_filter_level 99.5 \
-o <postVQSR_snp.vcf>

java <java_arguments> -jar <GATK_jar_file> \
-T VariantRecalibrator \
-R <reference_genome_fasta> \
-input <cohort_indel_vcf> \
-resource:mills,known=false,training=true,truth=true,prior=12.0
<Mills_and_1000G_gold_standard.indels.hg19.sites.vcf> \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
<dbsnp_138.hg19.vcf> \
-an DP -an QD -an FS -an SOR -an MQRankSum -an ReadPosRankSum \
-mode INDEL \
-tranche 100.0 -tranche 99.99 -tranche 99.98 -tranche 99.97 -tranche 99.96 \
-tranche 99.95 -tranche 99.94 -tranche 99.93 -tranche 99.92 -tranche 99.91 \
-tranche 99.90 -tranche 99.80 -tranche 99.70 -tranche 99.60 -tranche 99.50 \
-tranche 99.00 -tranche 98.00 -tranche 97.00 -tranche 96.00 -tranche 95.00 \
-tranche 94.00 -tranche 93.00 -tranche 92.00 -tranche 91.00 -tranche 90.00 \
-recalFile <indel.recalfile> \
-tranchesFile <indel.tranchesfile> \
-rscriptFile <indel.plotsfile> \
--maxGaussians 4

java <java_arguments> -jar <GATK_jar_file> \
-T ApplyRecalibration \
-R <reference_genome_fasta> \
-input <cohort_indel_vcf> \
-mode INDEL \
-recalFile <indel.recalfile> \
-tranchesFile <indel.tranchesfile> \
--ts_filter_level 99.0 \
-o <postVQSR_indel.vcf>
```

```
### Variant Hard-Filtration ###

java <java_arguments> -jar <GATK_jar_file> \
-T VariantFiltration \
-R <reference_genome_fasta> \
-V <cohort_other_vcf> \
--filterExpression "QD < 2.0" \
--filterName "OtherHardQD" \
--filterExpression "FS > 200.0" \
--filterName "OtherHardFS" \
--filterExpression "ReadPosRankSum < -20.0" \
--filterName "OtherHardReadPosRankSum" \
-o <cohort_other_filtered_vcf>
```

2.6.1.4 CombineVariants (GATK)

SNV, indel and mixed variant VCFs were merged following VQSR and hard filtering, respectively.

```
### Merge VCF Files ###

java <java_arguments> -jar <GATK_jar_file> \
-T CombineVariants \
-R <reference_genome_fasta> \
-V:snp <postVQSR_snp.vcf> \
-V:indel <postVQSR_indel.vcf> \
-V:other <cohort_other_filtered_vcf> \
-o <cohort_flagged_vcf> \
-assumeIdenticalSamples \
-genotypeMergeOptions PRIORITIZE \
-priority snp,indel,other
```

2.6.1.5 ValidateVariants (GATK)

Information and format within the merged VCF file were validated.

```
### Validate VCF File ###

java <java_arguments> -jar <GATK_jar_file> \
-T ValidateVariants \
-R <reference_genome_fasta> \
--dbsnp <dbsnp_138.hg19.vcf> \
--reference_window_stop 300 \
-V <cohort_flagged_vcf>
```

2.6.1.6 Hardy-Weinburg Equilibrium and Missingness Hard Filtering

Variant sites with a p-value falling under the threshold of 10^{-6} when testing for Hardy Weinburg Equilibrium exact test (Wigginton, Cutler and Abecasis, 2005), were flagged

for removal using *VCFtools - - hwe.* Variant sites missing greater than 10% of data were flagged for exclusion using *VCFtools - - maxmissing.*

Flagged variants were removed from further analysis using *VCFtools* parameters *--remove-filtered-all --recode --recode-INFO-all.*

```
### Removal of VQSR Flagged Variants ###

vcftools \
--vcf <cohort_flagged_vcf> \
--out <cohort_filtered_vcf> \
--remove-filtered-all \
--recode \
--recode-INFO-all
```

2.6.2 Variant filtration of Cohort 2

Software and corresponding versions used throughout variant filtration are listed in Table 2-17. Input files for analysis were sourced from the GATK Resource Bundle for reference genome hg19, available at: ftp://gsapubftp- anonymous@ftp.broadinstitute.org/bundle/hg19/ (Table 2-18).

Software	**Version**
Genome Analysis ToolKit (GATK)	3.8-0-ge9d806836
VCFtools	0.1.14-gcc-8.2.0-srywzy
Picard	2.20.2-SNAPSHOT

Table 2-17 Software used at variant filtration.

Input	**Version**
Reference Genome	UCSC hg19
VQSR Model Training Set	dbsnp 138 (hg19) hapmap 3.3 (hg19) omni 2.5 (hg19) 1000G Phase 1 High-Confidence SNPs (hg19) Mills Gold Standard Indels

Table 2-18 Input datasets used at variant filtration.

2.6.2.1 Split Cohort VCF by variant type

The cohort VCF file was split by variant type using *SelectVariants* (GATK) and output into distinct cohort SNV, indel and mixed VCF files.

```
### Split Variants by Variant Type ###

java <java_arguments> -jar <GATK_jar_file> \
-T SelectVariants \
-R <reference_genome_fasta> \
--variant <newly_genotyped_cohort_vcf> \
-L chr1 -L chr2 -L chr3 -L chr4 -L chr5 \
-L chr6 -L chr7 -L chr8 -L chr9 -L chr10 \
-L chr11 -L chr12 -L chr13 -L chr14 -L chr15 \
-L chr16 -L chr17 -L chr18 -L chr19 -L chr20 \
-L chr21 -L chr22 -L chrX -L chrY \
-o <cohort_snp_vcf> \
-selectType SNP

java <java_arguments> -jar <GATK_jar_file> \
-T SelectVariants \
-R <reference_genome_fasta> \
--variant <newly_genotyped_cohort_vcf> \
-L chr1 -L chr2 -L chr3 -L chr4 -L chr5 \
-L chr6 -L chr7 -L chr8 -L chr9 -L chr10 \
-L chr11 -L chr12 -L chr13 -L chr14 -L chr15 \
-L chr16 -L chr17 -L chr18 -L chr19 -L chr20 \
-L chr21 -L chr22 -L chrX -L chrY \
-o <cohort_indel_vcf> \
-selectType INDEL

java <java_arguments> -jar <GATK_jar_file> \
-T SelectVariants \
-R <reference_genome_fasta> \
--variant <newly_genotyped_cohort_vcf> \
-L chr1 -L chr2 -L chr3 -L chr4 -L chr5 \
-L chr6 -L chr7 -L chr8 -L chr9 -L chr10 \
-L chr11 -L chr12 -L chr13 -L chr14 -L chr15 \
-L chr16 -L chr17 -L chr18 -L chr19 -L chr20 \
-L chr21 -L chr22 -L chrX -L chrY \
-o <cohort_other_vcf> \
-xlSelectType SNP \
-xlSelectType INDEL
```

2.6.2.2 VQSR and Hard Filtering

VariantRecalibration (GATK) and *ApplyRecalibration* (GATK) were run independently on SNV and indel raw variants.

Mixed variants not falling into these variant types were hard filtered using GATK Best Practices for indel filtration, as too few variants fell into this category to train the Gaussian mixture model of VQSR (Table 2-16).

```
### VQSR ###

java <java_arguments> -jar <GATK_jar_file> \
-T VariantRecalibrator \
-R <reference_genome_fasta> \
-input <cohort_snp_vcf> \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
<hapmap_3.3.hg19.sites.vcf> \
-resource:omni,known=false,training=true,truth=true,prior=12.0
<1000G_omni2.5.hg19.sites.vcf> \
-resource:1000G,known=false,training=true,truth=false,prior=10.0
<1000G_phase1.snps.high_confidence.hg19.sites.vcf> \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
<dbsnp_138.hg19.vcf> \
-an DP -an QD -an FS -an SOR -an MQ -an MQRankSum -an ReadPosRankSum \
-mode SNP \
-tranche 100.0 -tranche 99.99 -tranche 99.98 -tranche 99.97 -tranche 99.96 \
-tranche 99.95 -tranche 99.94 -tranche 99.93 -tranche 99.92 -tranche 99.91 \
-tranche 99.90 -tranche 99.80 -tranche 99.70 -tranche 99.60 -tranche 99.50 \
-tranche 99.00 -tranche 98.00 -tranche 97.00 -tranche 96.00 -tranche 95.00 \
-tranche 94.00 -tranche 93.00 -tranche 92.00 -tranche 91.00 -tranche 90.00 \
-recalFile <snp.recalfile> \
-tranchesFile <snp.tranchesfile> \
-rscriptFile <snp.plotsfile> \
--maxGaussians 4

java <java_arguments> -jar <GATK_jar_file> \
-T ApplyRecalibration \
-R <reference_genome_fasta> \
-input <cohort_snp_vcf> \
-mode SNP \
-recalFile <snp.recalfile> \
-tranchesFile <snp.tranchesfile> \
--ts_filter_level 99.5 \
-o <postVQSR_snp.vcf>

java <java_arguments> -jar <GATK_jar_file> \
-T VariantRecalibrator \
-R <reference_genome_fasta> \
-input <cohort_indel_vcf> \
-resource:mills,known=false,training=true,truth=true,prior=12.0
<Mills_and_1000G_gold_standard.indels.hg19.sites.vcf> \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
<dbsnp_138.hg19.vcf> \
-an DP -an QD -an FS -an SOR -an MQRankSum -an ReadPosRankSum \
-mode INDEL \
-tranche 100.0 -tranche 99.99 -tranche 99.98 -tranche 99.97 -tranche 99.96 \
-tranche 99.95 -tranche 99.94 -tranche 99.93 -tranche 99.92 -tranche 99.91 \
-tranche 99.90 -tranche 99.80 -tranche 99.70 -tranche 99.60 -tranche 99.50 \
-tranche 99.00 -tranche 98.00 -tranche 97.00 -tranche 96.00 -tranche 95.00 \
-tranche 94.00 -tranche 93.00 -tranche 92.00 -tranche 91.00 -tranche 90.00 \
-recalFile <indel.recalfile> \
-tranchesFile <indel.tranchesfile> \
-rscriptFile <indel.plotsfile> \
--maxGaussians 4

java <java_arguments> -jar <GATK_jar_file> \
-T ApplyRecalibration \
-R <reference_genome_fasta> \
-input <cohort_indel_vcf> \
-mode INDEL \
-recalFile <indel.recalfile> \
-tranchesFile <indel.tranchesfile> \
--ts_filter_level 99.0 \
-o <postVQSR_indel.vcf>
```

```
### Variant Hard-Filtration ###

java <java_arguments> -jar <GATK_jar_file> \
-T VariantFiltration \
-R <reference_genome_fasta> \
-V <cohort_other_vcf> \
--filterExpression "QD < 2.0" \
--filterName "OtherHardQD" \
--filterExpression "FS > 200.0" \
--filterName "OtherHardFS" \
--filterExpression "ReadPosRankSum < -20.0" \
--filterName "OtherHardReadPosRankSum" \
-o <cohort_other_filtered_vcf>
```

2.6.2.3 CombineVariants (GATK)

SNV, indel and mixed variant VCFs were merged following VQSR and hard filtering, respectively.

```
### Merge VCF Files ###

java <java_arguments> -jar <GATK_jar_file> \
-T CombineVariants \
-R <reference_genome_fasta> \
-V:snp <postVQSR_snp.vcf> \
-V:indel <postVQSR_indel.vcf> \
-V:other <cohort_other_filtered_vcf> \
-o <cohort_flagged_vcf> \
-assumeIdenticalSamples \
-genotypeMergeOptions PRIORITIZE \
-priority snp,indel,other
```

2.6.2.4 ValidateVariants (GATK)

Information and format within the merged VCF file were validated.

```
### Validate VCF File ###

java <java_arguments> -jar <GATK_jar_file> \
-T ValidateVariants \
-R <reference_genome_fasta> \
--dbsnp <dbsnp_138.hg19.vcf> \
--reference_window_stop 300 \
-V <cohort_flagged_vcf>
```

2.6.2.5 VQSR Flagged Variant Filter

Variants sites flagged for removal during VQSR were removed from downstream analyses using *VCFtools* --remove-filtered-all, with --recode and --recode-INFO-all parameters.

```
### Removal of VQSR Flagged Variants ###

vcftools \
--vcf <cohort_flagged_vcf> \
--out <cohort_filtered_vcf> \
--remove-filtered-all \
--recode \
--recode-INFO-all
```

2.6.2.6 Hardy-Weinburg Equilibrium and Missingness Hard Filtering

Variant sites with a p-value falling under the threshold of 10^{-6} when testing for Hardy Weinburg Equilibrium exact test (Wigginton, Cutler and Abecasis, 2005), were flagged for removal using *VCFtools -- hwe.* Variant sites missing greater than 2% of data were flagged for exclusion using *VCFtools -- maxmissing.*

Flagged variants were removed from further analysis using *VCFtools* parameters

```
--remove-filtered-all --recode --recode-INFO-all.
```

2.6.2.7 Variant Evaluation

VariantEval (GATK) was run with parameters --evalModule CountVariants and --stratificationModule Sample, to simultaneously evaluate variant counts and generate an index for the newly created VCF.

CollectVariantCallingMetrics (Picard) was run to generate a more detailed evaluation report.

```
### Variant Evaluation ###

java <java_arguments> -jar <GATK_jar_file> \
-T VariantEval \
-R <reference_genome_fasta> \
-eval <cohort_filtered_vcf> \
--evalModule CountVariants \
--stratificationModule Sample \
-noEV \
-o <cohort_filtered.varianteval>

java <java_arguments> -jar <picard_jar_file> CollectVariantCallingMetrics \
INPUT= <cohort_filtered_vcf> \
OUTPUT= <cohort_filtered.varianteval.picardmetrics> \
DBSNP= <dbsnp_138.hg19.vcf>
```

2.6.3 Variant filtration of Cohort 3

Software and corresponding versions used throughout variant filtration are listed in Table 2-19.

Software	**Version**
BCFtools (Danecek *et al.*, 2021)	1.15
VCFtools	0.1.17

Table 2-19 Software used in variant filtration.

2.6.3.1 Indel Removal

Indel variants sites flagged for removal from downstream analyses using *VCFtools* --remove-indels --remove-filtered-all, with --recode and --recode-INFO-all parameters.

2.6.3.2 Hardy-Weinburg Equilibrium and Missingness Hard Filtering

Variant sites with a p-value falling under the threshold of 10^{-6} when testing for Hardy Weinburg Equilibrium exact test (Wigginton, Cutler and Abecasis, 2005), were flagged for removal using *VCFtools -- hwe.* Variant sites missing greater than 2% of data were flagged for exclusion using *VCFtools -- maxmissing.*

Flagged variants were removed from further analysis using *VCFtools* parameters -- *remove-filtered-all --recode --recode-INFO-all.*

2.6.3.3 Quality Flagged Variant Filter

In the absence of VQSR variants sites were flagged for removal using *BCFtools* with filtering thresholds set as defined in Table 2-20. These hard-filtering thresholds are set as outlined in the table using default thresholds recommended by GATK Best Practices.

Filter	Description	Threshold
QD	Variant quality / depth	< 2.0
MQ	Mapping Quality	< 40.0
FS	Phred-score Fisher's test p-value for strand bias	> 60.0
HaplotypeScore	Consistency of the site with haplotype	> 13.0
MQRankSum	Mapping quality of reference reads vs alternative reads	<-12.5
ReadPosRankSum	Distance of alternative allele from the end of the reads	< -8.0

Table 2-20 GATK recommended variant quality filters for SNVs.

2.6.3.4 Variant Evaluation

VCFtools --TsTv-summary was run to evaluate variant counts and measure transition-transversion ratios per individual. *VCFtools* --depth was run to evaluate mean depth of sequencing per individual.

2.7 Cohort-level QC

2.7.1 Cohort-level QC of Cohort 1

Software and corresponding versions used throughout cohort QC are listed in Table 2-21.

Software	Version
Plink (Purcell *et al.*, 2007)	plink-1.9-beta6.10-gcc-8.2.0-3uh4ocr
R studio	3.4.3 (plotting)/ 4.0.2(manipulation)
ggplot2	3.2.0
Peddy (Pedersen and Quinlan, 2017)	0.4.3
Htslib	htslib-1.9-gcc-8.2.0-7jwlitg
VCFtools	0.1.14
Python	python-3.7.0-gcc-8.2.0-g4ikncu
Tidyverse (Wickham *et al.*, 2019)	1.3.0

Table 2-21 Software used at cohort-level QC.

The filtered cohort-level VCF was converted to *plink* format using *plink - - make-bed.* Default sex, family IDs and parental information were updated using *plink - - update-sex*, *- - update-ids* and *- - update-parents,* respectively.

Pruning of variants for downstream analysis was carried out using the following *plink* parameters:

```
- -indep-pairwise 50 5 0.2
- -maf 0.01
```

The cohort-level QC-filtered VCF was prepared for input to *peddy* by zipping (*bgzip*) and indexing (*tabix*) (Li *et al.*, 2009). A standard input ped file was prepared from clinical data. *peddy* was run as standard (*- -p 4 - -plot*) with output files outlined in Table 2-25 generated for manual inspection, along with an interactive html report file (Pedersen and Quinlan, 2017).

2.7.1.1 Sex Check

Plink - - sex-check compared sex assignments against sex imputed from X-chromosome inbreeding coefficients.

To confirm sample identity, sex-check was performed on pruned variants, by evaluating X chromosome inbreeding coefficients as measured by an F-statistic (Table 2-22). The default F value threshold is <0.2 indicating a female call and values >0.8 indicating male assignment. In the case of this small cohort, the F threshold was reduced to values greater than 0.7 being sufficient for male gender assignment. A problem status occurs when there is discordance between the imputed sex and that input from phenotypic data.

Sex check designated 28 individuals as male and 14 individuals as female. This result excluded three individuals from analysis by identifying discordance between reported and imputed sex (Anderson *et al.*, 2010). Individuals are excluded from further analyses to eliminate downstream inaccuracies.

FID	IID	Reported Sex	Imputed Sex	Status	F-statistic
AS023	AS023C1	1	1	OK	0.843
AS023	AS023F	1	1	OK	0.746
AS023	AS023M	2	2	OK	-0.167
AS070	AS070C	1	1	OK	0.880
AS070	AS070M	2	2	OK	-0.082
AS075	AS075C	1	1	OK	0.835
AS075	AS075F	1	1	OK	0.874
AS108	AS108C1	2	1	PROBLEM	0.852
AS108	AS108C2	1	2	PROBLEM	-0.111
AS108	AS108F	1	2	PROBLEM	-0.115
AS108	AS108M	2	2	OK	-0.040
AS126	AS126C	1	1	OK	0.862
AS142	AS142C1	1	1	OK	0.871
AS157	AS157C1	1	1	OK	0.789
AS157	AS157F	1	1	OK	0.871
AS157	AS157M	2	2	OK	-0.120
AS190	AS190C	2	2	OK	0.021
AS198	AS198F	1	1	OK	0.850
AS198	AS198M	2	2	OK	0.027
AS217	AS217C1	1	1	OK	0.859
AS217	AS217F	1	1	OK	0.871
AS217	AS217M	2	2	OK	-0.153
AS218	AS218C1	1	1	OK	0.838
AS218	AS218F	1	1	OK	0.849
AS218	AS218M	2	2	OK	-0.113
AS306	AS306	1	1	OK	0.862
AS306	AS306F	1	1	OK	0.861
AS306	AS306M	2	2	OK	-0.163
AS310	AS310C1	1	1	OK	0.890
AS311	AS311C1A	1	1	OK	0.812

AS311	AS311C1B	1	1	OK	0.858
AS312	AS312	1	1	OK	0.867
AS314	AS314	1	1	OK	0.871
AS315	AS315	2	2	OK	-0.117
AS316	AS316C1	1	1	OK	0.893
AS319	AS319	1	1	OK	0.824
AS321	AS321	1	1	OK	0.881
AS322	AS322C1	2	2	OK	-0.124
AS420	AS420C1	1	1	OK	0.841
AS420	AS420C2	1	1	OK	0.853
AS420	AS420F	1	1	OK	0.880
AS420	AS420M	2	2	OK	-0.181

Table 2-22 F-statistics in imputation of sex from genomic data.

FID indicates the family ID with PID indicating the unique individual ID. F refers to father, M to mother and C1 and C2 to children 1 and 2, respectively. The F-statistic is derived from the inbreeding coefficient. F-statistic greater than 0.8 indicate male sex, with values less than 0.2 indicating female. PROBLEM in that status field for each sample shows discrepancy between the reported and imputed sex for that sample. OK indicated that no discrepancy in sex was detected.

Contradictions between clinically reported sex and sex inferred from genotypes were also highlighted by *peddy* (Figure 2-3) (Pedersen and Quinlan, 2017). Sex is estimated through genotype evaluation of the pseudo-autosomal regions of the X chromosome. With one X chromosomes, males would be expected to have no heterozygous genotype calls on the X chromosome. *peddy* measures sex as the ratio of heterozygous to homozygous genotypes in this region.

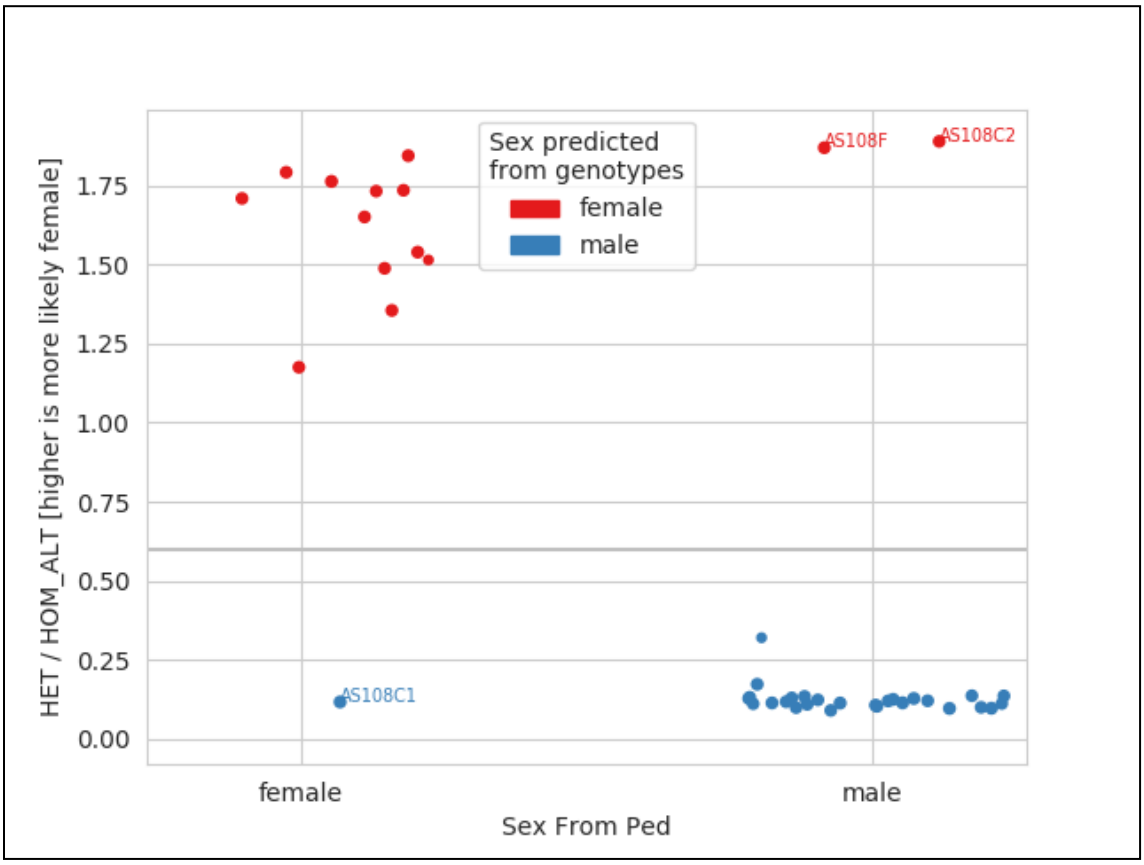


Figure 2-3 Cohort-level sex-check.

Genotype inferred sex is estimated from the proportion of heterozygous/homozygous-alternative genotypes on the X chromosome. The higher the proportion of such calls, the increased likelihood of female sex. Colours, red (female) and blue (male), illustrate the sex reported from clinical data.

2.7.1.2 Relatedness Confirmation

Plink - - genome was used to compute genome-wide identity by descent (IBD) estimates and report the proportion of IBD, i.e., PI_HAT= P(IBD=2) + 0.5*P(IBD=1).

Imputation of relatedness was carried out as further confirmation of sample identity and as verification of familial relationships, as well as identifying any potential duplicate samples. *Plink* estimates relatedness by calculating genome-wide estimates of identity by decent for each pair of individuals in the cohort. Identity by decent is an estimate of the number of alleles in the pair of individuals that are derived from the same ancestral chromosome. Unrelated individuals are expected to have a negligible PI_HAT estimate, while parent-child and sibling pairs are expected to have a PI_HAT estimate of 0.5. Figure 2-4 shows the distribution of PI_HAT across the cohort. As expected, most relationships show negligible PI_HAT values indicating lack of relatedness. A clustering at approximately 0.5 is also expected as this represent those individuals that are truly

related in the cohort. The relationship scoring 1.0 indicates a duplicated sample in the cohort, one of which was removed from further analyses.

Relatedness estimates were calculated to confirm true familial relationship. IBD calculations are also used to identify duplicates and distant relatives in the cohort. Five individuals in the cohort show discordance between reported relatedness and calculated relatedness. Three of these individuals had already been excluded from downstream analysis based on failing sex-check. Related individuals are excluded from PCA to eliminate overrepresentation of alleles in the population.

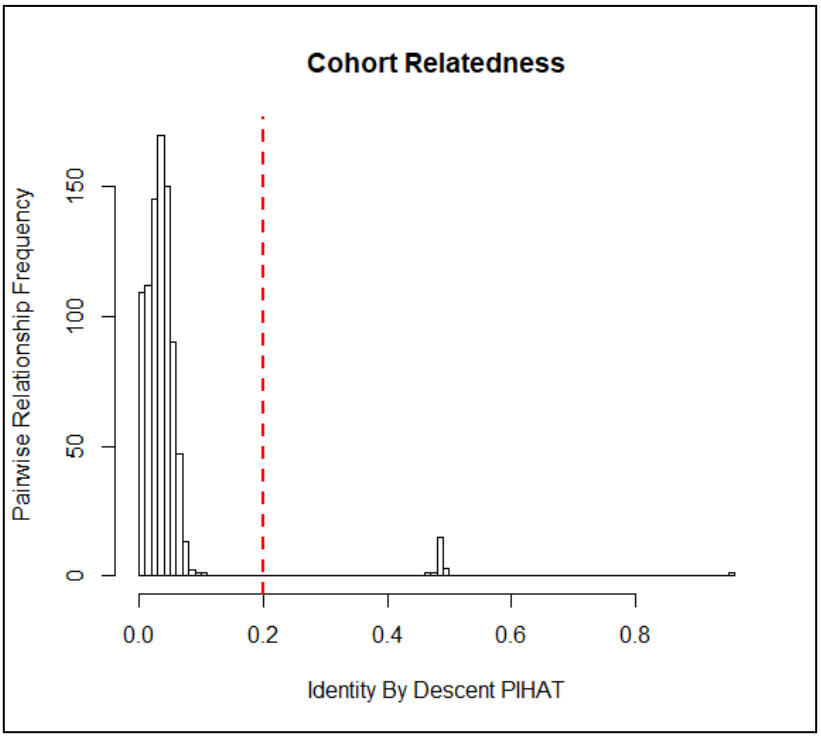


Figure 2-4 Cohort relatedness as measure by PI_HAT.

This plot illustrates the degree of relatedness within the cohort. Pairwise relationships are systematically investigated for PIHAT estimates of identity by decent. The dashed red line indicates cut-off for unrelated individuals in the cohort. Any relationships exceeding this cut-off are true related or duplicate samples. This cut-off is used to exclude related individuals from downstream PCA analyses.

Further investigation into the validity of the familial relationship was carried out in the comparison of expected IBD and IBD estimated (Figure 2-5). A linear relationship is expected to indicate correct familial relationship data. Deviation from this expectation indicates discordance between input family information and imputed IBD. This is evident in the outliers present in Figure 2-5.

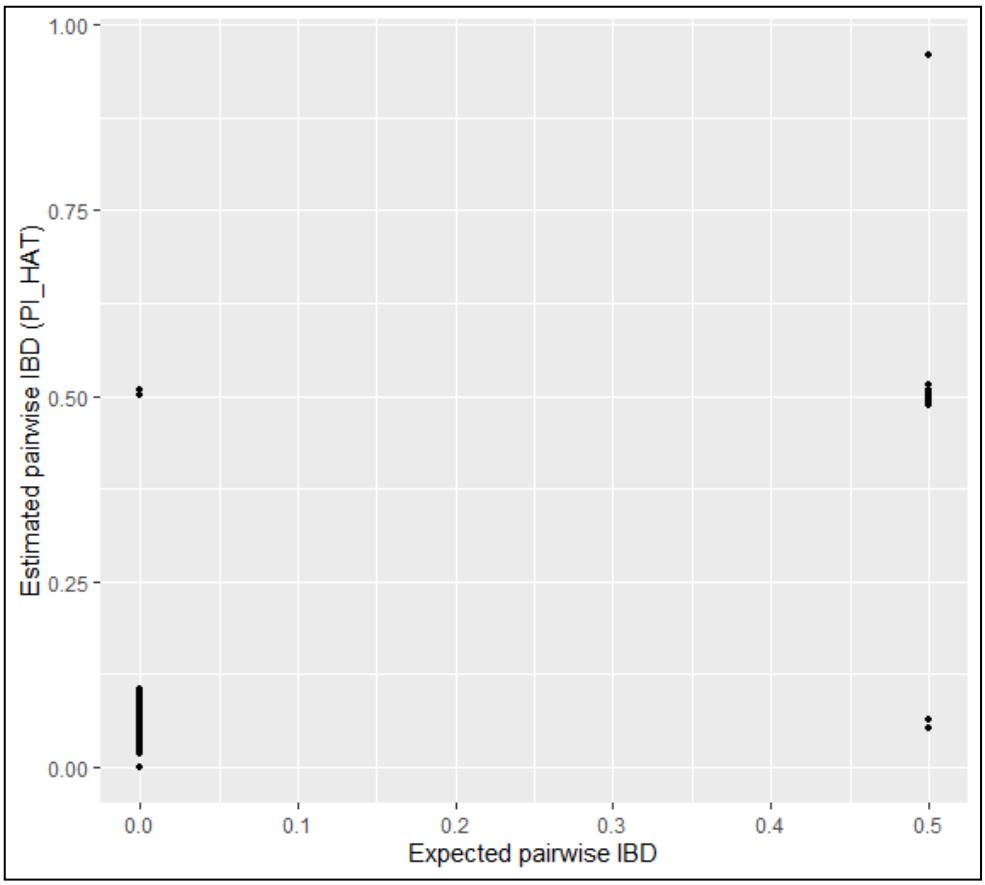


Figure 2-5 Expected IBD vs estimated IBD.

Expected pairwise IBD is derived from phenotypic data, while estimated pairwise IBD is calculated from genotypes. Deviations from expectation can be identified as points straying from linear.

Relatedness within the cohort was also imputed by *peddy* as Identity-by-State (Pedersen and Quinlan, 2017) (Figure 2-6). IBS0 reports the number of shared variant alleles, the number of variant sites at which individuals share 0. Errors are noted in Figure 2-6 where the colour of each relationship point, as specified in the legend, is not positioned where expected from the relationship specified as the coefficient of relatedness.

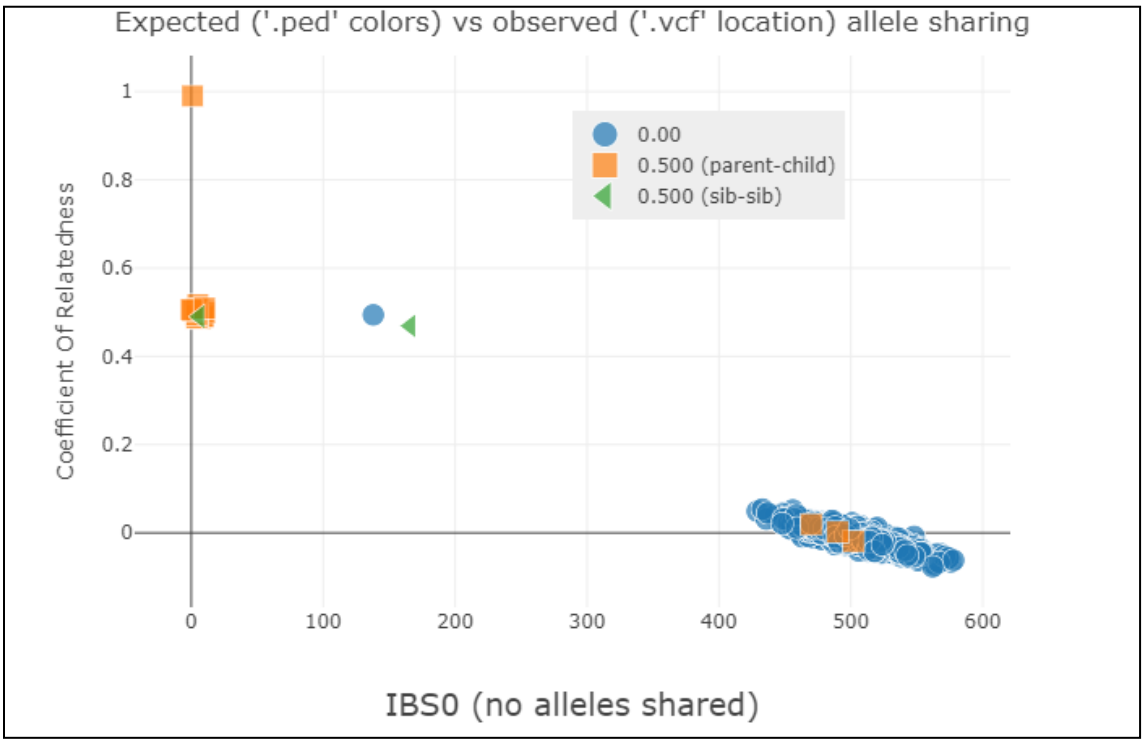


Figure 2-6 Relatedness inference.

IBS0 statistics vs IBS2 statistics plots the number of variant sites at which individuals share 0 alleles (x-axis) vs the coefficient of relatedness (y-axis). Colours represent the clinically reported relationships within the cohort as outlined in the key.

2.7.1.3 Identification of population substructures

Principal component analysis (PCA) was used to identify potential sample clustering and outliers. Principal components representing the greatest variance may reflect population substructures. Pruning of variants was carried out prior to PCA to ensure the variants on which the principal components are derived are common (MAF >0.01) are approximately independent.

Plink - - pca var wts was run to compute principal components within the cohort on unrelated individuals. Independence is determined for SNP window sizes of 50kb with the number of SNPs to shift window at each step of 5. Independent SNPs are filtered by removing one of a pair of SNPs within the window if the pairwise linkage disequilibrium is greater than an r^2 threshold 0.2. This ensures that the principal components are not computed to represent areas of local linkage disequilibrium. Duplicate and closely related individuals were removed prior to computation of principal components to reduce bias from over-representation of alleles (n=17 individuals removed with n=25 individuals remaining).

R Studio (plot) was used to plot the top two principal components and top five principal components. All samples in the cohort are reported to be of European ancestry. PCA of a total of 37,263 variants passing QC filters shows clustering of principal components with deviation likely reflecting sub-European population structures (Figure 2-7,Figure 2-8).

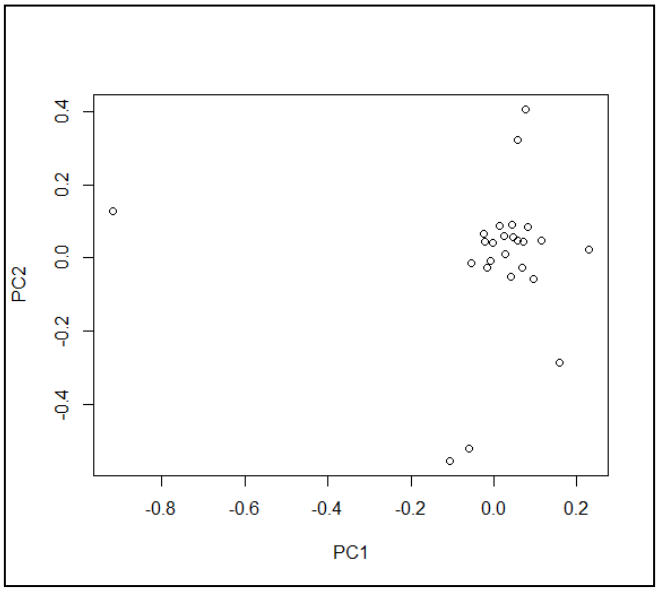


Figure 2-7 Principal component 1 and principal component 2.

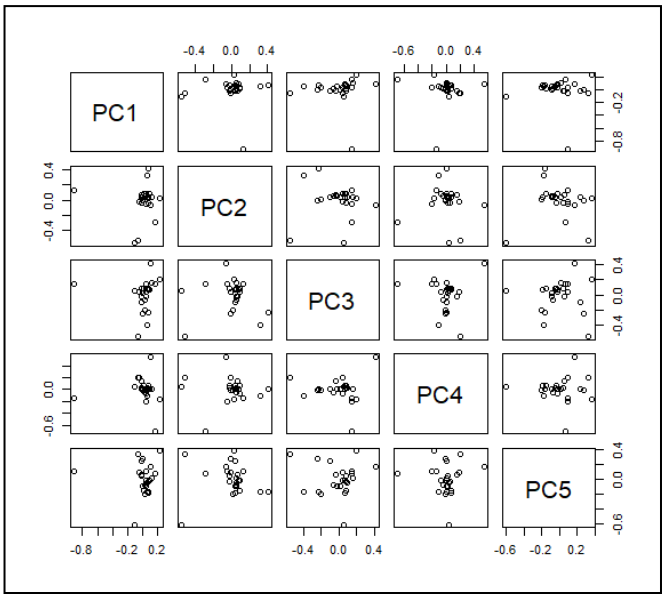


Figure 2-8 Top five principal components.

All samples within the cohort are of European ancestry as estimated by *peddy* (Pedersen and Quinlan, 2017). This randomised PCA is trained on 2,504 samples from the 1000 Genomes Project (Halko, Martinsson and Tropp, 2009; The 1000 Genomes Project

Consortium, 2015). Figure 2-9 presents four principal components of the cohort projected onto 1000G, and ancestry is predicted as European in all samples.

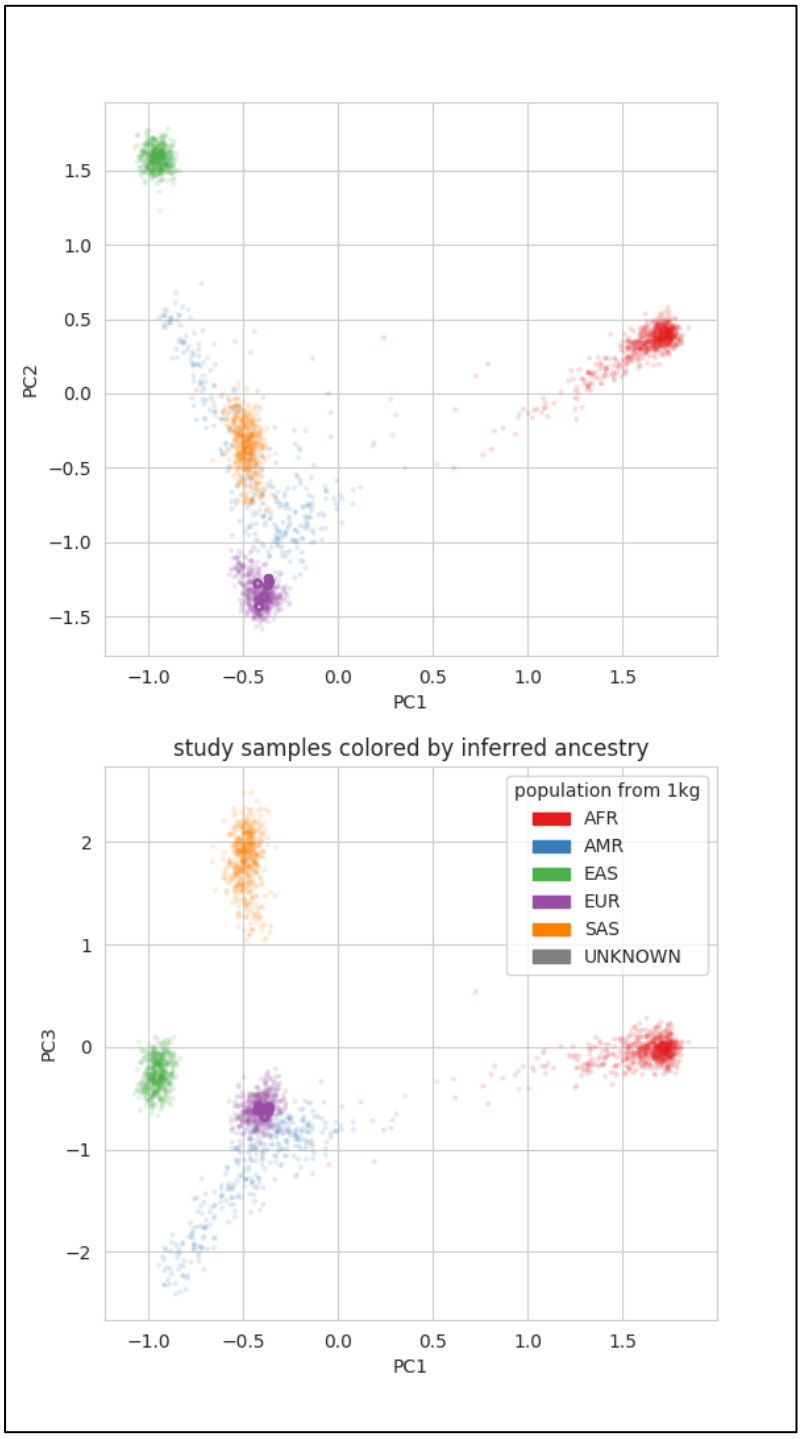


Figure 2-9 Ancestry evaluation through principal components 1,2 and 3.

The first three principal components are calculated during PCA of 1000 genomes dataset and the cohort under analysis and are plotted against each other. Each point on the graph represents an individual, with bold points representing the samples under investigation and faint points representing samples from the 1000G dataset. Samples cluster into 'super-populations,' with corresponding ancestry denoted by colour, outlined in the legend. The cohort under investigation are visible as bold purple points, clustering within the European population. Abbreviations: PC principal component, PCA principal component analysis, AFR African, AMR Ad Mixed American, EAS East Asian, EUR European, SAS South Asian.

2.7.1.4 Cohort QC in summary

Cohort–level QC of the WES dataset flagged seven individuals for errors in sample identity as outlined. Discordance in reported and imputed sex flagged 3 samples for removal from the cohort (Table 2-23). Following inspection of relationships, 1 sample was removed because of sample duplication and 4 samples were removed due to sample mix-up, as reinforced by their incorrect sex assignment (in n=3).

VCFtools --remove-indv was used to remove flagged individuals from the cohort VCF file with parameters *--recode and --recode-INFO-all.*
This removes the sample from the ID column of the VCF file. However, variants occurring in these individuals are not removed from the cohort VCF file. These variants remain present in the cohort file but are unassigned to an individual as the respective ID column has been removed. These now unassigned variants are removed using the *filter* function in tidyverse with specification *grepl('0/1|1/0|1/1' to* select out variants in a specific individual.

All individuals within families AS023 and AS108 (n=7) were removed from downstream analyses based on the above deviations from expectation. Confirmation of these sample mix-ups was obtained through an independently sequenced genotyping array. This indicates that sample misidentification did not occur during preparation of the sequencing run, but rather are the result of an error in labelling of the DNA stocks. These DNA stocks were subsequently destroyed to prevent future errors.

Flagged Sample	**Sex**	**Relatedness**
AS023C1	TRUE	Unrelated to AS023F and AS023M
AS023F	TRUE	Unrelated to AS023C1
AS023M	TRUE	Unrelated to AS023C1
AS108C1	FALSE	Unrelated to AS108M
AS108C2	FALSE	Duplicate sample of AS108F
AS108F	FALSE	Duplicate sample of AS108C2
AS108M	TRUE	Related to AS108F

Table 2-23 Samples flagged for removal by cohort-level QC checks.

The samples included in this table were removed based on discrepancies in clinically reported characteristics and genetically imputed characteristics.

Individual AS310 was excluded from downstream analyses based on consent (updates from clinical collaborator Professor Louise Gallagher, May 2021). Data was removed from the cohort dataset. No individuals were flagged for removal based on population stratification in these analyses.

2.7.2 Cohort-level QC of Cohort 2

The cohort-level QC-filtered VCF was prepared for input to *peddy* by zipping (*bgzip*) and indexing (*tabix*) (Li *et al.*, 2009)(Table 2-24). A standard input ped file was prepared from clinical data. *peddy* was run as standard with output files outlined in Table 2-25 generated for manual inspection, along with an interactive html report file (Pedersen and Quinlan, 2017).

Software	**Version**
Peddy	0.4.3
Samtools/htslib	htslib-1.9-gcc-8.2.0-7jwlitg
Python	python-3.7.0-gcc-8.2.0-g4ikncu

Table 2-24 Software used at cohort-level QC.

Suffix	**Context**	**Format**
.ped_check .ped_check.rel-difference	Discrepancies in ped-reported and genotype-inferred relationship	csv, png
.sex_check	Discrepancies in ped-reported and genotype-inferred sex	csv, png
.het_check	Samples with higher levels of HET calls	csv, png
.pca_check .background_pca	Ancestry prediction based on projection onto 1000G principal components	csv, png, json

Table 2-25 Report files generated by peddy cohort analysis tool.

2.7.2.1 Sex check

To confirm sample identity, contradictions between clinically reported sex and sex inferred from genotypes were highlighted by *peddy* (Pedersen and Quinlan, 2017). Sex is estimated through genotype evaluation of the pseudo-autosomal regions of the X chromosome. With one X chromosome, males would be expected to have no

heterozygote genotype calls on the X chromosome. *peddy* measures sex as the ratio of heterozygous to homozygous genotypes in this region (Pedersen and Quinlan, 2017). There are no discrepancies between clinically reported sex and sex inferred from genotypes across the cohort (Figure 2-10).

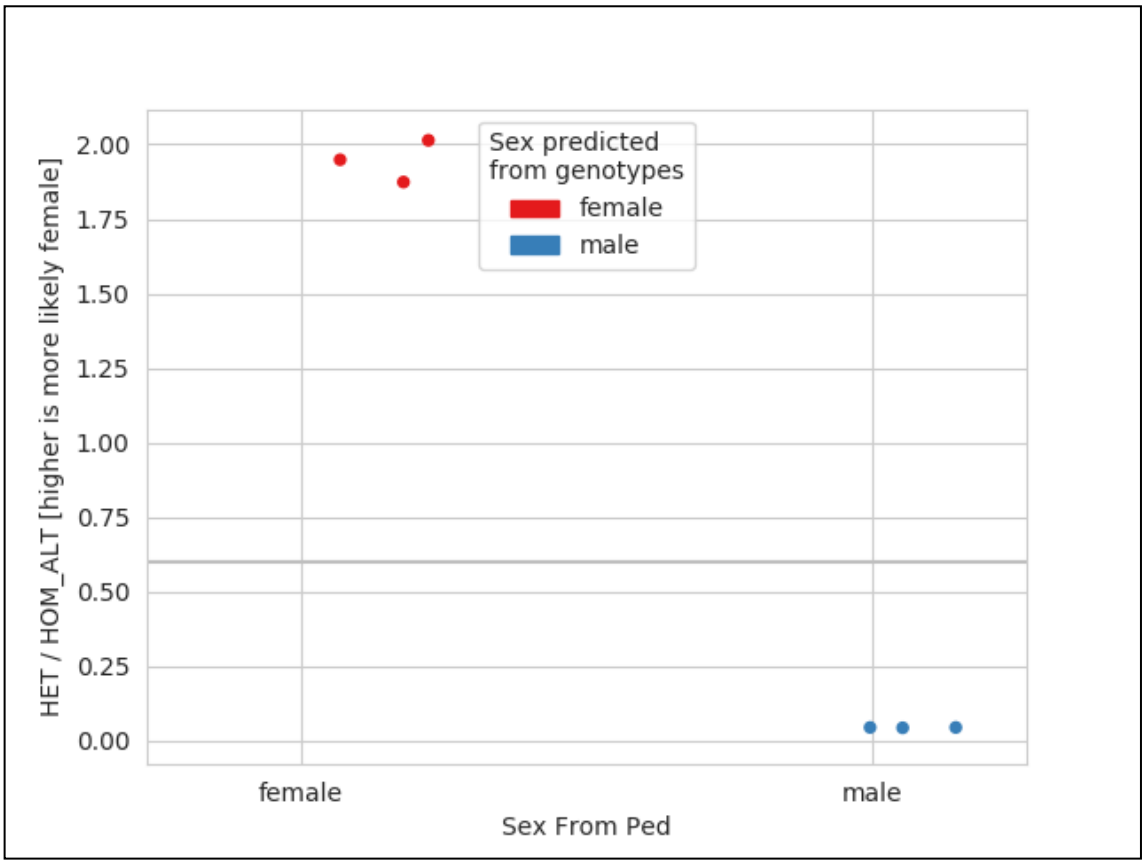


Figure 2-10 Cohort-level sex-check.

Genotype inferred sex is estimated from the proportion of heterozygous/homozygous-alternative genotypes on the X chromosome. The higher the proportion of such calls, the increased likelihood of female sex. Colours, red (female) and blue (male), illustrate the sex reported from clinical data.

2.7.2.2 Relatedness confirmation

There are no contradictions between self-reported relationships and relationships inferred from genotypes across the cohort (Figure 2-11). *peddy* runs a modification of the KING algorithm across a total of 23,556 variant sites (Manichaikul *et al.*, 2010; Pedersen and Quinlan, 2017). This relatedness inference generates and plots statistics IBS0 and IBS2. IBS0 represents the number of variant sites at which a pair of individuals shares 0 alleles, for example a site at which one individual is A/A and the other is G/G. IBS0 enables the differentiation between sibling-pair and parent-offspring relationships, not possible through traditional relatedness estimates as both relationships are

estimated at 0.5. IBS0 would be expected near 0 for parent-offspring pairs as sites not shared between parent and offspring would be Mendelian violations. In contrast, an IBS0 statistic greater than 0 indicates that not all variant sites are shared, as would be expected in sibling-pair relationships. IBS2 estimates the number of variant sites at which both samples share the same genotype (both alleles). In plotting IBS0 and IBS2 separation can be made between related and unrelated individuals, while also differentiating between sibling-pair and parent-offspring relationships.

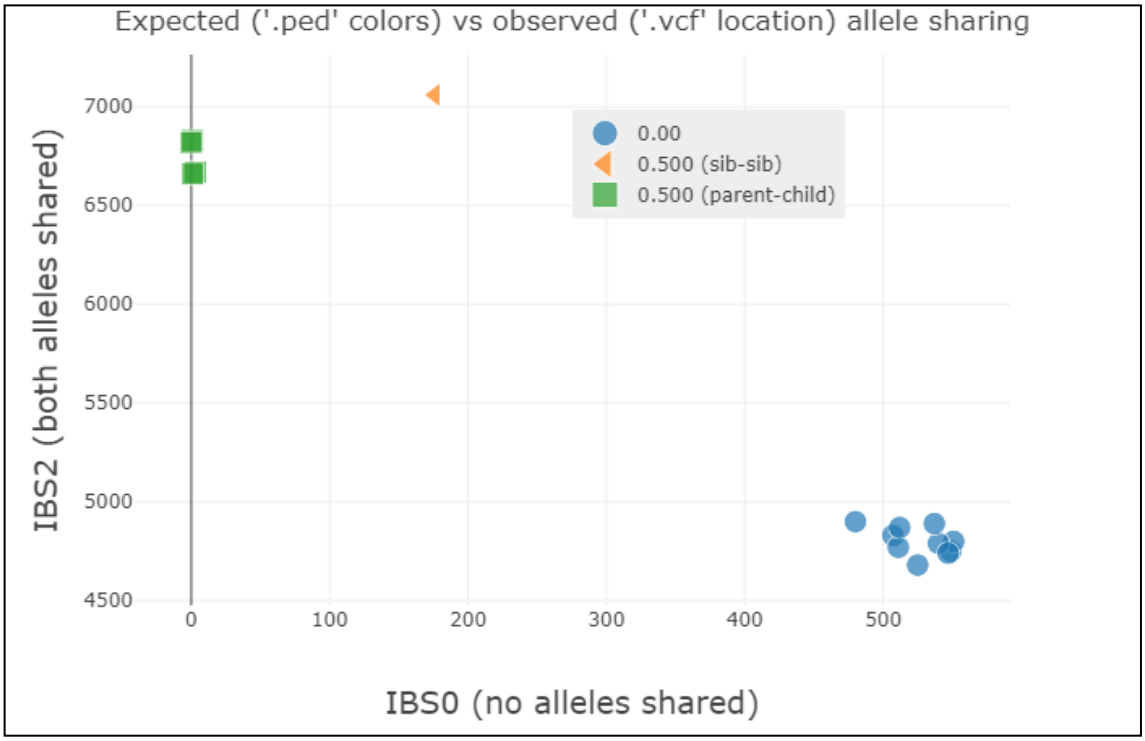


Figure 2-11 Relatedness inference.

IBS0 statistics vs IBS2 statistics plots the number of variant sites at which individuals share 0 alleles (x-axis) vs the number of variant sites with the same genotype (y-axis). Colours represent the clinically reported relationships within the cohort as outlined in the key.

2.7.2.3 Identification of population substructures

All samples within the cohort are of European ancestry as estimated by *peddy* (Pedersen and Quinlan, 2017). This randomised PCA is trained on 2,504 samples from the 1000 Genomes Project (Halko, Martinsson and Tropp, 2009; The 1000 Genomes Project Consortium, 2015). Four principal components of the cohort are projected onto 1000G, and ancestry is predicted (Figure 2-12).

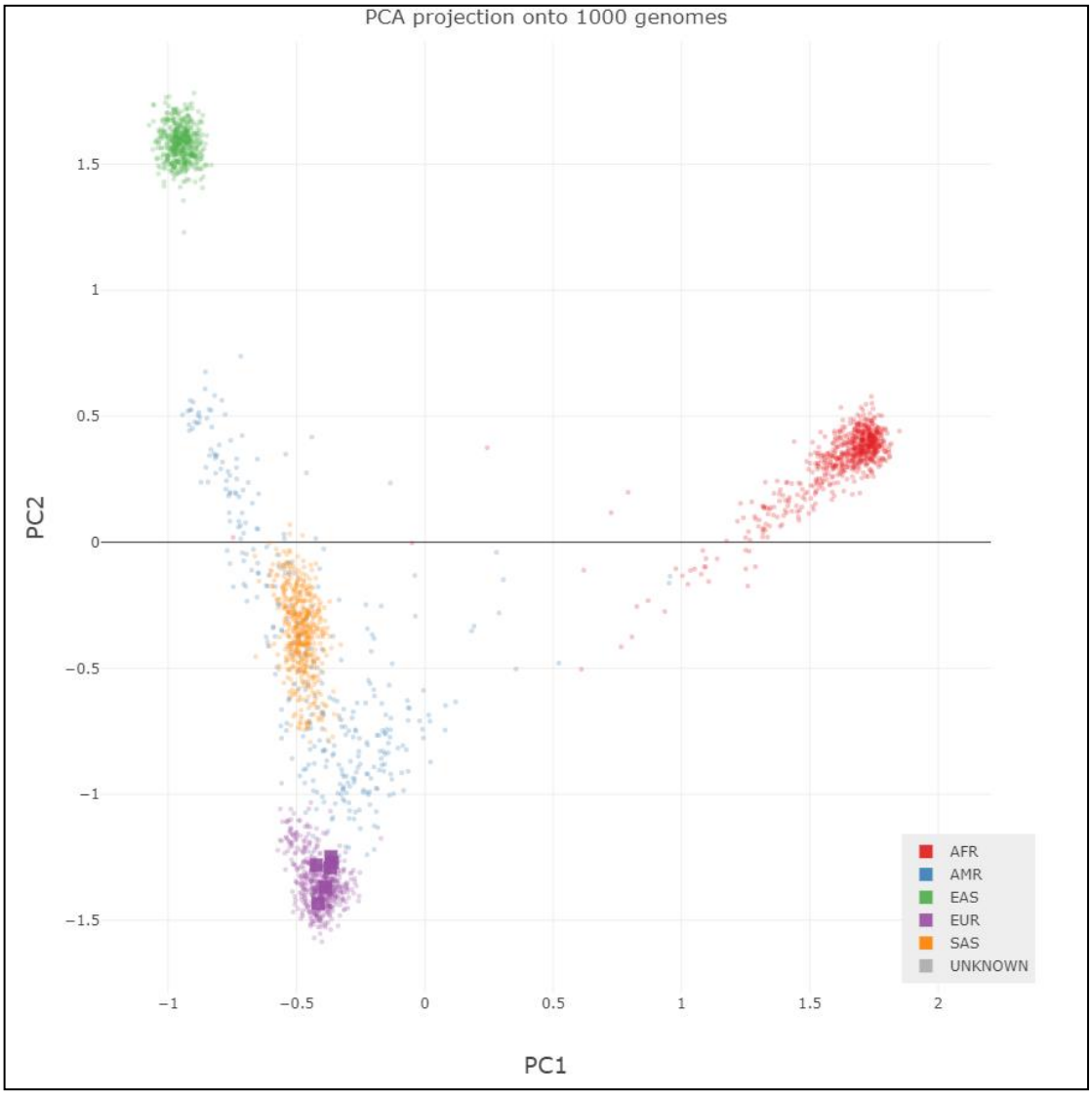


Figure 2-12 Ancestry evaluation through principal components 1 and 2.

The first two principal components are calculated during PCA of 1000 genomes dataset and the cohort under analysis and are plotted against each other. Each point on the graph represents an individual, with bold square points representing the samples under investigation and faint points representing samples from the 1000G dataset. Samples cluster into 'super-populations,' with corresponding ancestry denoted by colour, outlined in the legend. The cohort under investigation are visible as bold purple squares, clustering within the European population. Abbreviations: PC principal component, PCA principal component analysis, AFR African, AMR Ad Mixed American, EAS East Asian, EUR European, SAS South Asian.

2.7.3 Cohort-level QC of Cohort 3

Software and corresponding versions used throughout cohort QC of Cohort 3 are listed in Table 2-26.

Software	**Version**
Plink	1.07
R studio	4.0.3
ggplot2	3.3.3
Tidyverse	1.3.0

Table 2-26 Software used at cohort-level QC.

The genome-wide genotyping dataset of Cohort 3 was interrogated during cohort QC in *plink* format (bfile). Default sex, family IDs and parental information were updated using *plink - - update-sex*, *- - update-ids* and *- - update-parents,* respectively.

Pruning of variants for analysis was carried out using the following *plink* parameters:

```
--geno 0.2
--hwe 0.000001
```

2.7.3.1 Sex Check

Plink - - sex-check compared sex assignments against sex imputed from X-chromosome inbreeding coefficients.

Sex check was performed on the genome-wide genotype data of Cohort 3. To confirm sample identity, sex-check was performed by evaluating X chromosome inbreeding coefficients as measured by an F-statistic. The default F value threshold to determine imputed set is <0.2 indicating a female call and values >0.8 indicating male assignment. A problem status occurs when there is discordance between the imputed sex and that input from phenotypic data, otherwise the sample is passed by the sex-check.

Sex check designated 14 individuals as male and 15 individuals as female. No discordance between reported and imputed sex was identified and no individuals were excluded from further analyses.

2.7.3.2 Relatedness Confirmation

Plink - - genome was used to compute genome-wide identity by descent (IBD) estimates and report the proportion of IBD, i.e., PI_HAT= P(IBD=2) + 0.5*P(IBD=1).

Imputation of relatedness was carried out as further confirmation of sample identity and as verification of familial relationships, as well as identifying any potential duplicate samples. Plink estimates relatedness by calculating genome-wide estimates of identity by decent (IBD) for each pair of individuals in the cohort. IBD is an estimate of the number of alleles in the pair of individuals that are derived from the same ancestral chromosome.

Relatedness was imputed from genome-wide genotypes across the cohort. These variant sites were pruned to variant sites with an 80% genotyping rate and passing a Hardy-Weinburg exact test at a threshold of $p<=1e^{-06}$ (removing n=3,823 variant sites). Investigation into the validity of the familial relationship was carried out in the comparison of expected IBD and IBD estimated (Figure 2-13). A linear relationship is expected to indicate correct familial relationship data. Deviation from this expectation indicates discordance between input family information and imputed IBD.

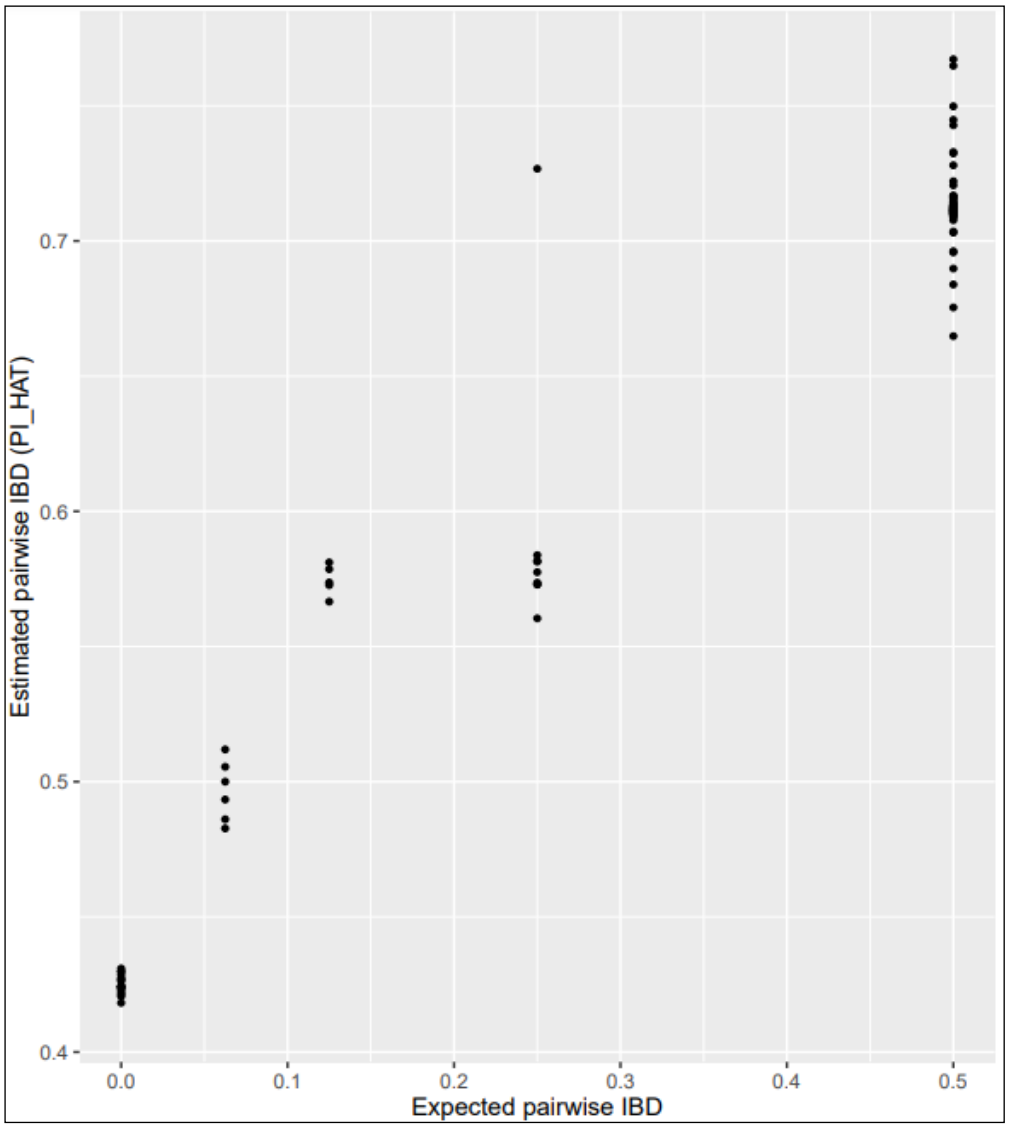


Figure 2-13 Expected IBD vs estimated IBD.

Expected pairwise IBD is derived from phenotypic data, while estimated pairwise IBD is calculated from genotypes. Deviations from expectation would be identified as points straying from a linear relationship. Pairwise relationships are present for this cohort of 29 individuals estimated from genome-wide genotyping.

2.7.4 Variant quality

The 23,556 variant sites interrogated by *peddy* were assessed for depth of coverage and rate of heterozygosity. Presented here are the variant filtration metrics for Cohort 2 indicating that sequenced samples fall within the expected range (Figure 2-14, Table 2-27). Deviations from this range would be indicative of potential sample contamination or consanguinity within the dataset.

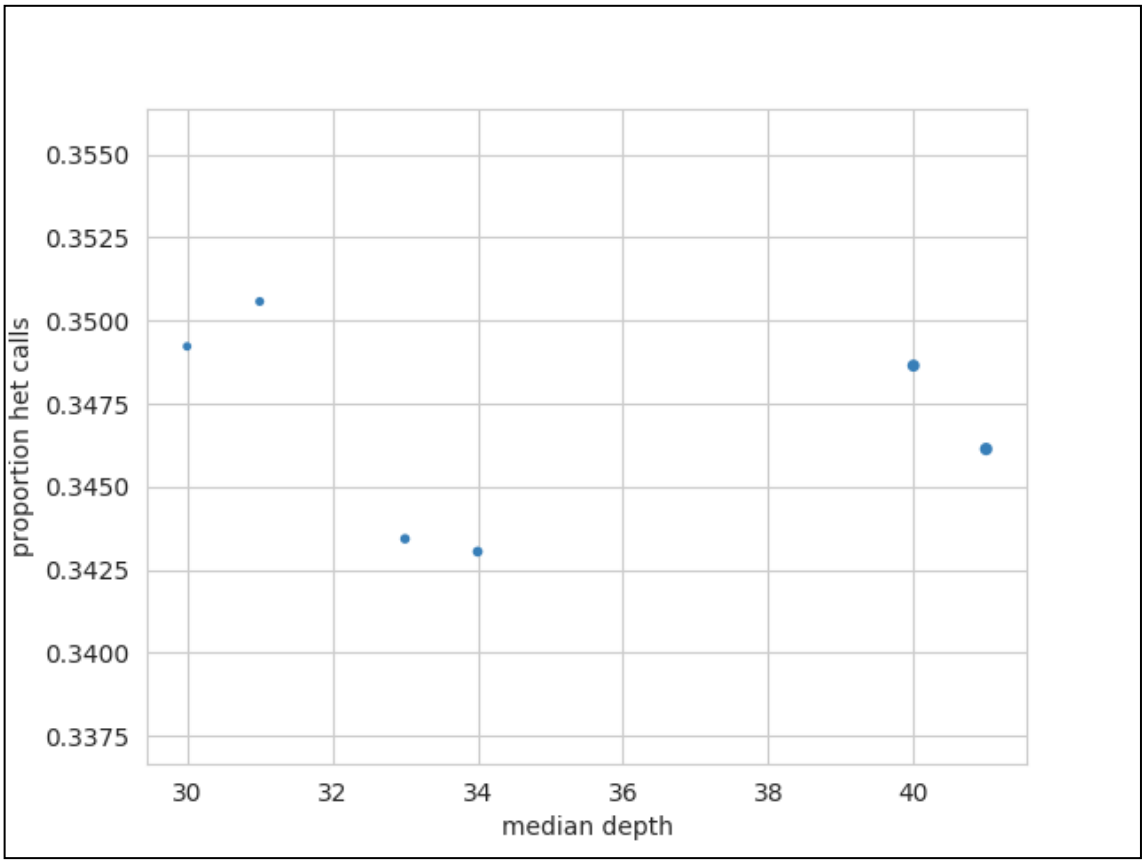


Figure 2-14 Rate of heterozygosity across samples.

Each individual in the cohort is represented as a blue point. All samples are reported as 'OK' as per peddy expected range for proportion of heterozygous calls across 23,556 variant sites (X-axis) and depth of coverage (y-axis).

Family ID	Sample ID	Mean Depth	Heterozygous Call Rate	Het Ratio	Inter-decile range of b-allele frequency
AS315	AS315	34.89	0.99	0.34	0.21
AS322	AS322C1	35.91	0.99	0.34	0.20
AS420	AS420F	42.83	1	0.35	0.19
AS420	AS420C2	42.03	1	0.35	0.19
AS420	AS420C1	33.26	1	0.35	0.22
AS420	AS420M	30.48	0.99	0.35	0.24

Table 2-27 Heterozygosity check.

Reported are QC measures pertaining to rates of heterozygosity at the annotated variant call sites. Mean Depth presents the mean depth of coverage for the annotated variant sites. The proportion of sites that were heterozygous is presented as Het Ratio. Inter-decile range of b-allele frequency is computed as the number of alternative allele sites as a proportion of reference and alterative variant sites and reported the difference between the 90th and the 10th percentile of the b-allele frequency. Large values of this measure are likely to indicate sample contamination.

2.8 Variant annotation

Variant annotation was aligned for annotation of the variant set of Cohort 1, Cohort 2 and Cohort 3 as follows (Table 2-28).

Software	**Version**
dbNSFP (Liu *et al.*, 2020)	4.0a (Cohort 1 and 2), 4.3a (Cohort 3)
Java	openjdk version "1.8.0_242"
Rstudio	version 4.0.2 (Cohort 1 and 2), version 4.0.3 (Cohort 3)
tidyverse	1.3.0

Table 2-28 Variant annotation software and versions.

2.8.1 dbNSFP annotation

Variants were annotated using dbNSFP (database of non-synonymous functional variants) through the java search_dbNSFP40a tool. dbNSFP compiles annotations from 29 prediction algorithms, nine conservation scores, allele frequencies from major population databases, including 1000 Genomes and gnomAD, as well as gene-based annotations of expression and interactions (Liu *et al.*, 2016). These annotations are applied to an input call-set in VCF format using the dbNSFP java database search tool.

Optional parameters *-p* (output existing VCF columns), *-v hg19/hg38* (specifying reference genome), and *-g* (include full gene annotation set) were applied. Cohort 1 and Cohort 2 have been aligned to hg19, while Cohort 3 has been aligned to GrCh38.

For a full outline of the variables included in the annotation refer to: https://sites.google.com/site/jpopgen/dbNSFP.

```
### dbNSFP annotation

#Set java memory requirements and set temporary directory to hold temporary files
generated

java              ///
-Xmx6G -Xmx6G          ///
-Djava.io.tmpdir= <tmp>              ///

# Run dbNSFP java search tool to run annotation

search_dbNSFP43a       ///

# Define input and output files
-i <input_vcf>              ///
-o <output_csv>          ///

# output existing VCF columns in annotated csv output file
-p   ///

# include the full gene annotation set
-g   ///

# specify the reference genome
-v hg38
```

2.8.2 Formatting and cleaning the dataset

Reading dbNSFP annotation file directly to R in the absence of a data manipulation tool results in widespread errors, because of incorrect parsing. Due to the large number of variables under analysis in the input file (n=483), default parsing is the most effective strategy to input the data into a workable R environment (version 4.0.2 (Cohort 1 and 2), version 4.0.3 (Cohort 3)). The *tidyverse* toolkit was used in this manipulation. Specifically, *readr (read_tsv)* function was used to parse the dataset.

```
read_tsv(“", na = c("."),
col_types = cols(
hg19_chr = col_character(),
hg18_chr = col_character(),
'#chr' = col_character(),

MutPred_score = col_skip(),
MIM_id = col_skip(),
.default = col_guess()
)
)
```

Parsing is designated by the first 1000 rows in the data frame. This results in chromosome variables being designated as doubles. However, this causes errors when chromosomes X and Y are evaluated. For this reason, default parsing is overridden to consider all chromosome names as characters.

As outlined above parsing overwriting default parsing option was specified for variables. With the reason for each outlined.

problems() was used to evaluate any parsing issues.
as_tibble() was used to convert from data frame to tibble for optimal manipulation.

2.9 Rare variant selection by allele frequency

Rare variants isolated based on minor allele frequencies collated in the Genome Aggregation Database (gnomAD), using the non-Finnish European cohort (n=7,718 individuals). Parameters used in rare variant filter are specified per cohort under investigation in Table 2-29. The annotated variant were filtered according to the workflow presented in Figure 2-15, and as detailed in 2.9 and 2.10.

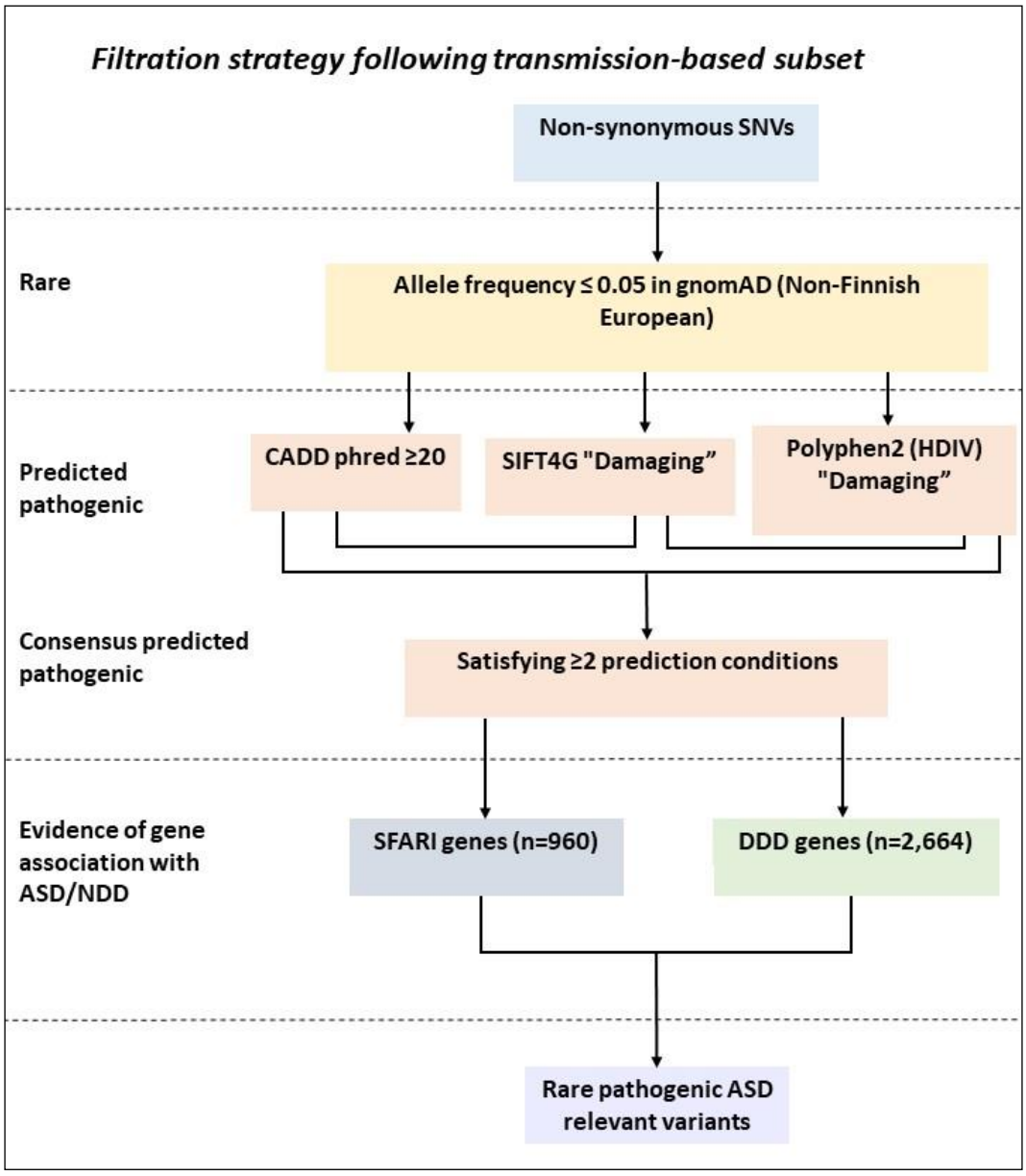


Figure 2-15 Flow of variant filtering.

Arrows show the direction of flow from each level of filtering (specified on the left). SFARI refers to Simons Foundation Autism Research Initiative Gene Module (Abrahams *et al.*, 2013). *DDD refers to the gene2phenotype database arising from the DDD study* (Wright *et al.*, 2015).

Cohort	Parameter	Description	Cut-Off	Reference
1	gnomAD_exomes_NFE_AF	Alternative allele frequency in the non-Finnish European gnomAD exome samples (56,885 samples)	Rare variants isolated as those appearing at an allele frequency of less than 5%	(Karczewski *et al.*, 2020)
2	gnomAD_genomes_NFE_AF <= 0.05	Alternative allele frequency in the non-Finnish European gnomAD genome samples (7,718 samples)	Rare variants isolated as those appearing at an allele frequency of less than 5%	(Karczewski *et al.*, 2020)
3	gnomAD_genomes_NFE_AF <= 0.05	Alternative allele frequency in the non-Finnish European gnomAD genome samples (gnomAD genome samples v3.1)	Rare variants isolated as those appearing at an allele frequency of less than 5%	(Karczewski *et al.*, 2020)

Table 2-29 Rare variant isolation parameters.

2.10 Pathogenic variant selection

Putatively pathogenic variants were classed as such when satisfying two or more of the conditions outlined in Table 2-30. CADD is an algorithm-based estimate of variant pathogenicity. The CADD phred-like score is phred-like rank score based on whole genome raw CADD scores. The larger the CADD-phred score the more likely the SNV annotated has damaging effect. The CADD-phred filter was set at variants with a score of >= 20. SIFT and Polyphen-2 were taken as candidate pathogenicity measures due to their widespread use in human genomics. Criteria for pathogenic as determined by SIFT was taken as “D,” indicating damaging, measured by SIFT4G_pred (SIFT 4G < 0.05).

In addition to these classifications, REVEL was considered for use in pathogenicity determination. REVEL variant scoring ranges from zero to one representing the proportion of trees in the random forest classifying the variant under investigation as pathogenic (Ioannidis *et al.*, 2016).

Parameter	**Description**	**Cut-Off**	**Reference**
CADD_phred >= 20	This is phred-like rank score based on whole genome CADD raw scores. The larger the score the more likely the SNP has damaging effect.	CADD phred-like score of greater than or equal to 20.	(Kircher *et al.*, 2014)
grepl('D', SIFT4G_pred)	If SIFT4G is < 0.05 the corresponding nsSNV is predicted as "D(amaging)"; otherwise, it is predicted as "T(olerated)".	D(amaging) or SIFT 4G < 0.05. SIFT 4G scores range from 0 to 1. The smaller the score the more likely the variant has damaging effect.	(Ng and Henikoff, 2003)

grepl('D', Polyphen2_HDIV_pred)	Polyphen2 score based on HumDiv, i.e., hdiv_prob. The score ranges from 0 to 1.	"D" ("probably damaging", HDIV score in [0.957,1]	(Adzhubei, Jordan and Sunyaev, 2013)

Table 2-30 Pathogenicity parameters.

```
## Pathogenic Variant Isolation Following dbNSFP Annotation

# Define output variant set
<PathogenicVariantSet> <-          ///

# Specify starting variant set
filter(<FullVariantSet>,  ///

# Specify conditions to be satisfied using OR arguments
CADD_phred >= 20 & grepl('D', SIFT4G_pred) |          ///
CADD_phred >= 20 & grepl('D', Polyphen2_HDIV_pred) |          ///
grepl('D', SIFT4G_pred) & grepl('D', Polyphen2_HDIV_pred))
```

2.11 Autism and neurodevelopmental-associated variant selection

Variants were further subset to those with relevant to neurodevelopmental conditions by subsetting to variants impacting the gene lists specified in Table 2-31. The largest curated gene list to date in collating a gene-list for autism is most the SFARI Gene database (Abrahams *et al.*, 2013). This regularly maintained resource presents evidence supporting the role of >1,000 genes in autism (currently n=1,045 genes as of February 2022 update), with use of a gene-phenotype association scoring system to represent the confidence of a given gene in autism. This gene scoring approach collates all available evidence supporting the relevance of the gene to autism risk and categorises each gene depending on the strength of evidence as gene scores of 1 (high confidence), 2 (strong confidence), 3 (suggestive evidence) and/or S (syndromic).

A similarly scored gene list of neurodevelopmental condition-relevant genes is DDD gene2phenotype gene list generated from the DDD study. This gene list is collated from variants identified in a cohort of children with severe and complex neurodevelopmental phenotypes (Wright *et al.*, 2015). DDD assigns genes with a level of certainty of association, given as “Definitive,” “Strong,” or “Limited.”

Variants were further subset based on confidence scoring within these databases. SFARI Genes with high confidence in autism-associated were designated as those with a gene score of 2 and/or syndromic. This subset results in a combined gene set of SFARI high confidence (n=393) and syndromic genes (n=126) of 510 genes, of 960 overall SFARI Genes. The DDD Gene Set was subset to those determined by the consortium with confirmed evidence of pathogenicity. This subsets to 1,648 genes of 2,664 genes overall.

Parameter	Description	Cut-Off	Reference
SFARI Gene	A maintained database of genes implicated in autism susceptibility	Ensembl GeneID present in SFARI-Gene_genes_08-07-2020release	(Abrahams *et al.*, 2013)
DDD	A curated list of genes associated with developmental conditions	OMIM GeneID present in DDG2P_8_9_2020	(Wright *et al.*, 2015)

Table 2-31 Autism and neurodevelopmental condition gene list filtration.

2.12 Filtering by genotype

Software	Version
RStudio	4.0.2 (Cohort 1 and 2), 4.0.3 (Cohort 3)
tidyverse	1.3.0

Table 2-32 Software and versions used in variant filtering by genotype.

Where joint genotyping was performed at cohort-level (Cohort 1 and Cohort 2), homozygosity in for alternative alleles was detected by filtering for variant sites with a 1/1 genotype (Table 2-32). Heterozygosity in these cohorts was detected by filtering for variant sites with either a 0/1 or 1/0 genotype. Homozygosity in Cohort 3 was detected by filtering for variant sites with a per individual allele count of two for the alternative allele. Heterozygosity was detected by filtering variant sites with an allele count of 1 for the alternative allele.

Across sample variant filtering by genotype was performed by *tidyverse filter* function using operator %in% and inverse operator *!%in%* on variant rsID.

2.13 Application of an evidence-based curation framework to aid gene discovery

2.13.1 Dataset under investigation

This study interrogates Cohort 2. To summarise, this dataset is comprised of non-synonymous SNVs arising from WGS of 6 individuals (3 probands and 3 unaffected relatives. Rare putative pathogenic variants with relevance to autism have been isolated as outlined in 2.2.

This filtering strategy isolates 91 genes (Supplemental Table 1) in which autism-relevant variants are occurring. Each of these genes are a candidate for curation in this study.

2.13.2 Evaluation of ClinGen curated genes

With the aim of applying the proposed gene curation strategy to otherwise uncurated genes, genes were excluded when curation records exist in the ClinGen Gene-Disease Validity database (Table 2-33).

Database	**Source**	**Date of export**	**Number of genes**	**Reference**
Clinical Genome Resource: Gene-Disease Clinical Validity Browser	https://www.clinicalgenome.org/curation-activities/gene-disease-validity/	2020-09-28	1,848	(Strande *et al.*, 2017)

Table 2-33 ClinGen gene-disease clinical validity dataset.

This table details the export of genes which have been curated through the ClinGen Gene-Disease Validity process, hereafter referred to as 'ClinGen gene list.' Detailed are the number of genes curated at the date of export specified. Note: the number of genes curated is subject to frequent increase as users submit new entries.

tidyverse functions *read_csv* and *filter* were used to cross-check the dataset under investigation with the ClinGen gene list (Table 2-34).

Software	**Version**
RStudio	4.0.2
tidyverse	1.3.0

Table 2-34 Software used in ClinGen gene exclusion.

2.13.3 Evaluation of number of reports relevant to autism

A targeted search was carried out to determine the presence/absence of literature related to each gene under investigation and its relevance to autism. Two approaches were taken to carry out this search as follows.

2.13.3.1 GeneCards search for autism-relevant reports

GeneCards®: The Human Gene Database is a searchable, integrative database providing information on human genes (Stelzer *et al.*, 2016). The publications tool within GeneCards provides titles of and links to research articles in PubMed, as associated via Novoseek, HGNC, Entrez Gene, UniProtKB, GAD, HMDB, and/or DrugBank. Genes were searched against this database using the search terms “autism.” Number of reports correct as of 2020-09-30.

2.13.3.2 SFARI Human Gene Module search for autism-relevant reports

The SFARI Gene Human Gene module is a thoroughly annotated and well-maintained list of genes that have been studied in the context of autism (Abrahams *et al.*, 2013). This database compiled autism-relevant reports as follows:

> *“Reports – This section includes citations for the studies connecting the gene to ASD. The reports table includes the following columns of information about each report: the type of report (Primary, Positive Association, Negative Association, and Support), its title, the author and year of the publication, whether the report was ASD-specific, and any associated disorders mentioned in the report. These articles are not necessarily limited to the field of autism research. We also include links to the PubMed abstracts of the reference articles.”*

Genes were sequentially searched against this database for autism-relevant reports and autism-specific reports. Number of reports correct as of Q2 2020 release.

2.13.4 Gene selection for curation by Schaaf *et al.* (2020) modified ClinGen curation framework

Following evaluation of ClinGen curated genes and evaluation of number of reports relevant to autism, as detailed above and summarised in Figure 2-16, candidate genes were selected for gene curation though the proposed curation framework (Schaaf *et al.*, 2020).

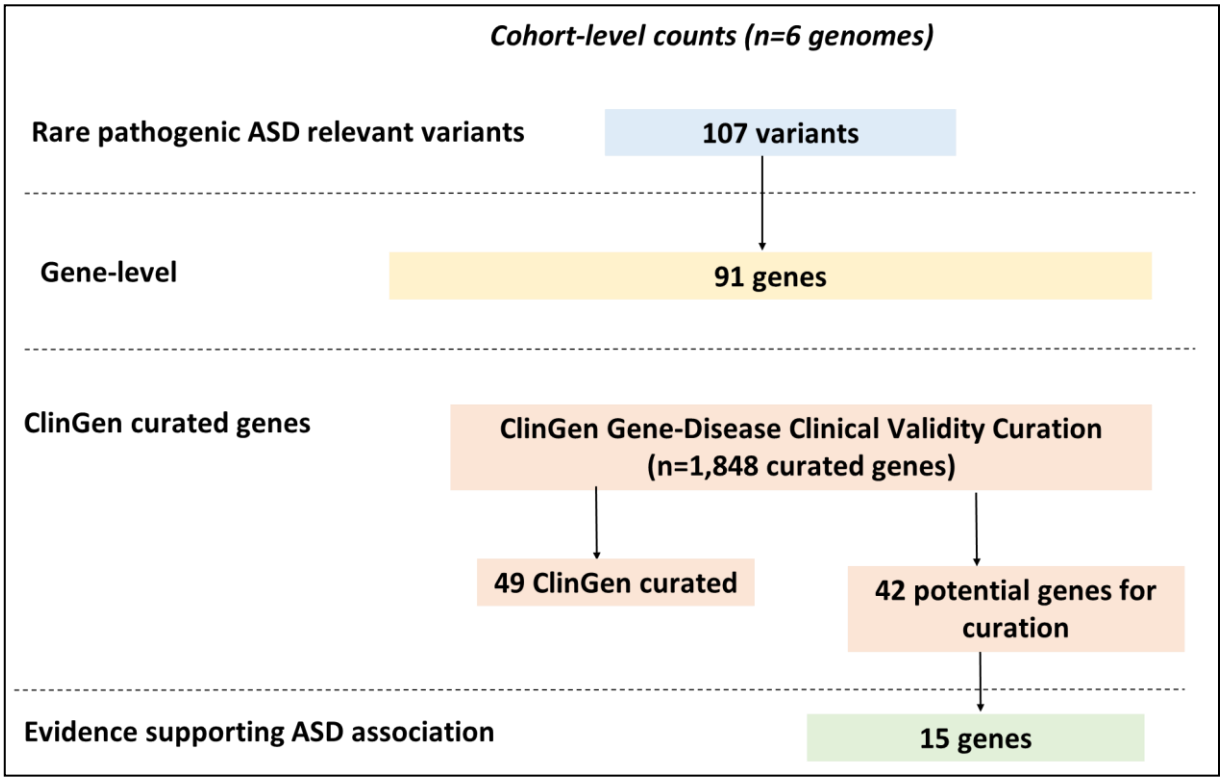


Figure 2-16 Gene selection for curation.

Genes arising from Cohort 2 outlined in Figure 4-2 were excluded from this analysis when already curated by ClinGen. Genes were selected for analysis when evidence of autism association is reported in the literature.

As demonstrated in Table 2-35, SFARI Gene was the effective method of report selection, referencing reports which were not identified by GeneCards. SFARI Gene reporting was further stratified based on reporting in an autism specific study versus a study non-specific to autism. Fifteen genes with evidence supporting autism association, and not previously curated through the ClinGen Gene-Disease validity process remain as candidates for interrogation in this study. Three genes had greater than five autism reports as evaluated by SFARI Human Gene Module and were prioritised for curation on the basis that a greater number of reports may yield a greater number of variants within each gene as candidates for curation. *NAV2*, *NINL* and *CACNA2D3* were selected for curation in this study.

Gene name ("genename" as assigned through dbNSFP 4.0a annotation)	**GeneCards search for autism-relevant reports**		**SFARI Human Gene Module search for autism-relevant reports**		
	Publications based on search on "autism" in GeneCards publication search	***Number of autism-relevant publications (GeneCards)***	***Presence in SFARI Human Gene Module 2.0 (Release Q2 2020)***	***Total number of relevant non-autism-specific reports mentioning gene***	***Number of autism-specific reports implicating the gene***
GRHL3	No		No		
AMPD1	Yes	2	Yes	3	3
NTRK1	Yes	2	Yes	8	1
ERCC6	No		No		
PAPSS2	No		No		
NAV2	No		Yes	9	6
SLC6A5	No		No		
LRP4	No		No		
C12orf57	No		Yes	11	1
CCDC65	No		No		
TRPV4	No		No		
EP400	No		Yes	6	5
FREM2	No		No		

TOGARAM1	No		No		
PYGL	No		No		
KIAA0586	No		No		
TRIP11	No		No		
MAP1A	Yes	1	Yes	3	3
CGNL1	No		Yes	4	4
DNAH9	No		No		
SCN4A	No		Yes	4	3
COMP	No		No		
ZC3H4	No		Yes	3	3
MED25	No		No		
STAMBP	No		No		
UNC80	No		Yes	5	2
COL4A4	No		No		
COL6A3	No		No		
SNX5	No		Yes	3	2
NINL	No		Yes	6	6
PLXNB1	No		Yes	3	3
CACNA2D3	No		Yes	7	6
TBCK	No		Yes	1	0
FAT4	No		No		

PLK4	No		No		
SKIV2L	No		No		
PKHD1	No		No		
FBXL4	No		No		
ADGRG6	No		No		
PLXNA4	Yes	1	Yes	4	3
HR	No		No		
CRB2	No		No		

Table 2-35 Evaluation of number of reports relevant to autism.

Genes included in this table are those arising from analysis on 6 individuals by WGS as detailed in Figure 4-2. This table outlines the number of reports obtained through two avenues of evaluation: GeneCard search and SFARI Human Gene Module search. Yes/No indicate the presence or absence of reports for each gene by the corresponding search method. Grey filled observations indicate N/A values where no reports are retrieved. Highlighted in bold are three genes for which the highest number of SFARI reports are available for interrogation.

2.13.4.1 Evaluation of gene constraint scores

The three genes selected for gene curation were evaluated for putative pathogenicity of variants occurring within these genes through gnomAD analysis estimating constraint (Karczewski *et al.*, 2020). Constraint is estimated by variant type, i.e., LoF and missense. Observed/expected (o/e) is a continuous measure of how tolerant a gene is to a certain class of variation. Low o/e values indicate the gene is under stronger selection for that class of variation than a gene with a higher value. 90% confidence interval (CI) is given for each o/e value. Z score given is the deviation of observed counts from the expected number. Positive Z scores indicate increased constraint. The closer pLI is to one, the more intolerant of protein-truncating variants the transcript is predicted to be.

2.14 Evaluating the inclusion of ACMG59 in autism and neurodevelopmental gene lists

This analysis was carried out using the software presented in Table 2-36.

Software	**Version**
RStudio	4.0.3
Tidyverse	1.3.0
ggvenn	0.1.8

Table 2-36 Software and versions used in evaluation of ACMG59 overlap.

ACMG59, from American College of Medical Genetics and Genomics (ACMG) (Version 2.0 06-04-2021 export), SFARI Gene (13-01-2021 release), DDD gene2phenotype (09-04-2021 export) and the autism gene panel list compiled in Identifying autism gene panels were queried for overlap. Nomenclature was aligned using HGNC Multi-Symbol Tool (Version: 2021-01-06 update) as already outlined. Overlap filtering was run by *tidyverse filter* function using operator %in% and inverse operator *!%in%* on HUGO aligned gene name.

This overlap of genes in the clinical gene-sets presented in is further quantified by jaccard similarity coefficient measured as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

```
# Define the jaccard similarity coefficient

 jaccard <- function(a, b) {
    intersection = length(intersect(a, b))
    union = length(a) + length(b) - intersection
    return (intersection/union)
}

# Compute the jaccard similarity coefficient

jaccard(a,b)
```

Chapter 3. An analysis strategy to isolate exonic rare pathogenic single nucleotide variants using next-generation sequence data.

Presentations arising from the contents of this chapter:

“Rare genetic variation in autism; an exome sequencing study.” Fiana Ní Ghrálaigh, Cathal Ormond, Elaine Kenny, Louise Gallagher & Lorna M. Lopez

Poster presented at the Irish Society for Human Genetics, September 2020 (Appendix III-V).

“Analysis Pipeline of Whole Genome Sequencing Data in Neurodevelopmental Disorders.” Fiana Ní Ghrálaigh, Niamh M. Ryan, Louise Gallagher, Lorna M. Lopez

Poster presented at the British Neuroscience Association Festival of Neuroscience, April 2019 (Appendix III-VI).

3.1 Abstract

Key hypothesis and key outcomes. The key aim of this chapter was to establish an analysis pipeline to isolate rare, putatively pathogenic SNVs with autism-relevance. This chapter details analysis of a WES cohort of 42 individuals, varied in family structure. This work has enabled the development of a variant interpretation strategy from align sequence reads to a filtered and relevant variant call-set. The key outcome of this chapter is an analysis pipeline, which was applied to the datasets analysed within this thesis. A further outcome of this chapter is a high-quality and robustly annotated set of rare putatively pathogenic variants with evidence for autism relevance. Further value will be gained from this variant set in the future upon combined analyses of these data with larger autism sequencing cohorts.

3.2 Introduction

3.2.1 Next-generation sequencing

NGS enables variant detection across many classes and sizes of variation, across the allele frequency spectrum. Management of NGS output requires application of tools and algorithms to manipulate the large-scale datasets generated. These tools and algorithms are used in combination in bioinformatic pipelines to translate raw data into interpretable variant callsets for downstream biological interpretation.

3.2.2 An introduction to GATK and the gVCF file format

NGS technologies output large raw read files (FASTQ files). These files require substantial computational power to process from raw sequencer generated reads (FASTQ files), through aligned reads that have been mapped to a reference genome (BAM files), to a readily interpretable called variant file (VCF files). VCF files can then be manipulated and interrogated for biological relevance. There is currently no gold standard in genome analysis pipeline, however GATK is a widely applied collection of bioinformatic tools and is robustly maintained and supported by the Broad Institute (Figure 3-1). The GATK Best Practices provide guidelines for effective use of the tool set, enabling manipulation of parameters to suit the data set under investigation, for example specification of target intervals or specification of reference genome. The GATK Best Practices Workflow also incorporates use of widely applied tools *Picard* and *BWA* (Li, 2013; Broad Institute, 2019). This multi-step analysis strategy guides bioinformatic analysis from raw sequence read to a variant call set for biological interpretation.

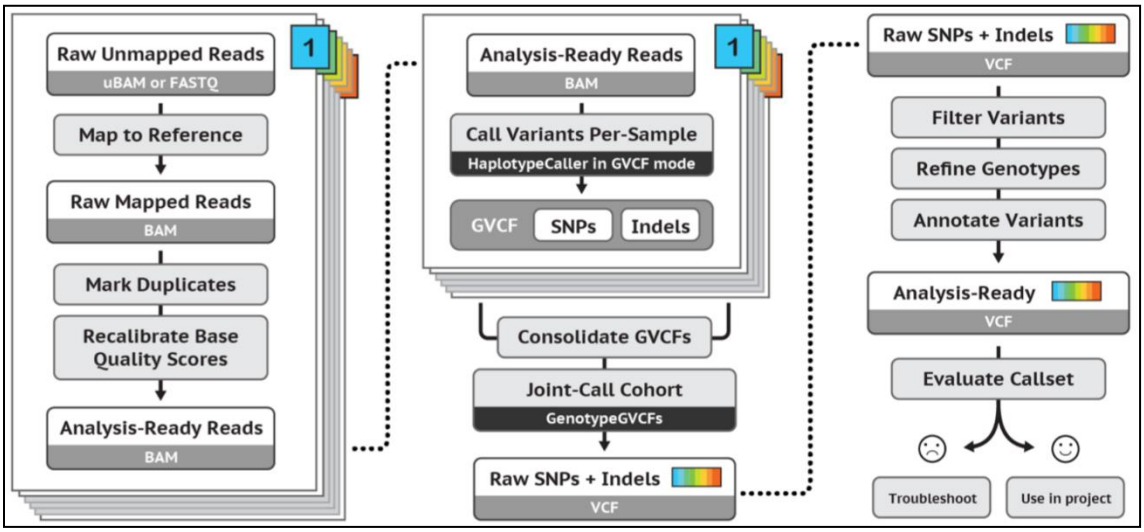


Figure 3-1 GATK best practices.

Workflow recommended from GATK. Taken from

https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Jointly genotyping is an approach to genotyping that determines genotypes at each locus with consideration of the other genotypes in the call-set. This approach is taken in GATK Best Practices. Joint genotyping requires the use of HaplotypeCaller to generate input genotype assignments. The gVCF file format details all variant sites in the genome whether reference (ref) or alternative (alt), as opposed to the traditional VCF file listing alternative variant sites only. Joint genotyping is a more time and computationally intensive approach to genotyping, however, it improves the detection of rare variants in the genome making it beneficial for use in a family-based study design where accurate and sensitive rare variant discovery is required.

3.2.3 The reference genome

Following the initial publication of the mapped human genome by the Human Genome Project in 2003, iterations of the reference human genome continue to emerge with improving quality enabled by advances in sequencing technologies. GrCh38, curated by the Genome Reference Consortium in 2013, is the most recent reference genome release with gaps and errors in the original Human Genome Project reference genome corrected by shotgun sequencing (Schneider *et al.*, 2017). hg19, also referred to as GrCh37, was released in 2009 by the Genome Reference Consortium and while it has been updated to GrCh38 it remains widely applied in genome sequencing studies.

GrCh38 and hg19 reference genomes are currently the most common genome builds to which human sequence data are aligned. Depending on the availability of existing datasets and analysis requirements of a study there may be reasons to use a previous genome build. These reasons include enabling use of legacy analysis pipelines developed on a previous genome build. Another reason is the need for consistency throughout dataset analysis. To achieve a cohesive dataset which can be cross analysed it is necessary to use the same reference genome build to avoid genomic coordinate discrepancies. Finally, annotation databases have lagged in their reannotation to GRCh38 coordinates, often making hg19 a more efficient reference genome for alignment.

3.2.4 Cohort-level QC measures

Sample identity is a major concern in the reproducibility and reliability of genomic datasets. Even with strict protocols for QC and the maintenance of a high standard of data control and data processing, errors are observed across datasets and cohorts. Specifically sample identity is key to robust linkage of genetic and corresponding phenotypic data. Large-scale cohort analyses frequently observe misidentification of samples. This can result from many possible errors such as pipetting errors at sample collection, DNA extraction and preparation or errors in library preparation, such as mislabelling of sequencer indices. Mismatches in genetic and phenotypic information may also be the result of errors at the time of phenotypic data collection.

There are several features of genomic datasets that may be used as QC measures to identify such errors. Most commonly, genotype imputed sex can be cross-referenced with sex reported at phenotyping. In family-studies, relatedness may be used as a means of QC by determining the degree of genomic sharing with that expected on the basis of reported familial relationship. In the case of relatedness assignment, it is important to consider possible true errors in relatedness that are not related to sample mix-up such as adoptive families or families for which parental relationships are not as expected. Deviations from expectation in these factors can be used to flag samples as discordant for reported phenotype, resulting in a need for further investigation and potential removal from the study.

3.2.5 Allele frequency annotation

The rate at which a particular variant occurs in the population is a key annotation to be made. Depending on the genetic architecture of the disease/disorder and the study approach used, variants will be isolated based on their absence or low frequency in a sample from the general population. Typically, common variants are defined as genetic variants with a minor allele frequency of greater than 0.05, meaning occurrence in over 5% of the population. Rare variants are defined as genetic variants of low frequency with a minor allele frequency of less than 0.05, meaning occurrence in less than 5% of the population. Very rare variants are classed as those occurring in less than 1% of the population. Rare variants are of key importance in many complex conditions, including autism, however by nature they are challenging to identify, and even more challenging to confidently associate with a phenotype. For example, a study of rare neurodevelopmental-associated variants, in this case CNVs, required a minimum of five cases of a rare variant within the cohort to enable robust stratification of their trait of interest, cognition (Kendall *et al.*, 2019). At a population frequency of less than 5%, rare variant identification requires large sample sizes to achieve variant identification. However, a family-based study design enriches for within family rare variants often giving power to association analysis (Glahn *et al.*, 2019).

Rarity is determined by the frequency of the less common allele, or minor allele (MAF) in the population. Population-based cohorts are used to estimate the MAF of a given variant within an unaffected population. These include 1000 Genomes (2,504 low coverage and exome sequence data) (The 1000 Genomes Project Consortium, 2015), Exome Aggregation Consortium (ExAC) (60,706 exomes) (Lek *et al.*, 2016), and most recently emerging as leader in the field, gnomAD (v3: 71,702 whole genomes mapped to GRCh38 reference genome) (Karczewski *et al.*, 2020). The estimation of ancestry of the individual from genotype is critical to determine the expected rarity of a particular variant. Population stratification is key to establishing the MAF of a given variant in the relevant population substructure.

3.2.6 Algorithm based approaches to measure predicted pathogenicity

Accurate pathogenicity prediction is essential to variant interpretation and putting variants in the context of their biological impact, however interpretation remains an enormous challenge. Bioinformatic tools and scoring algorithms have been developed with an aim to provide these variant annotations and prioritise variants that are impacting on human phenotypes.

The degree to which a variant may impact the carrier can be estimated by the change in genomic sequence that it causes. At a broad level, a variant may have a synonymous or non-synonymous change to the protein sequence. Synonymous refers to a change in genetic code that does not result in a change to the amino acid encoded. Non-synonymous refers to variant sites at which a variant is encoding a different amino acid to wild-type, potentially impacting protein function. Within the non-synonymous effect of variation, variant changes can be further categorised by the way in which protein sequence is disrupted. Missense variants cause a change in sequence that results in production of a different amino acid to wild type. Loss-of-function (LoF) variants are predicted to cause a complete disruption to the protein-coding gene in which it is found (Figure 3-2).

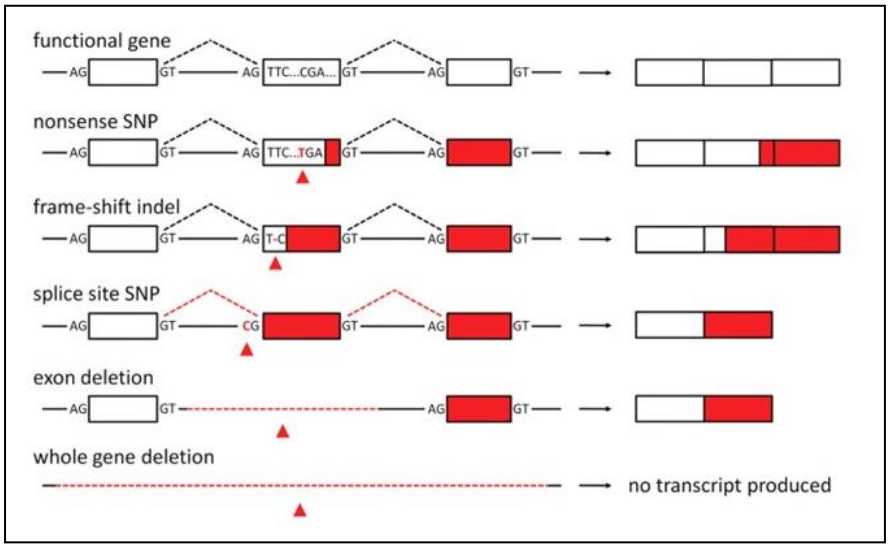


Figure 3-2 Classes of LoF variation affecting protein-coding regions.

Taken from MacArthur & Tyler-Smith (2010): "A model three-exon gene is shown both intact (top) and following the introduction of various types of LOF variant (red triangles). Effects on the transcript produced by the gene are shown at the right. LOF variants typically result in a loss of protein-coding functionality downstream of the variant (red boxes)."

Missense variants are far more abundant in the genome than LoF variants, with the majority having no harmful effect on gene function (Ronemus *et al.*, 2014). For this reason, it is common to eliminate missense variants, which are not predicted to be pathogenic by scoring algorithms, based on functionality and conservation, from analysis of putative autism variants.

A variety of bioinformatic statistical tools are available to determine the expected pathogenicity of a given variant (Ng and Henikoff, 2003; Adzhubei, Jordan and Sunyaev,

2013; Kircher *et al.*, 2014; Jagadeesh *et al.*, 2016). Each score is based on an individual algorithm, many considering a combination of variant genomic coordinates, amino acid consequence, base-pair change and conservation of the gene in which the variant is occurring.

Individual pathogenicity scoring approaches are typically used in a combined consensus approach, for example pathogenic by SIFT and PolyPhen-2 scoring. SIFT (Ng and Henikoff, 2003) and PolyPhen-2 (Adzhubei, Jordan and Sunyaev, 2013) both predict the effect of amino acid substitution on protein function. However, use of these algorithms comes with certain limitations. Firstly, SIFT and PolyPhen-2 are highly dependent on the protein sequence database that is used to retrieve homologous sequences. Secondly, there is incomplete coverage of the genome by these scores. By excluding variants reaching standards for pathogenic by these algorithms, there is potential to exclude true pathogenic variants solely on the basis that they are missing the appropriate scoring by individual algorithms.

With the high volume of potential variant annotation tools available, algorithms-based approaches have been created to incorporate multiple scoring measures and output a singular determination of pathogenicity. Most popular in the field, is CADD, Combined Annotation-Dependent Depletion (Rentzsch *et al.*, 2019).

In practice, CADD outputs variant specific raw scores and PHRED-scales scores, which are normalised to all potential SNVs in the genome (~9 billion). A PHRED-scaled CADD score of >10 indicates a raw CADD score occurring in the top 10% of reference SNVs, and a score of >20 indicates a raw CADD score in the top 1%. A variant cannot be deemed as pathogenic based on a blanket CADD score cut-off, similarly a variant cannot be deemed benign by having a CADD score below a cut-off value. Rather consideration is needed of phenotype severity, mode of inheritance and also resources for genomic interpretation of the output variant list (Rentzsch *et al.*, 2019). Instead, the top-ranked variants in a dataset should be further investigated in a way that is particular to the study design.

3.2.7 Database annotation – retrieval of known information from databases

Database search tools are available to retrieve existing variant information from a set of input coordinates or variant IDs. The advantages of using such annotation databases comes from the ability to carry out multiple levels of annotation in a fast and effective

way. However, the utility of these databases hinges of the maintenance of the tool and the frequency at which it is updated, specifically in line with release of new versions of individual annotation sources, e.g., CADD (Rentzsch *et al.*, 2019), Clinvar (Landrum *et al.*, 2018).

Some commonly used annotation databases include dbNSFP, Variant Effect Predictor (VEP) and Annovar. dbNSFP compiles annotations from 29 prediction algorithms, nine conservation scores, allele frequencies from major population databases, including 1000 Genomes and gnomAD, as well as gene-based annotations of expression and interactions (Liu *et al.*, 2016). These annotations are applied to an input call-set in VCF format using the dbNSFP java database search tool, as has been applied in this chapter.

3.2.8 Exonic variation in focus

Genetic variants associated with autism disrupt a wide variety of biological pathways and processes (De Rubeis *et al.*, 2014). Huge efforts have been made to understand these pathways and how they are disrupted in autism. Identifying pathways and processes showing an increased mutational burden in autism advances our understanding of autism aetiology. The role of non-coding variation in autism has been established, as has been introduced in 1.7.3 (Turner *et al.*, 2016; Brandler *et al.*, 2018). The interpretation of non-coding variation is challenging but despite this robust gene-phenotype associations have been made in neurodevelopmental conditions (Wright *et al.*, 2021). As cohort sample sizes increase and power to detect rare non-coding variation increases these variant classes are likely to uncover much of the rare genetic contribution to autism.

WES enables detection of rare variation within the protein-coding genome. Satterstrom *et al.* demonstrate the ability of whole-exome sequencing to identify rare autism-relevant variants when sample sizes are large in their association of 102 genes with autism (Satterstrom *et al.*, 2020). Compilation of gene-lists containing genes involved in a given process, are invaluable in establishing the process which a putative variant may be disrupting, and such gene lists are often consulted for membership when investigating the impact of a variant (Feliciano *et al.*, 2019).

Sequencing studies of autism cohorts can compile variants in affected individuals into gene lists against which rare variants may be searched, such as pathway based gene lists (Yuen *et al.*, 2015). Further to these autism-associated processes, genes associated with schizophrenia and ID may be informative to consider in analyses (Iossifov *et al.*, 2014). This due to the shared global gene expression pathways identified among some psychiatric conditions (Gandal *et al.*, 2018).

While these pathways and processes are frequently implicated in autism, these lists of gene do not constitute a clinically relevant gene-list. Consequently, these gene lists should not be considered a finite and exclusive list of genes to be used in genetic testing in autism. The establishment and maintenance of databases in which gene-level information is openly shared are crucial to progress the field. The largest curated gene list to date in collating a gene-list for autism is most the SFARI Gene database (Abrahams *et al.*, 2013). This regularly maintained resource presents evidence supporting the role of >1,000 genes in autism (currently n=1,045 genes as of February 2022 update), with use of a gene-phenotype association scoring system to represent the confidence of a given gene in autism. This gene scoring approach collates all available evidence supporting the relevance of the gene to autism risk and categorises each gene depending on the strength of evidence as gene scores of 1 (high confidence), 2 (strong confidence), 3 (suggestive evidence) and/or S (syndromic).

While this is an invaluable tool for use in research, the lack of a systematic evidenced based framework means this has limited applicability in clinical settings, such as diagnostic testing. A similarly scored gene list of neurodevelopmental condition-relevant genes is DDD gene2phenotype gene list generated from the DDD study. This gene list is collated from variants identified in a cohort of children with severe and complex neurodevelopmental phenotypes (Wright *et al.*, 2015). DDD assigns genes with a level of certainty of association, given as “Definitive,” “Strong,” or “Limited.”

3.2.9 Hypothesis and aims

No gold-standard pipeline currently exists for the isolation of rare exonic SNVs from NGS datasets. The research outlined within this chapter has yielded a strategy for the isolation of such variants in autism cohorts as is applied in Chapters 4 and 5. This pipeline has been informed by literature in the field and makes use of available databases to leverage existing information, including SFARI gene and DDD (Abrahams *et al.*, 2013; Wright *et al.*, 2015). Rare SNVs in autism-relevant genes are detectable by WES. This dataset is limited in its power to provide statistically significant rare variant autism associations. However, the data can be leveraged to build an analysis strategy for use in the identification of rare SNVs in autism. Putatively pathogenic autism-relevant SNVs may be identified through these analyses building evidence toward existing gene-phenotype association.

The aims of this chapter are:

1) to establish an analysis strategy for isolation of rare exonic pathogenic SNVs from NGS data.

2) to discover rare putatively pathogenic autism-relevant SNVs in a cohort of autistic individuals and their unaffected family members.

3.3 Results

3.3.1 Cohort in summary

This chapter describes analysis of WES of Cohort 1. Cohort 1 is a dataset of a total of 42 individuals. Ascertainment of this cohort is described in 2.1.1. The cohort is comprised of 23 putatively simplex cases of autism and their unaffected family members whose genotypes have been used, where available, to restrict to putatively pathogenic variation, as will be described in 3.3.5.

3.3.2 Low-confidence variant filtering

The initial variant call-set of a total of 127,842 variant sites comprised of SNVs, Indels and other variants as detailed in Table 3-1. Table 3-1 shows successful variant calling as demonstrated by the breakdown of variants by variant type. As expected, the largest proportion of variants called are SNVs. Mixed variants in these analyses account for single variant positions at which both SNVs and indels are occurring in the cohort. GATK Haplotype Caller is unable to call SVs or CNVs.

Variant Type	**Cohort Count**
SNV	117,101
Indel	10,310
Mixed	431

Table 3-1 Variant count by variant type.

Exact variant counts per variant type. SNV refers to single nucleotide variants; Indel refers to insertion deletion variants; Mixed refers to variant loci at which both SNVs and indels have been identified.

Towards filtering to high-confidence variants, the VQSR machine learning filtering approach was applied through GATK, as specified in 2.6.1.3. VQSR flagged a total of 17,283 SNV and indel variant sites likely to be sequencing artefacts. VQSR is not compatible with non-SNV and indel variant sites. These variants were instead hard-filtered using a more crude approach of hard-filter removal of variant sites on the basis of quality score normalised by read depth (Quality by Depth < 2.0; generic filtering recommendation), probability of strand bias at the site (Fisher Strand > 200.0; little to no strand bias at the site will be indicated by values close to 0) and position of reference versus allele positions within reads (Read Position Rank Sum Test < -20.0; a negative score indicates the alternative allele is found at the ends of reads more often than the

reference and a score close to zero indicates little difference between positions in the reads). A total of 110,559 variants were retained in the call-set.

Following VQSR of SNVs and Indels and hard-filtering based on sequencing metrics, variants were hard-filtered to isolate and remove those variants deviating from Hardy-Weinburg equilibrium (exact test $<10^{-6}$) and variant sites missing greater than 10% of data. A total of 106,590 (of n=110,559) high-confidence variants were retained in the call-set for downstream annotation. Variants remaining following these QC filters were counted and plotted by variant type. The variant calling and joint genotyping pipeline is effective across the exome, as shown by the variant counts by chromosome (Figure 3-3). The inconsistency in number of variants per chromosome is expected given the difference in chromosomal length and the variation in the number of probes targeting each chromosome.

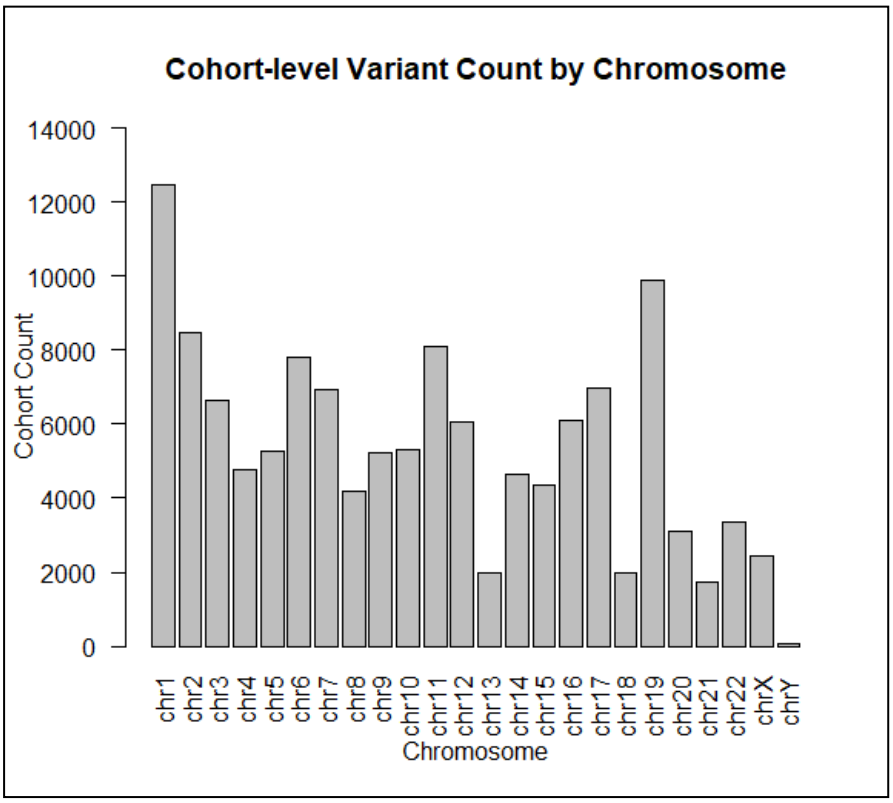


Figure 3-3 Cohort-level variant count by chromosome.

The bar chart shows counts of the overall variants (SNV, indels and mixed variant) called from the cohort and presents them across the genome per chromosome. Note that mitochondrial chromosomes variants were not included in these analyses. The inconsistency in number of variants per chromosome is expected given the difference in chromosomal length and the variation in the number of probes targeting each chromosome in the whole exome sequencing panel.

3.3.3 Variant annotation

Annotation by dbNSFP at the cohort-level for a total of 34 whole exomes passing QC. dbNSFP annotates all non-synonymous variation according to the specified parameters. In this case dbNSFP has annotated with optional parameters *-p* (output existing VCF columns), *-v hg19* (specifying reference genome), and *-g* (include full gene annotation set).

A total of 36,872 SNVs were found in annotation as non-synonymous variants. A total of 69,718 SNVs were not found and are excluded, and further analysis of this variant class falls outside the aim of this project. Further analysis of these excluded variant classes will be necessary to characterise the full variant set per individual. dbNSFP annotated each of the 37,148 counts of the non-synonymous SNVs identified with 511 variables.

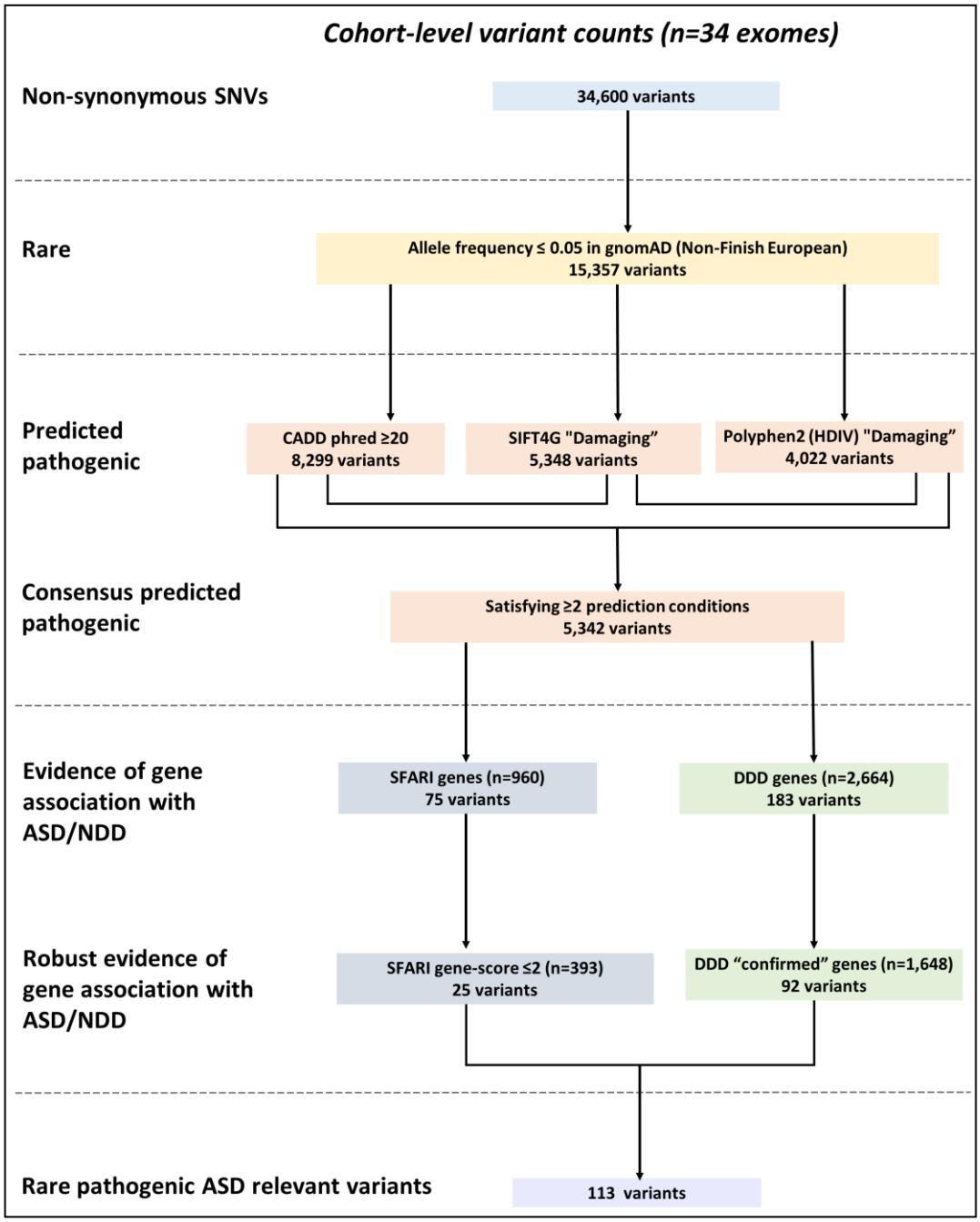


Figure 3-4 Flow of variant filtering with cohort-level variant counts.

Arrows show the direction of flow from each level of filtering (specified on the left). Rare variants are determined on the basis of their allele frequency reported in gnomAD, with an allele frequency of less than or equal to 0.05 indicating rarity. Predicted pathogenicity is determined by consensus scoring as pathogenic by CADD (Phred greater than or equal to 20), SIFT4G ("Damaging") and Polyphen-2 (HDIV "Damaging"). Gene-level autism associations are determined by gene membership in SFARI or DDD databases. SFARI refers to Simons Foundation Autism Research Initiative Gene Module (Abrahams *et al.*, 2013)*. DDD refers to the gene2phenotype database arising from the DDD study* (Wright *et al.*, 2015)*.*

3.3.4 Variant filtration

The workflow developed and applied in these analyses is presented in Figure 3-4. The full set of non-synonymous variants which has been annotated as described in 3.3.3 are subject to further subsetting on the bases of allele frequency, predicted pathogenicity and gene-level associations as follows.

3.3.4.1 Allele frequency filtration

The estimation of ancestry of the individual from genotype is critical to determine the expected rarity of a particular variant. Population stratification is key to establishing the MAF of a given variant in the relevant population substructure. PCA of the cohort advised of a European ancestry in all individuals within the cohort. To align with the parameter gnomAD_exomes_NFE_AF, was selected as the relevant allele frequency for filtration of this cohort. This quantifies the frequency which the alternative allele was observed in the non-Finnish European gnomAD exome cohort of 56,885 individuals. Variants were restricted to those observed in less than 5% of this population (<= 0.05). A total of 17,667 non-synonymous rare variants remains for further analysis. Common genetic variation, while putatively pathogenic in this complex condition, is beyond the scope of this analysis strategy.

3.3.4.2 Pathogenicity filtration

Due to the inconsistencies and differences in approach of pathogenicity prediction algorithms described earlier, a consensus scoring approach was taken to determine variant pathogenicity. Variants determined to be predicted pathogenic were required to be predicted pathogenic by greater than or equal to 2 of the specified condition as detailed below.

As previously described, CADD is an algorithm-based estimate of variant pathogenicity. The CADD phred-like score is phred-like rank score based on whole genome raw CADD scores. The larger the CADD-phred score the more likely the SNV annotated has damaging effect. The CADD-phred filter was set at variants with a score of >= 20. A total of 9,624 variants satisfy this criterion.

SIFT and Polyphen-2 were taken as candidate pathogenicity measures due to their widespread use in human genomics. Criteria for pathogenic as determined by SIFT was taken as “D,” indicating damaging, measured by SIFT4G_pred (SIFT 4G < 0.05). A total

of 6,182 variant were scored pathogenic. Polyphen2_HDIV_pred with a score of “D”, indicating damaging (HDIV score in [0.957,1]), isolated 4,638 pathogenic variants.

REVEL score filtering was applied to the annotated variant set at two levels >= 0.75 and >= 0.5. This strategy identified just 382 and 1,388 variants respectively, that satisfy this criterion. REVEL was excluded as a candidate pathogenicity filter score as it so greatly contradicted the other pathogenicity scores investigated. A total of 6,167 variants were designated as consensus predicted pathogenic (Figure 3-4).

3.3.4.3 Gene-set filtration

Rare predicted pathogenic variants were further filtered for those that occur in genes that have been associated with autism and neurodevelopmental conditions. The gene-sets used towards this goal were the SFARI Gene database (SFARI-Gene_genes_08-07-2020release_09-07-2020export) and the DDD gene2phenotype database (DDG2P_8_9_2020).

The annotated and filtered variant set were filtered for those associated with autism through SFARI by ENSEMBL gene IDs to avoid complications of filtering on gene names in the case that there are multiple gene names used. A total of 87 variants remaining when restricting variants to the 960 genes included in the database. Variants were further subset to higher confidence SFARI Gene variants, those with a gene-disease evidence score of one of two. This resulted in isolation of 29 variants. In parallel to this filtration the annotated and filtered variant set were restricted to those occurring only in DDD gene2phenotype genes, as identified by Online Mendelian Inheritance in Man (OMIM) ID for consistency across annotations. A total of 212 variants were isolated. This was further restricted to 111 variants when only those genes with a confirmed association with developmental conditions were included. Both variant sets arising from gene-level filtration were taken for downstream analysis, as outlined in Figure 3-4. Just 5 variants overlapped in these parallel gene lists filters, i.e., five variants isolated have evidence for phenotype association as determined by both SFARI and DDD. The variants isolated through this filtration strategy are shown graphically in Figure 3-5.

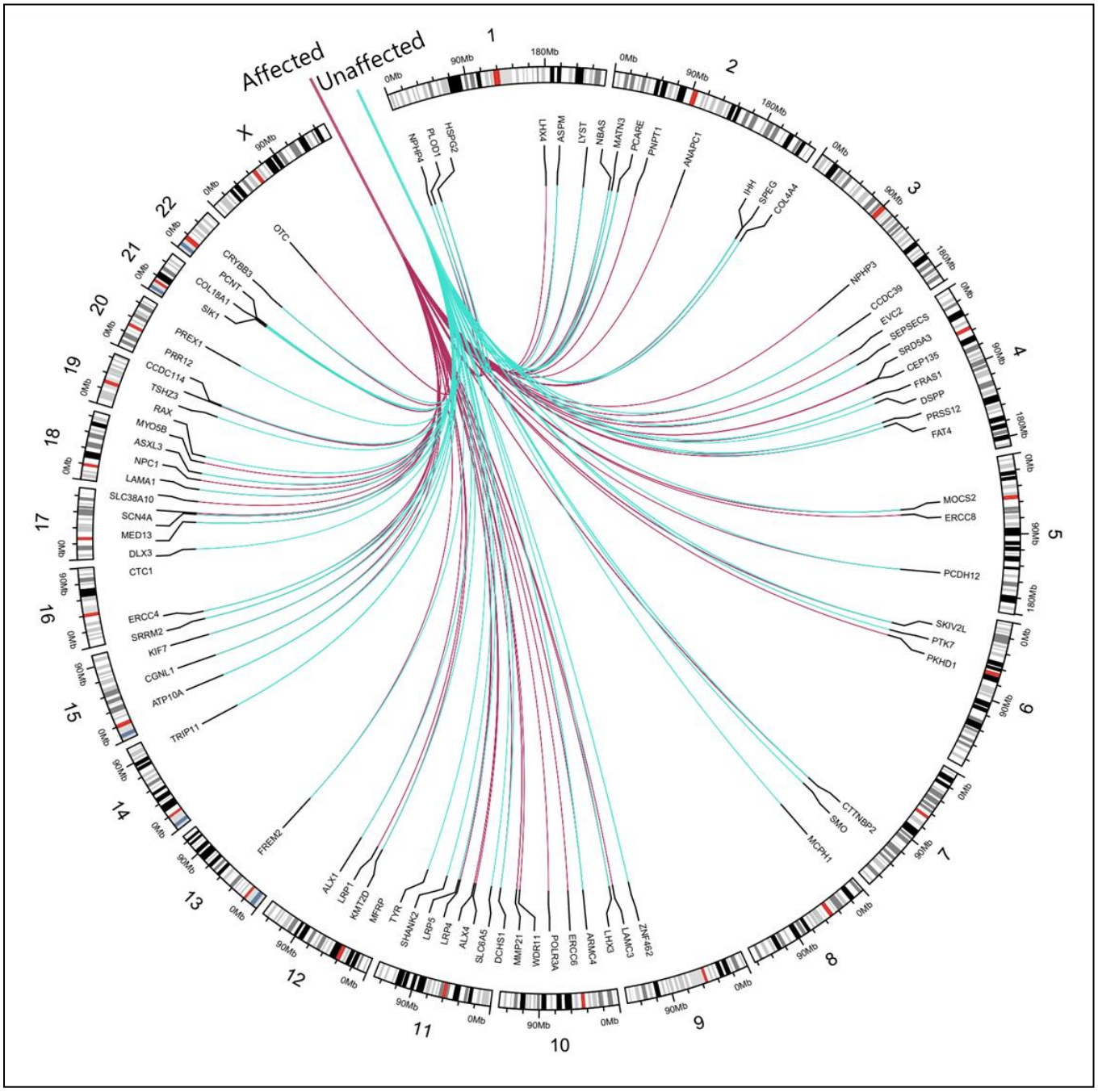


Figure 3-5 Spread of variation across genomic regions.

Chromosomes are shown around the outer track of the figure (1:22, X). The gene names are given on the inner track. These are the genes in which the rare pathogenic autism-relevant variants outlined in Figure 3-4 are located. Links are made in purple (affected n=103 variants) and blue (unaffected n=86 variants) between each gene and the respective affection status of the individual harbouring the variant. Affected denotes individuals with an autism diagnosis.

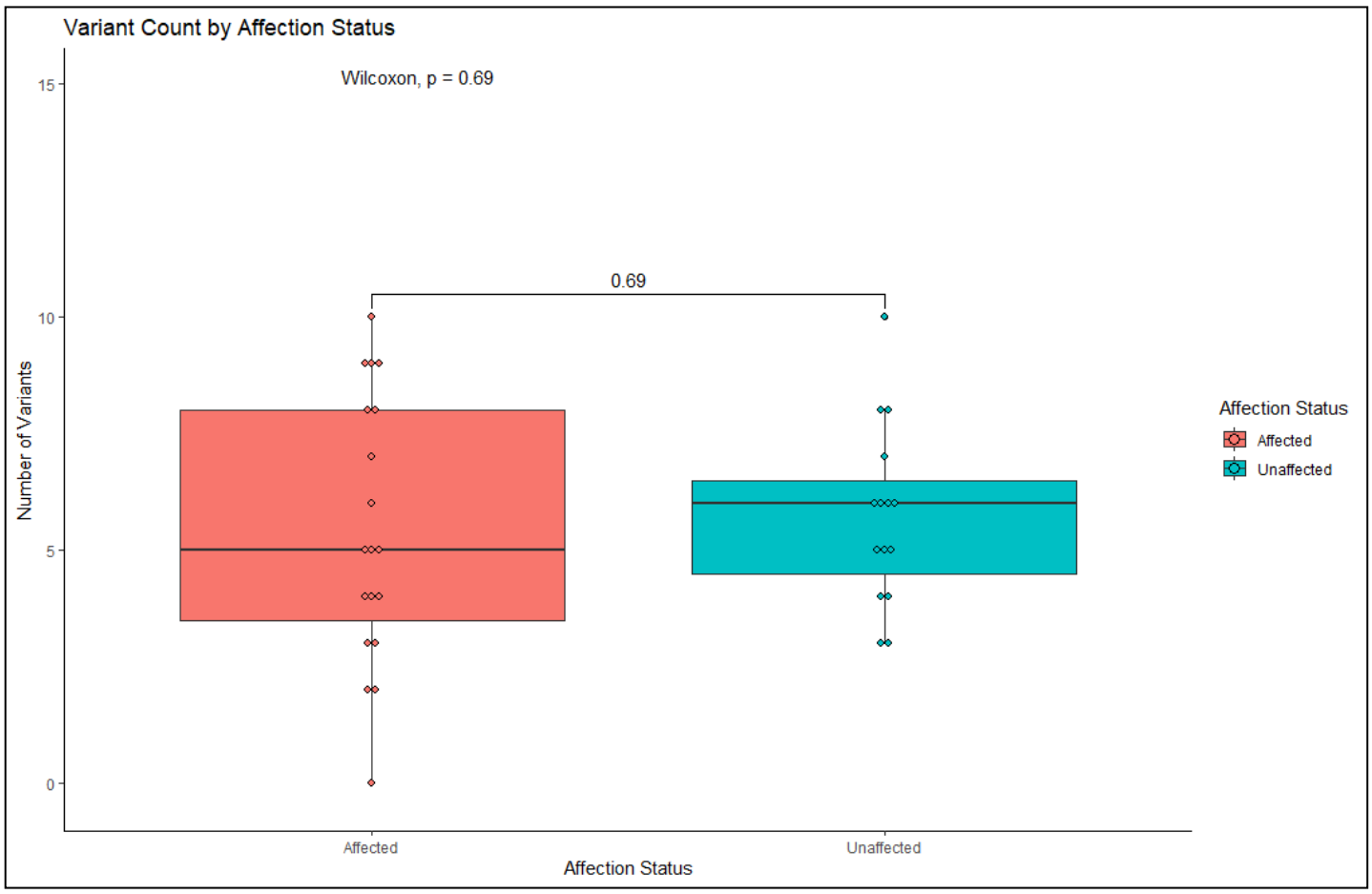


Figure 3-6 Per individual counts of variants under investigation.

Box plots are representative of affection status as detailed in the figure legend. Points on boxplot indicate each individual represented, against the number of rare putatively pathogenic variants occurring in the autism-relevant gene set investigated. Wilcoxon statistic was computed between groups with no significant difference in number of variants identified between groups.

The rare putatively pathogenic variants were subset to those occurring in affected individual and those occurring in unaffected family members. The number of variants is presented per individual in Figure 3-6. There is no statistically significant difference in the number of variants per individual dependant on affection status (Wilcoxon, p=0.69).

3.3.5 Trios in focus

This cohort comprises of singletons and trios. Here probands were selected for further investigation where both parent samples were also included in the cohort. Five probands and their unaffected family members, 16 individuals total, were subset from the cohort. Family structure within this cohort was used to subset the identified putatively pathogenic variants to those with added confidence in pathogenicity on the basis of their absence from unaffected family members. Variants in probands that are absent in unaffected family members were isolated from the variant subset of rare putatively pathogenic variants in autism-relevant genes where present and results are presented in Table 3-2.

Proband ID	**Homozygous *de novo* variant subset**	**Heterozygous *de novo* variant subset**	**Gene Name**	**Decipher Reported Phenotype**
AS157C1	0	0	N/A	N/A
AS217C1	1	0	*MATN3*	Multiple Epiphyseal Dysplasia Type 5
AS218C1	0	0	N/A	N/A
AS306	0	0	N/A	N/A
AS420C1	0	0	N/A	N/A

Table 3-2 Trios in focus.

Presented are the number of de novo variants isolated from 5 probands within the cohort, where both parent samples were available. Where applicable further detail is given on variants of interest identified. Phenotype is given as that specified in Decipher GrCh38.

3.3.6 Biological interpretation of variation impacting *MATN3*

Analysis of trio data within Cohort 1 yielded a single *de novo* variant of interest occurring in proband AS217C1. This variant, rs77245812 is a single base G>A change impacting position chr2:20202930, within the protein coding gene *MATN3*. The protein encoded by this gene forms a major component of the extracellular matrix of cartilage and has been reported to be involved in the formation of filamentous networks in the extracellular matrices of a variety of human tissues (Chapman *et al.*, 2001). Variation within *MATN3* has been associated with Multiple Epiphyseal Dysplasia 5, Osteoarthritis Susceptibility 2, and Spondyloepimetaphyseal Dysplasia, Borochowitz-Cormier-Daire Type, as reported by OMIM:602109 (Amberger *et al.*, 2015). Null mouse models for *MATN3* have

been generated by Ko *et al.* (2004) reporting no obvious skeletal malformations in homozygous mutant mice and the authors suggest redundancy in the matrilin family of proteins (Ko *et al.*, 2004). The single base change identified in within Cohort 1 results in an amino acid change of p.Thr303Met as estimated by HGVSp_SNPeff through dbNSFP. Following evaluation of the biological implication of variation within this gene, the reported autism phenotype of this proband cannot be accounted for by rs77245812. This gene is not included in SFARI gene, lacking autism-specific associations (Abrahams *et al.*, 2013).

3.4 Discussion

This chapter describes preliminary analyses of this cohort of WES data of 42 individuals. Three outputs are generated from the research outlined here. Firstly, a variant set of rare predicted pathogenic variants (n=135) occurring in genes with evidence of relevance to autism and neurodevelopmental conditions (Figure 3-5). Secondly, an analysis strategy has been developed for further application in discovery of rare coding autism-relevant variation. Finally, this analysis has resulted in the generation of robustly and uniformly processed and quality-controlled variant-call sets which will be combined with large whole-exome sequencing studies of autism, both in-house and internationally, bringing research value to these studies.

3.4.1 Variant-level QC

Genomic data processing was carried out as recommended by the Genome Analysis Tool-Kit (Van der Auwera *et al.*, 2013) (Figure 3-1) with parameters adapted from GitHub commit: cathaloruaidh/**WGSVariantFiltering.**

Reads failing the duplicate read filter were removed from analysis and Base Quality Score Recalibration (BQSR) was applied. Base quality scores are per-base estimates of error arising from sequencing machines. The scores represent the confidence that an individual base has been correctly called. Accurate variant calling is dependent on accuracy of base quality scoring. However, initial base quality scores are subject to technical errors and biases arising from sequencing. BQSR overcomes these errors by applying a machine-learning algorithm to model the inaccuracies and adjust the base quality scores to represent true base quality scores more accurately.

Following BAM processing variant calling was carried out using the GATK Haplotype Caller, calling SNVs and indels. Genotyping was carried out at cohort-level by jointly genotyping samples. The advantages to jointly genotyping, as opposed to genotyping at the sample-by-sample level are:

a) Improved sensitivity in the detection of rare variants
b) Improved distinction of homozygous reference variant site and missing variant sites
c) Greater ability to filter out false positives.

Rather than applying independent hard filters to the variant call-set, VQSR is applied with subsequent filtration on VQSLOD, i.e., the log of the odds ratio of the variant being true versus false under the model. This filtration method is advantageous as it considers

each variant annotation in combination to minimise the number of confident variants lost. Variants failing the Hardy-Weinburg Equilibrium filter were excluded from analysis on the basis that departure from Hardy-Weinburg equilibrium may indicate inaccurate genotyping. Variant sites with a high proportion of missing data are removed from analysis also. The analyses described in this chapter interpret the SNVs detected through this variant detection strategy only. Future analysis incorporating the mixed variant sites and indels summarised in Table 3-1, with the SNVs described will improve these analyses considering a more complete view of the exonic variants detected.

3.4.2 Cohort-level QC

Sample identity is critical in building and maintaining datasets, particularly clinical datasets due to the potential for results to be returned to individuals and inform medical and lifestyle decision making. QC measures such as those outlined in these analyses aim to highlight sample mix-up and discrepancies in biologically imputed characteristics against those reported.

Early identification of errors is key to avoidance of significant time and resource losses associated with processing of invalid data and the necessary re-analysis of datasets following eventual removal of invalid data points. In this chapter, two methods of cohort-level QC checks have been carried out. Both, *plink* and *peddy* approaches were successful in flagging samples for removal (Purcell *et al.*, 2007; Pedersen and Quinlan, 2017). *peddy* computes relatedness as IBS, identity by state, rather than more traditional IBD, identity by descent measures. In a family-based study, this statistic has the added benefit of differentiating between sibling-sibling and parent-child relationship, both having a coefficient of relatedness of 0.5 (Figure 2-4, Figure 2-5).

An important consideration when evaluating discordance in phenotypic and genomic data is the possibility of errors in reported affection status or other phenotyping measures. Schaaf *et al.* outline the importance of robust clinical phenotyping in studies of human disease (Schaaf *et al.*, 2020). Incorrect assignment of affection can have a detrimental impact on the integrity of the dataset. Inclusion of a neurodevelopmental-affected individual, who have been incorrectly assigned as unaffected or a control, may harbour a pathogenic neurodevelopmental-relevant variant which may then be considered as non-pathogenic due to the designation of the individual as unaffected. This may have further impacts on the cohort in the situation where the variant under consideration is shared with another affected individual within the cohort, however

presence of the variant in an unaffected individual weakens the evidence supporting the role of the variant in the neurodevelopmental phenotype. For this reason, clinical diagnosis of affected individuals has been validated by Prof. Louise Gallagher (Child & Adolescent Psychiatrist) prior to inclusion in the study and assignment of affection status (2.1.1).

3.4.3 Selection of rare variant parameters

These analyses focus of the discovery of variants that are rare in the population. Rarity is determined by the frequency of the alternative allele in the population. Large scale projects have been carried out to determine allele frequencies in the general human population. Key to the analysis of MAF is the ancestral background of each sample under investigation, to determine which alleles are rare given the presence of variant alleles in a similar genetic background. Importantly, large sample sizes are required to identify extremely rare variants. The largest collection is gnomAD, composed of a total of 125,748 human exomes and 15,708 human genomes, from (Karczewski *et al.*, 2020)

In determining variants that are rare in the population, MAF cut-off thresholds vary depending on study, typically using MAF <0.01 or MAF <0.05. In these analyses the allele frequency cut-off is set at 0.05 to ensure variants are rare but include as many rare variants as possible to avoid discarding potential variants of interest. In keeping these variants that are rare by population standards, while including those that are not necessarily “very rare 0.01” or “ultra-rare 0.001”. Given the small sample size of this cohort and that these sequences are already restricted to exonic variants, this less stringent allele frequency threshold still yields a manageable variant set for downstream interpretation.

Allele frequencies are determined in this cohort for the gnomAD non-Finnish European cohort, as annotated within dbNSFP version 4.0a (Liu *et al.*, 2020). Two factors were considered in the selection of datasets from which allele frequency would be determined. Firstly, the larger sample size in the gnomAD WES dataset when compared to the gnomAD WGS dataset. This larger sample size hosts data on a larger number of individuals, potentially making for more accurate estimations of the true population allele frequency of any given variant. For this reason, it is possible that the use of this dataset may be preferable to the smaller WGS dataset. However, it is important to note that the 1000 genomes dataset, which prior to the release of gnomAD was the favoured database

of allele frequencies, consistent of just 2,500 individuals, of which a subset is of European descent (Auton *et al.*, 2015).

The second consideration is the appropriate use of the genome dataset when analysing a WES dataset. Specifically, use of this dataset enables determination of non-coding variant allele frequencies. This is critical to determine allele frequencies of what is the largest proportion of variants that are called. Rare variants were isolated as those satisfying the parameter in the gnomAD WES dataset, due to the sample size of the gnomAD exome population enabling accurate allele frequency determination, in combination with the lack of requirement for allele frequency estimates outside of the exome.

3.4.4 Selection of pathogenicity predicting parameters

The number of variants identified through NGS is constantly growing and there is a huge need for analyses to assess variation tolerance and prioritise those which are candidates for causing disease/disorder. The approaches and tools used in this study, CADD, SIFT, Polyphen-2, while informative and widely applied, come with limitations. Functional prediction across bioinformatic tools is inconsistent (Niroula and Vihinen, 2019). At the molecular level, the disruption caused by a genetic variant can range from no disturbance to detrimental effects on genomic sites that are key to protein function. These effects are particularly difficult to categorise in the case of missense variants. Missense variants are far more abundant in the genome than LoF variants, with the majority having no harmful effect on gene function (Ronemus *et al.*, 2014). Specifically, half of the variants predicted to be deleterious correspond to nearly neutral variants, which have minimal clinical relevance, but they will be subject to purifying selection (Miosge *et al.*, 2015).

The consensus scoring approach, here where a variant is required to satisfy pathogenicity thresholds in two of three scoring to be deemed pathogenic, is an effort to overcome inconsistencies in pathogenicity prediction. However, pathogenicity filter algorithms REVEL and ClinPred when applied alone have been demonstrated to be well tuned and could be as useful when applied independently than a consensus scoring approach (Ioannidis *et al.*, 2016; Alirezaie *et al.*, 2018; Gunning *et al.*, 2021). The purpose of variant scoring is key to selection of pathogenicity prediction parameters. Predictors may perform well for evaluation of variants in a research setting, but when not as well in when applied in clinical variant assessment (Gunning *et al.*, 2021). A further consideration when scoring variant pathogenicity is the performance, as measured by

the specificity of the tool in the population under investigation (Niroula and Vihinen, 2019) (Figure 3-7). Taken together these limitations highlight the need to benchmark variant pathogenicity criteria against variants that are representative of the study.

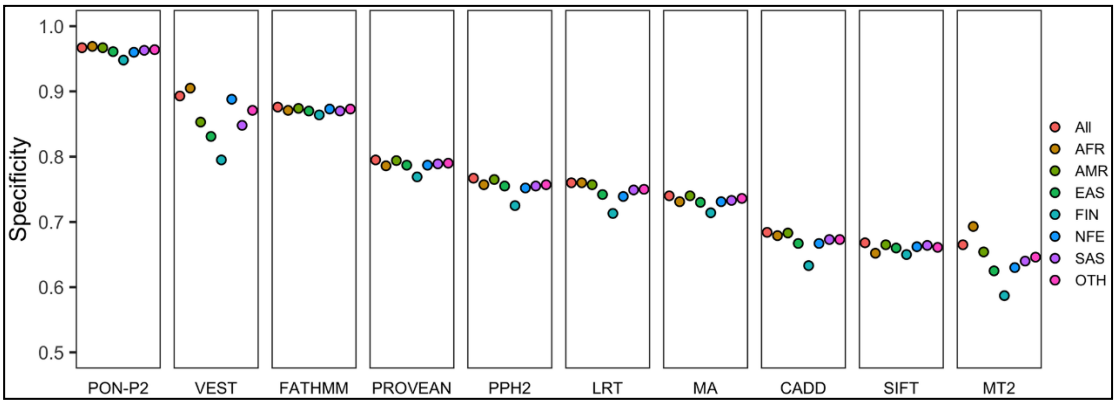


Figure 3-7 Performance of variant tolerance predictors for variants in ethnic groups.

Taken from Niroula et al. (2019): "Performance of variant tolerance predictors for variants in ethnic groups. Specificities of prediction tools for common variants (AF ≥1% and <25%) in different populations. AFR, African; AMR, American; EAS, East Asian; FIN, Finnish; NFE, non-Finnish European; OTH, Other; SAS, South Asian; MA, MutationAssessor; MT2, MutationTaster2; PPH2, PolyPhen-2. https://doi.org/10.1371/journal.pcbi.1006481.g002."

3.4.5 Relevance of variation identified

The analysis presented here applies a stringent variant filtration strategy to identify putatively pathogenic rare variation in genes with existing reports of association to autism and neurodevelopment. This pipeline is limited to the isolation of rare exonic SNVs, however NGS technologies enable additional classes of variation to be detected, with evidence supporting their involvement in the genetic basis of autism, such as CNVs, SVs and tandem repeat expansions as will be discussed later. There are future opportunities to explore these classes of variation in Cohort 1 and enable expansion of the understanding of the genetic basis of autism within this cohort. While variant discovery did not yield pathogenic SNVs with association to autism, expansion beyond this gene-set based variant filtration strategy will enable detection of more genetic variants which could be contributing to the phenotype.

Expanding beyond this filtration strategy may detect causative variation in the cohort, when unrestricted by the requirement to restrict analyses to genes with an existing gene-disease association reported. This gene-set based filtration step within the variant isolation strategy is a weakness, leaving many rare putatively pathogenic variants

uninterrogated. However, there is opportunity to overcome this in the future with an increase in sample size, achievable by analysis of this dataset in combination with other WES autism datasets, such as those described in Table 1-2. Large sample sizes give statistical power to enable gene-phenotype associations, while the small sample size of Cohort 1 enables only variant detection within known autism-associated genes.

3.4.6 Conclusion

This chapter describes analysis of a WES cohort of 42 individuals, varied in family structure. This work has enabled the development of a variant interpretation strategy from align sequence reads to a filtered and relevant variant call-set. No gold-standard pipeline currently exists for the isolation of rare exonic SNVs from NGS datasets. The research outlined within this chapter has yielded a strategy for the isolation of such variants in autism cohorts as is applied in Chapters 4 and 5.

Sample size and variability in family structure within this dataset limit the use of this cohort. This study is statistically underpowered to perform rare variant association testing and subsequently cannot be used to draw overall conclusions on the genomic basis of autism. However, analysis of this dataset has led to a high-quality and robustly annotated set of rare putatively pathogenic variants with evidence for autism relevance. Importantly, these findings have not be confirmed by Sanger sequencing and have not been validated to clinical genetic standards. Further value will be gained from this work in the future upon combined analyses of these data with larger autism sequencing cohorts.

Chapter 4. Evaluating gene-phenotype relationships through gene curation.

Presentations arising from the contents of this chapter:

“Application of an evidence-based curation framework to aid gene discovery: a pilot investigation in an autism family cohort.” Fiana Ní Ghrálaigh, Louise Gallagher & Lorna M. Lopez

Poster presented at the World Congress of Psychiatric Genetics, October 2020 (Appendix III-IV).

“Analysis Pipeline of Whole Genome Sequencing Data in Neurodevelopmental Disorders.” Fiana Ní Ghrálaigh, Niamh M. Ryan, Louise Gallagher, Lorna M. Lopez

Poster presented at the British Neuroscience Association Festival of Neuroscience, April 2019 (Appendix III-VI).

4.1 Abstract

Here, a gene-phenotype curation framework is in applied to three genes, *NAV2, NINL* and *CACNA2D3*. The dataset from which variant data is derived for curation is a set of rare putatively pathogenic exonic variants impacting autism-relevant genes are identified through WGS in a cohort of six individuals. All three genes achieved a classification of "Limited" by the Schaaf *et al.* (2020) gene curation framework, despite confidence in autism association supported by SFARI Gene.

4.2 Introduction

4.2.1 A family-based approach to identifying autism-associated variation

Penetrance refers to the number of cases harbouring a particular variant for which the phenotype is observed. Highly penetrant pathogenic variation by nature is not common in a healthy population as it would result in a high prevalence of a phenotype associated with reduced fecundity. Rather they can be expected to be and have been found to be rare in allele frequency. As specified in 3.4.3, rarity can be defined as a MAF of less than 5% in the population. Where cases aggregate within a family there is an expectation that the family harbours an enrichment of inherited, penetrant pathogenic variation.

Population-based studies require a variant to reach genome-wide significance in a large proportion of the unrelated affected individuals to be associated. In keeping with this, a variant will need to be sufficiently common to be identified as statistically associated. To identify rare variation in an unrelated cohort very large sample sizes are required to reach statistical association. Currently population-based studies are applied with an aim to identify common genetic variation. However, the effects of common variation are small and cannot explain observed patterns of heritability such as those seen in autism and other neurodevelopmental conditions. While autism is known to have a common component to its genetic basis, common variants are not expected to be the causative variation in multiplex families.

4.2.2 Dissecting gene-phenotype relationships

Disentangling gene-phenotype relationships in a complex condition faces many challenges. Until these challenges are overcome there is ambiguity in the degree of causation a variant is contributing to the condition. In the area of rare disease, including rare neurodevelopmental conditions, variant specific phenotypic data is crucial to

collating individual level information to reach sample sizes sufficient to gain insights into gene-disease relationships. One example of these resources is DECIPHER used by clinicians to share phenotype and genotype data of over 43,000 patients (Firth *et al.*, 2009). Phenotyping collated per variant locus has improved diagnosis of severe developmental conditions and has potential to inform on disrupted processes causing these phenotypes (Fitzpatrick and Firth, 2020).

Strategies and guidelines to streamline curation of gene-disease relationships are key to determination of pathogenicity of a variation relevant to a condition. At the early stages of variant discovery OMIM was and still is in certain diseases and disorders, a gold-standard resource to be use in variant interpretation (Amberger, Bocchini and Hamosh, 2011; Amberger *et al.*, 2019).

Further examples of successful resource development include the ClinGen framework. This framework outlines a standardised procedure, with specific criteria for assessment clinical validity and a quantitative approach, to collect evidence to support gene-disease association (Strande *et al.*, 2017). Expert gene curation panels can then systematically validate the gene-disease relationship. Another widely applied toolkit in genomic analysis comes from the ACMG, who maintain a set of standards and guidelines to adhere to in variant interpretation (Green *et al.*, 2013). A potential solution for disentangling gene-phenotype relationships comes from a proposed adaptation of the ClinGen gene curation framework for use in autism, accounting for the degree of certainty in autism diagnoses in studies reporting association and accounting for co-occurring diagnoses and well as incorporating genetic evidence, providing consistency throughout gene discovery (Schaaf *et al.*, 2020).

4.2.3 Hypothesis and aims

Gene curation in the context of autism can be used to quantify the evidence supporting gene-phenotype associations arising from sequencing studies. An evidence-based curation framework accounting for phenotypic heterogeneity has been proposed for use in autism. This chapter focuses on application of a gene curation framework in autism. The dataset from which variant data is derived for curation is a set of rare putatively pathogenic exonic variants impacting autism-relevant genes are identified through WGS in a cohort of six individuals. A subset of these impacted genes is selected for evaluation by gene curation through the framework.

The aims of this chapter are:

1) to apply an analysis strategy for isolation of rare exonic pathogenic SNVs from NGS data.

2) to isolate *de novo* variation in an autism-affected proband using a family-based approach to variant discovery.

3) to dissect gene-phenotype relationships through application of a gene curation framework.

4.3 Results

4.3.1 Cohort-level QC

This chapter describes analysis of WGS of Cohort 2. Cohort 2 is a dataset of six individuals and includes three autistic probands, as summarised in Table 4-1 and Table 4-2. Ascertainment of this cohort is described in 2.1.2. All samples are also included in Cohort 1. The pedigree in focus in this chapter is presented in Figure 4-1.

Cohort Overview	**N = 6**
Number of families	1 quad family 2 affected singletons
Male probands	N=1
Female probands	N=2

Table 4-1 Overview of Cohort 2.

Outlined are the family structures included in the cohort and proband counts by sex.

FID	**IID**	**Sex**	**Phenotype**	**Parental ID**
AS315	AS315C	F	Autism	N/A
AS322	AS322C1	F	ASD, ADHD	N/A
AS420	AS420C1	M	Autism, moderate ID, self-injurious behaviour, catatonia, dysmorphology	Father; AS420F Mother; AS420M
AS420	AS420C2	M	Unknown	Father; AS420F Mother; AS420M
AS420	AS420F	M	Unknown	N/A
AS420	AS420M	F	Unknown	N/A

Table 4-2 Cohort 2 phenotype, sex, and parental ID.

Outlined are reported relationships, sex and clinically validated phenotype for individuals analysed within Cohort 2.

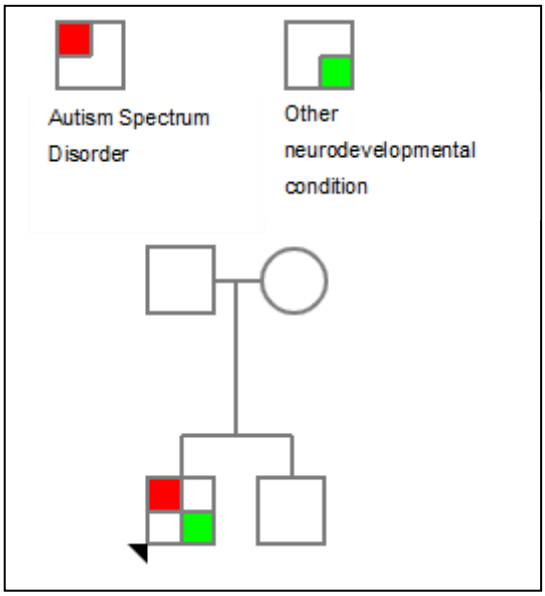


Figure 4-1 Pedigree AS420.

Red marking indicates affection in the proband. Affection here refers to the complex phenotype of autism regression in infancy, and co-occurring moderate intellectual disability, severe self-injurious behaviours, and catatonia over the course of development.

4.3.2 Variant filtration

The workflow developed and applied in these analyses is presented in Figure 4-2. The full set of non-synonymous variants which has been annotated as described in Materials and Methods 2.8, are subject to further subsetting on the bases of allele frequency, predicted pathogenicity and gene-level associations as follows.

Rare putatively pathogenic autism-relevant variants were isolated as detailed in Chapter 3, yielding 107 variants of relevance in six individuals (Figure 4-2). These variants were further subset to those occurring in affected individuals, (three unrelated probands) and unaffected individuals (family members of one proband) (Figure 4-3). A subset of the variants isolated through the framework outlined in Figure 4-2 have been curated through an evidence-based gene curation as follows within this chapter.

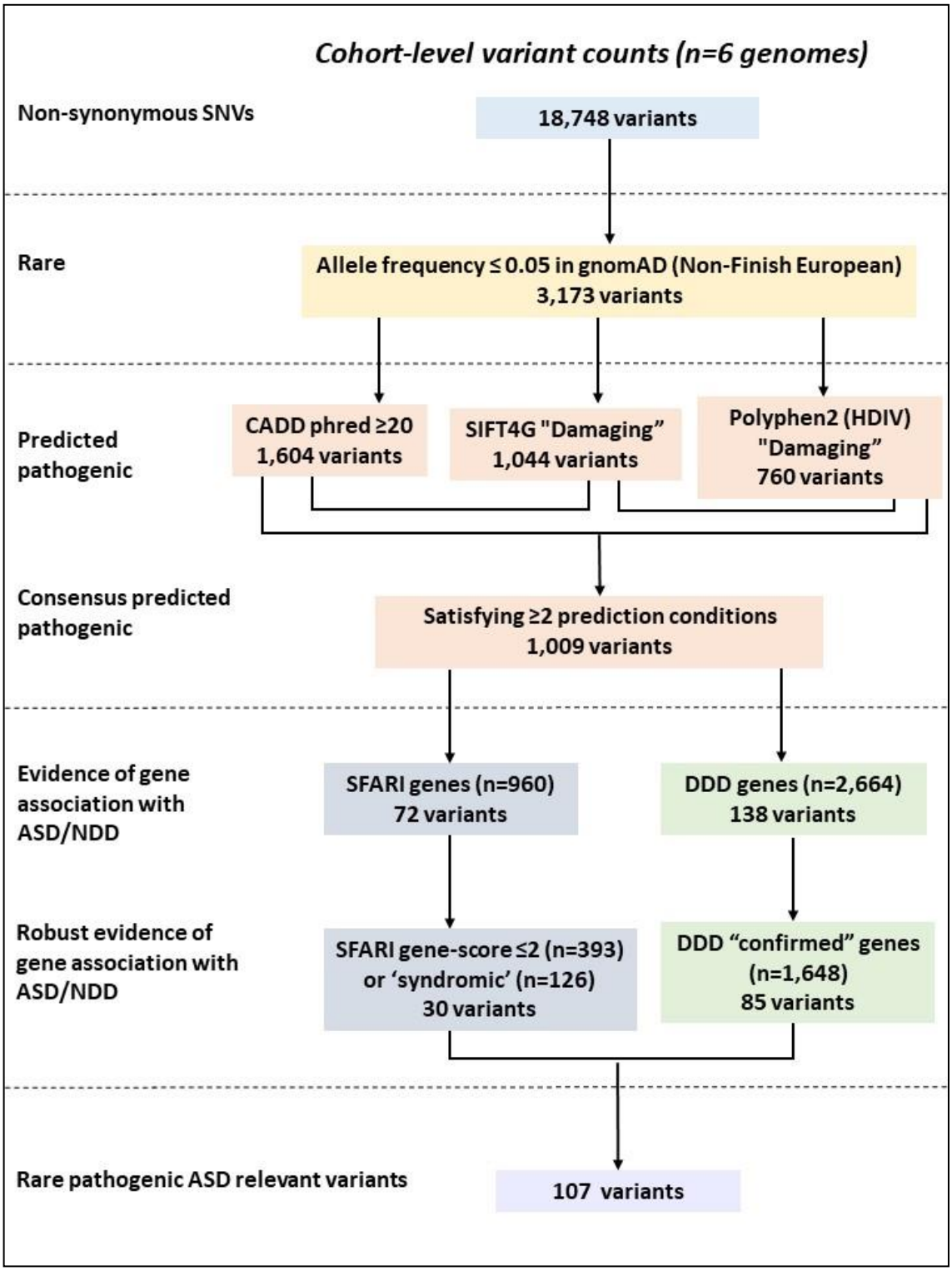


Figure 4-2 Flow of variant filtering with cohort-level variant counts.

Arrows show the direction of flow from each level of filtering (specified on the left). SFARI refers to Simons Foundation Autism Research Initiative Gene Module. DDD refers to the gene2phenotype database arising from the DDD study.

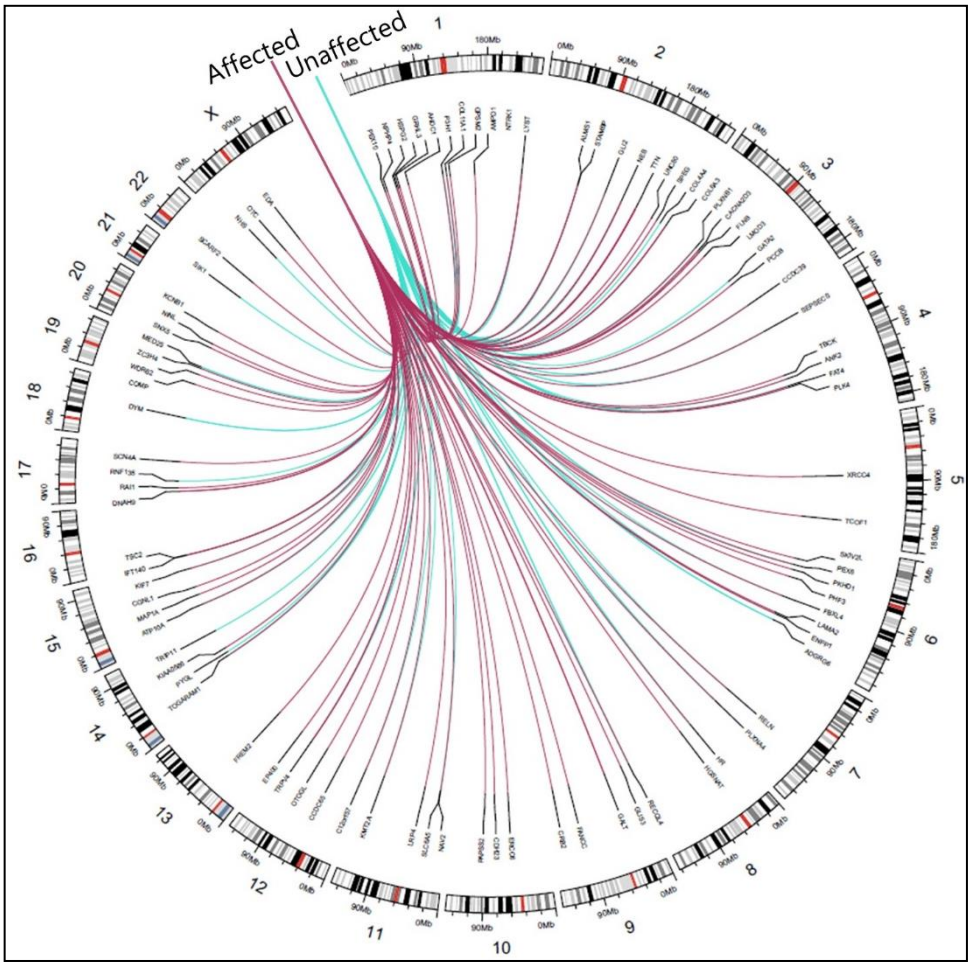


Figure 4-3 Spread of variation across genomic regions.

Chromosomes are shown around the outer track of the figure (1:22, X). The gene names are given on the inner track. These are the genes in which the rare pathogenic autism-relevant variants outlined in Fig.1 are located. Links are made in purple (affected; n=3 individuals, n=91 variants) and blue (unaffected; n=3 individuals, n=69 variants) between each gene and the respective affection status of the individual harbouring the variant. Affected denotes individuals with an autism diagnosis.

4.3.2.1 Isolation of de novo variation

This cohort of six individuals include a quad family with one proband, an unaffected sibling and unaffected family members (Figure 4-1). This structure enables the further subset of variants to those that are unique to the proband and may be of relevance in this sporadic case of autism. Due to the severity in the phenotype of this proband, as detailed in 2.1.2, it is hypothesised that rare highly penetrant variation is causative of the phenotype.

In addition to this, there is complexity in the phenotype of this individual with the proband experiencing regression in infancy, and co-occurring moderate ID, severe self-injurious behaviours, and catatonia over the course of development. Following this hypothesis, *de novo* rare predicted pathogenic variants were isolated for further interpretation, not restricted to those in autism and neurodevelopmental-relevant genes. Table 4-3 presents variant counts in this family context. No homozygous variants were isolated as unique to proband in this variant call-set.

Variant Call Set	**Variant Count**
Rare pathogenic variants	280 variants
Paternal inherited (of 271 paternal variants)	152 variants
Maternal inherited (of 267 maternal variants)	144 variants
Shared unaffected sibling variants (of 280 sibling variants)	153 variants
de novo variation unique to proband	0 variants

Table 4-3 Variant transmission in a quad family of autism.

Variation is here determined a variant locus at which an individual has a 0/1, 1/0 or 1/1 genotype, indicating heterozygous or homozygous alterative allele presence. Variant counts refer to number of variants in the proband variant call-set. Here the rare (gnomAD <5%,) predicted pathogenic (through CADD, SIFT4G and Polyphen-2) variant call set of family members are queried against that of the proband.

Homozygous variant sites in the proband where there is heterozygosity in both parents are isolated in Table 4-4 and Table 4-5. Candidate variants are excluded based on homozygosity in unaffected family members (i.e., paternal, maternal of sibling homozygosity).

Variant Call Set	Variant Count
Homozygous rare pathogenic variants	10 variants
Paternal homozygosity detected	1 variant
Maternal homozygosity detected	1 variant
Sibling homozygosity detected	4 variants
***de novo* variation unique to proband**	**4 variants**

Table 4-4 Recessive inherited homozygosity in an affected proband.

Homozygous variation is here determined a variant locus at which an individual has a 1/1 genotype. Variant counts refer to number of variants in the proband variant call-set. Here the rare (gnomAD <5%,) predicted pathogenic (through CADD, SIFT4G and Polyphen-2) variant call set of family members are queried against that of the proband.

Chromosome	ID	ref	alt	Gene	Ensembl Gene ID	Clinvar ID
chr2	rs116298748	G	A	*COL5A2*	ENSG00000204262	136944
chr7	rs146095374	C	A	*TYW1B*	ENSG00000277149	N/A
chr20	rs34396614	C	G	*MYLK2*	ENSG00000101306	36652
chrX	rs45557031	G	A	*C1GALT1C1*	ENSG00000171155	460285

Table 4-5 Homozygous proband variants in focus.

4.3.2.2 Biological interpretation of variants identified

Of the four homozygous variants identified in the proband under investigation, none impact genes that are included in SFARI Gene, indicating that none of these variants have existing evidence supporting association with autism. The genes identified have been implicated in a number of phenotypes that are unobserved in the proband including Ehlers-Danlos Syndrome (*COL5A2*), Cardiomyopathy (*MYLK2)* and Tn syndrome, a rare autoimmune disease (*C1GALT1C1)* (Richards *et al.*, 1998; Ju and Cummings, 2005).

TYW1B is a protein-coding gene encoding a component of the wybutosine biosynthesis pathway. Wybutosine is a hypermodified guanosine found in phenylalanine tRNA. A recent case report links a large-scale chromosomal rearrangement impacting a number of genes including *TYW1B,* with an MRD44-like phenotype which includes intellectual disability, microcephaly, finger anomalies, and facial dysmorphia (Córdova-Fletes *et al.*, 2022). This phenotype is not consistent with the complex phenotype observed in the proband investigated here, however the intellectual disability resulting from large-scale genomic changes in this region may be suggestive of a role of the single base change

identified here, rs146095374 impacting *TYW1B,* accounting for aspects of the neurodevelopmental phenotype reported.

4.3.3 Evaluating gene-phenotype relationships through gene curation

4.3.3.1 Gene selection for curation

Candidate genes curation through an evidence-based framework were selected from the dataset analysed within this chapter as outlined in Figure 4-4. Candidate genes are those in which putatively pathogenic variants in autism-associated genes were identified. Genes were excluded where ClinGen curation is completed (Figure 4-4). ClinGen curation refers to curation following the ClinGen guidelines rather than the modified ClinGen guidelines proposed by Schaaf *et al.* (2020).

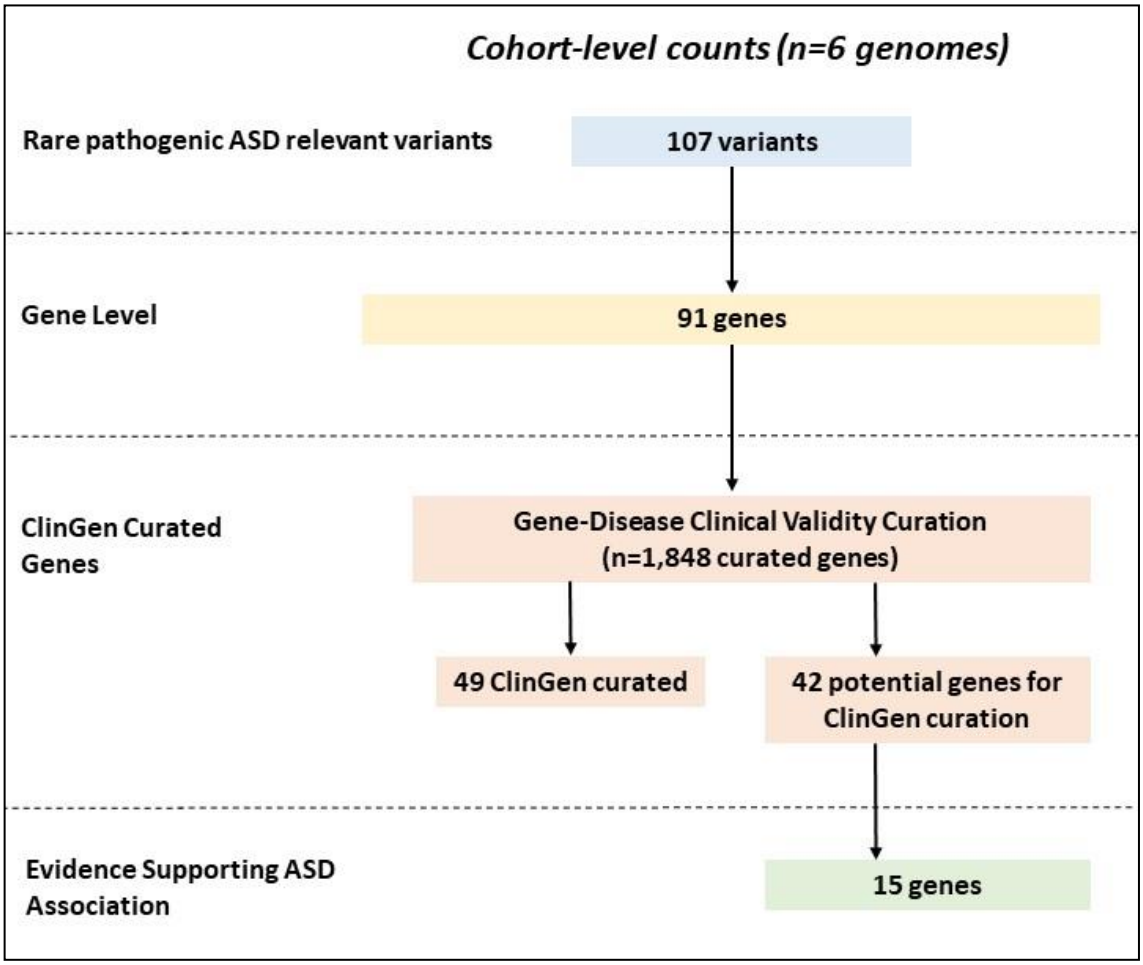


Figure 4-4 Gene selection for curation.

Classification of three genes with highest number of autism reports. Genes were excluded from this analysis when already curated by ClinGen. Genes were selected for analysis when evidence of autism association is reported in the literature.

Candidate genes were prioritised for gene curation based on the number of reported autism cases for which a variant in the gene had been associated. Selection of those with the highest number of reports lead to gene selection with the most evidence to feed into the gene curation framework. Two metrics of reports were used for this purpose: GeneCards publication search and SFARI Human Gene Module.

Three genes were selected for curation using by the Schaaf *et al.* (2020) framework. These genes were selected based on the highest number of autism specific reports of all candidate genes. *NAV2* (11p15.1, neuron navigator 2)*, NINL (*20p11.21*,* ninein Like) and *CACNA2D3 (*3p21.1-p14.3, calcium channel voltage-dependent alpha 2/delta subunit 3) are detailed in Table 4-6 along with gnomAD constraint scores measuring predicted tolerance of each gene to LoF and missense variation.

Gene Symbol	**Gene Name**	**Cytogenetic Location**	**ClinGen Curation**	**gnomAD Constraint Scores**	
				Loss-of-Function	**Missense**
CACNA2D3	Calcium channel, voltage-dependent, alpha 2/delta subunit 3	3p21.1-p14.3	Not curated	pLI = 0.58 o/e = 0.22 CI= 0.14 - 0.34	Z = 1.86 o/e = 0.78 CI= 0.72 - 0.84
NAV2	Neuron navigator 2	11p15.1	Not curated	pLI = 1 o/e = 0.16 CI= 0.11 - 0.25	Z = 1.39 o/e = 0.9 CI= 0.86 - 0.94
NINL	Ninein Like	20p11.21	Not curated	pLI = 0 o/e = 0.89 CI= 0.72 - 1.09	Z = -0.98 o/e = 1.1 CI= 1.04 - 1.16

Table 4-6 Constraint metrics are estimated based on expected vs observed SNVs identified within the gene.

These estimates are further broken down by variant type with LoF and missense variant constraint scored indicated in this table. Observed/expected (o/e) is a continuous measure of how tolerant a gene is to a certain class of variation. Low o/e values indicate the gene is under stronger selection for that class of variation than a gene with a higher value. 90% confidence interval (CI) is given for each o/e value. Z score given is the deviation of observed counts from the expected number. Positive Z scores indicate increased constraint. The closer pLI is to one, the more intolerant of protein-truncating variants the transcript is predicted to be.

4.3.3.2 Gene scoring

Gene scoring was carried out per Schaaf *et al.* modified ClinGen Gene-Disease Validity scoring guidelines. Detailed accounts of scoring are presented for the three genes selected *NAV2* (Table 7-1, Table 7-2), *NINL (*Table 7-3) and *CACNA2D3* (Table 7-4). Scoring was carried out for each reported autism case harbouring a variant in the specified gene. Variant characteristics were recorded from supplemental tables associated with the publication, denoted by author and year of publication in Table 4-7. A gene matrix was constructed following that proposed by Schaaf *et al.* where each variant score was summed to generate a gene-level classification.

Phenotype methods for each study and the quality score that they have been assigned here, through the modified ClinGen gene-disease curation framework are presented in Table 4-7. This table presents also details on consideration of cognitive ability of the study participants and highlights studies where robustness of autism diagnosis may be compromised by insufficient consideration of intellectual ability. This lack of consideration in phenotyping leads to ambiguity in assignment of an autism diagnosis and for this reason results in a downgrading in gene-disease association scoring as shown in Table 7-1.

Authors (Year): Title	**Phenotyping Method/ Notes:**	**Quality of Autism Phenotype Report**	**Cognitive Ability Cautionary Comment Assigned**
De Rubeis S, *et al.* (2014)	**Autism**: As part of the ASC, all subjects were diagnosed with “Autistic Disorder” as the primary phenotype (DSM-5). **Cognition**: No information provided.	High confidence.	No cautionary comment required.
Guo H, *et al.* (2018)	**Autism:** Autism diagnosed primarily according to DSM-IV/5 criteria, documenting additional co-occurring conditions where possible. **Cognition:** "[T]he majority of patients with severe DNMs (*de novo* mutations) and a cognitive assessment showed evidence of some form of intellectual impairment. Only TNRC6B, NCKAP1, and one of the two ZNF292 LGD DNMs occur in Autism patients with an IQ in the normal range."	High confidence.	Uncertainty regarding validity of Autism diagnosis considering insufficient information regarding intellectual ability.
Iossifov *et al.* (2014)	**Autism:** Simons Simplex Collection (SSC) - extensive autism phenotyping, including ADI-R, ADOS, cognitive testing, Vineland, SRS, SCQ (see https://www.sfari.org/resources/ssc-instruments/ for full phenotyping information). **Cognition:** No information provided; however, as part of the SSC, thorough cognitive testing was performed.	High confidence.	No cautionary comment required.
Leblond CS, *et al.* (2019)	**Autism:** Extensive autism phenotyping, including ASSQ, DISCO-10, DISCO-11. **Cognition**: WISC or WEIS, IQ (DISCO)	High confidence.	ID included in study evaluate case by case.

Lim ET, *et al.* (2017)	**Autism:** As part of the ASC, all subjects were diagnosed with “Autistic Disorder” as the primary phenotype (DSM-5). **Cognition:** No information provided.	High confidence.	No cautionary comment required.
O'Roak BJ, *et al.* (2012)	**Autism:** Simons Simplex Collection (SSC) - extensive autism phenotyping, including ADI-R, ADOS, cognitive testing, Vineland, SRS, SCQ (see https://www.sfari.org/resources/ssc-instruments/ for full phenotyping information). **Cognition:** Little information provided; however, as part of the SSC, thorough cognitive testing was performed.	High confidence.	No cautionary comment required.
Ruzzo EK, *et al.* (2019)	**Autism:** No specific information about phenotyping assessments provided - notes that "[s]tudy subjects were carefully selected from the Autism Genetic Resource Exchange (AGRE) and chosen from families including two or more individuals with autism (those with a "derived affected status" of "autism," "broad-spectrum," "nqa," "asd," or "spectrum.")" **Cognition:** No information provided	High confidence	No cautionary comment required
Sanders SJ, *et al.* (2012)	**Autism:** Simons Simplex Collection (SSC) - extensive autism phenotyping, including ADI-R, ADOS, cognitive testing, Vineland, SRS, SCQ (see https://www.sfari.org/resources/ssc-instruments/ for full phenotyping information. **Cognition:** Little information provided; however, as part of the SSC, thorough cognitive testing was performed.	High confidence.	No cautionary comment required.

Wang T, *et al.* (2016)	**Autism:** Autism diagnosed according to DSM-IV criteria. **Cognition:** No information available.	High confidence.	Uncertainty regarding validity of ASD diagnosis considering insufficient information regarding intellectual ability.
Wu H, *et al.* (2019)	**Autism:** Autism Clinical and Genetic Resources in China (ACGC) cohort. Diagnosed according to DSM-IV or DSM-V by experienced clinicians. In addition, co-occurring conditions including medical problems, such as epilepsy, gastrointestinal issues, and sleep disorders; developmental diagnoses, such as ID and language delay; and mental-health conditions, such as ADHD, obsessive-compulsive disorder, and depression, were collected for presenting patients. **Cognition:** No information provided for this individual.	High confidence.	ID included in study evaluate case by case.
Yuen RK *et al.* (2017)	**Autism:** Autism diagnosis of all participants must have met criteria on one or both diagnostic measures: ADI-R and ADOS or considered a clinical diagnosis when given by an expert clinician according to the DSM IV or V edition). **Cognition:** "Many participants were assessed with standardized measures of intelligence (IQ), language, and general adaptive function. 19.6% had scores within the range for ID (FSIQ < 70)."	High confidence.	No cautionary comment required.

Table 4-7 Evaluating phenotyping in sequencing studies of autism.

This tables presents autism studies in which variants were identified in one or more of the three genes under investigation: NINL, NAV2 or CACNA2D3. Phenotype methods are recorded for both autism and cognition as reported in respective study methods. Confidence scores and cautionary notes are assigned according to Schaaf et al. (2020) recommendations. Where a variant has been scored from these studies by Schaaf et al. confidence scores have been taken directly from those assigned in the guidelines study.

Variants reported in genes *NAV2, NINL* and *CACNA2D3* were individually scored according to the gene scoring matrix outlined by Schaaf *et al.* (2020) (Table 7-1-Table 7-4). Variant reports were identified as specified in 2.13.3. Individual variant data was extracted from the publication source specified per variant and collated per gene as shown in the respective matrices. Where available experimental evidence was evaluated using the experimental scoring matrix proposed by Schaaf *et al.* (2020) (Table 7-2). Variants scored were restricted to variants where the proband carried the pathogenic variant under investigation. For example the non-synonymous coding variant identified in *CACNA2D3* (Chr3(hg19):g. 54925398G>A, V629M (De Rubeis *et al.*, 2014)), is a variant which would otherwise be awarded default scoring of 2, autosomal dominant variant. However, while identified through a trio study of autism with the aim of isolating *de novo* variation, the proband carried the reference allele while the variant detected is a paternal variant. For this reason, several variant reports counted in Table 2-35 were excluded from scoring.

Key to the scoring matrix are variant specific details including gnomAD allele frequency and to determine rarity of the alternative allele in an unaffected population and protein coding consequence as an estimate of pathogenicity (gnomAD v2.1.1 October 2020). Mode of inheritance is recorded where parent genotype information is available.

Individual report scores were totalled and the sum of scores were used to designate classifications of each gene (Table 7-1-Table 7-4). All genes, despite enriching for those genes with the highest number of autism-specific reports, were designated as having limited evidence supporting their role in autism, as presented in Table 4-8. This limited classification is derived from the ClinGen protocol, justified as follows: *"There is limited evidence to support a causal role for this gene in this disease, such as: Fewer than three observations of variants with sufficient supporting evidence for disease causality OR Variants have been observed in probands, but none have sufficient evidence for disease causality. Limited experimental data supporting the gene-disease association"* (Gene-Disease Validity Standard Operating Procedures, Version 7 - ClinGen | Clinical Genome Resource, 2019).

Gene	SFARI Gene Score	Sum of scores	Classification
NAV2	Strong Candidate (SFARI Gene Score 2)	0.5	Limited
NINL	Strong Candidate (SFARI Gene Score 2)	4.5	Limited
CACNA2D3	High Confidence (SFARI Gene Score 1)	5.0	Limited

Table 4-8 Classification of three genes with highest number of autism reports.

The three genes with the highest number of autism reports (6 publications each) were selected for curation. Sum of scores represents the raw sum of genetic and experimental evidence towards autism based on Schaaf et al. framework with gene classification. SFARI Gene score is included here as a comparative score to that determined by the gene curation framework. SFARI Gene score of 1 or 2, as are assigned to all three genes, indicate support for autism association as determined by SFARI Gene curation.

The "limited" classification of these genes is a result of downgrading of variants identified. The most frequently applied downgrading justifications are outlined in Table 4-9. These downgrading classifications may be used to inform variant discovery pipelines, such as allele frequency thresholds in gnomAD or criteria for predicted gene disruption of missense variants. In addition, phenotypic data collected during cohort ascertainment should consider the impact of cognition scoring measures on downstream variant associations with autism.

Frequently applied variant downgrading	Rationale
de novo missense variant with suggested functional evidence	Limited evidence of disruption of gene function
Observed in gnomAD	Allele identified in an unaffected control cohort
Synonymous variant with no functional data provided	Unknown impact on gene function
WES/WGS not performed	Lack of confidence in sequence quality
Lack of confidence in ID/ cognition score	Autism phenotype in the presence of ID
Intronic variant	Unknown impact on gene function
Proband carries reference allele	Variant not clearly associated with autism
Inherited missense variant without functional evidence	No evidence of disruption of gene function

Table 4-9 Modified ClinGen downgrading frequently applied in variant scoring matrices.

4.4 Discussion

This chapter reports on gene-phenotype curation of a subset of genes identified through WGS of an autism cohort of six individuals. When restricted to affected individuals three genes in which rare putatively pathogenic autism-relevant SNVs were detected were selected for curation. These genes were prioritised for curation on the number of autism-reported variants impacting the genes. Curation was carried perform on *NAV2, NINL* and *CACNA2D3.* Each of these gene was classified as "Limited," scoring the gene-phenotype association (Table 4-8) (*Gene-Disease Validity Standard Operating Procedures, Version 7 - ClinGen | Clinical Genome Resource*, 2019). This implies that while these genes are implicated in the genetics of autism, their association is not restricted to an autism only phenotype. In the case of the cohort investigated two of three of the probands investigated are affected by both autism and a co-occurring neurodevelopmental phenotype.

Phenotypic heterogeneity has an impact on the power of genetic associations (Manchia *et al.*, 2013). This effect has been demonstrated in psychiatric genetics through GWAS studies. Specifically, a landmark GWAS study by Ripke *et al.* achieved 18% phenotypic variance explained by PGS (Ripke *et al.*, 2014). However, Stahl *et al.* reported a GWAS study with comparable sample size, achieved 8% of phenotypic variance explained by PGS (Stahl *et al.*, 2019). In addition to the underlying difference in genetic architecture of the phenotype, the difference in variance explained by these studies is contributed to by heterogeneity of the sample studied.

While lists of genes relevant to autism have been developed, for example two SFARI Gene and DDD gene2phenotype applied in these analyses, these lists are limited in their ability to dissect neurodevelopmental phenotypes presenting with autism (Abrahams *et al.*, 2013; Wright *et al.*, 2015; Myers, Challman, Bernier, *et al.*, 2020). Application of a formal evidence-based gene curation framework, such as that proposed by Schaaf *et al.*, accounts for these co-occurring diagnoses and provides consistency throughout gene discovery (Schaaf *et al.*, 2020). This framework was developed with psychiatrists with expertise in these phenotypes and unlike SFARI Gene or the standard ClinGen Gene-Disease curation, downgrades evidence of association with autism when the individuals for whom the gene has been associated has any ID.

As demonstrated in the classification of *NAV2, NINL* and *CACNA2D3*, this is a stringent approach resulting in three genes with multiple reports of association and SFARI scoring

of strong or high confidence (Scores 1 or 2), summing to a limited association with the autism phenotype (Table 4-8). Importantly, this interpretation comes from classification of just three genes and a wider classification of the full set of genetic variation associated with autism would be needed to determine the potential for this framework of classification. Furthermore, this framework is a labour-intensive process requiring comprehensive review of all literature reporting variation in the gene undergoing classification, as well as input from an expert panel and while the importance of gene curation is well-understood in the genomics community, this workload may not be feasible when considering the ~1,000 genes with some degree of evidence of association.

In addition, the results presented associate variation in four genes with the complex neurodevelopmental phenotype observed in a proband within the cohort studied. This *de novo* variation was isolated in an autism-affected proband using a family-based approach to variant discovery, identifying four variants impacting genes *COL5A2, TYW1B, MYLK2* and *C1GALT1C1.* Existing evidence does not link these single-base changes to the phenotype observed in this individual and these variants require functional analysis to robustly determine their contribution. Expanding beyond this filtration strategy may detect causative variation in the cohort. The pipeline applied here is limited to the isolation of rare exonic SNVs, however NGS technologies enable additional classes of variation to be detected, with evidence supporting their involvement in the genetic basis of autism, such as CNVs, SVs and tandem repeat expansions, as will be discussed later. There are future opportunities to explore these classes of variation in Cohort 2 and enable expansion of the understanding of the genetic basis of autism within this cohort. In addition, this analysis has focused on identification and interpretation of exonic variants only. Cohort 2 has undergone WGS enabling detection of non-coding variation which will be informative to the genomic basis of the individuals studied.

4.4.1 Conclusion

An autism gene curation framework was applied to three genes *NAV2, NINL* and *CACNA2D3* to dissect gene-phenotype associations with autism, each being scored as “Limited” association to autism despite literature suggesting confidence in the autism association. In addition, the analysis outlined in this chapter applies an analysis strategy for isolation of rare exonic pathogenic SNVs from WGS data and reports 107 variants in 91 genes with existing evidence of autism-association. *De novo* variation was isolated in an autism-affected proband using a family-based approach to variant discovery,

identifying four variants impacting genes *COL5A2, TYW1B, MYLK2* and *C1GALT1C.* These require functional analysis to robustly determine their contribution to autism.

Chapter 5. A pedigree driven approach to identify pathogenic variation in multiplex families of neurodevelopmental conditions.

Presentations arising from the contents of this chapter:

“Genomic syndromes in autism: Using whole genome sequencing to investigate multiplex families with autism and associated neurodevelopmental conditions.” Fiana Ní Ghrálaigh, Aoife Coghlan, Louise Gallagher & Lorna M. Lopez

Poster presented at Genomics of Rare Diseases (Wellcome Connecting Science), April 2022 (Appendix III-I).

5.1 Abstract

Here rare putatively pathogenic SNVs in genes with evidence supporting their role in autism are detected, using a family-based study design to evaluate variant transmission. The analyses outlined in this chapter follow the framework for calling and annotation of rare, exonic SNVs occurring in genes with existing evidence supporting autism association, outlined in Chapter 3. Variant detection and interpretation have been carried out on four multiplex pedigrees and putatively pathogenic variation detected is detailed within this chapter. These findings add to evidence supporting the involvement of genes *TTN, PSPH, RECQL4 NECTIN4 TSC2 TSHZ3 SLC26A2* and *FKTN* pending confirmation by Sanger sequencing.

5.2 Introduction

5.2.1 Enriching for penetrant inherited variation in multiplex pedigrees

A multiplex family has several affected members with the causative genetic variation likely to lie in variant sites that are shared between affected individuals. A family-based genomic study involves analysis of sequencing data in unaffected and affected family members.

Within the family-based study design comes innovative approaches to collecting and analysing family data, as has been summarised by Morris *et al.* (2015). These range from large population-based family-studies, for example a Utah registry which collates pedigree information on all state residents for decades, through to smaller family-based studies such as Wang *et al.* (2013) who analyse only independent probands from within families to inform on IQ differences in autistic individuals (Nelson *et al.*, 2013; Wang *et al.*, 2013; Morris *et al.*, 2015). Leveraging affection status informs on variant penetrance and narrows the search for causative variation within families. Studying large pedigrees of multiplex or extended families leads to a more homogenous causative variant set, due to the high degree of genetic sharing between related individuals. In contrast to a case-control or population-based study design, this causative variant set can be isolated in the absence of control genomes.

5.2.2 Using family structure to inform on mode of variant transmission

Families boost the ability to perform association and linkage studies effectively by enriching for a causative variant, as compared to population-based studies. This is particularly valuable in autism where a combination of genetic variants is likely to be causative of the complex phenotype observed (Antaki *et al.*, 2022). In a family-based design it is likely that within families a smaller number of genes contributing to the condition will be identified, than by a population-based design where a genome-wide association is performed. In practice by analysing individual genomes within families, focus can be put on a smaller number of contributing genes, as opposed to case-control studies where all genes are interrogated leaving limited power for gene discovery. Larger multiplex and extended families with multiple affected individuals further reduce the sample size required for rare variant identification by increasing the number of copies of a variant detected (Glahn *et al.*, 2019).

Ascertainment of pedigrees for genomic research, particularly extended pedigrees, is challenging, with added expense and time commitment required for identification, recruitment, and sample collection of whole pedigrees than those involved in the study of unrelated individuals in a case-control approach. Unaffected family member genotyping and phenotyping is as important as consideration of affected individuals for robust evaluation of variant transmission in families. For this reason, depending on phenotype and the underlying genetic architecture, it may be more efficient to take a case-control study design. However, when considering autism there is benefit to family-based ascertainment for genomic analysis as outlined in Chapter 1 and the informative, yet limited, number of loci identified through large-scale genome-wide association studies to date (Grove *et al.*, 2019). While ascertainment is costly, pedigree sequencing can be cost effective. Given that genetic relationships between family members are known, WGS where appropriate can be imputed for family members that have not be sequenced, decreasing the effective cost per sample (Glahn *et al.*, 2019).

5.2.3 Hypothesis and aims

Family structure enables mode of transmission of relevant genetic variation to be interrogated. Specifically, the contribution of *de novo* variation is smaller in multiplex families than the contribution in simplex families (Yoon *et al.*, 2021). Rare larger multiplex or extended pedigrees in contrast, are expected to have a burden of rare highly penetrant genetic variants that are causative of autism and co-occurring phenotypes. This chapter leverages the additional information available from studying extended pedigrees. In

particular the cohort study within this chapter is hypothesised to enrich for rare fully penetrant SNVs, aiding in the identification of autism-associated variants.

The analyses outlined in this chapter follow the framework for calling and annotation of rare, exonic SNVs occurring in genes with existing evidence supporting autism association. While this cohort is underpowered to include linkage analyses which are enabled by extended pedigrees, analysis is performed within this chapter on affected vs unaffected family members.

The aim of this chapter is to identify rare putatively pathogenic SNVs in genes with evidence supporting their role in autism, using a family-based study design to evaluate variant transmission.

5.3 Results

5.3.1 Cohort in summary

Cohort 3 is a family-based dataset of 29 individuals from 4 multiplex families, as presented in Figure 5-1. Ascertainment of this cohort is described in 2.1.3. This chapter describes analysis of WGS data analysis of this cohort. One sample, AS325C1, failed at WGS. One family was excluded based on unmet inclusion criteria following QC. A total of 28 samples remains in this cohort for analysis. In parallel genome-wide genotyping was performed on this cohort to provide QC checks prior to sequencing. This data, for 29 individuals was used in the cohort QC check described in 2.7.3. Based on affection within each family, the mode of transmission expected to be relevant was hypothesised independently for each family. The hypothesised mode of transmission is presented in Figure 5-1.

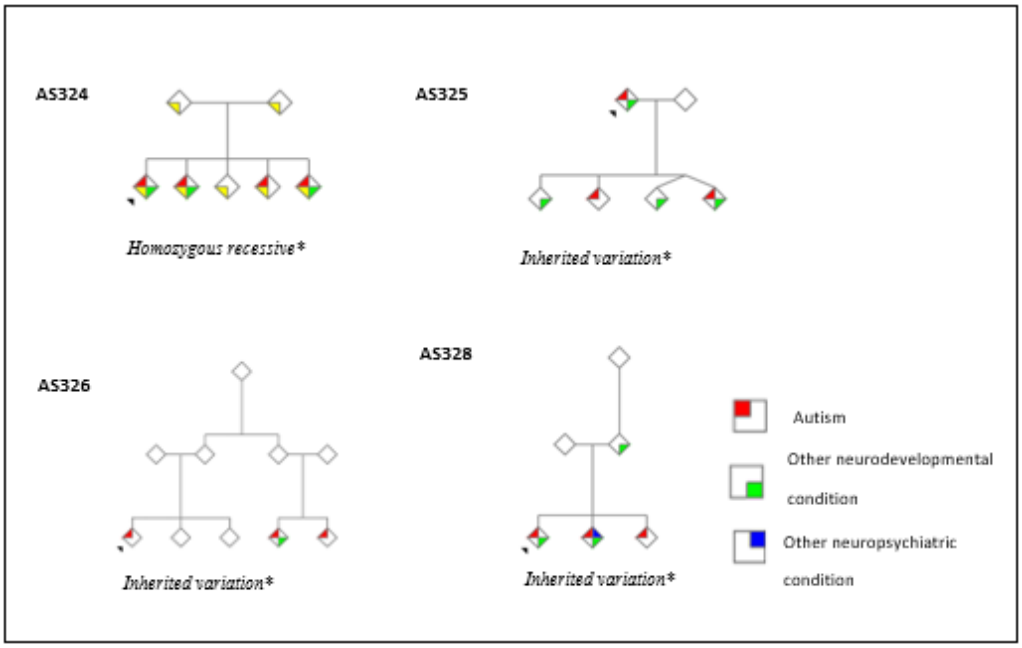


Figure 5-1 Cohort 3 in summary.

*Proposed mode of transmission for variant interpretation.} Presented are the pedigrees of the 4 families sequenced in this rare cohort. The key associated with affection and sequencing status is presented alongside. * Denotes the mode of variant transmission hypothesised to be relevant within each family.*

5.3.2 Variant QC

Variant evaluation was performed on WGS raw variant calls isolated by a Sention based pipeline within Genuity Science Pipeline Services (detailed in 2.5.3). Raw variant calls were restricted to SNVs for downstream analysis. SNVs were hard-filtered to isolate and remove those variants deviating from Hardy-Weinburg equilibrium (exact test $<10^{-6}$) and variant sites missing greater than 20% of data.

In the absence of GATK Best Practices to produce variant calls, VQSR variant filtering could not be applied to isolate a high-confidence variant call-set, as was performed in analysis of Cohort 1 and Cohort 2. In place of VQSR, variant hard filtering was performed in line with GATK recommendation to remove mapping errors and sequencing errors from the call set Table 5-1. High-confidence variants were retained in the call-set for downstream annotation.

Filter	**Description**	**Threshold**
QD	Variant quality / depth	< 2.0
MQ	Mapping Quality	< 40.0
FS	Phred-score Fisher's test p-value for strand bias	> 60.0
HaplotypeScore	Consistency of the site with haplotype	> 13.0
MQRankSum	Mapping quality of reference reads vs alternative reads	<-12.5
ReadPosRankSum	Distance of alternative allele from the end of the reads	< -8.0

Table 5-1 GATK recommended variant quality filters for SNVs.

Recommendations available at

https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants.

The Ts/Tv ratio was evaluated before and after variant QC to access data quality and the efficacy of variant quality filtering. Data quality is reported per sample to highlight any discrepancies which may have arisen due to the method of sample collection, i.e., DNA extracted from blood or saliva.

The expected Ts/Tv is 2.0-2.1 in human genome sequencing. The improvement in Ts/Tv ratio across all sample from outside of this range to within the range shows the need for low confidence variant call removal from the call-set to achieve a high confidence variant set (Table 5-2).

FID	IID	Raw variant call-set		High-quality genotypes	
		Mean Depth	Ts/Tv	Mean Depth	Ts/Tv
AS324	AS324C1	37.234	1.955	37.919	2.021
AS324	AS324C2	35.278	1.959	35.481	2.021
AS324	AS324C3	33.869	1.962	34.315	2.023
AS324	AS324C4	34.998	1.957	35.387	2.020
AS324	AS324C5	43.166	1.952	43.679	2.021
AS324	AS324F	32.253	1.960	32.513	2.019
AS324	AS324M	33.548	1.962	33.855	2.024
AS325	AS325C2	48.401	1.950	49.283	2.022
AS325	AS325C3	66.177	1.938	66.781	2.021
AS325	AS325C4	33.386	1.958	33.117	2.021
AS325	AS325F	39.771	1.950	39.523	2.020
AS325	AS325M	39.703	1.956	39.771	2.022
AS326	AS326C11	38.169	1.951	38.036	2.019
AS326	AS326C12	36.175	1.954	36.419	2.020
AS326	AS326C13	40.898	1.948	40.611	2.019

AS326	AS326C21	45.324	1.945	45.246	2.018
AS326	AS326C22	30.371	1.963	30.296	2.022
AS326	AS326F1	35.013	1.954	34.745	2.020
AS326	AS326F2	45.869	1.947	46.062	2.020
AS326	AS326GM	39.594	1.950	39.659	2.022
AS326	AS326M1	39.885	1.955	40.347	2.023
AS326	AS326M2	35.426	1.959	35.721	2.023
AS328	AS328C1	40.217	1.955	40.185	2.022
AS328	AS328C2	41.411	1.951	41.530	2.022
AS328	AS328C3	32.672	1.961	32.984	2.024
AS328	AS328F	33.969	1.957	33.770	2.022
AS328	AS328GM	42.008	1.952	42.278	2.023
AS328	AS328M	45.289	1.950	45.638	2.024

Table 5-2 Ts/Tv ratio evaluation of variant filtration.

Presented in the table are genotype variant transitions (Ts) and Transversions (Tv) across variant sites per individual within Cohort 3. Transitions are defined as a change of purine bases or pyrimidine bases, i.e. A with G or C with T. Transversion are defined as changes between purine and pyrimidine bases, i.e. A with C/T, C with G or G with T. The mean depth of coverage across variant sites is also reported per individual.

5.3.3 Variant annotation and filtration

Annotation by dbNSFP for the post-QC variant call-set is detailed in Table 5-2. dbNSFP annotates all non-synonymous variation according to the specified parameters. In this case dbNSFP has annotated with parameters “-p -g -v hg19” (Liu *et al.*, 2020). Following annotation, variants were filtered following the framework previously applied in Chapter 3 and Chapter 4. This has been summarised in Figure 5-2. This filtering strategy results in the isolation of rare, putatively pathogenic SNVs with evidence of association in autism. These variants are further subset on the basis of penetrance as determined by family genotypes, enabled by extended pedigrees analysed within this cohort as follows within this chapter.

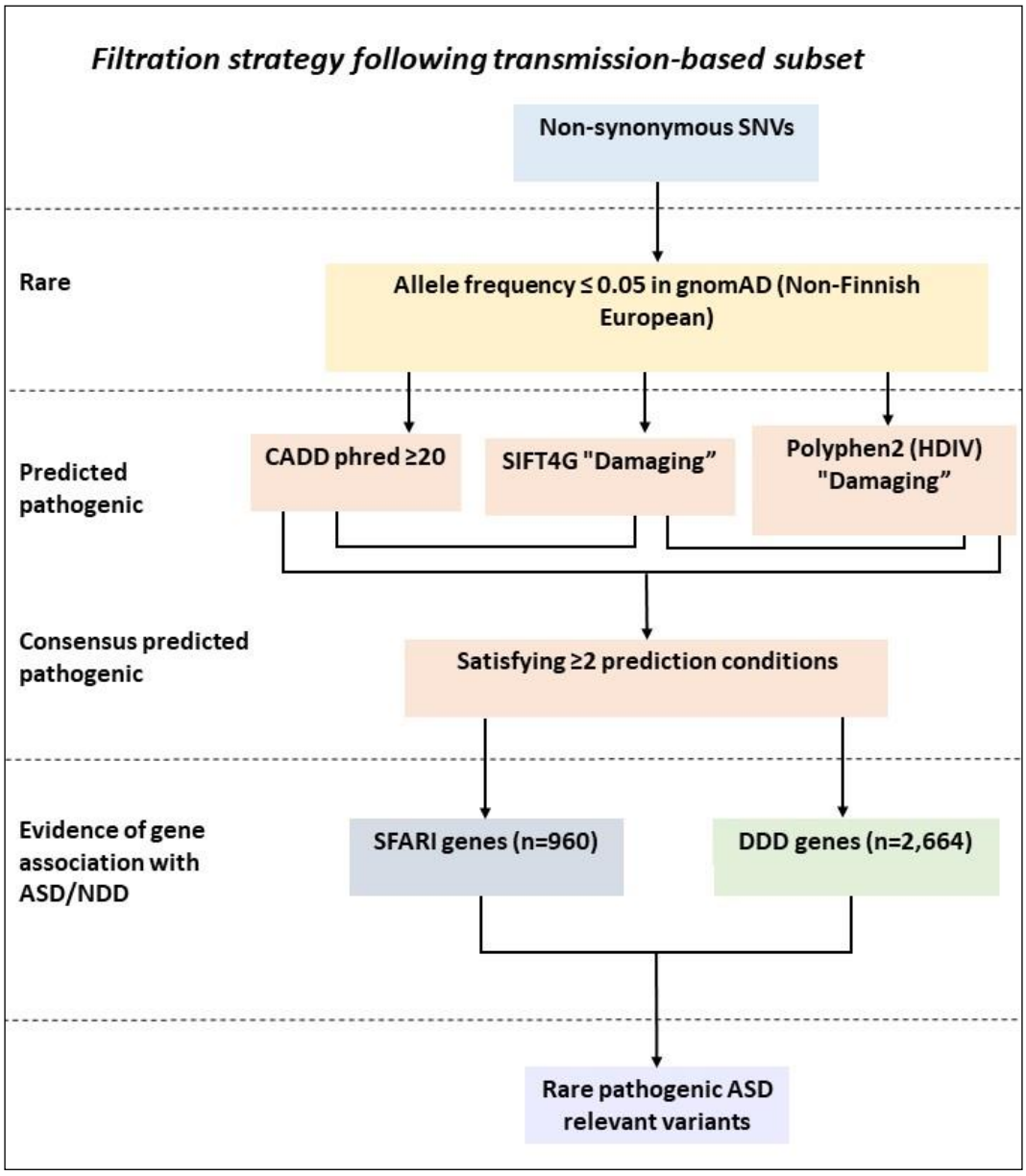


Figure 5-2 Flow of variant filtering.

Arrows show the direction of flow from each level of filtering (specified on the left). SFARI refers to Simons Foundation Autism Research Initiative Gene Module (Abrahams *et al.*, 2013). *DDD refers to the gene2phenotype database arising from the DDD study* (Wright *et al.*, 2015).

5.3.4 Pedigree AS324 variant isolation

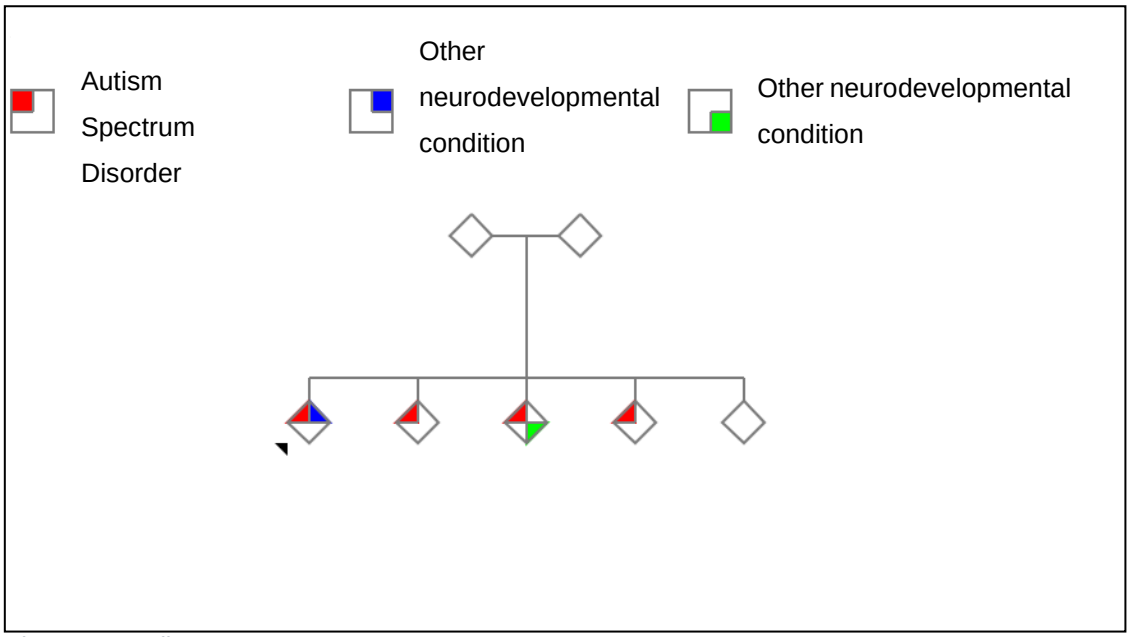


Figure 5-3 Pedigree AS324.

FID	IID	Non-synonymous SNVs	Rare (<0.05)	Consensus Predicted Pathogenic	Evidence of gene association with Autism/NDD
AS324	AS324C1	11,719	1,014	276	39
AS324	AS324C2	11,668	969	251	36
AS324	AS324C3	11,829	1,047	276	37
AS324	AS324C4	11,808	1,043	262	39
AS324	AS324C5	11,728	1,028	253	41
AS324	AS324F	11,591	1,023	258	37
AS324	AS324M	11,946	1,029	278	42

Table 5-3 Variant counts through variant prioritisation in pedigree AS324.

Variants were prioritised as outlined in Figure 5-2 and variant counts are presented here through prioritisation stages as filtered from left to right within the table. Non-synonymous SNVs are the complete set of high-confidence variants annotated by dbNSFP as non-synonymous. These variants were subset to rare variants with a MAF of <5% of the population. Variants were determined to be consensus predicted pathogenic when fulfilling two or more of SIFT, Polyphen-2 or CADD thresholds for classification. The resulting variants were

interrogated at gene-level for their inclusion in SFARI Gene or DDD gene2phenotype as a measure of relevance to autism and other neurodevelopmental conditions.

5.3.4.1 Hypothesised variant transmission

Homozygous recessive variant(s) in affected offspring, with heterozygosity in parents, and heterozygosity or homozygous wildtype allele(s) in unaffected sibling are hypothesised to contribute to neurodevelopmental conditions within this pedigree (Figure 5-3).

5.3.4.2 Variant report

Affected family members (AS324C1, AS324C2, AS324C3 and AS324C4) were interrogated for homozygous variation. Of the rare predicted pathogenic variants in genes with evidence supporting their role in autism, no variants were found to be homozygous and shared between all affected individuals. No homozygous variation within the variant set was carried by AS324C1 or AS324C2. One homozygous variant within was shared between individuals AS324C3 and AS324C4, rs55742743 (*TTN;* chr2). However, this variant is not shared between all affected family members within the pedigree.

A total of 6 variants in the variant set were identified as shared between all affected individuals in either a heterozygous or homozygous state. These variants are rs1800556 (*ACADS*; chr12), rs146665183 (*DLL4;* chr15), rs200546805 (*ANKRD11;* chr16), rs55742743 (*TTN;* chr2), rs1801208 (*WFS1;* chr4) and rs78008536 (*RELN;* chr7). None of these six candidate heterozygous variants were absent in all unaffected family members ruling out a heterozygous mode of pathogenicity of any single variant.

5.3.5 Pedigree AS325 variant isolation

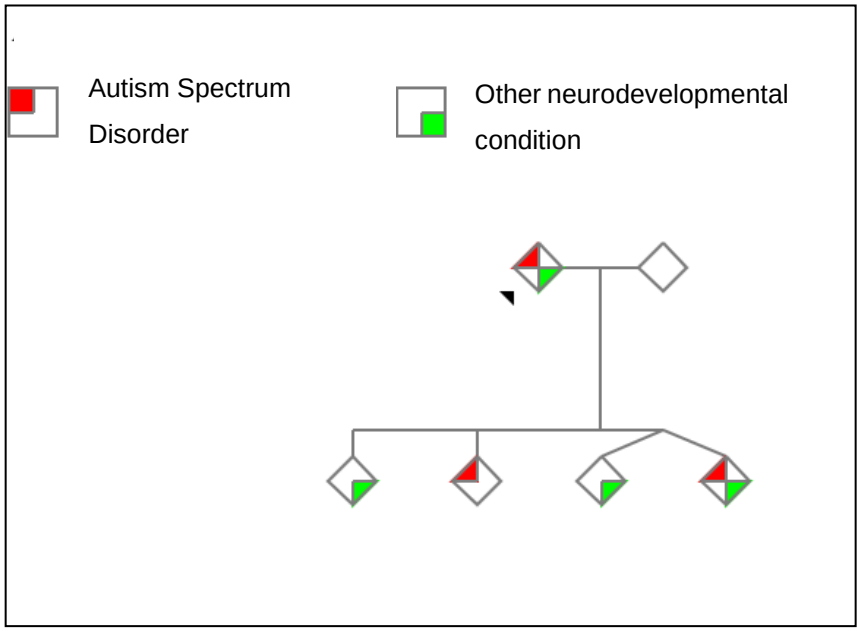


Figure 5-4 Pedigree AS325.

FID	IID	Non-synonymous SNVs	Rare (<0.05)	Consensus Predicted Pathogenic	Evidence of gene association with Autism/NDD
AS325	AS325C2	11,968	977	294	41
AS325	AS325C3	11,977	1,020	297	35
AS325	AS325C4	11,825	1,022	307	36
AS325	AS325F	11,712	925	277	36
AS325	AS325M	12,152	1,026	319	42

Table 5-4 Variant counts through variant prioritisation in pedigree AS325.

Variants were prioritised as outlined in Figure 5-2 and variant counts are presented here through prioritisation stages as filtered from left to right within the table. Non-synonymous SNVs are the complete set of high-confidence variants annotated by dbNSFP as non-synonymous. These variants were subset to rare variants with a MAF of <5% of the population. Variants were determined to be consensus predicted pathogenic when fulfilling two or more of SIFT, Polyphen-2 or CADD thresholds for classification. The resulting variants were interrogated at gene-level for their inclusion in SFARI Gene or DDD gene2phenotype as a measure of relevance to autism and other neurodevelopmental conditions.

5.3.5.1 Hypothesised variant transmission

Maternal inherited variant(s) present in all affected offspring and absent in the unaffected father are hypothesised to contribute to neurodevelopmental conditions within this

pedigree. Alternatively homozygous variant(s) present in all affected individuals with paternal heterozygous may be pathogenic (Figure 5-4).

5.3.5.2 Variant report

Initial investigation of the homozygous rare, predicted pathogenic autism-relevant variant set within this pedigree identified one candidate variant, rs1800328 (*OTC; chrX*). This homozygous variant was present in the maternal sample only and was not found to be shared with affected offspring within the pedigree.

Expanding beyond homozygous variation, this pedigree was interrogated for all variants within the variant set (homozygous and heterozygous) maternally inherited by all affected offspring and absent in the paternal genome. This search yielded 5 heterozygous candidate variants, rs77444104 (*NECTIN4*; chr1), rs1800729 (*TSC2*; chr16), rs61747224 (*TSHZ3;* chr19), rs78676079 (*SLC26A2*; chr5) and rs41277797 (*FKTN;* chr9).

5.3.6 Pedigree AS326 variant isolation

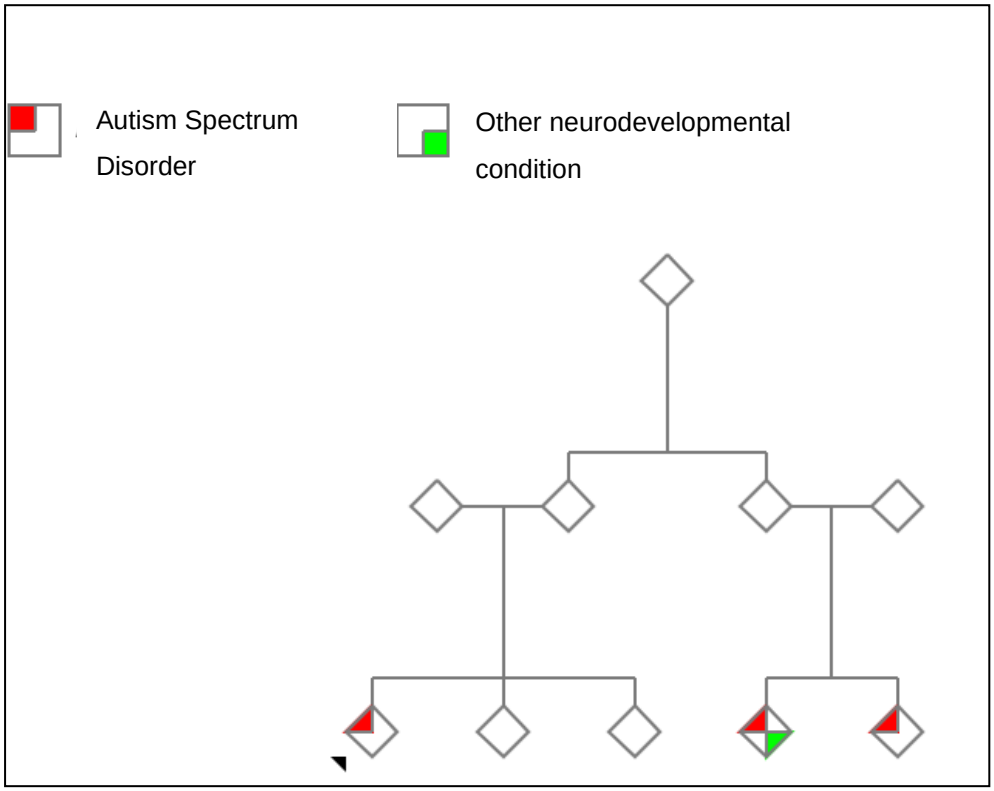


Figure 5-5 Pedigree AS326.

FID	IID	Non-synonymous SNVs	Rare (<0.05)	Consensus Predicted Pathogenic	Evidence of gene association with Autism/NDD
AS326	AS326C11	11,678	984	282	42
AS326	AS326C12	11,861	957	282	52
AS326	AS326C13	11,698	924	274	48
AS326	AS326C21	11,610	947	258	36
AS326	AS326C22	11,458	918	238	30
AS326	AS326F1	11,682	941	288	47
AS326	AS326F2	11,716	922	252	35

AS326	AS326GM	11,736	1,013	265	33
AS326	AS326M1	11,893	975	259	42
AS326	AS326M2	11,833	989	267	42

Table 5-5 Variant counts through variant prioritisation in pedigree AS326.

Variants were prioritised as outlined in Figure 5-2 and variant counts are presented here through prioritisation stages as filtered from left to right within the table. Non-synonymous SNVs are the complete set of high-confidence variants annotated by dbNSFP as non-synonymous. These variants were subset to rare variants with a MAF of <5% of the population. Variants were determined to be consensus predicted pathogenic when fulfilling two or more of SIFT, Polyphen-2 or CADD thresholds for classification. The resulting variants were interrogated at gene-level for their inclusion in SFARI Gene or DDD gene2phenotype as a measure of relevance to autism and other neurodevelopmental conditions.

5.3.6.1 Hypothesised variant transmission

Maternal inherited variant(s) with homozygosity in shared across affected individuals and absent from unaffected individuals are hypothesised to contribute to neurodevelopmental conditions within this extended pedigree (Figure 5-5).

5.3.6.2 Variant report

Initial investigation of the homozygous rare, predicted pathogenic autism-relevant variant set within this pedigree identified one shared variant, rs113964173 (*MYH11;* chr16), in AS326C11 and AS326C12. As this variant is shared by an affected and unaffected sibling it is not considered as being causative. Three homozygous variants were identified in the unaffected fathers within the pedigree (rs34144324 (*GRID2*; chr4*)* (AS326F1), rs753740777 (*POLA1*; chrX) (AS326F2) and rs1800273 (*DMD;* chrX) (AS326F2) and for this reason are not considered to be causative.

Evaluation of the complete variant set (homozygous and heterozygous) identified 5 heterozygous variants shared across all affected individuals (AS326C11, AS326C21 and AS326C22). These variants are rs146798796 (*TPP1*; chr11), rs149558764 (*LRP6*; chr12), rs116105292 (*TDO2*; chr4), rs1059582 (*HLA-DRB1*; chr6) and rs779037714 (*TRPV6*; chr7). None of these 5 variants were absent across all unaffected individuals in the pedigree.

5.3.7 Pedigree AS328 variant isolation

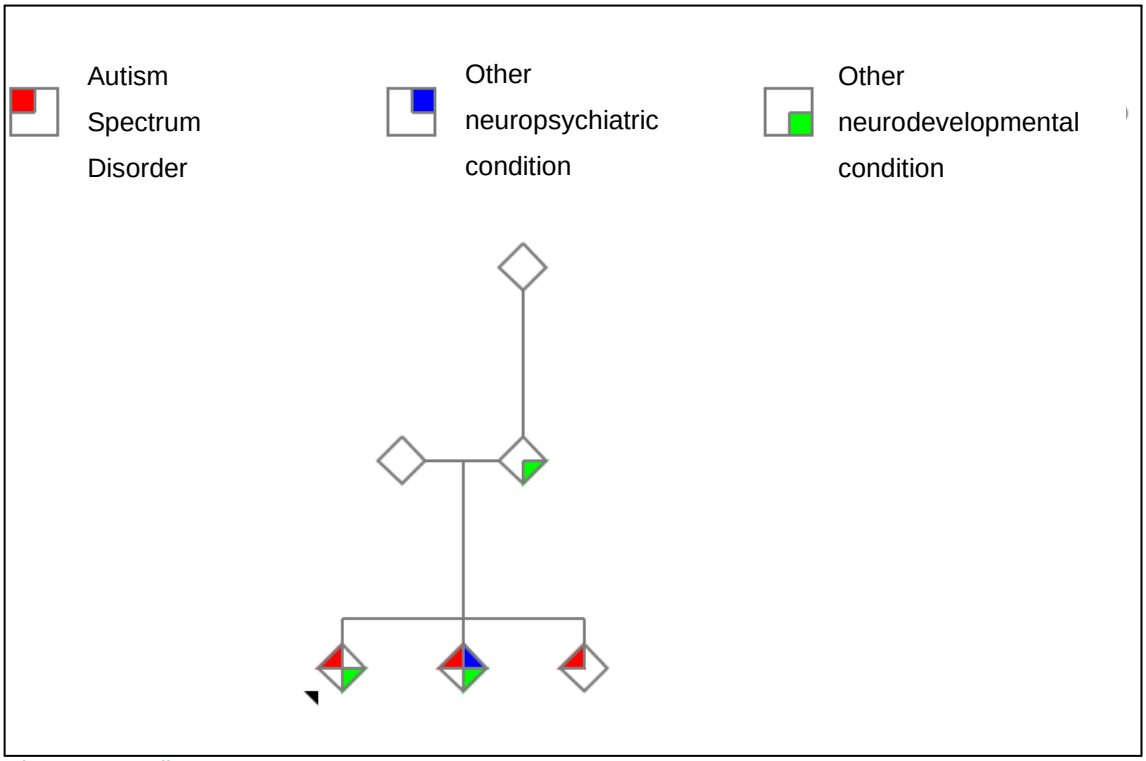


Figure 5-6 Pedigree AS328.

FID	IID	Non-synonymous SNVs	Rare (<0.05)	Consensus Predicted Pathogenic	Evidence of gene association with Autism/NDD
AS328	AS328C1	11,644	926	281	38
AS328	AS328C2	11,507	896	257	31
AS328	AS328C3	11,993	996	314	45
AS328	AS328F	11,824	958	267	40
AS328	AS328GM	12,053	1,033	308	53
AS328	AS328M	11,964	1,001	314	36

Table 5-6 Variant counts through variant prioritisation in pedigree AS328.

Variants were prioritised as outlined in Figure 5-2 and variant counts are presented here through prioritisation stages as filtered from left to right within the table. Non-synonymous SNVs are the complete set of high-confidence variants annotated by dbNSFP as non-synonymous. These variants were subset to rare variants with a MAF of <5% of the population. Variants were determined to be consensus predicted pathogenic when fulfilling two or more of SIFT, Polyphen-2 or CADD thresholds for classification. The resulting variants were interrogated at gene-level for their inclusion in SFARI Gene or DDD gene2phenotype as a measure of relevance to autism and other neurodevelopmental conditions.

5.3.7.1 Hypothesised mode of transmission

Maternal inherited variant(s) shared by all affected offspring are hypothesised to contribute to neurodevelopmental conditions within this pedigree. The causative variant(s) is expected to be heterozygous dominant in all affected individuals with wild type homozygosity in unaffected family members. Alternatively, the causative variant(s) may be homozygous in all affected individuals with heterozygosity in unaffected family members from whom a variant allele would be expected to have been transmitted (Figure 5-6).

5.3.7.2 Variant report

Initial investigation of the rare, predicted pathogenic autism-relevant variant set within this pedigree was carried out on maternal inherited variation shared between affected offspring. A total of 10 heterozygous variants within this call set were shared by all affected family members within the pedigree (AS328M, AS328C1, AS328C2 and AS328C3) (Table 5-7).

Of the 10 variants shared between affected family members presented in Table 5-7, 4 variants are not carried by unaffected members (AS328F and AS328GM). These variants are presented in Table 5-8.

dbSNP rsID	Chromosome	Position	Gene	Ref	Alt	Minor Allele Frequency
rs72648273	chr2	178,539,771	*TTN*	G	C	0.0043
rs199895260	chr2	178,589,803	*TTN*	C	T	0.0042
rs34144324	chr4	92,590,245	*GRID2*	C	T	0.0462
rs116105292	chr4	15,591,1563	*TDO2*	A	C	0.0430
rs1059582	chr6	32,584,240	*HLA-DRB1*	G	C	0.0094
rs2229792	chr6	33,163,724	*COL11A2*	G	A	0.0247
rs147304638	chr7	56,015,138	*PSPH*	G	A	0.0011
rs34293591	chr8	144,513,286	*RECQL4*	C	T	0.0275
rs118113109	chr12	52,568,196	*KRT74*	C	T	0.0193
rs200555745	chr16	88,716,580	*PIEZO1*	C	T	0.0009

Table 5-7 Variants shared between affected individuals in pedigree AS328.

Variant IDs are given by dbSNP rs IDs. Variant position is reported as 1-based with coordinates per GrCh38. Ref refers to the reference allele and alt refer to the alternative allele. Allele frequencies are reported from gnomAD v2.1.1 European (non-Finnish) combined exome and genome cohorts of high-quality genotypes, except for rs147304638 and rs34293591 where allele frequency is reported from dbSNP gnomAD European genomes report.

dbSNP rsID	Chromosome	Position	Gene	Ref	Alt	Minor Allele Frequency
rs72648273	chr2	178,539,771	*TTN*	G	C	0.0043
rs199895260	chr2	178,589,803	*TTN*	C	T	0.0042
rs147304638	chr7	56,015,138	*PSPH*	G	A	0.0011
rs34293591	chr8	144,513,286	*RECQL4*	C	T	0.0275

Table 5-8 Candidate causative variants in pedigree AS328.

Variant IDs are given by dbSNP rs IDs. Variant position is reported as 1-based with coordinates per GrCh38. Ref refers to the reference allele and alt refer to the alternative allele. Allele frequencies are reported from gnomAD v2.1.1 combined exome and genome cohorts of high-quality genotypes, except for rs147304638 and rs34293591 where allele frequency is reported from dbSNP gnomAD European genomes report.

Of note within this set of 4 candidate variants are missense variants rs72648273 and rs199895260 impacting the protein coding gene *TTN* on chromosome 2 as presented in Table 5-8. Both variants are maternally inherited by all affected offspring from affected AS328M. *TTN* encodes a large protein of striated muscle which plays a key role in muscle assembly, force transmission at the Z line of the sarcomere, and maintenance of resting tension in the I band region (Itoh-Satoh *et al.*, 2002). OMIM reports implication of variation in this gene in Cardiomyopathy, Muscular Dystrophy, Myofibrillar Myopathy With Early Respiratory Failure and Salih Myopathy (OMIM: 188840). In addition to these associations, SFARI gene collated 16 reports of autism with variation detected in *TTN* and designates this gene as a strong candidate gene causative of syndromic autism (Abrahams *et al.*, 2013). Single base missense variants in *TTN* have been identified in four unrelated probands from the Simon's Simplex Collection, adding evidence that rs72648273 and rs199895260 which are also missense variants within pedigree AS328 may be pathogenic (Iossifov *et al.*, 2012; O ' Roak *et al.*, 2012).

Homozygous rare predicted pathogenic variants were also investigated in autism-relevant genes identifying variation in the three affected offspring in this pedigree (AS328C1, AS328C2, AS328C3). Variant rs116105292 (*TDO2*, chr4) was identified as homozygous in AS328C1. Variant rs34144324 (*GRID2*, chr4) was identified as homozygous in AS324C3. These homozygous variants were unique to these individuals.

5.4 Discussion

5.4.1 Summary of results

This chapter describes the detection of rare, exonic SNVs in four pedigrees with multiple affected individuals. Family structure has been leveraged to hypothesise the mode of transmission of putatively pathogenic family, considering each pedigree family by family. This approach was successful in the isolation of predicted pathogenic variation in two of four families reported on here. This supports the use of a multiplex family approach in genomic studies of autism, particularly when sample size is limited and underpowered to perform linkage analyses.

However, the analysis pipeline applied here is limited to the isolation of rare exonic SNVs, while NGS technologies enable additional classes of variation to be detected, with evidence supporting their involvement in the genetic basis of autism, such as CNVs, SVs and tandem repeat expansions as will be discussed later. There are future opportunities to explore these classes of variation in Cohort 3 and there is potential to detect non-coding variation using the complete variant set sequenced by WGS.

5.4.1.1 Pedigree AS324

Variants were isolated by restricting to rare predicted pathogenic variation within the SFARI Gene (Abrahams *et al.*, 2013) and DDD gene2phenotype (Wright *et al.*, 2015) gene lists within this pedigree. Restricting variant discovery to the hypothesised mode of transmission identified variant rs55742743 in AS324C3 and AS324C4. This variant may be contributory to the neurodevelopmental phenotype of these individuals. As this variant was not identified in a homozygous state in the other affected individuals it can be determined that while the variant may be contributory in AS324C3 and AS324C4, it is likely not a single gene cause of the phenotype at family-level. Further investigation beyond this restrictive call set is necessary to identify novel variation, without existing evidence of association, which may be contributing to the burden of neurodevelopmental conditions within this pedigree.

5.4.1.2 Pedigree AS325

Restricting variant discovery to the hypothesised mode of transmission identified 5 candidate potentially contributory heterozygous variants shared between all affected family members. Further investigation through *in vitro/in vivo* functional evaluation is needed to determine

whether heterozygosity of these variants is likely to be pathogenic in these individuals. In addition, AS325C1 requires follow-up sequencing having failed.

5.4.1.3 Pedigree AS326

Restricting variant discovery to the hypothesised mode of transmission yielded no candidate variants. Further investigation beyond this restrictive call set is necessary to identify novel variation, without existing evidence of association, which may be contributing to the burden of neurodevelopmental conditions within this pedigree. Furthermore, variant rs113964173 shared by affected (AS326C11) and unaffected (AS326C12) siblings should be further investigated through *in vitro/in vivo* functional evaluation to determine relevance to the autism phenotype of AS326C11. Sex differences in the sibling carrying this variant may explain a variable manifestation of the condition resulting from a female protective effect in AS326C12.

5.4.1.4 Pedigree AS328

Restricting variant discovery to the hypothesised mode of transmission yielded 4 heterozygous candidate variants (Table 5-8). Of note within this set of 4 candidate variants are missense variants rs72648273 and rs199895260 impacting *TTN.* Both variants are maternally inherited by all affected offspring from affected AS328M. *TTN* has been classified by SFARI Gene as a gene score of 2S indicating strong evidence for implication in idiopathic autism as well as syndromic autism. Syndromic variant scoring by SFARI Gene classifies variants “that are associated with a substantial degree of increased risk and consistently linked to additional characteristics not required for an ASD diagnosis” (Abrahams *et al.*, 2013). Further investigation through *in vitro/in vivo* functional evaluation is needed to confirm pathogenicity of these heterozygous variants.

5.4.2 Conclusion

The variant filtration approach applied subset variants to those impacting genes with evidence for autism and neurodevelopmental condition association, as determined by presence in SFARI Gene and DDD gene2 phenotype. These gene lists were unrestricted in this filtration, i.e., genes were not subset to those with substantial evidence supporting the association as indicated by Gene score of 1 or 2 in SFARI (Abrahams *et al.*, 2013) or assigned “High Confidence” by DDD (Wright *et al.*, 2015). Functional analyses, for example the CRISPR/Cas9-induced mutagenesis of *DDX3X,* a monogenic neurodevelopmental cause, are required to robustly assign causation to the variants identified through these analyses

(Radford *et al.*, 2022). In efforts for discovery of gene-phenotype association, analysis should extend beyond this gene-set enabling variant detection across all genes and non-coding regions of the genome. The approach applied here is a first pass analysis which with increased sample size should be evaluated at a more widespread level across the genome or in parallel to large-scale WGS efforts.

Chapter 6. Determining the clinical utility of gene panels in autism; a study of diagnostic yield and relevance.

The contents of this chapter have been published in part as the following article:
Ní Ghrálaigh, F. *et al.* (2022) 'Brief Report: Evaluating the Diagnostic Yield of Commercial Gene Panels in Autism', Journal of Autism and Developmental Disorders 2022. Springer, pp. 1–5. doi: 10.1007/S10803-021-05417-7 (Appendix IV-II).

Presentations arising from the contents of this chapter:
"Determining the clinical utility of gene panels in autism; a study of diagnostic yield, relevance, and penetrance." Fiana Ní Ghrálaigh, Thomas Dinneen, Ellen McCarthy, Daniel N. Murphy, Louise Gallagher & Lorna M. Lopez
Poster presented at the World Congress of Psychiatric Genetics, October 2021 (Appendix III-II).

"Evaluating the diagnostic yield of commercial gene panels in autism." Fiana Ní Ghrálaigh, Ellen McCarthy, Daniel N. Murphy, Louise Gallagher & Lorna M. Lopez
Poster presented at the Irish Society for Human Genetics, September 2021 (Appendix III-III).

"A Search for Rare Variants in a Family-Based Study of ASD." Fiana Ní Ghrálaigh, Jessica E. Smith, Elaine Kenny Louise Gallagher & Lorna M. Lopez
Poster presented at the World Congress of Psychiatric Genetics, October 2018, and the Irish Society for Human Genetics, September 2018 (Appendix III-VII).

6.1 Abstract

This chapter aims to overcome challenges in translation of genomic findings to clinical application. The analysis performed in this chapter adds to this discussion of the heterogeneity of clinical sequencing tests, “gene panels,” marketed for application in autism by evaluating their clinical utility and considering gene selection. This analysis demonstrates the low diagnostic yield of autism gene panels currently. In addition, this chapter determines the clinical relevance of genes included within these panels. This work concludes that commercial gene panels marketed for autism are currently of limited clinical utility.

6.2 Introduction

Clinical genetic diagnosis is limited to the identification of rare causative variants for evaluation of symptomatic individuals at present. Diagnostic genetic testing in neurodevelopmental conditions and neuropsychiatric conditions is limited.

6.2.1 The benefits of genetic diagnosis in psychiatric conditions

The broad opportunity of precision medicine is to advance therapeutics. At the individual level there are benefits to receiving a genetic diagnosis in psychiatry. Understanding cause is of profound importance to individuals and families living with neurodevelopmental and neuropsychiatric conditions. The International Society of Psychiatric Genetics propose in their consensus statement on genetic testing that the “*identification of known pathogenic variants may help diagnose rare conditions that have important medical and psychiatric implications for individual patients and may inform family counselling*” (*Genetic Testing Statement | ISPG - International Society of Psychiatric Genetics*, no date). Specifically, genetic diagnosis establishes the primary etiology of clinical diagnoses. This may enable healthcare providers to provide genetic or reproductive counselling for affected individuals and their families. A genetic diagnosis give opportunity for provision of personalised medicine, such as provision of anticipatory medical guidance and treatment plans (Moeschler *et al.*, 2014).

A further benefit to receiving a genetic diagnosis may be the opportunity to take part in targeted research, such as variant specific clinical trials. While there are currently no genotype-guided precision therapies available for use in autism, many examples of treatments are available or in development for variant specific forms of epilepsy, another

neurodevelopmental condition. One such example is treatment of *SCN1A*-related epilepsy with an antisense oligonucleotide to block exon splicing. This treatment almost entirely prevents seizures and resulting death in mouse models and is currently in Phase I clinical trial (Carvill *et al.*, 2018; Han *et al.*, 2020). With the increasing number of genes associated with autism comes an increase in potential targets for treatment. Genomic discovery in autism will enable to discovery of molecular targets for autism, and potential precision medicine gene targets for rare syndromic causes of autism. Beyond clinical treatment strategies, a genetic diagnosis can be of great personal utility to an affected individual by enabling access to etiology-specific advocacy organisations for example 22q11Ireland (https://www.22q11ireland.org).

Pharmacogenomic testing is another stream of genetic testing in neuropsychiatric conditions; with the aim of predicting drug response rather than aiding diagnosis. This area of genomics research in neuropsychiatric conditions shows huge potential in medical management of conditions, with genotype guided therapy leading to better patient response (Bousman *et al.*, 2019). An example of success in translation of pharmacogenomics into the clinical setting is *CYP2D6* testing to identify under and rapid metabolisers of selective serotonin reuptake inhibitors in the context of treatment of major depressive disorder (Hall-Flavin *et al.*, 2013). Beyond this introduction to pharmacogenomics, unless otherwise specified, in the context of this thesis, genomics of psychiatric conditions refers only to genomics efforts to identify causative variation, and further discussion is beyond the scope of this thesis.

6.3 Genetic heterogeneity of autism

Genetic diagnosis in autism is limited by the ability to robustly determine the clinical relevance of putatively pathogenic genetic variation. Genomic research in autism is progressing quickly, enabled by advancements in NGS technologies and the subsequent establishment of large-scale sequencing cohorts and pedigree-based sequencing cohorts (Glahn *et al.*, 2019; Ní Ghrálaigh, Gallagher and Lopez, 2020). To date, more than 990 genes have been identified as having some link to autism (Abrahams *et al.*, 2013). Despite this progress, major challenges remain in the translation of findings from research to clinic (7.2).

At cohort-level, studies discovering “autism genes” are compounded by an apparent lack of specificity to autism. A candidate pathogenic variant may be evaluated, in most autism

cases, as being contributory to the genetic risk rather than being wholly causative of an individual's condition. For example, in individuals affected by both autism and ID, genes identified show relevance to both autism and other neurodevelopmental conditions (Myers, Challman, Bernier, *et al.*, 2020). For these reasons, the development of effective gene panels to aid autism diagnosis is extremely complicated. To date there are a number of curated gene lists for autism. These include genes involved biological pathways critical to brain development and function.

Despite progress, autism genomics has yet to reach the target of establishing a comprehensive gene list with clinical utility. This arises from the challenges with interpretation of the genetic and phenotypic heterogeneity of autism (Myers, Challman, Bernier, *et al.*, 2020), and the resulting the lack of a consensus in establishing a gene curation framework. As outlined in Chapter 4, approaches have been developed to address these challenges in gene-phenotype curation.

6.4 Interpreting the clinical relevance of genetic findings

Multi-disciplinary experts propose WES as a first-tier diagnostic test to be applied to the genomes of individuals affected by neurodevelopmental conditions (Srivastava *et al.*, 2019). The full potential for WGS in a clinical setting has yet to be determined; the additional costs and the technical demands of data processing and data storage, and data interpretation may not yet be justified for routine clinical use.

Targeted gene panels have been successfully developed for disease-specific use, such as in hereditary cancer (LaDuca *et al.*, 2019). The applicability of such gene panels in autism would at this time be extremely limited. At present no genes can be exclusively associated with autism, i.e. association in the absence of ID or other co-occurring neurodevelopmental conditions (Myers, Challman, Bernier, *et al.*, 2020).

Evaluation of the clinical implication of a given genetic variant is a further level of variant annotation. A key distinction currently, is clinical investigation in research vs. clinical settings, with the latter requiring use of accredited clinical molecular laboratories, technologies, and analysis strategies. Existing variant information for those variants, which are characterised to have clinical significance, are accessible through databases such as OMIM, ClinVar and Human Gene Mutation Database (HGMD) (McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, no date; Stenson *et al.*, 2017; Landrum *et al.*, 2018). Commercial genome analysis platforms

present a useful option to perform variant interpretation with minimal need for computational infrastructure and bioinformatic analysis expertise. Some examples include Complete Genomics, Agilent Alissa Interpret and SOPHiA DDM (Appendix III-V). These platforms benefit from the use of in-house algorithms to identify the variants most relevant to disease. However, patenting controls mean that there is often a lack of transparency in the methods of the variant ranking systems used by these platforms.

6.4.1 Hypothesis and aims

Despite the challenges in translation of genomic findings to clinical application, commercial gene panels are available and marketed for use in autism diagnosis. Hoang *et al.* (2018) evaluated many of these gene panels, clearly demonstrating their heterogeneity (Hoang, Buchanan and Scherer, 2018). Their survey shows large variability in the number of genes being tested by panels, lack of consensus in the genes selected for inclusion, as well as variability in the reporting of laboratory qualification and reporting protocols. The analysis performed in this chapter adds to this discussion of the heterogeneity of clinical sequencing tests, “gene panels,” marketed for application in autism by evaluating their clinical utility and considering gene selection.

The aims of this thesis chapter are:

1) Evaluate the diagnostic yield of commercial gene panels marketed for use in autism and determine the relevance gene selection for these panels.
2) Determine the overlap of ACMG59 genes and autism-related gene lists.

6.5 Results

6.5.1 Evaluating the diagnostic yield of commercial gene panels in autism

Here we estimate the clinical utility of commercial gene panels marketed for use in autism. Diagnostic yield, which is the proportion of cases interrogated for which a genetic cause can be determined, is a strong measure of the clinical utility of a sequencing technology.

6.5.1.1 Identifying autism gene panels

Commercial gene panels marketed for use in autism, collated through literature search and systematic searching (October 2020-January 2021), are presented in Table 6-1. Gene panels marketed for use in autism were identified and collated through the following approaches: web browser search (search terms *"autism gene panel"*, *"ASD gene panel"*, *"sequencing tests for autism spectrum disorder"*, *"gene panels for autism testing"* and *"autism genetic testing"*), gene panels analysed by Hoang *et al.* (2018) (Hoang, Buchanan and Scherer, 2018) and Genomics England PanelApp (search terms *"Autism", "ASD"*) (Martin *et al.*, 2019). Panels identified for which gene lists were not provided were excluded from analyses (CGC genetics "Autism" panel & Michigan Medicine "Autism/ Intellectual Disability Panels").

Gene Panel Provider	Source
Ambry Genetics	https://www.ambrygen.com/providers/genetic-testing/62/neurology/autismnext
Asper Neurogenetics	https://www.asperbio.com/asper-neurogenetics/autism-spectrum-disorders-ngs-panel/
Blueprint Genetics	https://blueprintgenetics.com/tests/panels/neurology/autism-spectrum-disorders-panel/
Center for Human Genetics	https://www.ncbi.nlm.nih.gov/gtr/tests/529181/
Centogene	https://www.centogene.com/science/centopedia/syndromic-autism-gene-panel.html
Centogene	https://www.centogene.com/diagnostics/ngs-panels/neurology.html
EGL Genetics	https://www.egl-eurofins.com/tests/MM021
Fulgent Genetics	https://www.fulgentgenetics.com/Autism
GeneDx	https://www.genedx.com/test-catalog/available-tests/autismid-xpanded-panel/
GENETAQ	http://genetaq.com/en/catalogue/test/autism
Genomics England PanelApp	https://panelapp.genomicsengland.co.uk/panels/657/
Greenwood Genetic Centre	https://www.ggc.org/test-finder-item/syndromic-autism-sequencing-panel
GX Sciences	https://www.gxsciences.com/genetic-testing-autism-s/202.htm
MNG Laboratories	https://mnglabs.com/tests/NGS325/comprehensive-intellectual-disability-autism-ngs-panel-and-copy-number-analysis-mtdna
Munroe-Meyer Institute	https://www.unmc.edu/mmi/geneticslab/_documents/gene-lists/genelist-p-autism-v3-117.pdf
Prevention Genetics	https://www.preventiongenetics.com/testInfo?val=Autism+Spectrum+Disorders+%28ASD%29+Panel
Reference Laboratory Genomics	https://www.ncbi.nlm.nih.gov/gtr/tests/559901/overview/
Sema4	https://sema4.com/products/test-catalog/comprehensive-autism-spectrum-disorders-panel-228/

Table 6-1 Source of autism-relevant gene panels investigated.

Outlined are the company names for 18 targeted gene panels marketed for use in autism with the sources at which gene lists and descriptions were obtained (collated October 2020-January 2021).

6.5.1.2 Refining gene lists

Gene lists corresponding to each of the targeted gene panels presented, were collated, and refined. Each gene panel identified provided a list of genes targeted by the probes. By nature, these gene lists arise from a variety of sources and were compiled at varying times. For this reason, gene lists were run through HUGO Gene Nomenclature Committee (HGNC) Multi-Symbol Tool (Version: 2021-01-06 update). Where the gene symbol reported by the provider is an approved gene symbol in HGNC, it is used in analyses. Where the gene symbol is no longer approved by HGNC, it was updated to the approved gene symbol given by HGNC. A small number of deviations occurred that could not be resolved, which resulted in the removal genes from the analyses. Gene counts reported in these analyses reflect these updates.

6.5.1.3 Estimating the diagnostic yield of commercial gene panels in autism

Diagnostic yield, which is the proportion of cases interrogated for which a genetic cause can be determined, is a strong measure of the clinical utility of a sequencing technology. Feliciano *et al.* (2019) estimate the diagnostic yield of WES to be 10.4% in the initial 457 families enrolled in the SPARK cohort (Feliciano *et al.*, 2019). A 'likely pathogenic' variant, a variant with greater than 90% certainty of being disease causing, was identified in a further 3.4% of families studied. This estimate comes from the identification of a variant that fulfils either the 'likely pathogenic' or 'pathogenic' criteria, according to ACMG standards (Richards *et al.*, 2015).

To determine the clinical utility of each autism gene panel, variants meeting 'likely pathogenic' or 'pathogenic' criteria in the SPARK cohort can be limited to those within the gene set of each panel, respectively. In doing so, we ask how many of the pathogenic variants identified by Feliciano *et al.* would have been identified in the SPARK cohort with application of an autism gene panel, instead of application of WES.

Clinically relevant variants, as identified and characterised by WES in the Simon's Powering Autism Research Knowledge (SPARK) cohort, were used to determine the clinical utility of each panel. Variants included in these analyses are those reported in Feliciano *et al.* (2019), comprising inherited and *de novo* SNVs, indels and CNVs (Feliciano *et al.*, 2019). Reported chromosomal abnormalities were not included. Gene lists were assembled to include those for which clinically relevant SNVs and indels could be defined and those that fall within the boundaries of clinically relevant CNVs. While

targeted gene panels lack the ability to define CNV boundaries, genes within these variants will appear as deleted or duplicated, thus a variant site will be detected. For this reason, this class of variation has been considered in these analyses where it would otherwise be excluded.

6.5.1.4 Determining and reporting diagnostic yield

Diagnostic yield was calculated as the proportion of individuals with relevant variants that would have been identified in the SPARK WES cohort if using the gene lists for each gene panel. Diagnostic yield was determined by cross-referencing the gene list of each gene panel with the lists of implicated genes in the SPARK cohort.

The number of individuals in the cohort was taken as 472 affected individuals (465 offspring and seven parents) as detailed by Feliciano *et al.* (2019). In keeping with this study, 13 individuals, those in families self-reporting a genetic diagnosis were not included in the estimates of diagnostic yield. With this justification, diagnostic yield was calculated as the number of individuals with a relevant variant, as a percentage of the total cohort of 459 affected individuals without a genetic diagnosis.

The number of individuals for which a clinically relevant finding would have been identified by using each targeted gene panel is reported for both pathogenic and probable pathogenic variants, as assigned by Feliciano *et al.* (2019). The diagnostic yield of each gene panel, estimated with respect to Feliciano *et al.* (2019) analyses, is presented in Table 6-4. The diagnostic yields range from 0.22% to 10.02%, with most gene panels achieving a diagnostic yield below 3%.

6.5.1.5 Determining and reporting correlation

SFARI Gene is a database (all gene scores and genetic categories) of genes implicated in autism susceptibility (Version: 2021-01-13 release) (Table 6-2,Table 6-3). Each panel was assessed for overlap with SFARI Gene to determine the proportion of genes included on commercial panels that have known relevance to autism. Where necessary, the SFARI Gene list (n=1,003) was updated to HGNC approved gene symbols (n=5) and genes with symbol mismatch (n=3) were removed. The number of genes targeted by each panel that overlap with SFARI Gene are estimated as a percentage of the total genes in the panel. SFARI Gene was subset to high-confidence autism-associated

genes, assigned as such based on SFARI Gene scoring of 1 or 2. Percentage overlap was calculated for the subset of high-confidence genes and presented.

Software	Version
RStudio	4.0.3
Tidyverse	1.3.0

Table 6-2 Software versions used in analyses.

Input Data	Version	Source
SFARI Gene list	01-13-2021 release	https://gene.sfari.org/database/human-gene/
Gene panel gene list	Up to date as of January 2021	As specified in 6.5.1.1.
Clinically relevant variant set	As published	Feliciano *et al. (2019)*

Table 6-3 Data input files with sources and versions used in analyses.

The degree of overlap of gene lists of each gene panel, with SFARI Gene is presented in Table 6-4 The overlap is expressed as the percentage of genes interrogated by each panel that are also included in SFARI Gene. Most genes included in these gene panels have some relevance to autism, illustrated by the inclusion of a large proportion the panel-specific genes in the SFARI Gene database (Abrahams *et al.*, 2013). SFARI Gene is a collated list of genes for which there is evidence of association with autism and is used here as an arbitrary measure of 'relevance' of genes included with autism. The Genomics England PanelApp (Autism Version 0.2) was used as a positive control in the analysis. Its gene list is derived from SFARI Gene, reflected in the 100% overlap with the database. Conversely, Gx Sciences (Developmental Nutrigenomic Panel) has an overlap of just 15.15% of genes with those in SFARI Gene, reflecting the more specific intended application of this gene panel (nutrigenomics rather than diagnostics). SFARI Gene was subset to high-confidence autism-associated genes, assigned as such based on SFARI Gene scoring of 1 or 2 (Table 6-4).

Pearson's product-moment correlation was computed with 16 degrees freedom for diagnostic yield and number of genes targeted and for diagnostic yield against

percentage overlap with SFARI Gene (all genes). Diagnostic yield of the gene panels and size of the panel were found to be positively correlated, ($r = 0.82$, $p = 3.033e-05$), indicating an increased number of genes per gene panel enables detection of a clinically relevant variant in a greater number of individuals. No significant correlation between percentage overlap with SFARI Gene and diagnostic yield was detected.

Service provider	Panel name	Number of genes targeted	Percentage overlap with SFARI Gene		Diagnostic yield in SPARK
			SFARI Gene All Genes	SFARI High Confidence Genes (Scores 1 and 2)	
Ambry Genetics	AutismNext Panel	72	87.5%	76.39%	2.61%
Asper Neurogenetics	Autism Spectrum Disorders NGS Panel	76	88.16%	71.05%	2.83% (0.22%)
Blueprint Genetics	Autism Spectrum Disorders Panel	75	45.33%	36%	1.53% (0.44%)
Center for Human Genetics	Autism Spectrum Disorder 53-Gene Panel	53	84.91%	45.28%	1.96% (0.22%)
Centogene	Syndromic Autism Gene Panel	50	88%	76%	2.4% (0.22%)
Centogene	Intellectual Disability Panel	599	43.41%	24.54%	5.23% (1.31%)
EGL Genetics	Autism Spectrum Disorders Tier 2 Panel	62	74.19%	66.13%	2.18%
Fulgent Genetics	Autism NGS Panel	121	76.86%	55.37%	4.36% (0.44%)

GeneDx	Autism/ID Xpanded Panel	2641	20.64%	10.98%	10.02% (3.49%)
GENETAQ	Autism	27	92.59%	66.67%	1.53%
Genomics England PanelApp	Autism (Version 0.20)	733	100%	42.7%	7.63% (1.96%)
Greenwood Genetic Centre	Syndromic Autism Sequencing Panel	83	80.72%	69.88%	3.05%
GX Sciences	Developmental Nutrigenomic Panel	33	15.15%	0%	0.22%
MNG Laboratories	Comprehensive Disability/Autism Panel	1345	19.85%	12.04%	6.1% (1.3%)
Munroe-Meyer Institute	Autism/Intellectual Disability/Multiple Anomalies Panel	117	55.56%	41.88%	2.4% (0.22%)
Prevention Genetics	Autism Spectrum Disorders Panel	170	95.29%	90.59%	6.32% (0.44%)
Reference Laboratory Genomics	Autism Spectrum Disorders (Expanded Panel)	77	77.92%	64.94%	3.05% (0.44%)
Sema4	Comprehensive Autism Spectrum Disorders Panel (228)	228	57.46%	43.42%	4.79% (0.87%)

Table 6-4: Diagnostic yield of gene panels marketed for use in autism.

Presented are gene panels relevant to autism. Diagnostic yield of gene panels marketed for use in autism. Presented are gene panels relevant to autism. The number of genes present in each gene panel are correct as of January 2021. Gene lists were updated to HGNC approved gene symbols where necessary. Percentage overlap with SFARI is estimated as the proportion of genes within each respective gene list appearing in SFARI Gene (01-13-2021 release). This overlap is presented for both the complete SFARI Gene gene lists and the High Confidence SFARI Genes only (Scores 1 and 2). Diagnostic yield is estimated as the number of individuals for which a genetic cause of autism was identified as a proportion of those investigated (459 affected individuals for which no genetic diagnosis was previously reported). Pathogenic variation is considered as variants listed in Feliciano et al. (2019). Variants considered are de novo and inherited SNVs, indel variants, and CNVs. Diagnostic yield of pathogenic variation is listed, with the additional diagnostic yield achieved by inclusion of probable pathogenic variants listed in brackets alongside.

6.5.2 Evaluating the inclusion of ACMG59 in autism and neurodevelopmental condition gene lists

6.5.2.1 Determining the overlap of autism gene lists with ACMG59

The ACMG has published recommendations for reporting genetic variation from clinical exome and genome sequencing (Miller *et al.*, 2021). Within these genes, 59 at time of analysis, variants that may be pathogenic have been characterised in ClinVar (Landrum *et al.*, 2018). These variants are recommended for reporting as they are likely to be informative to the individual carrying the variant, and potentially their family members. Variants identified within these genes are likely to be secondary findings, i.e., unrelated to the primary purpose of performing the test. However, these genes may be of neurodevelopmental relevance and subsequently may be the target of genomic interrogation in these individuals in a research setting.

Here three autism-relevant genes lists (SFARI Gene, DDD gene2phenotype and the genes targeted by autism gene panels) were interrogated for overlap with the ACMG gene list of clinically actionable genes (Table 6-6) (Figure 6-1). Six genes, included in all autism gene lists investigated are found on the ACMG59 gene list (*PTEN, TSC1, TSC2, BRCA2, FBN1* and *SMAD4*). Furthermore, three of these six genes (*PTEN, TSC1* and *TSC2*) are scored as “High Confidence” for autism-association and two of the six genes (*FBN1* and *SMAD4*) are scored as “Strong Candidate” for autism-association.

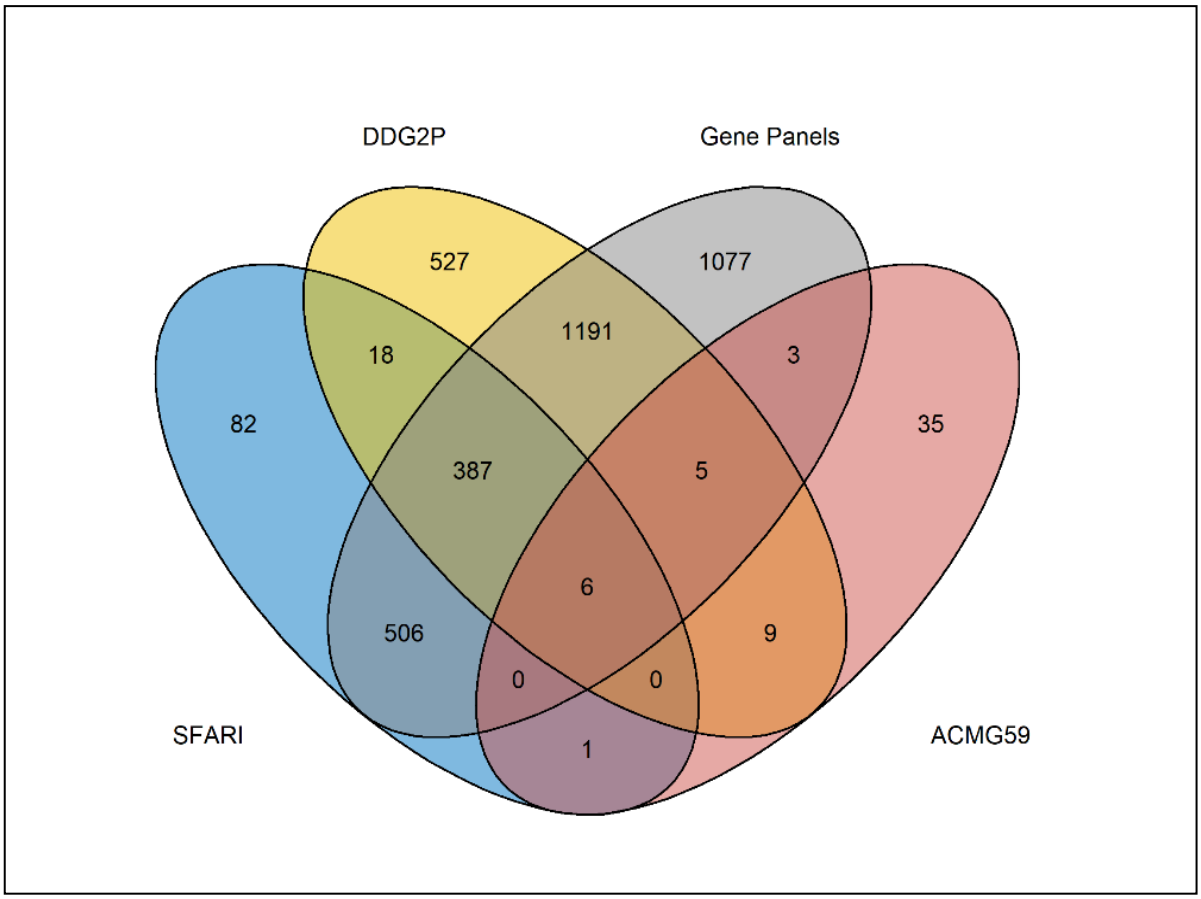


Figure 6-1 The overlap of autism gene lists with ACMG59.

Differentiated by colour are three autism-relevant gene lists (Blue-SFARI Gene; Yellow- DDD ND Gene to Phenotype gene lists; Grey- Collated gene list arising from commercial gene panels marketed for use in autism). The overlap is shown between these gene lists, and between each gene list and ACMG59, in red. These gene lists are further detailed in Table 6-4. Counts represent the number of genes within each category of overlap.

The autism gene list with the most substantial overlap with ACMG59 is DDD gene2phenotype, notably also the gene panel targeting the largest number of genes (n=2,426). Of particular interest are the three genes interrogated by gene panels and featuring on the ACMG59 gene lists, which are not interrogated by SFARI Gene or DD2GP. Given that these genes are not interrogated by either SFARI Gene or DDD gene2phenotype, the evidence supporting the inclusion of these genes in the gene panels is of concern and will be discussed later.

This overlap of genes in the clinical gene-sets presented in Figure 6-1 is further quantified by jaccard similarity coefficient, presented in

Table 6-5, measured as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Gene Set	Gene Set	Jaccard Similarity Coefficient
SFARI	ACMG59	0.006610009
SFARI	Gene Panels	0.2496529
SFARI	DDDG2P	0.1363184
DDDG2P	ACMG59	0.008090615
DDDG2P	Gene Panels	0.3663823
Gene Panels	ACMG59	0.003941441

Table 6-5 Jaccard Similarity Coefficients of Clinical Gene set Overlaps. Presented are pair-wise jaccard similarity coefficients for the clinical gene sets under investigation. The overlap is shown between these gene lists, and between each gene list and ACMG59, in red. These gene lists are further detailed in Table 6-4.

Gene List	Description	Number of Genes	Source/Reference	Release/Export
Query Gene Set				
ACMG59	The ACMG has published recommendations for reporting incidental findings in the exons of these genes	59	https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/ (Kalia *et al.*, 2017)	Version 2.0 06-04-2021 export
Autism Gene Set				
SFARI	SFARI Gene is a well-maintained database for the autism research community. This gene list collates genes implicated in autism susceptibility.	1,000	https://gene.sfari.org/ (Abrahams *et al.*, 2013) Gene lists has been refined as detailed in 6.5.1.2.	13-01-2021 release
DDD gene2phenotype	This gene lists are the neurodevelopmental condition "ND" gene list arising from the DDD study.	2,426	https://www.deciphergenomics.org/ddd/ddgenes (Wright *et al.*, 2015)	09-04-2021 export
Gene Panels	This gene lists includes all genes targeted by 18 autism commercial gene panels.	3,500	This gene list is compiled and refined as detailed in this chapter. (Ní Ghrálaigh *et al.*, 2022)	Current as of January 2021

Table 6-6 Description of gene lists used.

The column gene list lists the shortened name of each list as it is referred to within these analyses. Description gives context to the relevance of each gene list to these analyses. Columns 3-5 give further detail on the gene lists. Note that all gene symbols within these gene lists are HGNC approved.

Gene Symbol (HGNC)	Gene Description (GeneCards)	Interrogated by	Disease Phenotype (Coriell Institute)	GeneReviews®	SFARI Gene Score
ACTA2	Actin Alpha 2, Smooth Muscle	DDD	Marfan syndrome, Loeys-Dietz syndromes, and familial thoracic aortic aneurysms and dissections	(Milewicz and Regalado, 1993; Dietz, 2017; Loeys, 2017)	Not scored
ATP7B	ATPase Copper Transporting Beta	Gene Panels	Wilson disease	(Weiss, 1993)	Not scored
BRCA1	BRCA1 DNA Repair Associated	DDD	Hereditary breast and ovarian cancer	(Petrucelli, Daly and Pal, 1993)	Not scored
BRCA2	BRCA2 DNA Repair Associated	Gene Panels, SFARI, DDD	Hereditary breast and ovarian cancer	(Petrucelli, Daly and Pal, 1993)	3
DSP	Desmoplakin	DDD	Arrhythmogenic right ventricular cardiomyopathy	(McNally, MacLeod and Dellefave-Castillo, 1993)	Not scored
FBN1	Fibrillin 1	Gene Panels, SFARI, DDD	Marfan syndrome, Loeys-Dietz syndromes, and familial thoracic aortic aneurysms and dissections	(Milewicz and Regalado, 1993; Dietz, 2017; Loeys, 2017)	2
KCNQ1	Potassium Voltage-Gated Channel Subfamily Q Member 1	DDD	Romano-Ward long-QT syndrome types 1, 2, and 3, Brugada syndrome	(Yunis and Bhonsale, 2020)	Not scored

LDLR	Low Density Lipoprotein Receptor	SFARI	Familial hypercholesterolemia	(Youngblom, Pariani and Knowles, 1993)	3
LMNA	Lamin A/C	Gene Panels, DDD	Hypertrophic cardiomyopathy, dilated cardiomyopathy	(Cirino and Ho, 1993)	Not scored
MYH11	Myosin Heavy Chain 11	DDD	Marfan syndrome, Loeys-Dietz syndromes, and familial thoracic aortic aneurysms and dissections	(Milewicz and Regalado, 1993; Dietz, 2017; Loeys, 2017)	Not scored
OTC	Ornithine Transcarbamylase	Gene Panels, DDD	Ornithine transcarbamylase deficiency	(Lichter-Konecki *et al.*, 2016)	Not scored
PMS2	PMS1 Homolog 2, Mismatch Repair System Component	DDD	Lynch syndrome	(Idos and Valle, 2004)	Not scored
PTEN	Phosphatase And Tensin Homolog	Gene Panels, SFARI, DDD	PTEN hamartoma tumour syndrome	(Mester, 2016)	1
RET	Ret Proto-Oncogene	Gene Panels, DDD	Multiple endocrine neoplasia type 2, Familial medullary thyroid cancer	(Eng, 1993)	Not scored
RYR1	Ryanodine Receptor 1	DDD	Malignant hyperthermia susceptibility	(Allen, 1994)	Not scored
RYR2	Ryanodine Receptor 2	Gene Panels	Catecholaminergic polymorphic ventricular tachycardia	(Maragna and Napolitano, 2018)	Not scored
SDHD	Succinate Dehydrogenase Complex Subunit D	Gene Panels	Hereditary paragangliomapheochromocytoma syndrome	(Else, Greenberg and Fishbein, 1993)	Not scored

SMAD3	SMAD Family Member 3	DDD	Marfan syndrome, Loeys-Dietz syndromes, and familial thoracic aortic aneurysms and dissections	(Milewicz and Regalado, 1993; Dietz, 2017; Loeys, 2017)	Not scored
SMAD4	SMAD Family Member 4	Gene Panels, SFARI, DDD	Juvenile polyposis	(Hussain and Church, 2020)	2
TGFBR1	Transforming Growth Factor Beta Receptor 1	Gene Panels, DDD	Marfan syndrome, Loeys-Dietz syndromes, and familial thoracic aortic aneurysms and dissections	(Milewicz and Regalado, 1993; Dietz, 2017; Loeys, 2017)	Not scored
TGFBR2	Transforming Growth Factor Beta Receptor 2	Gene Panels, DDD	Marfan syndrome, Loeys-Dietz syndromes, and familial thoracic aortic aneurysms and dissections	(Milewicz and Regalado, 1993; Dietz, 2017; Loeys, 2017)	Not scored
TSC1	TSC Complex Subunit 1	Gene Panels, SFARI, DDD	Tuberous sclerosis complex	(Northrup *et al.*, 1993)	1
TSC2	TSC Complex Subunit 2	Gene Panels, SFARI, DDD	Tuberous sclerosis complex	(Northrup *et al.*, 1993)	1
WT1	WT1 Transcription Factor	DDD	WT1-related Wilms tumour	(Dome and Huff, 1993)	Not scored

Table 6-7 Clinical relevance of overlapping genes.

This table details and expands on genes included on the ACMG59 gene lists of clinically actionable genes, which are also interrogated by one or more autism-relevant gene lists. All gene symbols included are HGNC approved, gene descriptions are sourced from GeneCards (https://www.genecards.org/). Disease phenotype associated with each gene is specified as per Coriell Institute for Medical Research (https://www.coriell.org/). Literature documenting the gene-disease association and clinical handling of genetic variation with each gene is referenced as Gene Reviews (a point-of-care resource for clinicians, providing clinically relevant and medically actionable information for inherited conditions). Where scored, the SFARI Gene score (13-01-2021 release) is given as a measure of the level of evidence supporting a genes association with autism. Genes are scored on a scale of 1-3 based on the number of autism reports of variation within the gene. SFARI Gene scores are interpreted as 1; High Confidence, 2; Strong Candidate; 3, Suggestive Evidence. Where genes are "Not scored" by SFARI Gene, no reports of autism have been associated with variation within the gene, as per the SFARI Gene database.

6.5.2.2 Determining the clinical relevance of overlapping gene

The clinical relevance of genes overlapping ACMG59 are detailed in Table 6-7. The genes detailed in the table have evidence of neurodevelopmental association and may be targeted in the future or are currently targeted by commercial genetic testing strategies in autism and other neurodevelopmental conditions. The disease phenotype associated with the ACMG59 gene is given in Table 6-7 alongside the SFARI Gene Score. As shown, 17/24 overlapping genes have not been scored by SFARI Gene. Of those that have, two genes have been assigned SFARI Gene scores of three indicating a lack of robust evidence supporting the association.

6.6 Discussion

6.6.1 A lower number of targeted genes on commercial gene panels is associated with reduced detection of clinically relevant variants.

Considering the low diagnostic yield of the gene panels that were investigated, we can infer that, while the gene selection for inclusion in autism gene panels is evidence-based, these gene lists are not extensive enough to justify use in genetic diagnosis in the context of autism, a complex trait for which hundreds of genes have been associated (Table 6-4). The GeneDx "Autism/ID Xpanded Panel" represents the autism gene panel with the highest number of individuals for which a genetic diagnosis would have been obtained with its application (10.02%). This diagnostic yield is comparable to that of WES, 10.4% (Feliciano *et al.*, 2019) and that of chromosomal microarray sequencing with a median diagnostic yield of 8.1% (Savatt and Myers, 2021). However, important to note is that this gene panel targets many more genes (n=2,641) than some of the smaller gene panels, for example GENETAQ "Autism" panel (n=27), with a diagnostic yield of just 1.53%. The positive correlation of diagnostic yield associated with inclusion of a larger number of genes, reflects well the complex genetic architecture of autism and the number of loci expected to be associated. Critically, it must be communicated to healthcare providers ordering these diagnostic tests, that if a targeted gene panel test has a negative result for detection of pathogenic variation, one cannot conclude that a causative variant is not present. Rather, it is more likely that genetic causes have been missed due to the absence of the gene of interest from that panel.

This raises the question whether autism is an appropriate candidate for the development of commercial gene panels, which are limited due to the size of the gene panel, the cost and current knowledge of the genetic basis of autism. This suggests that sequencing technologies with a broader coverage, such as WGS may be more effective. Balancing the reduction in costs associated with restriction of the proportion of the genome covered, with the potential benefit of sequencing the entire genome must be considered based on the genetic architecture underlying the condition. Following developments in WGS, particularly with the advent of long-read sequencing, this technology has the potential to cover up to 100% of the human genome. When restricting variant discovery to coding regions only, just 1% of the human genome is explored. Targeted sequencing, including the use of gene panels or clinical exome sequencing, presents the opportunity to significantly reduce costs associated with genetic sequencing, challenges associated with variant interpretation and limitations associated with large data storage. However, these benefits come at the cost of restricting variant discovery to a miniscule proportion of the genome, depending on the number of genes targeted (Table 1-1).

Expanding beyond targeted autism genes, WGS presents the opportunity to explore more of the human genome and, ultimately, to further increase the diagnostic yield in autism (Yuen *et al.*, 2015). Progress in non-coding variant annotation and interpretation, accompanied by a decrease in sequencing costs, may further popularize the clinical use of WGS. Currently, WES is proposed as the first-tier diagnostic test for neurodevelopmental conditions (Srivastava *et al.*, 2019). Recent advances in data analysis have led to a capability of CNV calling by WES, eliminating the need for CMA entirely. This has enabled WES with CNV calling to achieve a high diagnostic yield in neuropsychiatric conditions. The diagnostic yield in autism using clinical exome sequencing has been estimated at 6.1% in autism (20% overall yield in neurodevelopmental conditions) (Martinez-Granero *et al.*, 2021). Genotyping chips have limited clinical utility for rare genetic variation of SNVs and should not be used to guide health decisions without validation (Mn *et al.*, 2021).

6.6.2 Challenges in the handling of genetic findings

6.6.2.1 Reporting of secondary findings

Consideration is needed of the potential to uncover of incidental genetic findings when analysing genomic research data. Incidental findings here refer to clinically relevant and potentially clinically actionable findings, unrelated to the primary purpose of performing the test. This discussion comes with several factors to consider, specifically ethical consideration. This is particularly considered in the context of 'actionable' findings such as variants in ACMG59 genes. Secondary findings found in a research setting require clinical validation of the findings and clinical reporting, resulting in communication preferably from a genetic counsellor, all considered only where the participant has consented for such findings to be communicated back.

Investigation of the overlap of autism-relevant genes with the ACMG59 list of gene in which variants in the exons of the genes may be clinically actionable for the individual, highlights that consideration must be made to the uncovering of secondary findings within these genes which may have impacts on an individual's life, unrelated to autism (Figure 6-1). Six genes, included in all autism gene lists investigated are found on the ACMG59 gene list (*PTEN, TSC1, TSC2, BRCA2, FBN1* and *SMAD4*). Furthermore, three of these six genes (*PTEN, TSC1* and *TSC2*) are scored as "High Confidence" for autism-association and two of the six genes (*FBN1* and *SMAD4*) are scored as "Strong Candidate" for autism-association.

While these genes show strong evidence supporting the need to sequence and interrogate them to determine the genetic basis of autism, there is risk of identification of secondary actionable genetic variation in a research context. With this risk comes a need for sensitivity in data handling and strict informed consent at study enrolment. A strategy for clinical validation and return of result is required should a putatively pathogenic variant be identified. Secondary findings found in a research setting require appropriate clinical validation of the findings and clinical reporting, and access to follow up genetic counselling which may not be universally available, as is the case in Ireland with major gaps in resourcing for provision of genetic advice (Lynch and Borg, 2016). This requires the appropriate consent for return of incidental findings. While these variants are “individually rare, they are collectively common.” This means the likelihood of identifying a rare secondary genetic finding in both large-scale and smaller research cohort is not uncommon, as shown in a 2.7% diagnostic yield of ACMG59 pathogenic variation in a cohort of 101 epilepsy patients (Benson *et al.*, 2020).

These genes discussed show evidence for autism association, a benefit to their inclusion when determining the genetic basis of autism, which should be weighed against the risk, or reward depending on the preferences of the individual, of identification of secondary genetic finding. However, 17 of the 24 autism-relevant genes overlapping with ACMG59 show lack of evidence of autism association, as determined by number of reports in the SFARI Gene database. Notably most of these genes are overlapping by the gene list arising from the Deciphering Developmental Disorder study, a gene list collated from severely affected neurodevelopmental cohort with atypical presentations, but likely relevant also to autism. Of note are the three genes (*SDHD, ATP7B* and *RYR2)* interrogated by commercial gene panels marketed for use in autism. These genes are “Not scored” by SFARI Gene, potentially indicating insufficient evidence for justification in targeted autism sequencing in any case.

PheWAS analysis has been performed to understand pleiotropic effects of rare variation in the ACMG59 genes on psychiatric phenotypes. This approach did not identify any ACMG59 genes that are significantly enriched with rare deleterious variants that confer risk for psychiatric conditions, showing a lack of association between psychiatric conditions and incidental findings in these medically actionable genes (Feng *et al.*, 2022). These findings suggest there may be little benefit of inclusion of these genes within autism gene panels, while a relevant finding brings healthcare challenges, requiring feedback and onward referral for further medical investigation where appropriate. Further follow-up may be required with a need for routine surveillance, for example pathogenic variation in *PTEN* where breast or bowel screening might be indicated.

Important to this discussion is that four genes (*BMPR1A, SMAD4, ATP7B*, and *OTC*) were added to the ACMG59 gene list in the update from version 1 to version 2 (Richards et al., 2015; Kalia et al., 2017). Three of these four genes are autism-relevant genes (Table 6-7), which may have been included in targeted autism gene panels developed prior to update of ACMG59. Importantly ACMG59 is not a definite collection of genes in which clinically actionable pathogenic variation may occur. In the time since these analyses were carried out a recent iteration of ACMG, ACMG73 has been released (Miller *et al.*, 2021). As new genomic findings emerge this list will continue to be expanded. Van der School *et al.* (2022) report variation in many genes not included on the ACMG list but with evidence of association to a disorder for which disease manifestation could be influenced, suggesting that such a list is arbitrary at best (van der Schoot *et al.*, 2022).

6.6.3 Ethical considerations

6.6.3.1 Re-analysis of variation

Alongside, the consideration of incidental findings comes the responsibility of re-analysis as functional prediction tools improve and study sizes increase (Deignan *et al.*, 2019). These advancements may lead to different interpretation of a variant's significance, or to availability of new information supporting its relevance. There is currently no legal requirement to recontact patients as new genetic findings emerge and there are major barriers to doing so, including procedures for re-analysis and re-contact, consenting and clinical resources (Carrieri *et al.*, 2019; David *et al.*, 2019). Progress is being made to facilitate clinical utilisation of changing variant classification and gene-disease relationships as they emerge. One such example is ClinGen GenomeConnect, a patient registry with capacity to trigger re-analysis as variants are updated in ClinVar (Savatt *et al.*, 2018). This service can supplement laboratory and clinician efforts to keep patients informed about their genetic testing results and expands patient-centred data sharing.

6.6.3.2 Risk communication in psychiatric genetics

Psychiatric genetic testing is available through direct-to-consumer (DTC) testing despite the limitations, gaps in knowledge and ethical complications discussed already. If the technology is made available to individuals, it is difficult to prevent its premature application, due to the huge benefits to a genetic diagnosis already mentioned. Substantial individual stress arises in the communication of genetic findings, whether identified in a clinical setting or population-

based DTC testing. Among the risks of clinical interpretation of DTC genetic tests, particularly SNP chip technologies, are false positives, in particular where rare variants are under investigation (Mn *et al.*, 2021). Another risk associated with DTC genetic testing is false reassurance arising for tests being less thorough than a customer realises. This is true for example in *BRCA1* and *BRCA2* variant testing for which some DTC tests only analyse a subset of potential variants, thus missing most of the variants associated and providing false reassurance to ~80% of individuals with a disease-causing variant (Rebbeck *et al.*, 2018). Unclear meaning of disease-causing variation is also a concern when DTC genetic tests are used in a population cohort, i.e. outside the context of symptoms or family history of the related disease (Wright *et al.*, 2019).

Risk communication becomes complicated with variants of small to moderate impact on outcomes (Eeltink *et al.*, 2021). Adding to the discussion on genetic counselling, we must also consider the outcome of a PGS in psychiatry, specifically just because an individual's risk has been identified does not mean the outcome can be changed. Elements of consent become difficult when communicating around PGS and potential impact, or lack of impact, on an individual's life. Clinician guidance is crucial to overcome unrealistic expectations of the results a genetic test may deliver, to ensure that consent remains valid (R. Horton and Lucassen, 2019). Key to these efforts are genetic counselling where possible, and clinician training in genomic literacy (JI *et al.*, 2018). Complex nomenclature, changing penetrance estimates, changing variant annotations, complex data formats and modest effect sizes are among the complexities associated with return of psychiatric genetic results. Tools have been developed to bridge the gap between genomic expertise and the treating clinician, for example GenoPred, a tool for converting PGS to an absolute scale risk (Pain *et al.*, 2021). Primers with the aim of translating the vast amounts of genomic literature into directly relevant key messages are also convenient for bridging this gap. These strategies, among many others, in the translation of genomic technologies into clinical settings are key to accompany technological advances in genomic medicine (R. H. Horton and Lucassen, 2019).

6.6.4 Conclusion

Gene panels have potential for clinical utility provided the relevant expertise and infrastructure for variant interpretation are available and cost effective. However, current evidence does not support their applicability in autism (Buxbaum *et al.*, 2020; Myers, Challman, Martin, *et al.*, 2020). Achieving the goal of a comprehensive autism gene panel will require uniform robust phenotyping to account for the heterogeneity in autism presentation, as discussed in Chapter 4. Consideration must be made of the inclusion of genes in which pathogenic variation, when

detected, is clinically actionable; the benefits of their inclusion weighed against the clinical management of their identification. To conclude, evaluation of the diagnostic yield of commercial gene panels marketed for autism determines that they are currently of limited clinical utility.

Chapter 7. General Discussion and Future Directions

7.1 Overview of aims and findings

7.1.1 Strengths and weaknesses of this research

The variant discovery performed in Cohorts 1, 2 and 3 here applies a stringent variant filtration strategy to identify putatively pathogenic rare variation in genes with existing reports of association to autism and neurodevelopment. This pipeline is limited to the isolation of rare exonic SNVs, however NGS technologies enable additional classes of variation to be detected, with evidence supporting their involvement in the genetic basis of autism, such as CNVs, SVs and tandem repeat expansions. There are future opportunities to explore these classes of variation in Cohort 1, 2 and 3 and enable expansion of the understanding of the genetic basis of autism within this cohort. While variant discovery did not yield pathogenic SNVs with association to autism, expansion beyond this gene-set based variant filtration strategy will enable detection of more genetic variants which could be contributing to the phenotype.

Expanding beyond this filtration strategy may detect causative variation in the cohort, when unrestricted by the requirement to restrict analyses to genes with an existing gene-disease association reported. This gene-set based filtration step within the variant isolation strategy is a weakness, leaving many rare putatively pathogenic variants uninterrogated. However, there is opportunity to overcome this in the future with an increase in sample size, achievable by analysis of this dataset in combination with other WES and WGS autism datasets, such as those described in Table 1-2. Large sample sizes give statistical power to enable gene-phenotype associations, while the small sample sizes studied in this thesis enable only variant detection within known autism-associated genes.

The strength of this work comes from the relevance of this research in the translation of autism genomic findings to clinically impactful knowledge, as detailed in the sections to follow. While clinically relevant variation identification was limited in this thesis, the knowledge gained through evaluation of clinical gene panels and gene curation strategies may be applied in both research and clinical settings.

7.1.2 An analysis strategy to isolate exonic rare pathogenic SNVs using next-generation sequence data.

This analysis aimed to establish a strategy for isolation of rare exonic pathogenic SNVs from NGS data. In doing so, this work also aimed to discover rare putatively pathogenic autism-relevant SNVs in a cohort of autism-affected individuals and their unaffected family members.

The strategy described in this study outputs high-confidence SNV calls that are rare by MAF, as estimated by gnomAD, and pathogenic, as predicted by consensus scoring of CADD, SIFT and PolyPhen-2 (Ng and Henikoff, 2003; Adzhubei, Jordan and Sunyaev, 2013; Karczewski *et al.*, 2019; Rentzsch *et al.*, 2019). Power for statistical associations of rare variant burden is limited by the small sample sizes of Cohort 1, Cohort 2, and Cohort 3. In the absence of power, variant discovery is restricted to genes with existing evidence for autism associations, specifically SFARI Gene and DDD gene2phenotype (Abrahams *et al.*, 2013; Wright *et al.*, 2015). This approach is restrictive, limiting variant discovery to a small subset of variation. In future, this cohort will contribute to larger sequencing efforts in autism and enabled by greater sample sizes will contribute to building understanding of the genetic basis of autism.

7.1.3 Evaluating gene-phenotype relationships through gene curation; a WGS study in autism.

This analysis aimed to dissect gene-phenotype relationships through application of an autism gene curation framework. To achieve this, rare exonic pathogenic SNVs from WGS data were identified by applying the analysis strategy outlined in Chapter 3.

In consideration of genes for curation, genes were enriched for those with a high level of evidence supporting autism association. Despite this all genes evaluated were classified as having a limited gene-phenotype association. Just one of the three evaluated genes had experimental evidence supporting an autism association highlighting the need for *in vitro* and *in vivo* functional studies alongside predictive genomic analysis to build robust evidence for gene-phenotype associations. Gene curation through the proposed framework accounts for the degree of certainty in autism diagnoses in studies reporting association and accounts for co-occurring diagnoses. This strategy, if applied widely, will provide consistency throughout gene discovery, and ultimately aid in the translation of genomic findings to the clinic.

7.1.4 A pedigree driven approach to identify pathogenic variation in multiplex families of neurodevelopmental conditions.

The aim of this analysis was to identify rare putatively pathogenic SNVs in genes with evidence supporting their role in autism, using a family-based study design to evaluate variant transmission. This analysis analysed WGS data from a rare cohort of multiplex and extended pedigrees. Mode of pathogenic variant transmission was hypothesised based on reported affection status through family structure. Of the four families investigated putatively pathogenic

heterozygous variation predicted to be causative, was detected in three families. One of these families harboured two predicted pathogenic variants impacting *TTN* in all affected individuals and absent from unaffected family members. Accompanied by the existing level of evidence supporting variation in this gene as a single gene cause of autism, this is the likely rare variant cause of affection within this family.

The approach applied is a first pass analysis which with increased sample size should be evaluated at a more widespread level across the genome and in parallel to large-scale WGS efforts.

7.1.5 Determining the clinical utility of gene panels in autism; a study of diagnostic yield and relevance.

This analysis aimed to evaluate the diagnostic yield of commercial gene panels marketed for use in autism and determine the relevance gene selection for these panels. This work also aimed to determine the overlap of ACMG59 genes and autism-relevant gene lists.

Current evidence does not support the applicability of targeted gene panels in autism (Buxbaum *et al.*, 2020; Myers, Challman, Bernier, *et al.*, 2020). Evaluation of the diagnostic yield of the autism gene panels, through secondary analysis of the SPARK WES cohort, determined that they are currently of limited clinical utility. Gene selection for inclusion in autism gene panels was found to be evidence-based, as indicated by the proportion of known autism genes included in these targeted sequencing panels. However, no panel was extensive enough to justify use in genetic diagnosis in the context of autism, a complex trait for which hundreds of genes have been associated. Analysis was performed on the overlap between ACMG59, a gene list of medically actionable genes, with genes association with autism and neurodevelopmental conditions. ACMG59 was also investigated for overlap with the genes included in the autism gene panels evaluated for diagnostic yield. This found a substantial number of genes which are associated with autism, which are also recommended to be reported on should a pathogenic variant be detected. These findings should impact the decision to apply these targeted autism gene panels, in their current form, in a clinical or research setting.

Importantly the limited scope for detection of putatively pathogenic variation to aid autism diagnosis should be considered by clinicians when discussing and consenting for this genetic testing. As highlighted by the overlap of these genes with ACMG59, pleotropic effects of the genes targeted for sequencing should be considered when ordering these genetic tests, as

well as preparation for appropriate follow-up should a variant in a clinically actionable gene be detected. Achieving the goal of a clinically valuable autism gene panel requires a comprehensive gene list of genes robustly associated with autism, specifically autism in the absence of other neurodevelopmental conditions. Application of a formal evidence-based gene curation framework, such as that evaluated in Chapter 4 works towards this goal.

7.2 Future Directions: Translating variant discovery from research to clinic

Variant discovery is key to building understanding of the biology underlying neurodevelopmental conditions and their etiology. There is huge potential for this translation to be integrated across healthcare systems. At this scale, success has been demonstrated in variant discovery and diagnosis in rare disease in routine healthcare system, through WGS within the UKBiobank population cohort (Turro *et al.*, 2020). Genetic diagnosis can subsequently facilitate variant-specific medical decision making. This has been demonstrated in the context of rare disease by Bhatia *et al.,* reporting changes in treatment in 27% of patients following a genetic diagnosis from whole exome or WGS (Bhatia *et al.*, 2021).

However, barriers must be addressed to enable the translation of genomic findings to clinically informative biomarkers in autism and neurodevelopmental conditions. Heterogeneity is the cause for these barriers, as has been described throughout this thesis at the genotype and phenotype level. Heterogeneity must be accounted for when considering the role and relevance of a putatively pathogenic variant. Informed prediction, while reliable in some conditions, is impacted by the heterogeneity of neuropsychiatric and neurodevelopmental conditions (Nunes, Trappenberg and Alda, 2020). At case-level a family-based study design can control for some levels of heterogeneity from the shared genetic and environmental effect between family members, enabling variant discovery. However, at condition-level future work on variant discovery must evaluate and limit biases influencing variant discovery to achieve the full potential of genomic data and to ensure broad translatability of findings to the clinic.

7.2.1 Phenotypic biases

Autism cohorts are biased by their mode of ascertainment. Clinically ascertained cohorts select for more severely affected individual than a population-based study design. DDD from which the DDD gene2phenotype gene list was derived recruits only the most severe neurodevelopmental-affected individuals, presenting with more complex phenotypes (Wright *et al.*, 2015). In contrast, SPARK is biased towards less severe autism phenotypes by the sample collection strategy used (Feliciano *et al.*, 2019). The individuals participating have

given saliva samples for sequencing, a method of collection which can be difficult for those with neurodevelopmental challenges potentially excluding these individuals from the study. These biases are reflected in the genetic architecture of these cohorts and the genetic findings emerging from their study. Specifically, it would be expected to discover an enrichment of rare highly penetrant variants in the DDD study when compared to SPARK. Conversely, it may be expected to discover higher polygenic burden in the SPARK cohort, indicating a higher burden of common genetic variants of lower effect sizes resulting in the autism phenotype. While both approaches yield impactful variant discovery in autism and neurodevelopmental conditions, consideration must be given to the penetrance of these variants when translating into the clinic.

7.2.2 Ancestral population biases

Research with broad applicability is currently challenging to apply to non-European populations. Most genomes sampled in research studies come from European ancestral populations. 23andMe is currently the most ancestral diverse genomic cohort. While not exempt from the biases affecting all population-based cohorts (73% European), this cohort holds genetic information on some of the largest cohorts of Africa in the world (Dr Sarah Laskey 23andMe, WCPG 2021). Discoveries made in underrepresented non-European populations have potential to impact healthcare broadly, beyond the ancestral population in which they are identified. For example, genetic sequencing applied to a cohort of ~15,000 African American individuals resulted in detection of variation in *EXOC3L1*, associating the protein product of the gene as a key facilitator of lipid receptor trafficking, giving insight to the biology underpinning cholesterol levels generally (Lanktree *et al.*, 2015). Working towards genomic research with broad applicability requires ancestry-based reference genomes enabling robust assessment of allele frequency and adaptation of widely applied analysis strategies to analyse non-European samples most accurately.

7.2.3 Sex biases

Autism is diagnosed 3-4 times more in males than females (Loomes, Hull and Mandy, 2017). Phenotypically autism varies in presentation in male and female individuals with males more likely to receive a diagnosis of autism than females (Kreiser and White, 2013). However, this difference does not account completely for the higher ratio observed. A Female Protective Effect has been proposed for autism whereby females can accumulate more risk than males before being affected by autism. Evidence for FPE in autism comes from both common and rare variation (Antaki *et al.*, 2022). Namely an increased burden of rare *de novo* variation in autistic females cases has been observed widely (Sanders *et al.*, 2015). A similar trend is

seen for the increased burden of common variant polygenic risk in affected females when compared to affected males (Antaki *et al.*, 2022). In the family context, Wigdor *et al.* provide evidence supporting a FPE against autism in the context of common, inherited variation (Wigdor *et al.*, 2022). Under FPE, more siblings of female autism cases are affected compared to siblings of male cases. Wigdor *et al.* show evidence of FPE in both affected and unaffected members of autism-impacted families with mothers of autistic children carrying on average more common, inherited genetic risk for autism than fathers (Wigdor *et al.*, 2022).

7.3 Future Directions: Maximising potential through data integration

A major challenge facing autism genomics is the integration of all aspects of its genomic basis, including both common and rare variation. regulatory effects, and epigenetic modifications (Figure 7-1). In order to maximise the data already existing in the field, best practices must be set out in the optimum analysis of all of these factors to best illustrate the genomic architecture of autism. As the research outlined in this thesis progressed, recent developments in genomics have impacted and will continue to benefit the field of autism genomics. Progress has been made in the annotation of non-coding variants as summarised in 7.3.1. The complete human genome has been sequenced, enabled by long-read sequencing technologies, as summarised in 7.3.2. While not investigated within this thesis, progress has been made on the characterisation of the contribution of common variation in autism which has resulted from increased sample sizes and improvement in methods of association (7.3.3).

7.3.1 Integrating the coding and non-coding genome

WGS enables sequencing of non-coding variation, undetected by targeted sequencing and WES. Investigations to date into the contribution of this variation in these genomic regions have associated genetic variants with autism and neurodevelopmental conditions (Turner *et al.*, 2016; Brandler *et al.*, 2018; Wright *et al.*, 2021). Work is underway to overcome challenges in discovery and interpretation of non-coding variation, where variant consequence is not as readily predicted as protein-coding variation. Collaborative efforts towards this goal, such as the generation of recommendations for clinical interpretation of non-coding variant by Ellingford *et al.*, will be key to maximising the potential of WGS datasets in the future (Ellingford *et al.*, 2021).

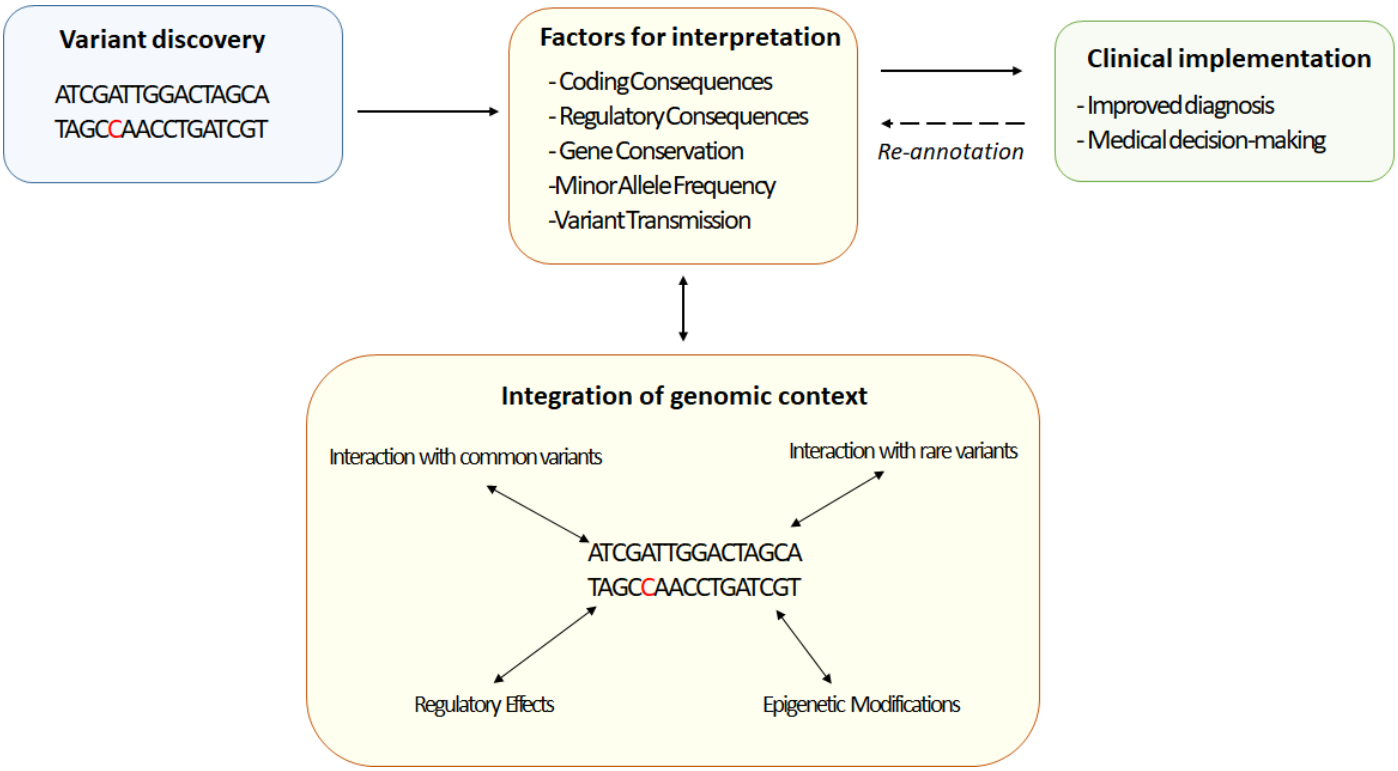


Figure 7-1 Pathway from sequencing to clinical implementation.

Outlined are the main stages of autism gene discovery; from variant discovery (blue), through genomic data analysis (yellow), to accurate translation for meaningful diagnosis (green). Re-annotation refers to regular re-analysis of genetic diagnosis, as additional variants reach significant association with autism. The variant highlighted in red, here a SNV, represents any variant type detectable through application of genomic technologies (Table 1-1). Epigenetic modifications include methylation changes, histone modification or microRNA dysregulations (reviewed (Eshraghi *et al.*, 2018)*. Research is ongoing to integrate genomic variants with other variation within an individual's genome, as described by* (McGuire *et al.*, 2020)*.*

7.3.2 Integrating classes of variation

The complete set of genomic variation is known to contribute to the genetic basis of autism. Enabled by WGS, repeat expansions, CNVs, SVs have been implicated in autism (Pinto *et al.*, 2010; Brandler *et al.*, 2018; Mitra *et al.*, 2021). Long-read sequencing enables sequencing of genomic regions inaccessible through NGS (Ebbert *et al.*, 2019). As well expanding coverage of genome sequencing, long-read sequencing will enable detection of variant classes otherwise challenging to robustly quantify, such as repeat expansions, large SVs, and chromosomal rearrangements. In 2022, the Telomere-to-Telomere consortium completed sequencing of the complete human genome, sequencing 8% more of the human genome than the previous iteration (Nurk *et al.*, 2022). With expanded coverage of the genome and added power of variant class detection, long-read sequencing application in autism is already in progress and will continue to contribute to understanding of the genetic basis of autism (Begum *et al.*, 2021; Noyes *et al.*, 2022).

7.3.3 Integrating rare and common variation

The role of rare variation has been highlighted throughout this thesis in the context of autism. Autism-associated rare variants were identified using exonic WGS data in 15% of autistic individuals from the latest MSSNG cohort and 17% of probands in the Simon's Simplex Collection (Trost *et al.*, 2022). Rare variation has been implicated across complex traits and there is potential for rare variant discovery at large-scale from well powered GWAS by WGS in the future (Wainschtein *et al.*, 2022). Common genetic variation, while not investigated in the research outlined in this thesis, is likely to account for a proportion of the unexplained heritability of autism. Enabled by increased sample size, common variant discovery will progress building on the association made to date (Grove *et al.*, 2019).

Common variant discovery will lead to improved polygenic scoring. PGS has huge potential benefit to psychiatric genetics as a clinical predictor, of which there are currently very few available. As discussed earlier, in the context of genetic diagnoses of rare variation, PGS has potential to have benefit in both prediction of disease status and prediction of treatment response, and this potential has been demonstrated in psychiatric conditions (Lewis and Vassos, 2020). This potential has already been demonstrated by the success of translating a high PGS for schizophrenia as a predictor of poor lithium response in the treatment of bipolar disorder (Amare *et al.*, 2018).

However, PGS has major limitations acting as a barrier to their translation to clinical settings. PGS is currently not useful to guide diagnosis at the individual level (So, Sham and Valencia, 2017). Clinical application of PGS is currently limited to interpretation of the extremes of the normal distribution of risk (Khera *et al.*, 2018). A combined approach to evaluation of rare and common variation will provide a more wholistic view of the genetic basis of autism. A "Genomic Risk Score" derived from a "Rare Variant Risk Score" and "Common Variant Risk Score" through multivariable regression showed a 40% improvement in predictive accuracy for autism than common or rare variant scoring alone (Antaki *et al.*, 2022). Integration of variation across the allele frequency spectrum will enable comprehensive understanding of the genetic basis of autism.

7.4 Conclusion

Genomic technologies have accelerated research progress in autism genomics and promises to further transform our understanding of the genetic basis of this neurodevelopmental condition. This thesis introduces the current evidence for the genetic basis of autism, presents the progress of large-scale studies to date and highlights the potential of genomic technologies. This thesis outlines building an analysis pipeline to evaluate rare genetic variation in the context of autism, applying a gene curation strategy to dissect gene-phenotype associations, discovery of rare variants in a rare cohort of multiplex extended family affected by neurodevelopmental conditions and finally examines the clinical utility of targeted gene panel sequencing in autism.

Together this work describes rare variant discovery and evaluation of its clinical utility in autism with an aim to contribute to the greater goal of understanding the genetic basis of autism. This thesis supports the importance of identifying rare genetic variants in family-based studies. Genomics is central to personalised medicine and is a key feature of the future healthcare. Autism genomics has potential to improve our biological understanding of neurodevelopmental conditions, to aid diagnosis and to inform medical decision-making in the future.

References

Abrahams, B. S. *et al.* (2013) 'SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs).', *Molecular autism*. BioMed Central, 4(1), p. 36. doi: 10.1186/2040-2392-4-36.

Abrahams, B. S. and Geschwind, D. H. (2008) 'Advances in autism genetics: on the threshold of a new neurobiology', *Nature Reviews Genetics*. Nature Publishing Group, 9(5), pp. 341–355. doi: 10.1038/nrg2346.

Adzhubei, I., Jordan, D. M. and Sunyaev, S. R. (2013) 'Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2', *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*. NIH Public Access, 0 7, p. Unit7.20. doi: 10.1002/0471142905.HG0720S76.

Alirezaie, N. *et al.* (2018) 'ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants', *American journal of human genetics*. Am J Hum Genet, 103(4), pp. 474–483. doi: 10.1016/J.AJHG.2018.08.005.

Allen, G. C. (1994) 'Malignant hyperthermia susceptibility', *Anesthesiology Clinics of North America*. University of Washington, Seattle, pp. 513–535. doi: 10.1111/j.1365-2044.1979.tb04865.x.

Amare, A. T. *et al.* (2018) 'Association of polygenic score for schizophrenia and HLA antigen and inflammation genes with response to lithium in bipolar affective disorder: A genome-wide association study', *JAMA Psychiatry*. American Medical Association, 75(1), pp. 65–74. doi: 10.1001/jamapsychiatry.2017.3433.

Amberger, J., Bocchini, C. and Hamosh, A. (2011) 'A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®)', *Hum Mutat*, 32. doi: 10.1002/humu.21466.

Amberger, J. S. *et al.* (2015) 'OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders', *Nucleic Acids Research*. Oxford University Press, 43(D1), pp. D789–D798. doi: 10.1093/nar/gku1205.

Amberger, J. S. *et al.* (2019) 'OMIM.org: Leveraging knowledge across phenotype-gene relationships', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D1038–D1043. doi: 10.1093/nar/gky1151.

American Psychiatric Association (2013) *Diagnostic and statistical manual of mental disorders (5th ed.)*.

An, J. Y. *et al.* (2014) 'Towards a molecular characterization of autism spectrum disorders: An exome sequencing and systems approach', *Translational Psychiatry*. Nature Publishing Group, 4(6). doi: 10.1038/tp.2014.38.

An, J. Y. *et al.* (2018) 'Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder', *Science*. American Association for the Advancement of Science, 362(6420). doi: 10.1126/science.aat6576.

An, J. Y. and Claudianos, C. (2016) 'Genetic heterogeneity in autism: From single gene to a pathway perspective', *Neuroscience and Biobehavioral Reviews*. Elsevier Ltd, pp. 442–453. doi: 10.1016/j.neubiorev.2016.06.013.

Anderson, C. A. *et al.* (2010) 'Data quality control in genetic case-control association studies', *Nature Protocols*. Nature Publishing Group, 5(9), pp. 1564–1573. doi: 10.1038/nprot.2010.116.

Anderson, D. K. *et al.* (2007) 'Patterns of growth in verbal abilities among children with autism spectrum disorder', *Journal of consulting and clinical psychology*. J Consult Clin Psychol, 75(4), pp. 594–604. doi: 10.1037/0022-006X.75.4.594.

Antaki, D. *et al.* (2022) 'A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex', *Nature Genetics 2022*. Nature Publishing Group, pp. 1–9. doi: 10.1038/s41588-022-01064-5.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*. Nature Publishing Group, pp. 68–74. doi: 10.1038/nature15393.

Van der Auwera, G. A. *et al.* (2013) 'From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline', in *Current Protocols in Bioinformatics*.

Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43.
Bai, D. *et al.* (2019) 'Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort', *JAMA Psychiatry*. doi: 10.1001/jamapsychiatry.2019.1411.
Bainbridge, M. N. *et al.* (2011) 'Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities', *Genome Biol*, 12. doi: 10.1186/gb-2011-12-7-r68.
Baird, G. *et al.* (2006) 'Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP)', *The Lancet*, 368(9531), pp. 210–215. doi: 10.1016/S0140-6736(06)69041-7.
Barton, K. S. *et al.* (2018) 'Pathways from autism spectrum disorder diagnosis to genetic testing', *Genetics in Medicine*. Nature Publishing Group, 20(7), pp. 737–744. doi: 10.1038/gim.2017.166.
Begum, G. *et al.* (2021) 'Long-Read Sequencing Improves the Detection of Structural Variations Impacting Complex Non-Coding Elements of the Genome', *International Journal of Molecular Sciences*. Multidisciplinary Digital Publishing Institute (MDPI), 22(4), pp. 1–14. doi: 10.3390/IJMS22042060.
Benson, K. A. *et al.* (2020) 'A comparison of genomic diagnostics in adults and children with epilepsy and comorbid intellectual disability', *European Journal of Human Genetics*. Springer Nature, pp. 1–12. doi: 10.1038/s41431-020-0610-3.
Betancur, C. (2011) 'Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting', *Brain Research*. Elsevier, 1380, pp. 42–77. doi: 10.1016/J.BRAINRES.2010.11.078.
Bhatia, N. S. *et al.* (2021) 'Singapore Undiagnosed Disease Program: Genomic Analysis aids Diagnosis and Clinical Management', *Archives of disease in childhood*. Arch Dis Child, 106(1), pp. 31–37. doi: 10.1136/ARCHDISCHILD-2020-319180.
Botstein, D. and Risch, N. (2003) 'Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease', *Nature Genetics*, 33(3s), pp. 228–237. doi: 10.1038/ng1090.
Bousman, C. A. *et al.* (2019) 'Pharmacogenetic tests and depressive symptom remission: A meta-analysis of randomized controlled trials', *Pharmacogenomics*. Future Medicine Ltd., 20(1), pp. 37–47. doi: 10.2217/pgs-2018-0142.
Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) 'Leading Edge Perspective An Expanded View of Complex Traits: From Polygenic to Omnigenic'. doi: 10.1016/j.cell.2017.05.038.
Brandler, W. M. *et al.* (2018) 'Paternally inherited cis-regulatory structural variants are associated with autism.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 360(6386), pp. 327–331. doi: 10.1126/science.aan2261.
Broad Institute (2019) 'Picard toolkit', *Broad Institute, GitHub repository*, \url{http:
Buxbaum, J. D. *et al.* (2012) 'The Autism Sequencing Consortium: Large-Scale, High-Throughput Sequencing in Autism Spectrum Disorders', *Neuron*, 76(6), pp. 1052–1056. doi: 10.1016/j.neuron.2012.12.008.
Buxbaum, J. D. *et al.* (2014) 'The Autism Simplex Collection: an international, expertly phenotyped autism sample for genetic and phenotypic analyses', *Molecular Autism*. BioMed Central, 5(1), p. 34. doi: 10.1186/2040-2392-5-34.
Buxbaum, J. D. *et al.* (2020) 'Not All Autism Genes Are Created Equal: A Response to Myers et al.', *American Journal of Human Genetics*. Cell Press, pp. 1000–1003. doi: 10.1016/j.ajhg.2020.09.013.
Callaghan, D. B. *et al.* (2019) 'Whole genome sequencing and variant discovery in the ASPIRE autism spectrum disorder cohort', *Clinical Genetics*, p. cge.13556. doi: 10.1111/cge.13556.
Carrieri, D. *et al.* (2019) 'Recontacting patients in clinical genetics services: recommendations of the European Society of Human Genetics', *European Journal of Human Genetics*. Nature Publishing Group, 27(2), pp. 169–182. doi: 10.1038/s41431-018-0285-1.
Carroll, L. S. and Owen, M. J. (2009) 'Genetic overlap between autism, schizophrenia and bipolar disorder', *Genome Medicine*, 1(10), p. 102. doi: 10.1186/gm102.
Carvill, G. L. *et al.* (2018) 'Aberrant Inclusion of a Poison Exon Causes Dravet Syndrome and

Related SCN1A-Associated Genetic Epilepsies', *American Journal of Human Genetics*. Cell Press, 103(6), pp. 1022–1029. doi: 10.1016/j.ajhg.2018.10.023.
Casey, J. P. *et al.* (2012) 'A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder', *Human Genetics*, 131(4), pp. 565–579. doi: 10.1007/s00439-011-1094-6.
Chakrabarti, S. and Fombonne, E. (2001) 'Pervasive developmental disorders in preschool children.', *JAMA*, 285(24), pp. 3093–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11427137 (Accessed: 31 January 2019).
Chapman, K. L. *et al.* (2001) 'Mutations in the region encoding the von Willebrand factor A domain of matrilin-3 are associated with multiple epiphyseal dysplasia', *Nature Genetics*, 28(4), pp. 393–396. doi: 10.1038/ng573.
Cirino, A. L. and Ho, C. (1993) *Hypertrophic Cardiomyopathy Overview*, *GeneReviews*®. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301725 (Accessed: 27 April 2021).
Córdova-Fletes, C. *et al.* (2022) 'A chromoanagenesis-driven ultra-complex t(5;7;21)dn truncates neurodevelopmental genes in a disabled boy as revealed by whole-genome sequencing', *European journal of medical genetics*. Eur J Med Genet, 65(10). doi: 10.1016/J.EJMG.2022.104579.
Creese, B. *et al.* (2019) 'Examining the association between genetic liability for schizophrenia and psychotic symptoms in Alzheimer's disease', *Translational Psychiatry*. Nature Publishing Group, 9(1). doi: 10.1038/s41398-019-0592-5.
D. Kashef-Haghighi et al. (2016) 'Random Forest Model for Identifying LCL-derived Mutations Enabling the Use of LCL DNA in Rare De Novo Variant Studies', *ASHG*.
D'Abate, L. *et al.* (2019) 'Predictive impact of rare genomic copy number variations in siblings of individuals with autism spectrum disorders', *Nature Communications*, 10(1), p. 5519. doi: 10.1038/s41467-019-13380-2.
Danecek, P. *et al.* (2011) 'The variant call format and VCFtools', *Bioinformatics*. Oxford Academic, 27(15), pp. 2156–2158. doi: 10.1093/BIOINFORMATICS/BTR330.
Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*. Gigascience, 10(2). doi: 10.1093/GIGASCIENCE/GIAB008.
David, K. L. *et al.* (2019) 'Patient re-contact after revision of genomic test results: points to consider—a statement of the American College of Medical Genetics and Genomics (ACMG)', *Genetics in Medicine*. Nature Publishing Group, 21(4), pp. 769–771. doi: 10.1038/s41436-018-0391-z.
Deignan, J. L. *et al.* (2019) 'Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the American College of Medical Genetics and Genomics (ACMG)', *GENETICS in MEDICINE*. Nature Publishing Group, 21(6), pp. 1267–1270. doi: 10.1038/s41436.
Dietz, H. (2017) *Gene Reviews: Marfan Syndrome*, *GeneReviews*®. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301510 (Accessed: 27 April 2021).
Dome, J. S. and Huff, V. (1993) *Wilms Tumor Predisposition*, *GeneReviews*®. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301471 (Accessed: 27 April 2021).
Ebbert, M. T. W. *et al.* (2019) 'Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight', *Genome Biology*. BioMed Central Ltd., 20(1), pp. 1–23. doi: 10.1186/s13059-019-1707-2.
Eeltink, E. *et al.* (2021) 'Polygenic risk scores for genetic counseling in psychiatry: Lessons learned from other fields of medicine', *Neuroscience & Biobehavioral Reviews*. Pergamon, 121, pp. 119–127. doi: 10.1016/J.NEUBIOREV.2020.11.021.
Ellingford, J. M. *et al.* (2021) 'Recommendations for clinical interpretation of variants found in non-coding regions of the genome', *medRxiv*. Cold Spring Harbor Laboratory Press, p. 2021.12.28.21267792. doi: 10.1101/2021.12.28.21267792.
Else, T., Greenberg, S. and Fishbein, L. (1993) *Hereditary Paraganglioma-Pheochromocytoma Syndromes*, *GeneReviews*®. University of Washington, Seattle.

Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301715 (Accessed: 27 April 2021).
Eng, C. (1993) *Multiple Endocrine Neoplasia Type 2*, *GeneReviews®*. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301434 (Accessed: 27 April 2021).
Eshraghi, A. A. *et al.* (2018) 'Epigenetics and Autism Spectrum Disorder: Is There a Correlation?', *Frontiers in Cellular Neuroscience*. Frontiers, 12, p. 78. doi: 10.3389/fncel.2018.00078.
Feliciano, P. *et al.* (2019) 'Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes', *npj Genomic Medicine*. Nature Publishing Group, 4(1), pp. 1–14. doi: 10.1038/s41525-019-0093-8.
Feng, Y.-C. A. *et al.* (2022) 'Psychiatric manifestations of rare variation in medically actionable genes: a PheWAS approach', *BMC genomics*. BMC Genomics, 23(1). doi: 10.1186/S12864-022-08600-X.
Firth, H. V. *et al.* (2009) 'DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources', *The American Journal of Human Genetics*. Elsevier, 84(4), pp. 524–533. doi: 10.1016/J.AJHG.2009.03.010.
Fitzpatrick, D. R. and Firth, H. V. (2020) 'Genomically Aided Diagnosis of Severe Developmental Disorders', *Annual review of genomics and human genetics*. Annu Rev Genomics Hum Genet, 21, pp. 327–349. doi: 10.1146/ANNUREV-GENOM-120919-082329.
Fombonne, E. (2003) 'Epidemiological surveys of autism and other pervasive developmental disorders: An update', *Journal of Autism and Developmental Disorders*, pp. 365–382. doi: 10.1023/A:1025054610557.
Gandal, M. J. *et al.* (2018) 'Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 359(6376), pp. 693–697. doi: 10.1126/science.aad6469.
Gardner, E. J. *et al.* (2019) 'Contribution of retrotransposition to developmental disorders', *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–10. doi: 10.1038/s41467-019-12520-y.
Gaugler, T. *et al.* (2014) 'Most genetic risk for autism resides with common variation.', *Nature genetics*. NIH Public Access, 46(8), pp. 881–5. doi: 10.1038/ng.3039.
Gene-Disease Validity Standard Operating Procedures, Version 7 - ClinGen | Clinical Genome Resource (no date). Available at: https://clinicalgenome.org/docs/summary-of-updates-to-the-clingen-gene-clinical-validity-curation-sop-version-7/ (Accessed: 2 February 2023).
Genetic Testing Statement | ISPG - International Society of Psychiatric Genetics (no date) *2013*. Available at: https://ispg.net/genetic-testing-statement/ (Accessed: 3 March 2021).
Geschwind, D. H. and Flint, J. (2015) 'Genetics and genomics of psychiatric disease', *Science*. American Association for the Advancement of Science, 349(6255), pp. 1489–1494. doi: 10.1126/SCIENCE.AAA8954.
Gilissen, C. *et al.* (2014) 'Genome sequencing identifies major causes of severe intellectual disability', *Nature*, 511(7509), pp. 344–347. doi: 10.1038/nature13394.
Glahn, D. C. *et al.* (2019) 'Rediscovering the value of families for psychiatric genetics research', *Molecular Psychiatry*. Nature Publishing Group, pp. 523–535. doi: 10.1038/s41380-018-0073-x.
Green, R. C. *et al.* (2013) 'American College of Medical Genetics and Genomics: ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing', *Genet Med*, 15. doi: 10.1038/gim.2013.73.
Grove, J. *et al.* (2019) 'Identification of common genetic risk variants for autism spectrum disorder', *Nature Genetics*. Nature Publishing Group, 51(3), pp. 431–444. doi: 10.1038/s41588-019-0344-8.
Gunning, A. C. *et al.* (2021) 'Assessing performance of pathogenicity predictors using clinically relevant variant datasets', *Journal of Medical Genetics*. BMJ Publishing Group Ltd, 58(8), pp. 547–555. doi: 10.1136/JMEDGENET-2020-107003.
Halko, N., Martinsson, P.-G. and Tropp, J. A. (2009) 'Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions', *SIAM Review*, 53(2), pp. 217–288. Available at: http://arxiv.org/abs/0909.4061 (Accessed: 10 June 2020).

Hall-Flavin, D. K. *et al.* (2013) 'Utility of integrated pharmacogenomic testing to support the treatment of major depressive disorder in a psychiatric outpatient setting', *Pharmacogenetics and Genomics*. Pharmacogenet Genomics, 23(10), pp. 535–548. doi: 10.1097/FPC.0b013e3283649b9a.
Han, Z. *et al.* (2020) 'Antisense oligonucleotides increase Scn1a expression and reduce seizures and SUDEP incidence in a mouse model of Dravet syndrome', *Science Translational Medicine*. American Association for the Advancement of Science, 12(558). doi: 10.1126/SCITRANSLMED.AAZ6100.
Heyne, H. O. *et al.* (2018) 'De novo variants in neurodevelopmental disorders with epilepsy', *Nature Genetics*. Nature Publishing Group, 50(7), pp. 1048–1053. doi: 10.1038/s41588-018-0143-7.
Hoang, N., Buchanan, J. A. and Scherer, S. W. (2018) 'Heterogeneity in clinical sequencing tests marketed for autism spectrum disorders', *npj Genomic Medicine*, 3(1), p. 27. doi: 10.1038/s41525-018-0066-3.
Horton, R. H. and Lucassen, A. M. (2019) 'Recent developments in genetic/genomic medicine', *Clinical Science (London, England : 1979)*. Portland Press Ltd, 133(5), p. 697. doi: 10.1042/CS20180436.
Horton, R. and Lucassen, A. (2019) 'Consent and Autonomy in the Genomics Era', *Current genetic medicine reports*. Curr Genet Med Rep, 7(2), pp. 85–91. doi: 10.1007/S40142-019-00164-9.
Hussain, T. and Church, J. M. (2020) 'Juvenile polyposis syndrome', *Clinical Case Reports*. Wiley-Blackwell Publishing Ltd, 8(1), pp. 92–95. doi: 10.1002/ccr3.2616.
Iafrate, A. J. *et al.* (2004) 'Detection of large-scale variation in the human genome', *Nature Genetics*, 36(9), pp. 949–951. doi: 10.1038/ng1416.
Idos, G. and Valle, L. (2004) 'Lynch Syndrome. 2004 Feb 5 [Updated 2018 Apr 12]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2019.', *GeneReviews(®)*. University of Washington, Seattle. Available at: https://www.ncbi.nlm.nih.gov/books/NBK1211/ (Accessed: 27 April 2021).
Ioannidis, N. M. *et al.* (2016) 'REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants', *American Journal of Human Genetics*. Elsevier, 99(4), p. 877. doi: 10.1016/J.AJHG.2016.08.016.
Iossifov, I. *et al.* (2012) 'De novo gene disruptions in children on the autistic spectrum', *Neuron*. Neuron, 74(2), pp. 285–299. doi: 10.1016/J.NEURON.2012.04.009.
Iossifov, I. *et al.* (2014) 'The contribution of de novo coding mutations to autism spectrum disorder', *Nature*, 515(7526), pp. 216–221. doi: 10.1038/nature13908.
Itoh-Satoh, M. *et al.* (2002) 'Titin Mutations as the Molecular Basis for Dilated Cardiomyopathy', *Biochemical and Biophysical Research Communications*. Academic Press, 291(2), pp. 385–393. doi: 10.1006/BBRC.2002.6448.
Jagadeesh, K. A. *et al.* (2016) 'M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity', *Nature Genetics*. Nature Publishing Group, 48(12), pp. 1581–1586. doi: 10.1038/ng.3703.
Jeste, S. S. and Geschwind, D. H. (2014) 'Disentangling the heterogeneity of autism spectrum disorder through genetic findings', *Nature reviews. Neurology*. NIH Public Access, 10(2), p. 74. doi: 10.1038/NRNEUROL.2013.278.
JI, N. *et al.* (2018) 'What Should a Psychiatrist Know About Genetics? Review and Recommendations From the Residency Education Committee of the International Society of Psychiatric Genetics', *The Journal of clinical psychiatry*. J Clin Psychiatry, 80(1). doi: 10.4088/JCP.17NR12046.
Ju, T. and Cummings, R. D. (2005) 'Protein glycosylation: chaperone mutation in Tn syndrome', *Nature*. Nature, 437(7063), p. 1252. doi: 10.1038/4371252A.
Kalia, S. S. *et al.* (2017) 'Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics', *Genetics in Medicine*. Nature Publishing Group, 19(2), pp. 249–255. doi: 10.1038/gim.2016.190.

Karczewski, K. J. *et al.* (2019) 'Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes', *bioRxiv*. Cold Spring Harbor Laboratory, p. 531210. doi: 10.1101/531210.
Karczewski, K. J. *et al.* (2020) 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*. Nature Research, 581(7809), pp. 434–443. doi: 10.1038/s41586-020-2308-7.
Kas, M. J. *et al.* (2014) 'Assessing behavioural and cognitive domains of autism spectrum disorders in rodents: current status and future perspectives', *Psychopharmacology*. Psychopharmacology (Berl), 231(6), pp. 1125–1146. doi: 10.1007/S00213-013-3268-5.
Kendall, K. M. *et al.* (2019) 'Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: Analysis of the UK Biobank', *British Journal of Psychiatry*. Cambridge University Press, 214(5), pp. 297–304. doi: 10.1192/bjp.2018.301.
Khera, A. V. *et al.* (2018) 'Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations', *Nature Genetics*. Nature Publishing Group, pp. 1219–1224. doi: 10.1038/s41588-018-0183-z.
Kircher, M. *et al.* (2014) 'A general framework for estimating the relative pathogenicity of human genetic variants.', *Nature genetics*. NIH Public Access, 46(3), pp. 310–5. doi: 10.1038/ng.2892.
Klei, L. *et al.* (2012) 'Common genetic variants, acting additively, are a major source of risk for autism', *Molecular Autism*. BioMed Central, 3(1), p. 9. doi: 10.1186/2040-2392-3-9.
Ko, Y. *et al.* (2004) 'Matrilin-3 Is Dispensable for Mouse Skeletal Growth and Development', *Molecular and Cellular Biology*. American Society for Microbiology, 24(4), pp. 1691–1699. doi: 10.1128/MCB.24.4.1691-1699.2004/ASSET/57D66A0D-540B-4AF8-A5D6-89712BFB118E/ASSETS/GRAPHIC/ZMB0040413090007.JPEG.
Kreiser, N. L. and White, S. W. (2013) 'ASD in Females: Are We Overstating the Gender Difference in Diagnosis?', *Clinical Child and Family Psychology Review 2013 17:1*. Springer, 17(1), pp. 67–84. doi: 10.1007/S10567-013-0148-9.
De La Torre-Ubieta, L. *et al.* (2016) 'Advancing the understanding of autism disease mechanisms through genetics', *Nature Medicine 2016 22:4*. Nature Publishing Group, 22(4), pp. 345–361. doi: 10.1038/nm.4071.
LaDuca, H. *et al.* (2019) 'A clinical guide to hereditary cancer panel testing: evaluation of gene-specific cancer associations and sensitivity of genetic testing criteria in a cohort of 165,000 high-risk patients', *Genetics in Medicine 2019 22:2*. Nature Publishing Group, 22(2), pp. 407–415. doi: 10.1038/s41436-019-0633-8.
Lai, M.-C. *et al.* (2019) 'Prevalence of co-occurring mental health diagnoses in the autism population: a systematic review and meta-analysis', *The Lancet Psychiatry*. Elsevier, 6(10), pp. 819–829. doi: 10.1016/S2215-0366(19)30289-5.
Landrum, M. J. *et al.* (2014) 'ClinVar: Public archive of relationships among sequence variation and human phenotype', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1113.
Landrum, M. J. *et al.* (2018) 'ClinVar: Improving access to variant interpretations and supporting evidence', *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D1062–D1067. doi: 10.1093/nar/gkx1153.
Lanktree, M. B. *et al.* (2015) 'Genetic meta-analysis of 15,901 African Americans identifies variation in EXOC3L1 is associated with HDL concentration', *Journal of Lipid Research*. American Society for Biochemistry and Molecular Biology, 56(9), p. 1781. doi: 10.1194/JLR.P059477.
Lauritsen, M. B., Pedersen, C. B. and Mortensen, P. B. (2005) 'Effects of familial risk factors and place of birth on the risk of autism: a nationwide register-based study', *Journal of Child Psychology and Psychiatry*. Wiley/Blackwell (10.1111), 46(9), pp. 963–971. doi: 10.1111/j.1469-7610.2004.00391.x.
Leblond, C. S. *et al.* (2019) 'Both rare and common genetic variants contribute to autism in the Faroe Islands', *npj Genomic Medicine*. Nature Publishing Group, 4(1), p. 1. doi: 10.1038/s41525-018-0075-2.
Lee, P. H. *et al.* (2019) 'Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders', *Cell*. Elsevier, 179(7), pp. 1469-1482.e11. doi:

10.1016/J.CELL.2019.11.020.
Lee, S. and Gleeson, J. G. (2020) 'Closing in on Mechanisms of Open Neural Tube Defects', *Trends in Neurosciences*. Elsevier Ltd, pp. 519–532. doi: 10.1016/j.tins.2020.04.009.
Lee, S. H. *et al.* (2013) 'Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs', *Nature Genetics*, 45(9), pp. 984–994. doi: 10.1038/ng.2711.
Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*. Nature Publishing Group, 536(7616), pp. 285–291. doi: 10.1038/nature19057.
Leppa, V. M. *et al.* (2016) 'Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families', *The American Journal of Human Genetics*. Cell Press, 99(3), pp. 540–554. doi: 10.1016/J.AJHG.2016.06.036.
Levy, D. *et al.* (2011) 'Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders', *Neuron*. Neuron, 70(5), pp. 886–897. doi: 10.1016/j.neuron.2011.05.015.
Lewis, C. M. and Vassos, E. (2020) 'Polygenic risk scores: From research tools to clinical instruments', *Genome Medicine*. BioMed Central Ltd., p. 44. doi: 10.1186/s13073-020-00742-5.
Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.
Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM'. Available at: http://arxiv.org/abs/1303.3997 (Accessed: 25 July 2019).
Lichter-Konecki, U. *et al.* (2016) *Ornithine Transcarbamylase Deficiency [Last Update 2016 April 14]*, *GeneReviews*®. University of Washington, Seattle. Available at: https://www.ncbi.nlm.nih.gov/books/NBK154378/ (Accessed: 27 April 2021).
Liu, X. *et al.* (2016) 'dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs', *Human Mutation*. John Wiley and Sons Inc., 37(3), pp. 235–241. doi: 10.1002/humu.22932.
Liu, X. *et al.* (2020) 'dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs', *Genome Medicine*. BioMed Central Ltd, 12(1), pp. 1–8. doi: 10.1186/S13073-020-00803-9/FIGURES/4.
Liu, X., Li, Y. I. and Pritchard, J. K. (2019) 'Trans Effects on Gene Expression Can Drive Omnigenic Inheritance.', *Cell*. Elsevier, 177(4), pp. 1022-1034.e6. doi: 10.1016/j.cell.2019.04.014.
Loeys, B. L. (2017) 'Loeys-Dietz Syndrome', in *Aneurysms-Osteoarthritis Syndrome: SMAD3 Gene Mutations*. Elsevier Inc., pp. 55–61. doi: 10.1016/B978-0-12-802708-0.00007-7.
Loomes, R., Hull, L. and Mandy, W. P. L. (2017) 'What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis', *Journal of the American Academy of Child & Adolescent Psychiatry*. Elsevier, 56(6), pp. 466–474. doi: 10.1016/J.JAAC.2017.03.013.
Lord, C. *et al.* (2020) 'Autism spectrum disorder', *Nature Reviews Disease Primers*. Nature Research, 6(1), pp. 1–23. doi: 10.1038/s41572-019-0138-4.
Lynch, S. A. and Borg, I. (2016) 'Wide disparity of clinical genetics services and EU rare disease research funding across Europe', *Journal of community genetics*. J Community Genet, 7(2), pp. 119–126. doi: 10.1007/S12687-015-0256-Y.
Manchia, M. *et al.* (2013) 'The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases', *PLoS ONE*. Edited by A. Reif. Public Library of Science, 8(10), p. e76295. doi: 10.1371/journal.pone.0076295.
Manichaikul, A. *et al.* (2010) 'Robust relationship inference in genome-wide association studies', 26(22), pp. 2867–2873. doi: 10.1093/bioinformatics/btq559.
Maragna, R. and Napolitano, C. (2018) 'Catecholaminergic Polymorphic Ventricular Tachycardia', in *Cardiac and Vascular Biology*. Springer Science and Business Media Deutschland GmbH, pp. 231–256. doi: 10.1007/978-3-319-77812-9_10.
Martin, A. R. *et al.* (2019) 'PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels', *Nature Genetics*. Nature Publishing Group, pp. 1560–1565. doi: 10.1038/s41588-019-0528-2.
Martinez-Granero, F. *et al.* (2021) 'Comparison of the diagnostic yield of aCGH and genome-

wide sequencing across different neurodevelopmental disorders', *npj Genomic Medicine*. Nature Research, 6(1), pp. 1–12. doi: 10.1038/s41525-021-00188-7.
McGuire, A. L. *et al.* (2020) 'The road ahead in genetics and genomics', *Nature Reviews Genetics*. Nature Research, pp. 1–16. doi: 10.1038/s41576-020-0272-6.
McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, M. (no date) *Online Mendelian Inheritance in Man, OMIM®*. Available at: https://omim.org/.
McNally, E., MacLeod, H. and Dellefave-Castillo, L. (1993) *Arrhythmogenic Right Ventricular Cardiomyopathy*, *GeneReviews®*. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301310 (Accessed: 27 April 2021).
MENDELIAN.CO (2019) *Autism Spectrum Disorders: Definition and top 150+ rare diseases related to them. | MENDELIAN.CO*. Available at: https://www.mendelian.co/autism-spectrum-disorders-150-rare-diseases-related (Accessed: 3 April 2019).
Mester, J. L. (2016) 'PTEN hamartoma tumor syndrome', in *Intestinal Polyposis Syndromes: Diagnosis and Management*. Springer International Publishing, pp. 87–100. doi: 10.1007/978-3-319-28103-2_7.
Miga, K. H. *et al.* (2020) 'Telomere-to-telomere assembly of a complete human X chromosome', *Nature*. Nature Research, pp. 1–9. doi: 10.1038/s41586-020-2547-7.
Milewicz, D. M. and Regalado, E. (1993) *Heritable Thoracic Aortic Disease Overview*, *GeneReviews®*. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301299 (Accessed: 27 April 2021).
Miller, D. T. *et al.* (2021) 'ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG)', *Genetics in Medicine 2021 23:8*. Nature Publishing Group, 23(8), pp. 1381–1390. doi: 10.1038/s41436-021-01172-3.
Miosge, L. A. *et al.* (2015) 'Comparison of predicted and actual consequences of missense mutations', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 112(37), pp. E5189–E5198. doi: 10.1073/pnas.1511585112.
Mitra, I. *et al.* (2021) 'Patterns of de novo tandem repeat mutations and their role in autism', *Nature 2020 589:7841*. Nature Publishing Group, 589(7841), pp. 246–250. doi: 10.1038/s41586-020-03078-7.
Mn, W. *et al.* (2021) 'Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation', *BMJ (Clinical research ed.)*. NLM (Medline), 372, p. n214. doi: 10.1136/bmj.n214.
Moeschler, J. B. *et al.* (2014) 'Comprehensive Evaluation of the Child With Intellectual Disability or Global Developmental Delays', *Pediatrics*. American Academy of Pediatrics, 134(3), pp. e903–e918. doi: 10.1542/PEDS.2014-1839.
Morris, N. *et al.* (2015) 'Novel approaches to the analysis of family data in genetic epidemiology', *Frontiers in Genetics*. Frontiers Research Foundation, 5(FEB), p. 27. doi: 10.3389/FGENE.2015.00027/BIBTEX.
Mousavi, N. *et al.* (2019) 'Profiling the genome-wide landscape of tandem repeat expansions', *Nucleic Acids Research*. Oxford University Press, 47(15), p. 90. doi: 10.1093/nar/gkz501.
Myers, S. M., Challman, T. D., Bernier, R., *et al.* (2020) 'Insufficient Evidence for "Autism-Specific" Genes', *American Journal of Human Genetics*. Cell Press, 106(5), pp. 587–595. doi: 10.1016/j.ajhg.2020.04.004.
Myers, S. M., Challman, T. D., Martin, C. L., *et al.* (2020) 'Response to Buxbaum et al.', *American Journal of Human Genetics*. Cell Press, p. 1004. doi: 10.1016/j.ajhg.2020.09.012.
Nelson, Q. *et al.* (2013) 'A population-based analysis of clustering identifies a strong genetic contribution to lethal prostate cancer', *Frontiers in Genetics*. Frontiers, 4(AUG), p. 152. doi: 10.3389/FGENE.2013.00152/BIBTEX.
Ng, P. C. and Henikoff, S. (2003) 'SIFT: Predicting amino acid changes that affect protein function.', *Nucleic acids research*. Oxford University Press, 31(13), pp. 3812–4. doi: 10.1093/nar/gkg509.
Ní Ghrálaigh, F. *et al.* (2022) 'Brief Report: Evaluating the Diagnostic Yield of Commercial Gene Panels in Autism', *Journal of Autism and Developmental Disorders*. Springer, pp. 1–5. doi: 10.1007/S10803-021-05417-7/TABLES/1.

Ní Ghrálaigh, F., Gallagher, L. and Lopez, L. M. (2020) 'Autism spectrum disorder genomics: The progress and potential of genomic technologies', *Genomics*, 112(6). doi: 10.1016/j.ygeno.2020.09.022.
Niroula, A. and Vihinen, M. (2019) 'How good are pathogenicity predictors in detecting benign variants?', *PLOS Computational Biology*. Public Library of Science, 15(2), p. e1006481. doi: 10.1371/JOURNAL.PCBI.1006481.
Northrup, H. *et al.* (1993) *Tuberous Sclerosis Complex-GeneReviews®*, *GeneReviews®*. University of Washington, Seattle. Available at: https://www.ncbi.nlm.nih.gov/books/NBK1220/ (Accessed: 27 April 2021).
Noyes, M. D. *et al.* (2022) 'Familial long-read sequencing increases yield of de novo mutations', *The American Journal of Human Genetics*. Cell Press, 109(4), pp. 631–646. doi: 10.1016/J.AJHG.2022.02.014.
Nunes, A., Trappenberg, T. and Alda, M. (2020) 'The definition and measurement of heterogeneity', *Translational Psychiatry*. Springer Nature, p. 299. doi: 10.1038/s41398-020-00986-0.
Nurk, S. *et al.* (2022) 'The complete sequence of a human genome', *Science*. American Association for the Advancement of Science, 376(6588), pp. 44–53. doi: 10.1126/SCIENCE.ABJ6987/SUPPL_FILE/SCIENCE.ABJ6987_MDAR_REPRODUCIBILITY_CHECKLIST.PDF.
O ' Roak, B. J. *et al.* (2012) 'Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations', *Nature*. Nature, 485(7397), pp. 246–250. doi: 10.1038/NATURE10989.
Ozonoff, S. *et al.* (2011) 'Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study.', *Pediatrics*. American Academy of Pediatrics, 128(3), pp. e488-95. doi: 10.1542/peds.2010-2825.
Pain, O. *et al.* (2021) 'A Tool for Translating Polygenic Scores onto the Absolute Scale Using Summary Statistics', *medRxiv*. Cold Spring Harbor Laboratory Press, p. 2021.04.16.21255481. doi: 10.1101/2021.04.16.21255481.
Pedersen, B. S. and Quinlan, A. R. (2017) 'Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy', *American Journal of Human Genetics*. Cell Press, 100(3), pp. 406–413. doi: 10.1016/j.ajhg.2017.01.017.
Pedersen, C. B. *et al.* (2018) 'The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders', *Molecular Psychiatry*. Nature Publishing Group, 23(1), pp. 6–14. doi: 10.1038/mp.2017.196.
Petrucelli, N., Daly, M. B. and Pal, T. (1993) *BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer*, *GeneReviews®*. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301425 (Accessed: 27 April 2021).
Pinto, D. *et al.* (2010) 'Functional impact of global rare copy number variation in autism spectrum disorders', *Nature*, 466(7304), pp. 368–372. doi: 10.1038/nature09146.
Psychiatric GWAS Consortium Bipolar Disorder Working Group, P. *et al.* (2011) 'Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4.', *Nature genetics*. Inserm, 43(10), pp. 977–83. doi: 10.1038/ng.943.
Purcell, S. *et al.* (2007) 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *American Journal of Human Genetics*. Elsevier, 81(3), p. 559. doi: 10.1086/519795.
Radford, E. J. *et al.* (2022) 'Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation', *medRxiv*. Cold Spring Harbor Laboratory Press, p. 2022.06.10.22276179. doi: 10.1101/2022.06.10.22276179.
Rebbeck, T. R. *et al.* (2018) 'Mutational spectrum in a worldwide study of 29,700 families with BRCA1 or BRCA2 mutations', *Human Mutation*. John Wiley and Sons Inc., 39(5), pp. 593–620. doi: 10.1002/humu.23406.
Rentzsch, P. *et al.* (2019) 'CADD: Predicting the deleteriousness of variants throughout the human genome', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D886–D894. doi: 10.1093/nar/gky1016.
Richards, A. J. *et al.* (1998) 'A single base mutation in COL5A2 causes Ehlers-Danlos

syndrome type II', *Journal of medical genetics*. J Med Genet, 35(10), pp. 846–848. doi: 10.1136/JMG.35.10.846.
Richards, S. *et al.* (2015) 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.', *Genetics in medicine : official journal of the American College of Medical Genetics*. NIH Public Access, 17(5), pp. 405–24. doi: 10.1038/gim.2015.30.
Ripke, S. *et al.* (2014) 'Biological insights from 108 schizophrenia-associated genetic loci', *Nature*. Nature Publishing Group, 511(7510), pp. 421–427. doi: 10.1038/nature13595.
Ronemus, M. *et al.* (2014) *The role of de novo mutations in the genetics of autism spectrum disorders*. doi: 10.1038/nrg3585.
De Rubeis, S. *et al.* (2014) 'Synaptic, transcriptional and chromatin genes disrupted in autism', *Nature*, 515(7526), pp. 209–215. doi: 10.1038/nature13772.
Ruzzo, E. K. *et al.* (2019) 'Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks', *Cell*. Cell Press, 178(4), pp. 850-866.e26. doi: 10.1016/j.cell.2019.07.015.
Sanders, S. J. *et al.* (2011) 'Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism', *Neuron*. Neuron, 70(5), pp. 863–885. doi: 10.1016/j.neuron.2011.05.002.
Sanders, S. J. *et al.* (2015) 'Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci', *Neuron*. Cell Press, 87(6), pp. 1215–1233. doi: 10.1016/j.neuron.2015.09.016.
Sandin, S. *et al.* (2017) 'The Heritability of Autism Spectrum Disorder', *JAMA*. American Medical Association, 318(12), p. 1182. doi: 10.1001/jama.2017.12141.
Satterstrom, F. K. *et al.* (2018) 'ASD and ADHD have a similar burden of rare protein-truncating variants', *bioRxiv*. Cold Spring Harbor Laboratory, p. 277707. doi: 10.1101/277707.
Satterstrom, F. K. *et al.* (2020) 'Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism', *Cell*. Cell Press, 180(3), pp. 568-584.e23. doi: 10.1016/j.cell.2019.12.036.
Savatt, J. M. *et al.* (2018) 'ClinGen's GenomeConnect registry enables patient-centered data sharing', *Human Mutation*. John Wiley and Sons Inc., 39(11), pp. 1668–1676. doi: 10.1002/humu.23633.
Savatt, J. M. and Myers, S. M. (2021) 'Genetic Testing in Neurodevelopmental Disorders', *Frontiers in Pediatrics*. Frontiers Media SA, 9, p. 526779. doi: 10.3389/FPED.2021.526779.
Schaaf, C. P. *et al.* (2020) 'A framework for an evidence-based gene list relevant to autism spectrum disorder', *Nature Reviews Genetics*. Nature Publishing Group, pp. 1–10. doi: 10.1038/s41576-020-0231-2.
Schizophrenia Working Group of the Psychiatric Genomics Consortium, S. W. G. of the P. G. *et al.* (2014) 'Biological insights from 108 schizophrenia-associated genetic loci.', *Nature*. Europe PMC Funders, 511(7510), pp. 421–7. doi: 10.1038/nature13595.
Schneider, V. A. *et al.* (2017) 'Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly', *Genome Research*. Cold Spring Harbor Laboratory Press, 27(5), pp. 849–864. doi: 10.1101/GR.213611.116.
van der Schoot, V. *et al.* (2022) 'Lessons learned from unsolicited findings in clinical exome sequencing of 16,482 individuals', *European journal of human genetics : EJHG*. Eur J Hum Genet, 30(2), pp. 170–177. doi: 10.1038/S41431-021-00964-0.
Sebat, J. *et al.* (2004) 'Large-Scale Copy Number Polymorphism in the Human Genome', *Science*, 305(5683), pp. 525–528. doi: 10.1126/science.1098918.
Sebat, J. *et al.* (2007) 'Strong Association of De Novo Copy Number Mutations with Autism', *Science*, 316(5823), pp. 445–449. doi: 10.1126/science.1138659.
So, H. C., Sham, P. C. and Valencia, A. (2017) 'Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits', *Bioinformatics*. Oxford Academic, 33(6), pp. 886–892. doi: 10.1093/BIOINFORMATICS/BTW745.
Srivastava, S. *et al.* (2019) 'Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental

disorders', *Genetics in Medicine*. Nature Publishing Group, p. 1. doi: 10.1038/s41436-019-0554-6.
Stahl, E. A. *et al.* (2019) 'Genome-wide association study identifies 30 loci associated with bipolar disorder', *Nature Genetics*. Nature Publishing Group, 51(5), pp. 793–803. doi: 10.1038/s41588-019-0397-8.
Stelzer, G. *et al.* (2016) 'The GeneCards suite: From gene data mining to disease genome sequence analyses', *Current Protocols in Bioinformatics*. John Wiley and Sons Inc., 2016(1), pp. 1.30.1-1.30.33. doi: 10.1002/cpbi.5.
Stenson, P. D. *et al.* (2017) 'The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies', *Human Genetics*. Springer Verlag, pp. 665–677. doi: 10.1007/s00439-017-1779-6.
Steward, C. A. *et al.* (2019) 'Re-annotation of 191 developmental and epileptic encephalopathy-associated genes unmasks de novo variants in SCN1A', *npj Genomic Medicine*. Nature Research, 4(1). doi: 10.1038/s41525-019-0106-7.
Strande, N. T. *et al.* (2017) 'Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource', *American Journal of Human Genetics*. Cell Press, 100(6), pp. 895–906. doi: 10.1016/j.ajhg.2017.04.015.
Tammimies, K. *et al.* (2015) 'Molecular Diagnostic Yield of Chromosomal Microarray Analysis and Whole-Exome Sequencing in Children With Autism Spectrum Disorder', *JAMA*. American Medical Association, 314(9), p. 895. doi: 10.1001/jama.2015.10078.
Tansey, K. E. *et al.* (2016) 'Common alleles contribute to schizophrenia in CNV carriers', *Molecular Psychiatry*, 21, pp. 1085–1089. doi: 10.1038/mp.2015.143.
The 1000 Genomes Project Consortium, T. 1000 G. P. (2015) 'A global reference for human genetic variation', *Nature*. Nature Publishing Group, 526(7571), pp. 68–74. doi: 10.1038/nature15393.
The Autism Genome Project Consortium, T. A. G. P. *et al.* (2007) 'Mapping autism risk loci using genetic linkage and chromosomal rearrangements', *Nature Genetics*. Nature Publishing Group, 39(3), pp. 319–328. doi: 10.1038/ng1985.
The UK10K Consortium, T. U. (2015) 'The UK10K project identifies rare variants in health and disease', *Nature*. Nature Publishing Group, 526(7571), pp. 82–90. doi: 10.1038/nature14962.
Tick, B. *et al.* (2016) 'Heritability of autism spectrum disorders: a meta-analysis of twin studies', *Journal of Child Psychology and Psychiatry*, 57(5), pp. 585–595. doi: 10.1111/jcpp.12499.
Trost, B. *et al.* (2022) 'Genomic architecture of Autism Spectrum Disorder from comprehensive whole-genome sequence annotation', *medRxiv*. Cold Spring Harbor Laboratory Press, 11, p. 2022.05.05.22274031. doi: 10.1101/2022.05.05.22274031.
Turner, T. N. *et al.* (2016) 'Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA', *The American Journal of Human Genetics*, 98(1), pp. 58–74. doi: 10.1016/j.ajhg.2015.11.023.
Turner, T. N. *et al.* (2017) 'Genomic Patterns of De Novo Mutation in Simplex Autism', *Cell*. Cell Press, 171(3), pp. 710-722.e12. doi: 10.1016/j.cell.2017.08.047.
Turner, T. N. *et al.* (2019) 'Sex-Based Analysis of De Novo Variants in Neurodevelopmental Disorders', *The American Journal of Human Genetics*. Elsevier, 0(0). doi: 10.1016/j.ajhg.2019.11.003.
Turro, E. *et al.* (2020) 'Whole-genome sequencing of patients with rare diseases in a national health system', *Nature*. Nature Research, 583(7814), pp. 96–102. doi: 10.1038/s41586-020-2434-2.
Veenstra-VanderWeele, J., Christian, S. L. and Cook, Jr., E. H. (2004) 'AUTISM AS A PARADIGMATIC COMPLEX GENETIC DISORDER', *Annual Review of Genomics and Human Genetics*. Annual Reviews, 5(1), pp. 379–405. doi: 10.1146/annurev.genom.5.061903.180050.
Vorstman, J. A. S. *et al.* (2013) 'No evidence that common genetic risk variation is shared between schizophrenia and autism', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 162(1), pp. 55–60. doi: 10.1002/ajmg.b.32121.
Wainschtein, P. *et al.* (2022) 'Assessing the contribution of rare variants to complex trait

heritability from whole-genome sequence data', *Nature Genetics 2022 54:3*. Nature Publishing Group, 54(3), pp. 263–273. doi: 10.1038/s41588-021-00997-7.
Wang, H. Z. *et al.* (2013) 'New insights into the genetic mechanism of IQ in autism spectrum disorders', *Frontiers in Genetics*. Frontiers, 0, p. 195. doi: 10.3389/FGENE.2013.00195.
Weiss, K. H. (1993) *Wilson Disease*, *GeneReviews*®. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20301685 (Accessed: 27 April 2021).
Werling, D. M. *et al.* (2018) 'An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder', *Nature Genetics*. Nature Publishing Group, 50(5), pp. 727–736. doi: 10.1038/s41588-018-0107-y.
Werling, D. M. and Geschwind, D. H. (2015) 'Recurrence rates provide evidence for sex-differential, familial genetic liability for autism spectrum disorders in multiplex families and twins', *Molecular Autism*. BioMed Central, 6(1). doi: 10.1186/S13229-015-0004-5.
Wickham, H. *et al.* (2019) 'Welcome to the Tidyverse', *Journal of Open Source Software*. The Open Journal, 4(43), p. 1686. doi: 10.21105/JOSS.01686.
Wigdor, E. M. *et al.* (2022) 'The female protective effect against autism spectrum disorder', *Cell Genomics*. Elsevier, 2(6), p. 100134. doi: 10.1016/J.XGEN.2022.100134.
Wigginton, J. E., Cutler, D. J. and Abecasis, G. R. (2005) 'A Note on Exact Tests of Hardy-Weinberg Equilibrium', *The American Journal of Human Genetics*, 76(5), pp. 887–893. doi: 10.1086/429864.
Woodbury-Smith, M. *et al.* (2017) 'Variable phenotype expression in a family segregating microdeletions of the NRXN1 and MBD5 autism spectrum disorder susceptibility genes', *npj Genomic Medicine*. Nature Publishing Group, 2(1), pp. 1–8. doi: 10.1038/s41525-017-0020-9.
Wray & Visscher (2008) *Estimating Trait Heritability*. Available at: https://www.nature.com/scitable/topicpage/estimating-trait-heritability-46889.
Wright, C. F. *et al.* (2015) 'Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data.', *Lancet (London, England)*. Elsevier, 385(9975), pp. 1305–14. doi: 10.1016/S0140-6736(14)61705-0.
Wright, C. F. *et al.* (2019) 'Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting', *American Journal of Human Genetics*. Cell Press, 104(2), pp. 275–286. doi: 10.1016/j.ajhg.2018.12.015.
Wright, C. F. *et al.* (2021) 'Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms', *American journal of human genetics*. Am J Hum Genet, 108(6), pp. 1083–1094. doi: 10.1016/J.AJHG.2021.04.025.
Yoon, S. *et al.* (2021) 'Rates of contributory de novo mutation in high and low-risk autism families', *Communications Biology 2021 4:1*. Nature Publishing Group, 4(1), pp. 1–10. doi: 10.1038/s42003-021-02533-z.
Youngblom, E., Pariani, M. and Knowles, J. W. (1993) *Familial Hypercholesterolemia*, *GeneReviews*®. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24404629 (Accessed: 27 April 2021).
Yuen, R. K. *et al.* (2016) 'Genome-wide characteristics of de novo mutations in autism', *npj Genomic Medicine*, p. 1. doi: 10.1038/npjgenmed.2016.27.
Yuen, R. K. C. *et al.* (2015) 'Whole-genome sequencing of quartet families with autism spectrum disorder', *Nature Medicine*, 21(2), pp. 185–191. doi: 10.1038/nm.3792.
Yuen, R. K. C. *et al.* (2017) 'Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder', *Nature Neuroscience*. Nature Publishing Group, 20(4), pp. 602–611. doi: 10.1038/nn.4524.
Yunis, A. and Bhonsale, A. (2020) 'Long QT syndrome', in *Cardiac Electrophysiology: Clinical Case Review*. Springer International Publishing, pp. 215–218. doi: 10.1007/978-3-030-28533-3_52.
Zarrei, M. *et al.* (2019) 'A large data resource of genomic copy number variation across neurodevelopmental disorders', *npj Genomic Medicine*. Nature Publishing Group, 4(1), pp. 1–13. doi: 10.1038/s41525-019-0098-3.
Zhou, J. *et al.* (2019) 'Whole-genome deep-learning analysis identifies contribution of

noncoding mutations to autism risk', *Nature Genetics*. Nature Publishing Group, 51(6), pp. 973–980. doi: 10.1038/s41588-019-0420-0.

Appendices

Appendix I-I: Irish Molecular Genetics Study in Autism REC: 2020-01 List 1 (17).

SJH/TUH Research Ethics Committee Secretariat
email: researchethics@tuh.ie

Mr Richard O'Conaill,
St James's Hospital,
James' Street,
Dublin 8

21st January 2020

Re: Irish Molecular Genetics Study in Autism

REC: 2020-01 List 1 (17)
(Please quote reference on all correspondence)

Date of Valid Submission to REC: 10.12.2019
Date of Ethical Review: 13.01.2020

Dear Mr O'Conaill,

Thank you for your correspondence in which you sent in an amendment to the above named study.

The Chairman has reviewed your response on behalf of the Committee, and gives approval for this study to proceed.

Documents Reviewed:

- Amendment Request, dated 10.12.2019
- SAF
- PIL/CF

Applicants must submit an annual report for ongoing projects and an end of project report upon completion of the study. It is the responsibility of the researcher/research team to ensure all aspects of the study are executed in compliance with the General Data Protection regulation (GDPR), Health Research Regulations and the Data Protection Act 2018.

Yours sincerely,

Sadhbh O'Neill

REC Officer – Dr Sadhbh O'Neill
SJH/TUH Research Ethics Committee

The SJH/TUH Joint Research and Ethics Committee operates in compliance with and is constituted in accordance with the European Communities (Clinical Trials on Medicinal Products for Human Use) Regulations 2004 & ICH GCP guidelines.

Appendix I-II: Genomics of Neurodevelopmental Disorders (Reference number BSRESC-2021-2402328).

MAYNOOTH UNIVERSITY RESEARCH ETHICS COMMITTEE
MAYNOOTH UNIVERSITY,
MAYNOOTH, CO. KILDARE, IRELAND

Dr Carol Barrett
Secretary to Maynooth University Research Ethics Committee

30 June 2021

Dr Lorna Lopez
Department of Biology
Maynooth University

Dear Lorna,

The Biomedical and Life Sciences Research Ethics Sub-Committee has reviewed the application for project title: **Genomics of Neurodevelopmental Disorders (Reference number BSRESC-2021-2402328)** and we would like to inform you that ethical approval has been granted.

Any deviations from the project details submitted to the ethics committee will require further evaluation. This ethical approval will expire on 31/07/2022.

Kind Regards,

Dr Carol Barrett
Secretary,
Maynooth University Research Ethics Committee

Reference Number
BSRESC-2021-2402328

Gene symbols as annotated by dbNSFP v4.0a		
ADGRG6	*HGSNAT*	*RECQL4*
AHDC1	*HR*	*RELN*
ALMS1	*HSPG2*	*RNF135*
AMPD1	*IFT140*	*SCARF2*
ANK2	*KCNB1*	*SCN4A*
ATP10A	*KIAA0586*	*SEPSECS*
C12orf57	*KIF7*	*SIK1*
CACNA2D3	*KMT2A*	*SKIV2L*
CCDC39	*LAMA2*	*SLC6A5*
CCDC65	*LMOD3*	*SNX5*
CDH23	*LRP4*	*SPEG*
CGNL1	*LYST*	*STAMBP*
COL11A1	*MAP1A*	*TBCK*
COL4A4	*MED25*	*TCOF1*
COL6A3	*NAV2*	*TOGARAM1*
COMP	*NEB*	*TRIP11*
CRB2	*NHS*	*TRPV4*
DNAH9	*NINL*	*TSC2*
DYM	*NPHP4*	*TTN*
EDA	*NTRK1*	*UNC80*
ENPP1	*OTC*	*WDR62*
EP400	*OTOGL*	*XRCC4*
ERCC6	*P3H1*	*ZC3H4*
FANCC	*PAPSS2*	
FAT4	*PCCB*	
FBXL4	*PEX10*	
FLNB	*PEX6*	
FREM2	*PHF3*	
GALT	*PKHD1*	
GATA2	*PLK4*	
GLI2	*PLXNA4*	
GLIS3	*PLXNB1*	
GPSM2	*PYGL*	
GRHL3	*RAI1*	

Supplemental Table 1 Candidate autism-relevant genes for curation arising from variant discovery in Cohort 4.

Authors (Year): Title	**Reported Case Details**	**Reported Variant Information** *(Variants checked in gnomAD (v2.1.1) in Oct. 2020)*	**Evidence Type**	**Suggested Points Per Case Default/(Range)**	**Final Score (incorporating genetic evidence, phenotype quality, expert input)**	**Notes (justification for score)**
Wang T, *et al.* (2016)	ID: SKLMG_M08710 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr11(GRCh37): g.20099182G>A, NM_001111018.1] Impact: splice-donor gnomAD: Not present Inheritance: Unknown	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Downgraded WES/WGS not performed
Wang T, *et al.* (2016)	ID: SKLMG_M26848 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr11(GRCh37): g.20067213G>A, NM_001111018.1] Impact: missense gnomAD: 1.03e-4 Inheritance: maternal inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)
Wang T, *et al.* (2016)	ID: SKLMG_M21717 Sex:	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing	Autosomal Dominant -> Other variant type	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without

	Phenotype: Autism	Variant reported: [Chr11(GRCh37): g.19970605G>A, NM_001111018.1] Impact: missense gnomAD: 1.37e-4 Inheritance: maternal inherited	not predicted/proven null			functional evidence, observed in gnomAD (score reduced to 0)
Wang T, *et al.* (2016)	ID: SKLMG_M26778 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr11(GRCh37): g.19970605G>A, NM_001111018.1] Impact: missense gnomAD: 1.37e-4 Inheritance: maternal inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)
Wang T, *et al.* (2016)	ID: SKLMG_M19652 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr11(GRCh37): g.20077419C>T, NM_001111018.1] Impact: missense gnomAD: 3.18e-5 Inheritance: maternal inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)

Wang T, *et al.* (2016)	ID: SKLMG_M01623 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr11(GRCh37): g.20067213G>A, NM_001111018.1] Impact: missense gnomAD: 1.03e-4 Inheritance: paternal inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)
Wang T, *et al.* (2016)	ID: SKLMG_M17623 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr11(GRCh37): g.20065735G>A, NM_001111018.1] Impact: missense gnomAD: 1.63e-4 Inheritance: paternal inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)
Wang T, *et al.* (2016)	ID: SKLMG_M26826 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr11(GRCh37): g.20125237C>T, NM_001111018.1] Impact: missense	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)

		gnomAD: 7.95e-6 Inheritance: paternal inherited				
Wang T, *et al.* (2016)	ID: SKLMG_M23133 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr11(GRCh37): g.20136247G>A, NM_001111018.1] Impact: missense gnomAD: 3.58e-5 Inheritance: paternal inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)
Sanders SJ, *et al.* (2012)	ID: 12241.p1 Sex: Female Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr11(hg19): g.20139742G>A, NM_145117, D2410N] Impact: Missense (DAMAGING *Warning! Low confidence.) gnomAD: 1.19e-5 Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0.5	Default score downgraded for genetic evidence: *de novo* missense variant with suggested functional evidence, observed in gnomAD (score reduced to 0.5)
O'Roak BJ, *et al.* (2012)	ID: 11459.p1 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr11(hg19): g.20119143C>Y] Impact: coding-synonymous gnomAD: Not present Inheritance: *de novo* (also lists father as parent of origin)	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	N/A	Not scored: unknown impact of synonymous variants; no functional data provided.

Lim ET, *et al.* (2017)	ID: 37434 Sex: Male Phenotype: Autism	Genotyping Method: WES; resequencing of post-zygotic mutations (PZMs) Variant reported: [Chr11(hg19), g. 20117286C>T, L1983] Impact: SYNONYMOUS_CODING gnomAD: 7.95e-5 Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	N/A	Not scored: unknown impact of synonymous variants; no functional data provided.
Lim ET, *et al.* (2017)	ID: NP053 Sex: Female Phenotype: Autism	Genotyping Method: WES; resequencing of post-zygotic mutations (PZMs) Variant reported: [Chr11(hg19), g. 19854079G>A, R35H] Impact: missense gnomAD: 4.24e-5 Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0	Default score downgraded for genetic evidence: de novo missense variant without functional evidence, observed in gnomAD (score reduced to 0)
Iossifov *et al.* (2014)	ID: 11397 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr11(hg19): g. 11:20057523:C:A] Impact: synonymous gnomAD: Not present Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	N/A	Not scored: unknown impact of synonymous variants; no functional data provided.
Iossifov *et al.* (2014)	ID: 12389 Sex: Male	Genotyping Method: WES Variant reported:	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0	Default score downgraded for genetic evidence: *de novo*

	Phenotype: Autism	[Chr11(hg19): g. 11:19914032:C:T] Impact: Missense gnomAD: 6.02e-5 Inheritance: *de novo*				missense variant without functional evidence, observed in gnomAD (score reduced to 0)
Iossifov *et al.* (2014)	ID: 14179 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr11(hg19): g. 11:20104534:G:C] Impact: intron gnomAD: Not present Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0	No points awarded - intronic variant. No evidence to suggest variant is pathogenic
Iossifov *et al.* (2014)	ID: 14604 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr11(hg19): g. 11:20117286:C:T] Impact: synonymous gnomAD: 7.95e-5 Inheritance: *de novo* (father parent of origin)	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0	Not scored: unknown impact of synonymous variants; no functional data provided.
Iossifov *et al.* (2014)	ID: 11459 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr11(hg19): g. 11:20119143:C:T] Impact: synonymous gnomAD: Not present Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0	Not scored: unknown impact of synonymous variants; no functional data provided.

Iossifov *et al.* (2014)	ID: 12241 Sex: Female Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr11(hg19): g. 11:20139742: G: A] Impact: missense gnomAD: 1.19e-5 Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0	Default score downgraded for genetic evidence: *de novo* missense variant without functional evidence, observed in gnomAD (score reduced to 0)
Guo H, *et al.* (2018)	ID: M08710 Sex: Male Phenotype: Autism	Variant scored in PMID: 27824329			N/A	Variant already scored

Table 7-1 Genetic evidence matrix for curation of NAV2 gene-phenotype relationship.

This scoring matrix follows the template proposed by Schaaf et al. (2020). Variant-level information was compiled from the publications specified for each report of a proband carrying a variant in the NAV2. Sequencing method, variant coordinates, genomic impact, gnomAD allele frequency and mode of inheritance are reported for each incidence of the variant.

Authors (Year): Title	Findings presented	Genotype information of model organism	Quality of the data presented	Evidence Type	Suggested Points Per Report Default/(Range)
Peeters PJ, *et al.* (2004)	General impaired acuity of several sensory systems (olfactory, auditory, visual and pain sensation) which in case of the visual system was corroborated by the morphological observation of hypoplasia of the optic nerve.	Wild-type, heterozygote, and homozygote unc53H2 mutant mice	Low confidence	Non-human model organism	2(0-4)

Table 7-2 Experimental evidence matrix for curation of NAV2 gene-phenotype relationship.

Presented is the gene scoring matrix for experimental evidence supporting the roles of NAV2 in autism. Scoring and justification are given following Schaaf et al.(2020) modified ClinGen curation framework. The evidence reported in this table is taken together with the variant-level evidence outlined in Table 7-1, to give an overall gene-disease curation classification.

Authors (Year): Title	**Reported Case Details**	**Reported Variant Information *(variants checked in gnomAD (v2.1.1) in Oct. 2020)***	**Evidence Type**	**Suggested Points Per Case Default/(Range)**	**Final Score (incorporating genetic evidence, phenotype quality, expert input)**	**Notes (justification for score)**
De Rubeis S, *et al.* (2014)	ID: AC02-1141-01 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr20(hg19), g.25477488A>C, NM_025176] Impact: Intronic gnomAD: Not present Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0	No points awarded - intronic variant that does not occur in a canonical splice site. No evidence to suggest variant is pathogenic
Iossifov I *et al.* (2014)	ID: 12036.p1 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr20(hg19): 25459809T>A, p. K651X] Impact: LoF_nonsense gnomAD: Not present Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	2	Default score applied: WES identifies a *de novo* nonsense variant does not present in gnomAD; high quality autism phenotyping
Leblond CS, *et al.* (2019)	ID: PN400119 Sex: Female Phenotype: Autism	Genotyping Method: Illumina SNP array (>4.3million SNPs) and WES Variant reported: 248KB [Chr20(hg19), g.	Paternally inherited deletion	N/A - Deletion involving more than one gene	0	Downgraded for lack of confidence in ID. Deletion also spans other genes Borderline to mild ID (Full-scale IQ [FSIQ] 50–85)

		25388358_ 25604606del] Impact: deletion gnomAD: NA Inheritance: *de novo*				
Ruzzo EK, *et al.* (2019)	ID: iHART2459 Sex: Male Phenotype: Autism	Genotyping Method: WGS Variant reported: [Chr20(hg19), g.25443018>A, ,] Impact: splice site donor-PTV gnomAD: not present Inheritance: maternal	Autosomal Dominant -> Predicted/Proven null variant	1.5/ (0-2)	1.5	Default awarded- splice site variant in protein truncating
Wu H, *et al.* (2019)	ID: GX0389.p1 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr20(hg19), g. 25479032G>A, p.Q698X, NM_025176] Impact: exonic stopgain gnomAD: Inheritance: paternal	Autosomal Dominant -> Other variant type not predicted/proven null	1.5/ (0-2)	1	Downgraded because of cognition score. Cognition score not presented for this individual

Table 7-3 Genetic evidence matrix for curation of NINL gene-phenotype relationship.

Note that no experimental evidence supports this gene-disease association and not experimental evidence contributes to the overall ClinGen association score reported.

Authors (Year): Title	**Reported Case Details**	**Reported Variant Information *(variants checked in gnomAD (v2.1.1) in Oct. 2020)***	**Evidence Type**	**Suggested Points Per Case Default/(Range)**	**Final Score (incorporating genetic evidence, phenotype quality, expert input)**	**Notes (justification for score)**
C Yuen RK *et al.* (2017)	ID: AU045514 Sex: Unknown Phenotype: Autism	Genotyping Method: WGS (Complete Genomics), validated with Sanger sequencing Variant reported: [Chr3(GRCh37): g. 54420739_ 54420740A>T] Impact: Splice site variant gnomAD: Not present. Inheritance: Inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0.5	Default awarded
De Rubeis S, *et al.* (2014)	ID: UK10K_SKUSE5080203 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr3(hg19): g.54872646G>T, E508X] Impact: LoF_nonsense gnomAD: Not present. Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	2	Default awarded
De Rubeis S, *et al.* (2014)	ID: 09C96031 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 55038892A>C, NM_018398] Impact: Intron gnomAD: 2.85e-5 Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	0	No points awarded - intronic variant present in gnomAD No evidence to suggest variant is pathogenic

De Rubeis S, *et al.* (2014)	ID: 10C114435 Sex: Male Phenotype: Father	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54925398G>A, V629M] Impact: NON_SYNONYMOUS_CODING gnomAD: Inheritance: Paternal variant (not in proband)		Not scored		Proband carries reference allele
De Rubeis S, *et al.* (2014)	ID: 10C114435 Sex: Male Phenotype: Father	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54930795G>A, D662N] Impact: NON_SYNONYMOUS_CODING gnomAD: Inheritance: Paternal variant (not in proband)		Not scored		Proband carries reference allele
De Rubeis S, *et al.* (2014)	ID: 09C91623 Sex: Female Phenotype: Mother	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54596896G>A, R110H] Impact: NON_SYNONYMOUS_CODING gnomAD: Inheritance: Maternal variant (not in proband)		Not scored		Proband carries reference allele
De Rubeis S, *et al.* (2014)	ID: DEASD_0336_001 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54913067G>A, R477Q] Impact:	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without

		NON_SYNONYMOUS_CODING gnomAD: 2.93e-5 Inheritance: Paternal inheritance				functional evidence, observed in gnomAD (score reduced to 0)
De Rubeis S, *et al.* (2014)	ID: 08C79339 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54922021G>A, A604T] Impact: NON_SYNONYMOUS_CODING gnomAD: 5.7e-5 Inheritance: Maternal inheritance	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)
De Rubeis S, *et al.* (2014)	ID: NDAR_INVWJ720YXQ_wes1 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54930795G>A, D662N] Impact: NON_SYNONYMOUS_CODING gnomAD: 1.6e-5 Inheritance: Paternal inheritance	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: inherited missense variant without functional evidence, observed in gnomAD (score reduced to 0)
De Rubeis S, *et al.* (2014)	ID: 10C107584 Sex: Male Phenotype: Father	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54913048A>G, R471G] Impact: NON_SYNONYMOUS_CODING gnomAD: Inheritance:		Not scored		Proband carries reference allele
De Rubeis S, *et al.* (2014)	ID: 10C108003 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54925422C>G, R637G] Impact:	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0.5	

		NON_SYNONYMOUS_CODING gnomAD: Not present Inheritance: Maternally inherited				
De Rubeis S, *et al.* (2014)	ID: DEautism_0077_600 Sex: Female Phenotype: Mother	Genotyping Method: WES Variant reported: [Chr3(hg19): g. 54922020C>G, D603E] Impact: NON_SYNONYMOUS_CODING gnomAD: Inheritance:		Not scored		Proband carries reference allele
Iossifov I *et al.* (2014)	ID: 13526.p1 Sex: Male Phenotype: Autism	Genotyping Method: WES Variant reported: [Chr3(hg19): 54921984A>G] Impact: LoF_3splice gnomAD: Not present Inheritance: *de novo*	Autosomal Dominant -> Variant is *de novo*	2/ (0-3)	2	
Guo *et al.* (2018	ID: SD0023.p1 Sex: Female Phenotype: Autism	Genotyping Method: Targeted sequencing of 211 autism candidate genes (Phase II-2); single-molecule molecular inversion probes Variant Reported: [Chr3(GRCh37): g. 54850898G>A, NM_018398: exon14, p. Arg3568Trp] Impact: splice-donor gnomAD: Not present Inheritance: Maternally inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: WES/WGS not used (-0.5)

Guo *et al.* (2018)	ID: M23096 Sex: Male Phenotype: Autism	Genotyping Method: Targeted sequencing of 211 autism candidate genes (Phase I); single-molecule molecular inversion probes Variant Reported: [Chr3(GRCh37): g. 54930847C>T, NM_018398: exon26, p.A773V] Impact: Missense gnomAD: Not present Inheritance: Maternally inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: WES/WGS not used (-0.5)
Guo *et al.* (2018)	ID: M08461 Sex: Female Phenotype: Autism	Genotyping Method: Targeted sequencing of 211 autism candidate genes (Phase I); single-molecule molecular inversion probes Variant Reported: [Chr3(GRCh37): g. 54930847C>T, NM_018398: exon26, p.A773V] Impact: Missense gnomAD: Not present Inheritance: Maternally inherited	Autosomal Dominant -> Other variant type not predicted/proven null	NOT SCORED	0	Variant already scored
Wang T, *et al.* (2016)	ID: SKLMG_M23096 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr3(GRCh37): g.54930847C>T, p. Ala773Val] Impact: Missense	Autosomal Dominant -> Other variant type not predicted/proven null	NOT SCORED	0	Variant already scored

		gnomAD: Inheritance: Maternally inherited				
Wang T, *et al.* (2016)	ID: SKLMG_M08461 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr3(GRCh37): g.54930847C>T, p. Ala773Val] Impact: Missense gnomAD: Inheritance: Maternally inherited	Autosomal Dominant -> Other variant type not predicted/proven null	NOT SCORED	0	Variant already scored
Wang T, *et al.* (2016)	ID: SKLMG_M23110 Sex: Phenotype: Autism	Genotyping Method: MIP-based resequencing, validated with PCR and Sanger sequencing Variant reported: [Chr3(GRCh37): g.54604066G>A, p. Ala275Thr] Impact: Missense gnomAD: Not present Inheritance: Paternal inherited	Autosomal Dominant -> Other variant type not predicted/proven null	0.5/ (0-1.5)	0	Default score downgraded for genetic evidence: WES/WGS not used (-0.5)

Table 7-4 Genetic evidence matrix for curation of CACNA2D3 gene-disease relationship.

Duplicate samples were identified in Guo et al. (2018) and Wang T, et al. (2016). Only one score has been taken per participant. Note that no experimental evidence supports this gene-disease association and not experimental evidence contributes to the overall ClinGen association score reported.

Appendix III-I: Genomic syndromes in autism: Using whole genome sequencing to investigate multiplex families with autism and associated neurodevelopmental conditions. Poster presented at Genomics of Rare Diseases (Wellcome Connecting Science), April 2022.

Genomic syndromes in autism:Using whole genome sequencing to investigate multiplex families with autism and associated neurodevelopmental conditions

Fiana Ní Ghrálaigh,[1,2] Aoife Coghlan,[1] Louise Gallagher[2] & Lorna M. Lopez[1]

1 Department of Biology, Maynooth University, Maynooth, Co. Kildare, Ireland.
2 Department of Psychiatry, Trinity College Dublin, Dublin, Ireland.

Contact: fiana.nighralaigh.2020@mumail.ie, lorna.lopez@mu.ie

Maynooth University
National University of Ireland Maynooth

Background & Aim

- Autism is a prevalent neurodevelopmental condition, highly heterogeneous in both genotype and phenotype.
- Autism, while common in its prevalence of 1% of the population, is evidenced to be rare in cause.
- Rare genetic causes of autism range from single nucleotide variants to copy number variants.
- Leveraging family structure enables mode of transmission of relevant genetic variation to be interrogated.
- Rare larger multiplex or extended pedigrees are expected to have a burden of rare highly penetrant genetic variants that are causative of autism and co-occurring phenotypes.

The aim of this study is to identify rare genetic variants associated with autism and other neurodevelopmental conditions within families.

Methods

1. Criteria for family inclusion are an autism-affected proband with two or more additional family members affected by a neurodevelopmental or neuropsychiatric condition.
2. Whole genome sequence data was analysed from n=33 individuals from n=5 multiplex families (Figure 1).
3. Sequencing was carried out to 30X coverage and raw data was processed by Genuity Science Pipeline Service (GSPS), a Sentieon based pipeline.
4. Hard filtering was carried out to remove deviations from Hardy-Weinburg Equilibrium (p-value $< 10^{-6}$) and variant sites exceeding a missingness threshold of 2%.
5. Variant annotation was carried out using dbNSFP(1).
6. Rare (<1% in gnomAD) non-synonymous single nucleotide variants were isolated and pathogenicity was predicted taking consensus of CADD, SIFT and Polyphen scoring(2-5).

Evaluating Family Structure

- Isolation of putative pathogenic variants is carried out taking a family-specific approach.
- Ascertainment of families for the study has enriched for rare variants, likely to be inherited.
- The mode of putative pathogenic variant transmission expected to be relevant from phenotype evaluation is outlined per family in Figure 1.

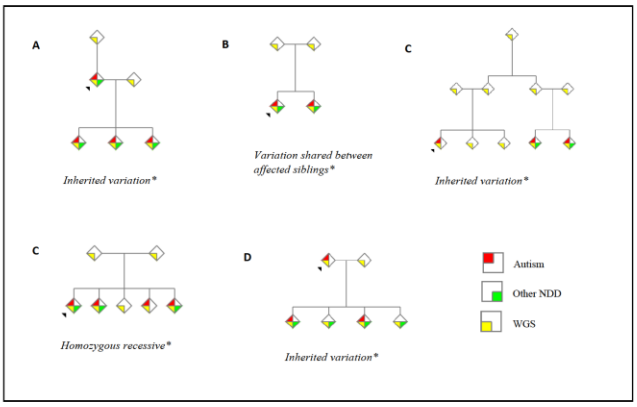


Figure 1. Proposed mode of transmission for variant interpretation. Presented are the pedigrees of the n=5 families sequenced in this rare cohort. The key associated with affection and sequencing status is presented alongside. * denotes the mode of variant transmission hypothesised to be relevant within each family.

Variant Counts: One family in focus

- Each family is evaluated for relevant genetic variation.
- Here we report variant counts for the six person three generation family, presented as Family A in Figure 1, for which inherited variation is suspected to be relevant in autism causation.
- Following annotation n=19,757 non-synonymous SNVs were identified within the family for further interrogation.
- n=1,114 of these variants were found to be rare (<1% in gnomAD). n=376 rare variants satisfy consensus pathogenicity criteria, as outlined in Figure 2.
- Functional evaluation of rare putative pathogenic variants will be carried out in the context of family structure.

Future Directions

- Here we propose our analysis strategy to isolate variants in this pedigree-based study.
- Putative pathogenic variants will be subset to those occurring in genes with evidence for relevant gene-phenotype association (SFARI Gene(6) and the Deciphering Developmental Disorders gene2phenotype dataset(7)).

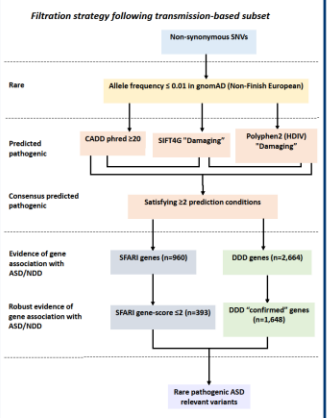


Figure 2. Variant filtration strategy for isolation of autism-relevant variation in a family-specific context.

- These findings will inform on rare genetic causes of autism and will subsequently improve understanding of the biology of autism.

References

1 Liu *et al.* 2020, Genome Medicine
2 Karczewski *et al.* 2020, Nature
3 Rentzsch *et al.* 2019, Nucleic Acids Res.
4 Ng *et al.* 2003, Nucleic Acids Res.
5 Adzhubei *et al.* 2010, Nat Methods
6 Abrahams *et al.* 2013, Mol. Autism
7 Wright *et al.* 2015, Lancet

Acknowledgements

Science Foundation Ireland sfi For what's next

This publication has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant No. 15/SIRG/3324.

We would like to thank Genuity Science (Ireland) Limited for its support for this project.

Appendix III-II: Determining the clinical utility of gene panels in autism; a study of diagnostic yield, relevance, and penetrance. Poster presented at World Congress of Psychiatric Genetics, October 2021.

Determining the clinical utility of gene panels in autism; a study of diagnostic yield, relevance, and penetrance

Fiana Ní Ghrálaigh,[1,2] Thomas Dinneen,[2] Ellen McCarthy,[1] Daniel N. Murphy,[1] Louise Gallagher[2] & Lorna M. Lopez[1]

1 Department of Biology, Maynooth University, Maynooth, Co. Kildare, Ireland.
2 Department of Psychiatry, Trinity College Dublin, Dublin, Ireland.

Contact: fiana.nighralaigh.2020@mumail.ie, lorna.lopez@mu.ie

Maynooth University
National University of Ireland Maynooth

Background & Aim

- Autism is a prevalent neurodevelopmental condition, highly heterogeneous in both genotype and phenotype.
- A genetic diagnosis of autism may allow for genetic counselling, opportunity to participate in targeted research or receive anticipatory medical advice.
- Genetic diagnosis in autism is limited by the ability to robustly determine the relevance of putative pathogenic genetic variation [1, 2].
- As a result, the development of effective gene panels to aid autism diagnosis is challenging.
- Despite this, commercial gene panels are available and marketed for use in autism [3].

The aim of this study is to estimate the clinical utility of 18 commercial gene panels in autism through analysis of diagnostic yield and clinical relevance.

Methods

1. The diagnostic yield of each panel was determined through secondary analyses of clinically relevant variation identified and characterised by Simon's Powering Autism Research Knowledge. The dataset analysed here arises from whole exome sequencing of 459 affected individuals, for whom no genetic diagnosis was previously reported [4].
2. The relevance of the genes included in each panel was quantified as the proportion of targeted genes with evidence supporting autism association [5].
3. Three autism-relevant gene sets were interrogated for the presence of ACMG59 clinically actionable genes [6].

Estimating Diagnostic Yield

- Diagnostic yield is estimated as the number of individuals for which a genetic cause of autism was identified, as a proportion of those investigated.
- Pathogenic variation is considered as variants listed in Feliciano *et al.* (2019). Variants considered are *de novo* and inherited single nucleotide variants (SNVs), insertion-deletion (indels) variants and copy number variants (CNVs). Reported chromosomal abnormalities were not included.
- Gene lists were assembled to include those for which clinically relevant SNVs and indels could be defined and those that fall within the boundaries of clinically relevant CNVs.
- Diagnostic yield of pathogenic variation is outlined in Table 1, with the diagnostic yield of probable pathogenic variants listed in brackets alongside

Service provider	Panel name	Genes targeted	Overlap SFARI Gene	Diagnostic yield in SPARK
Ambry Genetics	AutismNext Panel	72	87.5%	2.61%
Asper Neurogenetics	Autism Spectrum Disorders NGS Panel	76	88.16%	2.83% (0.22%)
Blueprint Genetics	Autism Spectrum Disorders Panel	75	45.33%	1.53% (0.44%)
Center for Human Genetics	Autism Spectrum Disorder 53-Gene Panel	53	84.91%	1.96% (0.22%)
Centogene	Syndromic Autism Gene Panel	50	88%	2.4% (0.22%)
Centogene	Intellectual Disability Panel	599	43.41%	5.23% (1.31%)
EGL Genetics	Autism Spectrum Disorders Tier 2 Panel	62	74.19%	2.18%
Fulgent Genetics	Autism NGS Panel	121	76.86%	4.36% (0.44%)
GeneDx	Autism/ID Xpanded Panel	2641	20.64%	10.02% (3.49%)
GENETAQ	Autism	27	92.59%	1.53%
Genomics England PanelApp	Autism (Version 0.20)	733	100%	7.63% (1.96%)
Greenwood Genetic Centre	Syndromic Autism Sequencing Panel	83	80.72%	3.05%
GX Sciences	Developmental Nutrigenomic Panel	33	15.15%	0.22%
MNG Laboratories	Comprehensive Disability/Autism Panel	1345	19.85%	6.1% (1.3%)
Munroe-Meyer Institute	Autism/Intellectual Disability/Multiple Anomalies Panel	117	55.56%	2.4% (0.22%)
Prevention Genetics	Autism Spectrum Disorders Panel	170	95.29%	6.32% (0.44%)
Reference Laboratory Genomics	Autism Spectrum Disorders (Expanded Panel)	77	77.92%	3.05% (0.44%)
Sema4	Comprehensive Autism Spectrum Disorders Panel	228	57.46%	4.79% (0.87%)

Table 1. Diagnostic yield of gene panels marketed for use in autism. Presented are diagnostic yield estimates of gene panels relevant to autism as estimated by secondary analysis of Feliciano *et al.* 2019. The number of genes present in each gene panel are correct as of January 2021. Percentage overlap with SFARI Gene is estimated as the proportion of genes within each respective gene list appearing in SFARI Gene (01-13-2021 release) [5].

Relevance to Autism

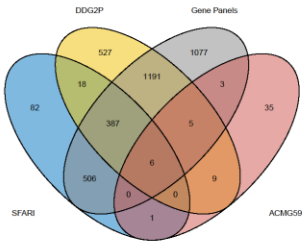


Figure 1 Relevance of gene panels genes to autism. Differentiated by colour are three autism-relevant gene lists (Blue-SFARI Gene [5]; Yellow- Deciphering Developmental Disorders ND gene2phenotype [7]; Grey- genes targeted by gene panels marketed for use in autism). The overlap is shown between these gene lists, and between each gene list and ACMG59 [6], in red. Counts represent the number of genes within each category of overlap.

- Six genes, included in all of the autism-relevant gene lists are found included in the ACMG59 list of genes in which variants in the exons may be clinically actionable (*PTEN, TSC1, TSC2, BRCA2, FBN1* and *SMAD4*).
- Three of these six genes (*PTEN, TSC1* and *TSC2*) are scored as "High Confidence" for autism-association and two of the six genes (*FBN1* and *SMAD4*) are scored as "Strong Candidate" for autism-association, as determined by number of reports in the SFARI Gene database [5].
- However, 17 of the 24 autism relevant genes overlapping with ACMG59 lack robust evidence of autism association.

Summary & Future Directions

- Here we estimate low diagnostic yields of the gene panels investigated (ranging from 0.22% to 10.02%).
- We recognise gene selection for inclusion in autism panels is variable in relevance (15.15% to 100% overlap with SFARI Gene).
- Together this suggests gene panels are currently of limited clinical utility and not extensive enough to justify use in autism diagnosis.
- Overlap of autism-relevant genes with ACMG59 genes, highlights that consideration should be made to potential identification of secondary findings when applying these panels.
- *We will estimate the rates and penetrance of rare variation in genes targeted by these panels by investigating carrier phenotypes in the UKBiobank whole-exome sequenced cohort (n= 300k).*

Code Available

https://github.com/FianaNG/autism_gene_panels.

References

1 Myers *et al.* 2020, American Journal of Human Genetics
2 Schaaf *et al.* 2020, Nature Reviews Genetics
3 Hoang *et al.* 2018, *npj* Genomic Medicine
4 Feliciano *et al.* 2019, *npj* Genomic Medicine
5 Abrahams *et al.* 2013, Molecular Autism
6 Kalia *et al.* 2017, Genetics in Medicine
7 Wright *et al.* 2015, Lancet

Acknowledgements

Science Foundation Ireland For what's next

This publication has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant No. 15/SIRG/3324.

This study is currently under peer review.

Appendix III-III: Evaluating the diagnostic yield of commercial gene panels in autism. Poster presented at the Irish Society for Human Genetics, September 2021.

Evaluating the diagnostic yield of commercial gene panels in autism

Fiana Ní Ghrálaigh,[1,2] Ellen McCarthy,[1] Daniel N. Murphy,[1] Louise Gallagher[2] & Lorna M. Lopez[1]

1 Department of Biology, Maynooth University, Maynooth, Co. Kildare, Ireland.
2 Department of Psychiatry, Trinity College Dublin, Dublin, Ireland.

Contact: fiana.nighralaigh.2020@mumail.ie, lorna.lopez@mu.ie

Maynooth University
National University of Ireland Maynooth

Background & Aim

- Autism is a prevalent neurodevelopmental condition, highly heterogeneous in both genotype and phenotype.
- A genetic diagnosis of autism may allow for genetic counselling, opportunity to participate in targeted research or receive anticipatory medical advice.
- Genetic diagnosis in autism is limited by the ability to robustly determine the relevance of putative pathogenic genetic variation [1, 2].
- As a result, the development of effective gene panels to aid autism diagnosis is challenging.
- Despite this, commercial gene panels are available and marketed for use in autism [3].

The aim of this study is to estimate the clinical utility of 18 commercial gene panels in autism through analysis of diagnostic yield and clinical relevance.

Methods

1. The diagnostic yield of each panel was determined through secondary analyses of clinically relevant variation identified and characterised by Simon's Powering Autism Research Knowledge. The dataset analysed here arises from whole exome sequencing of 459 affected individuals for whom no genetic diagnosis was previously reported [4].
2. The relevance of the genes included in each panel was quantified as the proportion of targeted genes with evidence supporting autism association [5].
3. Three autism-relevant gene sets were interrogated for the presence of ACMG59 clinically actionable genes [6].

Estimating Diagnostic Yield

- Diagnostic yield is estimated as the number of individuals for which a genetic cause of autism was identified as a proportion of those investigated.
- Pathogenic variation is considered as variants listed in Feliciano *et al.* (2019). Variants considered are *de novo* and inherited single nucleotide variants (SNVs), insertion-deletion (indels) variants and copy number variants (CNVs). Reported chromosomal abnormalities were not included.
- Gene lists were assembled to include those for which clinically relevant SNVs and indels could be defined and those that fall within the boundaries of clinically relevant CNVs.
- Diagnostic yield of pathogenic variation is outlined in Table 1, with the diagnostic yield of probable pathogenic variants listed in brackets alongside

Service provider	Panel name	Genes targeted	Overlap SFARI Gene	Diagnostic yield in SPARK
Ambry Genetics	AutismNext Panel	72	87.5%	2.61%
Asper Neurogenetics	Autism Spectrum Disorders NGS Panel	76	88.16%	2.83% (0.22%)
Blueprint Genetics	Autism Spectrum Disorders Panel	75	45.33%	1.53% (0.44%)
Center for Human Genetics	Autism Spectrum Disorder 53-Gene Panel	53	84.91%	1.96% (0.22%)
Centogene	Syndromic Autism Gene Panel	50	88%	2.4% (0.22%)
Centogene	Intellectual Disability Panel	599	43.41%	5.23% (1.31%)
EGL Genetics	Autism Spectrum Disorders Tier 2 Panel	62	74.19%	2.18%
Fulgent Genetics	Autism NGS Panel	121	76.86%	4.36% (0.44%)
GeneDx	Autism/ID Xpanded Panel	2641	20.64%	10.02% (3.49%)
GENETAQ	Autism	27	92.59%	1.53%
Genomics England PanelApp	Autism (Version 0.20)	733	100%	7.63% (1.96%)
Greenwood Genetic Centre	Syndromic Autism Sequencing Panel	83	80.72%	3.05%
GX Sciences	Developmental Nutrigenomic Panel	33	15.15%	0.22%
MNG Laboratories	Comprehensive Disability/Autism Panel	1345	19.85%	6.1% (1.3%)
Munroe-Meyer Institute	Autism/Intellectual Disability/Multiple Anomalies Panel	117	55.56%	2.4% (0.22%)
Prevention Genetics	Autism Spectrum Disorders Panel	170	95.29%	6.32% (0.44%)
Reference Laboratory Genomics	Autism Spectrum Disorders (Expanded Panel)	77	77.92%	3.05% (0.44%)
Sema4	Comprehensive Autism Spectrum Disorders Panel	228	57.46%	4.79% (0.87%)

Table 1. Diagnostic yield of gene panels marketed for use in autism. Presented are diagnostic yield estimates of gene panels relevant to autism as estimated by secondary analysis of Feliciano *et al.* 2019. The number of genes present in each gene panel are correct as of January 2021. Percentage overlap with SFARI Gene is estimated as the proportion of genes within each respective gene list appearing in SFARI Gene (01-13-2021 release) [5].

Relevance to Autism

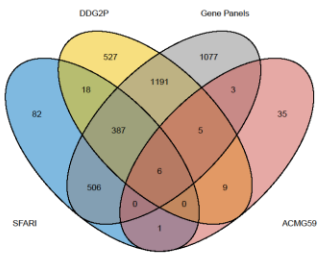


Figure 1 Relevance of gene panels genes to autism. Differentiated by colour are three autism-relevant gene lists (Blue-SFARI Gene [5]; Yellow- Deciphering Developmental Disorders ND gene2phenotype [7]; Grey- genes targeted by gene panels marketed for use in autism). The overlap is shown between these gene lists, and between each gene list and ACMG59 [6], in red. Counts represent the number of genes within each category of overlap.

- Six genes, included in all of the autism-relevant gene lists are found included in the ACMG59 list of genes in which variants in the exons may be clinically actionable (*PTEN, TSC1, TSC2, BRCA2, FBN1* and *SMAD4*).
- Three of these six genes (*PTEN, TSC1* and *TSC2*) are scored as "High Confidence" for autism-association and two of the six genes (*FBN1* and *SMAD4*) are scored as "Strong Candidate" for autism-association, as determined by number of reports in the SFARI Gene database [5].
- However, 17 of the 24 autism relevant genes overlapping with ACMG59 lack robust evidence of autism association.

Summary & Future Directions

- Here we estimate low diagnostic yields of the gene panels investigated (ranging from 0.22% to 10.02%).
- We recognise gene selection for inclusion in autism panels is relevant (43.41% to 100% overlap with SFARI Gene).
- Together this suggests gene panels developed are currently of limited clinical utility and not extensive enough to justify use in autism diagnosis.
- Overlap of autism-relevant genes with ACMG59 genes, highlights that consideration should be made to potential identification of secondary findings when applying these panels and, due to the limited evidence supporting their autism-association, the value added by their inclusion in these gene panels.

References

1 Myers *et al.* 2020, American Journal of Human Genetics
2 Schaaf *et al.* 2020, Nature Reviews Genetics
3 Hoang *et al.* 2018, *npj* Genomic Medicine
4 Feliciano *et al.* 2019, *npj* Genomic Medicine
5 Abrahams *et al.* 2013, Molecular Autism
6 Richards *et al.* 2015, Genetics in Medicine
7 Wright *et al.* 2015, Lancet

Data Available

Code available at https://github.com/FianaNG/autism_gene_panels.

This study is currently under peer review.

Acknowledgements

Science Foundation Ireland SFI For what's next

This publication has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant No. 15/SIRG/3324.

Appendix III-IV: Application of an evidence-based curation framework to aid gene discovery: a pilot investigation in an autism family cohort. Poster presented at the World Congress of Psychiatric Genetics, October 2020.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Application of an evidence-based curation framework to aid gene discovery: a pilot investigation in an autism family cohort

Fiana Ní Ghrálaigh *[1,2], Louise Gallagher [1] & Lorna M. Lopez *[1,2]

Background

- Rare genetic variants, both inherited and *de novo*, typically have higher effect sizes and are more penetrant than common variants in the population.
- There is need for consensus in the evaluation of evidence supporting association of a gene with autism.
- A recently published roadmap by Schaaf *et al.*, (2020) proposes the use of a modified ClinGen framework for curation of a gene-list associated with autism, with potential for use in a clinical setting [1].

Aim: ***Investigate rare genetic variants and their association with autism in a family-based genome-sequencing cohort, through application of the proposed evidence-based framework.***

Methods

- WGS (n=6) was carried out on Illumina NovaSeq6000 and processed following Genome Analysis Tool-Kit (GATK) Best Practices [2].
- Predicted pathogenic ASD-relevant rare variants are selected through dbNSFP annotation [3], as detailed in Fig.1.

Results

Variant Discovery

a

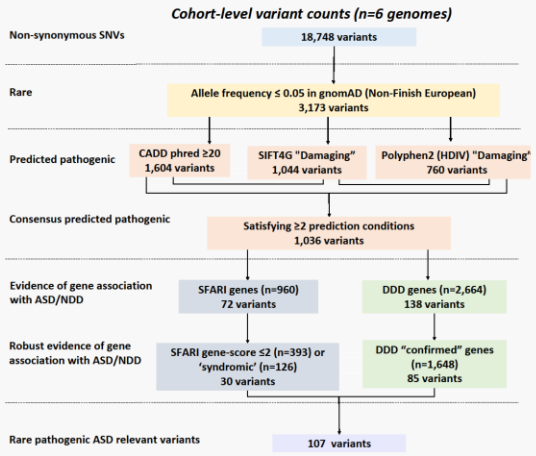


Fig.1 Flow of variant filtering with cohort-level variant counts. Arrows show the direction of flow from each level of filtering (specified on the left). SFARI refers to Simons Foundation Autism Research Initiative Gene Module [4]. DDD refers to the gene2phenotype database arising from the Deciphering Developmental Disorders study [5].

b

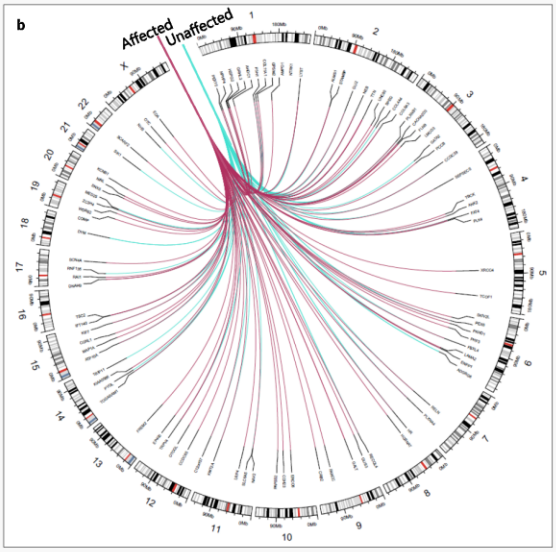
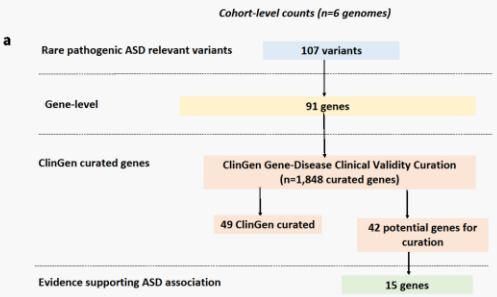


Fig.2 Spread of variation across genomic regions. Chromosomes are shown around the outer track of the figure (1:22, X). The gene names are given on the inner track. These are the genes in which the rare pathogenic ASD relevant variants outlined in Fig.1 are located. Links are made in purple (affected n=91 variants) and blue (unaffected n=69 variants) between each gene and the respective affection status of the individual harbouring the variant. Affected denotes individuals with an autism diagnosis.

Gene Curation

a



b

Gene	Sum of scores	Calculated classification
NAV2	0.5	Limited
NINL	4.5	Limited
CACNA2D3	5.0	Limited

Fig.3 a Gene selection for curation. Genes (Fig.2) were excluded from this analysis when already curated by ClinGen. Genes were selected for analysis when evidence of ASD association is reported in the literature. **b Classification of three genes with highest number of autism reports.** The three genes with the highest number of ASD reports (6 publications each) were selected for curation. Sum of scores represents the raw sum of genetic and experimental evidence towards autism based on Schaaf *et al.* framework with gene classification.

Discussion

- Many genes arising from this analysis have been curated by ClinGen Gene-Disease Validity framework.
- SFARI Gene module hosts much of the information needed to curate the genetic evidence for gene association within this framework.
- Phenotype curation of the already compiled and maintained SFARI gene module would add confidence to gene classification within the database.
- Less stringent filtering cut-offs are needed to identify the complete set of putatively pathogenic variants within this cohort.
- Note: Putative variants will be validated and replicated to confirm their relevance in autism. Curated genes have not be subject to expert panel review or submitted to ClinGen.

References

[1] Schaaf et al., (2020) *Nature Reviews Genetics*
[2] Van der Auwera *et al.*, (2013) *Curr Protoc Bioinformatics*
[3] Liu et al., (2016) *Human Mutation*
[4] Simon's Foundation (2018) *SFARI Human Gene Module*
[5] Wright *et al.* (2015) *The Lancet*

Affiliations

*Contact:
fiana.nighralaigh.2020@mumail.ie
lorna.lopez@mu.ie
1: Department of Psychiatry, Trinity College Dublin
2: Department of Biology, National University of Ireland Maynooth

Acknowledgements

TrinSeq
ELDAbiotech
Maynooth University National University of Ireland Maynooth
Science Foundation Ireland sfi For what's next

Appendix III-V: Rare genetic variation in autism; an exome sequencing study. Poster presented at the Irish Society for Human Genetics, September 2020.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Maynooth University
National University of Ireland Maynooth

Rare genetic variation in autism; an exome sequencing study

Fiana Ní Ghrálaigh *(1,2), Cathal Ormond (1), Elaine Kenny (1), Louise Gallagher (1) & Lorna M. Lopez *(1,2)

Background

- Rare genetic variants, both inherited and *de novo*, typically have higher effect sizes and are more penetrant than common variants in the population.
- Whole exome sequencing (WES) facilitates simultaneous investigation of many classes of variation in the coding genome, across the allele frequency spectrum.
- Causal variants aggregating in families with multiple affected individuals typically have a larger affect than variants in sporadic cases of autism [1].

- *Aim: apply WES to a cohort of 34 individuals in families affected with autism and other neurodevelopmental disorders, aiming to identify rare pathogenic variants.*

Methods

- WES was carried out using the Nextera Rapid Capture Exome (v1.2) on Illumina NovaSeq6000 and
- Data has been analysed following Genome Analysis Tool-Kit (GATK) Best Practices.
- Predicted pathogenic rare variants are selected through dbNSFP annotation.
- Putatively pathogenic variants are filtered through known autism-associated gene sets, including SFARI [2] and DDD [3] gene lists.

Results

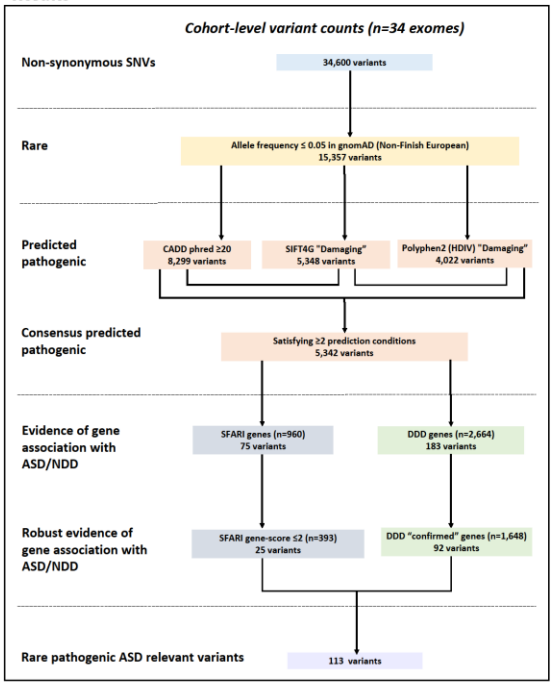


Fig.1 Flow of variant filtering with cohort-level variant counts. Arrows show the direction of flow from each level of filtering (specified on the left). SFARI refers to Simons Foundation Autism Research Initiative Gene Module [2]. DDD refers to the gene2phenotype database arising from the Deciphering Developmental Disorders study [3].

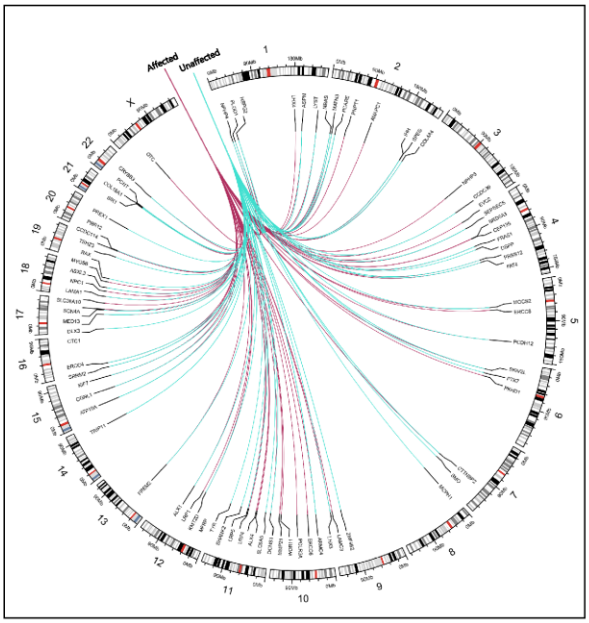


Fig.2 Spread of variation across genomic regions. Chromosomes are shown around the outer track of the figure (1:22, X). The gene names are given on the inner track. These are the genes in which the rare pathogenic ASD relevant variants outlined in Fig.1 are located. Links are made in purple (affected n=103) and blue (unaffected n=86) between each gene and the respective affection status of the individual harbouring the variant. Affected denotes individuals with an autism diagnosis.

Discussion

- Rare pathogenic ASD relevant variants isolated from the cohort occur across the genome.
- Rare pathogenic ASD relevant variants are harboured in unaffected family members also.
- Less stringent analyses are needed to identify the complete set of putatively pathogenic variants within this cohort.
- Putative variants will be validated and replicated to confirm.

References

[1] Ott, J. *et al.* (2011) *Nature Reviews Genetics*
[2] Simon's Foundation (2018) *SFARI Human Gene Module*
[3] Wright *et al.* (2015) *The Lancet*

Affiliations

*Contact:
fiana.nighralaigh.2020@mumail.ie
lorna.lopez@mu.ie
1: Department of Psychiatry, Trinity College Dublin
2: Department of Biology, National University of Ireland Maynooth

Acknowledgements

ELDAbiotech
TrinSeq
Science Foundation Ireland
SFI For what's next

Appendix III-VI: Analysis Pipeline of Whole Genome Sequencing Data in Neurodevelopmental Disorders. Poster presented at British Neuroscience Association Festival of Neuroscience, April 2019.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Analysis Pipeline of Whole Genome Sequencing Data in Neurodevelopmental Disorders

Fiana Ní Ghrálaigh[1], Niamh M. Ryan[1], Louise Gallagher[1], Lorna M. Lopez[1]

Background

Neurodevelopmental disorders (NDDs) such as autism, ADHD and epilepsy are highly heritable complex traits. Common variants associated with NDDs tend to have low effect sizes. Rare variants, both inherited and arising spontaneously, often have a higher effect size and are more penetrant than common variants. Taking a family study approach allows for the analysis of variant transmission from parent to offspring and allows interrogation of *de novo* variation [1]. Whole genome sequencing simultaneously investigates all classes of coding and non-coding variants across the allele frequency spectrum.

Data available for analysis:

1. Whole genome sequencing data of 100 individuals from multiplex families affected with autism and other neurodevelopmental disorders
   (Expected- 30X coverage on Illumina NovaSeq)
2. Pedigree structure data (Sample outlined in Fig.1)
3. Phenotype data in particular ADOS, ADI, medical history and IQ measures

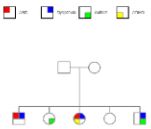


Figure 1 Sample Pedigree for Analysis. Pedigree of family with multiple affected offspring to unaffected parents. Legend outlines affection status of each individual.

Aim: The aim of this study is to identify rare genetic variants by applying whole genome sequencing technology in families with multiple affected individuals with autism and other neurodevelopmental diagnoses.

Methods

a) Quality Control

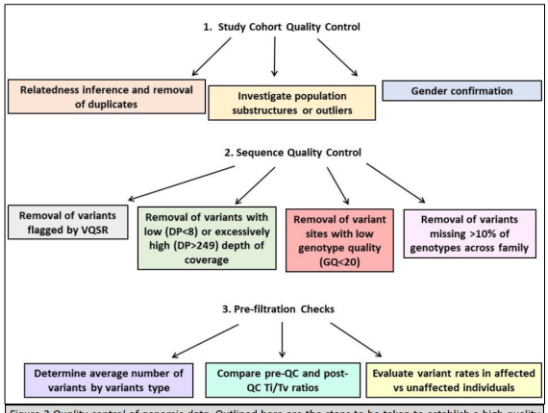


Figure 2 Quality control of genomic data. Outlined here are the steps to be taken to establish a high quality working data set. These steps will be carried out in three broad stages: study cohort quality control; sequencing data quality control; and pre-filtration checks.
VQSR: Variant Quality Score Recalibration Ti/Tv ratio: ratio of transition (Ti) to transversion (Tv) variants

Variant filtration will be carried out as per Fig. 3. This filtration mechanism will restrict variants to those that are predicted to be pathogenic and are occurring in, or affecting expression of, genes that have previously been associated with neurodevelopmental disorders.

b) Variant Filtration

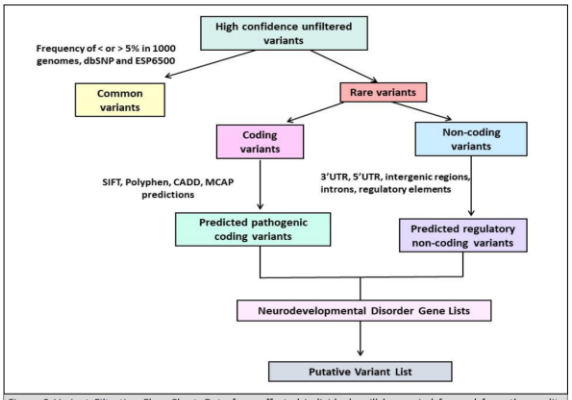


Figure 3 Variant Filtration Flow Chart. Data from affected individuals will be carried forward from the quality control pipeline (Fig. 2) to this variant filtration pipeline. Variants will be restricted as per flow diagram with each arrow representing a filtration step. Putative variants from this pipeline will be restricted to genes with previous evidence supporting their implication in neurodevelopmental disorder risk.

The Neurodevelopmental Disorder Gene List specified in Fig. 3 mirrors that applied by the SPARK Consortium [2]. This gene set includes genes that fall in the following categories:

SFARI gene score ≤ 2 [3]	Genes highly expressed in the brain [8]
Deciphering Developmental Disorders genes [4]	Transcript regulator GO:0006355
Post-synaptic density genes [5]	Chromatin modifier GO:0016569
Embryonic highly expressed genes [6]	Nervous system development GO:0007399
M2,M3,M16,M13 gene co-expression modules [7]	Nerve Impulse GO:0019227, GO:0019226 and GO:0050890
Brain specific expressed gene [8]	Neuron projection GO:0043005

Discussion

This analysis pipeline is expected to yield a set of high confidence putative coding and non-coding variants contributing to the genetic risk of neurodevelopmental disorders.

The putative pathogenic variants will be further analysed using:
- Functional interpretation
- Intra-family transmission analysis
- Gene network analysis
- Evaluation of recurrence within cohort

Impact: Variants identified through this analysis pipeline will provide supporting evidence for gene association with neurodevelopmental disorders. This study will explore the clinical utility of whole genome sequencing in neurodevelopmental disorders. Diagnostic yields up to 42.4% in autism have been achieved in whole genome sequencing studies to date, highlighting the potential of this study design to identify rare genetic risk factors in neurodevelopmental disorders [8].

References

[1] Ott *et al.* (2011)
[2] Feliciano *et al.* (2019)
[3] Simon's Foundation (2018)
[4] Wright et al. (2015)
[5] Bayes *et al.* (2011)
[6] Iossifov *et al.* (2014)
[7] Parikshak *et al.* (2013)
[8] Yuen *et al.* (2015)

Affiliations

*Email: nighralf@tcd.ie lorna.lopez@tcd.ie

1 Trinity College Dublin, School of Medicine, Department of Psychiatry
Trinity Centre for Health Science,
St. James Hospital, Dublin 8, Ireland

Acknowledgements

Science Foundation Ireland SFI For what's next

Appendix III-VII: A Search for Rare Variants in a Family-Based Study of ASD. Poster presented at the World Congress of Psychiatric Genetics, October 2018, and the Irish Society for Human Genetics, September 2018.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

A Search for Rare Variants in a Family-Based Study of Autism Spectrum Disorders

Fiana Ní Ghrálaigh[1], Jessica E. Smith[1], Elaine Kenny[1], Louise Gallagher[1], Lorna M. Lopez[1]

Background

- Genomic studies have identified thousands of genetic variants associated with Autism Spectrum Disorders (ASD) [1]
- Causal variants aggregating in families with multiple affected individuals typically have a larger affect than variants in sporadic cases of complex traits [2]
- Next Generation Sequencing (NGS) technologies generate data that allow near complete evaluation of the genetic variation of an individual, and therefore are a tool for analysis of variant transmission through families
- *Aim: To identify rare de novo variants contributing to the clustering of ASD within a family with 4 ASD-affected individuals (Fig.1)*

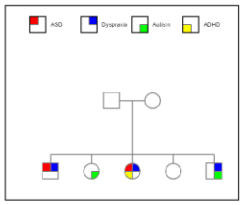


Figure 1: Pedigree of family with multiple affected offspring to unaffected parents. Legend outlines affection status of each individual. ASD and Autism diagnoses are according to DSM-V (as measured by ASOS-2 and ADI-R)

Methods

- Libraries were prepared from saliva DNA samples of parents and all affected offspring, and captured using Whole Exome Solution xGen® Lockdown® Probes by SOPHiA GENETICS
- Libraries were run on Illumina HiSeq 4000 (2x250) achieving an average coverage of 83X with 99.36% of reads successfully mapped (Fig.2)

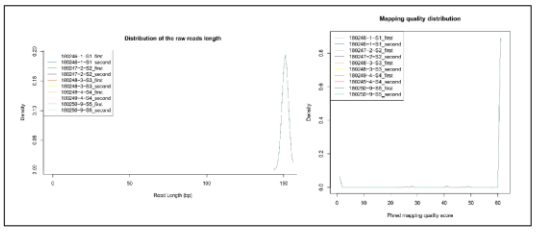


Figure 2: A) Distribution of raw read lengths. This curve illustrates long, uniform read lengths that were obtained from sequencing across all samples. B) Mapping quality distribution. A single defined peak can be seen at a Phred Score of 60, indicating high quality mapping across all samples

- Analysis was carried out on SOPHiA DDM®, an artificial intelligence software for visualisation and interpretation of sequencing data (Fig.3)

Figure 3: Workflow of SOPHiA DDM® filtration specifications. Retained variants specify those variants meeting SOPHiA criteria for confident variant calls. Variant fraction was set at 25%, that is a minimum of 25% of reads supporting the variant count, taking base Phred scores into account. Common variants were filtered out by eliminating variants appearing in over 1% of the population as measured in ExAC and gnomAD cohorts. SIFT, Polyphen and MutationTaster values were set at 0.02 (inverse represented in figure), 0.95 and 0.9 respectively to isolate variants with predicted pathogenicity. *De novo* variants were specified by selecting variants appearing in less than 2 individuals in the sample.

Results

The rare *de novo* likely pathogenic variants resulting from the variant filter specified (Fig.3) were prioritised according to SOPHiA DDM® pathogenicity scores and ACMG criteria [3]. Featured below are high priority variants identified in this analysis with potential implications in ASD (Table 1). The *dock3* variant is shown in detail in Fig.4

Gene	Variant Type	dbSNP	Affected Individual	Protein Function
med12	SNV	rs746104301	C1	Transcriptional co-activation of CD8K8
tyro3	INDEL	NA	C1	RNA binding interaction with FMR1
dock3	INDEL	NA	C2	Induces axonal outgrowth in CNS
chst15	INDEL	NA	C3	RNA binding interaction with FMR1
pclo	INDEL	rs758399155	C3	Component of presynaptic cytoskeletal matrix

Table 1: Rare *de novo* pathogenic variants. The table illustrates the highest ranked variants resulting from the analysis. The gene containing each variant is specified in column 1 and the accession number (dbSNP) in column 3 where available. Variant type is given as either INDEL insertion/deletion, or SNV single nucleotide variant. Affected individual refers to Figure 1 with C1-C5 representing children 1 to 5 (left to right).

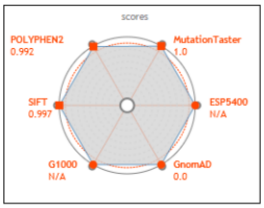


Figure 4: *dock3* INDEL in detail. Shown in the spider diagram are values indicating expected pathogenicity of the variant and the frequency of the variant in unaffected populations. SIFT score is represented here as the inverse of the standard SIFT score

Discussion

Here we show:

- The utility of SOPHiA DDM® as a platform for variant interpretation and a method of prioritising variants for investigation by functional studies
- A subset of rare *de novo* variants predicted to be implicated in ASD

Note: this data is preliminary and requires validation

References

[1] Simon's Foundation (2018) *SFARI Human Gene Module*
[2] Ott, J. *et al.* (2011) *Nature Reviews Genetics*
[3]Richards, S. *et al.* (2015) *Genetics in medicine : official journal of the American College of Medical Genetics.*

Affiliations

*Email: nighralf@tcd.ie lorna.lopez@tcd.ie
1 Trinity College Dublin, School of Medicine, Department of Psychiatry
Trinity Translational Medicine Institute, Trinity Centre for Health Science, St. James Hospital, Dublin 8, Ireland

Acknowledgements

Science Foundation Ireland sfi For what's next
wellcome
ELDAbiotech

Appendix IV: Research Articles

Appendix IV-I: Ní Ghrálaigh, F., Gallagher, L. and Lopez, L. M. (2020) 'Autism spectrum disorder genomics: The progress and potential of genomic technologies', Genomics, 112(6). doi: 10.1016/j.ygeno.2020.09.022.

Genomics 112 (2020) 5136–5142

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

Autism spectrum disorder genomics: The progress and potential of genomic technologies

Fiana Ní Ghrálaigh[a,b], Louise Gallagher[a], Lorna M. Lopez[a,b,*]

[a] Department of Psychiatry, Trinity College Dublin, Ireland
[b] Department of Biology, Maynooth University, Ireland

ARTICLE INFO

Keywords:
Autism
Genomics
Rare
Variation
Whole-genome
Sequencing
ASD

ABSTRACT

Genomic technologies have accelerated research progress in autism spectrum disorder (ASD) genomics and promises to further transform our understanding of the genetic basis of this neurodevelopmental disorder. Here we review the current evidence for the genetic basis of ASD, present the progress of large-scale studies to date and highlight the potential of genomic technologies. In particular, we discuss evidence for the importance of identifying rare genetic variants in family-based studies. Genomics is a key feature of future healthcare and our review illustrates it's potential to improve our biological understanding of neurodevelopmental disorders, and to ultimately aid ASD diagnosis, inform medical decision making and establish genomics as central to personalised medicine.

1. The genetic basis of ASD

Autism Spectrum Disorder (ASD) is a prevalent neurodevelopmental disorder occurring in around 1% of individuals in a population [1]. The condition manifests as restrictive repetitive behaviours and social communication deficits across a phenotypic spectrum [2]. ASD is a highly heritable complex trait. The heritability of ASD measures the genomic variation contributing to the phenotype and in ASD has been estimated at ~80–90% [3,4]. The genetic risk of ASD is contributed to by both rare and common genetic variants, and as yet the majority of the genetic risk remains unexplained [5]. Rare variants refer to those occurring at less than 5% of the population and very rare variants occur at a minor allele frequency of less than 1%. Common genetics variants typically refer to genetic variants with a minor allele frequency of greater than 5%. Rare variants, particularly those occurring *de novo*, have the potential to occur at higher effect sizes than common variants. The larger effect size of rare variants is in line with the hypothesis that variants of a higher effect sizes have a more detrimental effect on brain development resulting in the early-life manifestation of the autistic phenotype, when compared to neuropsychiatric disorders most commonly arising later in life, such as schizophrenia and psychosis.

In this review, we aim to inform the reader on state-of-the-art ASD genomics research. Our focus is on the application of genomic sequencing technologies to search for these genetic variants in extensive sample collections that have transformed our understanding of ASD genomics. We review cutting-edge research that use genomic sequencing methods, bioinformatic processing and clinical implementation for improved diagnosis and medical decision-making in ASD and other neurodevelopmental disorders. We explain the value of genomic sequencing technologies and highlight what they can achieve for neurodevelopmental and neuropsychiatric disorders.

2. Sequencing technologies have advanced the identification of rare variants

Genomic sequencing, specifically whole exome sequencing and whole genome sequencing, has transformed variant discovery. These technologies give the opportunity for more widespread and in-depth genomic analysis than older techniques, such as microarray studies and candidate gene studies, have allowed. Table 1 lists the next-generation sequencing technologies that can identify single nucleotide variants and insertion-deletion variants, as well as larger genomic hits, including structural or copy number variants, across the allele frequency spectrum. In the past decade, sequencing technologies have stretched from covering select points across to genome to cover up to 100%, when sequenced at high coverage with *de novo* assembly (Table 1) [6]. Higher coverage whole genome sequencing results in more precise variant calls across the coding and non-coding regions of the genome.

These advances in genomic technologies and decreasing costs have enabled large sequencing cohorts (Table 2), allowing key strides to be

* Corresponding author at: Department of Psychiatry, Trinity Centre for Health Sciences, St. James Hospital, Dublin 8, Ireland.
E-mail address: lorna.lopez@tcd.ie (L.M. Lopez).

https://doi.org/10.1016/j.ygeno.2020.09.022
Received 17 June 2020; Received in revised form 1 September 2020; Accepted 8 September 2020
Available online 15 September 2020
0888-7543/

Table 1
Genomic technologies compared.

	Exome sequencing		Whole genome sequencing	
	Clinical exome sequencing	Whole exome sequencing	Short-read	Long-read
% Genome covered	~0.5%	~1%	~90%	Potential for up to 100%
Types of variant detected	SNVs Indels CNVs (limited)	SNVs Indels CNVs (limited) SVs (limited) Mitochondrial	SNVs Indels CNVs SVs Mitochondrial Repeat expansions (including tandem repeats [69,70])	SNVs Indels CNVs SVs Mitochondrial Repeat expansions Complex SVs Haplotype phased variants Methylation
Diagnostic yield in ASD	Limited application	31% [62]	42.4% [42]	Not yet available
Cost estimate	€37.19[a]	€79.33[b]	€1239.50[c]	€918[d]

Outlined are four key sequencing technologies with potential for use to identify rare ASD genetic variants. Note that these costs are estimates and do not include library preparation costs, barcodes, access fees, labour, VAT, service provider, data processing and data storage and other associated sequencing costs.
Estimates a, b and c, are based on sequencing with Illumina NovaSeq S4 flowcell (2 × 150) up to 3000Gb/flowcell.
Acronyms; SNV single nucleotide variant, Indel insertion deletion, CNV copy number variant, SV structural variant.
[a] SOPHiA GENETICS Clinical Exome Solution (12 Mb covering ~4500 genes (2.5Gb/sample/800 samples/flowcell)).
[b] Illumina Nextera Rapid Capture Exome (37 Mb (8Gb/sample/375 samples/flowcell)).
[c] WGS (120Gb/sample/24 sample/flowcell).
[d] Oxford Nanopore Technologies (60×; 1 sample/flow cell/180GB) Sequencing metrics: https://nanoporetech.com/accuracy.

made in the field of ASD genomics. Large-scale analyses of these cohorts have identified hundreds of ASD-associated genetic variants across the genome. For example, discovery of rare variants, particularly rare CNVs, affecting *SHANK3*and *NRXN1* among other genes, implicated synaptic transmission and plasticity in ASD neurobiology [7]. Extending beyond variant discovery, combining rare variant analysis with single-cell investigation in the developing human cortex showed enriched expression of particular ASD-associated genes in maturing and mature excitatory and inhibitory neurons from mid-fetal development, and helped to validate the role of these genes in neuronal communication and regulation of gene expression [8]. Impactful findings such as these, suggest great potential for advancing our understanding of ASD neurobiology through rare variant discovery.

3. Common genetic variants have been challenging to associate with autism

The search for common genetic variants has been less successful than that in more typically adult-onset neuropsychiatric disorders, in particular schizophrenia (~7% of variance on the liability scale) [9] and bipolar disorder (~2.5% of variance on the liability scale) [10–12]. The largest study to date investigating common genetic variants in ASD, using genome-wide genotyping, provides evidence for statistically significant association of the first common risk variants with ASD. A Genome Wide Association Study (GWAS) was carried out on 18,381 ASD cases and 27,969 controls. While this sample size is large in terms of ASD, it is smaller than that of other traits such as schizophrenia with 36.989 cases or bipolar disorder with 20,352 cases [9,10]. Five loci showed significant association with ASD alone and seven further loci were identified upon analysis of schizophrenia, depression and educational attainment together [13]. Polygenic risk, measured by a polygenic risk score (PRS), is the combined impact of common variants on the probability of a phenotype. In ASD this explains just 2.5% of the observed variance in risk [13]. The lower yield of common variant loci in ASD may be because of a greater relative contribution of rare genetic variants than common variants in the genetic architecture of ASD [14]. However, the current smaller sample sizes in GWAS of ASD fail to validate this hypothesis.

4. Heterogeneity in the genetic architecture of ASD

ASD displays a high level of heterogeneity across a phenotypic spectrum, both between individuals and within the same individual throughout the lifespan. It is estimated that around 10% of individuals affected with ASD have a syndromal form of the condition, for which each single ASD risk gene accounts for at most 1% of overall cases on average [15]. Rare disorders often manifest with an underlying autistic phenotype [16]. These syndromes are frequently caused by highly penetrant variants in single genes, such as Fragile X syndrome, MIM:30024 (*FMR1*), and Tuberous Sclerosis Complex, MIM: 613254 (TSC2) (reviewed in Betancur, [17]). These syndromal forms of ASD are frequently associated with intellectual disability and developmental delay, suggesting that ASD may only form part of the overall behavioural phenotype of the syndrome.

ASD cases that do not fall into clinically defined syndromes appear to have more complex genetic architecture and various models of risk have been suggested to encompass this. The polygenic model, strongly supported in schizophrenia [18], proposes that multiple loci, each contributing a small effect, accumulate to surpass a threshold of disease liability. In contrast, Boyle et al. proposed the omnigenic model [19,20]. This model suggests that all genes expressed in disease-relevant cells have the ability to influence pathogenesis, through their interference with the expression of "core genes". In that, it may be hypothesised that most of the heritability of ASD could be explained by the effect of variation on genes outside of the core ASD pathways.

Understanding gene regulation is critical to parsing out the relative contribution of common and rare variants to ASD heritability. Whichever model is most appropriate in describing its architecture, it is clear that rare genetic variants are crucial to understanding ASD.

Further to heterogeneity in the genetic architecture among ASD cases, there is heterogeneity, both genetically and clinically, between males and females. Males are more frequently affected with ASD than females [21]. Although factors such as hormonal sex differences, sex-specific epigenetic factors and genetic factors related to sex chromosomes have been hypothesised to play a role in this bias, the biological basis remains unclear. A large-scale family study interrogating *de novo* variants in ASD reinforces the importance of evaluation of the X chromosome, identifying 5 of 7 genes replicated in the study are located on the X chromosome [22]. Together with the evidence of sex biases of autosomal genes, this study highlights the potential for genomic studies

Table 2
Key ASD genomics cohorts.

Cohort	Size of cohort	study design	Dataset	Reference
Australian National Autism Consortium	48 cases and 80 parent controls	Simplex & multiplex	WES	71
Autism Genetic Resource Exchange (AGRE)	> 1700 families	Simplex & multiplex	Genome-wide genotyping	https://www.autismspeaks.org/agre 41
Autism Sequencing Consortium (ASC)	12,772 individuals	Case control, simplex	WES	8,72
Deciphering Developmental Disorders (DDD)	12,000 individuals and their parents	Simplex	Genotyping & WES	https://www.ddduk.org/ 25,47
iHART	2308 individuals (from 493 AGRE families)	Quad & multiplex	WGS	33,
iPsych Danish Cohort	16,146 cases (Genotyping) and 4811 cases (WES)	Case control	Genome-wide genotyping (PsychChip array from Illumina) & WES	8,74,75
MSSNG	11,312 individuals (4258 families)	Simplex & multiplex	WGS	https://research.mss.ng/ 31,34,42,43,76
Simons Foundation Powering Autism Research for Knowledge (SPARK)	27,615 individuals (genotyping & WES) and 400 quad families (WGS)	Simplex & multiplex	Genome-wide genotyping (Illumina InfiniumCoreBeanno 24), WES & WGS	https://sparkforautism.org/ 40
Simons Simplex Collection (SSC)	8975 individuals (WGS) 2517 families (WES) and 10,220 individuals (Genotyping)	Quartet (phase 1–3) & Trio/ Incomplete family data (Phase 4)	Genome-wide genotyping, WES & WGS	https://www.sfari.org/resource/simons-simplex-collection 8,32,34,77–80
The Autism Genome Project's consortium including TASC and AGRE samples)	7917 individuals (1492 families)	Simplex & multiplex	Genome-wide genotyping (1.0 K SNP array and 400 microsatellite marker panel)	37,81
The Autism Simplex Collection (TASC)	5444 individuals (1719 families)	Simplex	Genome-wide genotyping (Illumina 1 M SNP) & WES	26,82
The Psychiatric Genetics Consortium	18,381 cases	Case control	Genome-wide genotyping	13

Featured in the table are large-scale ASD cohorts used in genomic studies to date. Note that there is significant overlap of samples between these cohorts, for example, the MSSNG cohort includes samples from both AGRE and TASC. These details are subject to frequent update. Reference refers to the original research article/website linked to the cohort and research studies cited in this review that analyse these cohorts.

to elucidate this phenomenon.

5. Rare variants disrupt gene function, dosage and regulation in ASD

Current whole genome and exome sequencing technologies enable investigation of most genomic variant classes (Table 1). The consequences of such variants in the genome occur to varying effects with different degrees of penetrance, as outlined below.

5.1. Gene disruption

Gene disruption refers to the disturbance of gene expression and the impact of variation on overall gene function. The consequence of a genetic variant can be detrimental to gene function or can have little effect depending on the variant in question and the overall genome environment (Fig. 1). Genes disrupted in ASD often include those related to brain development, post-synaptic density, nerve impulse and neuron projection [23]. Much focus lies on the importance of loss of function variants and damaging missense variants in the evaluation of genetic variation on ASD. In particular, variants impacting evolutionarily conserved genes to the detriment of crucial cellular processes.

Another mechanism of gene disruption is gene rearrangement, encompassing translocations, inversions and large-scale insertions and deletions. Although varying between studies, the estimated rate of large variants in ASD is approximately 5–10% [24]. A recent study implicates rare retro-transposition derived disruption in neurodevelopmental disorders through trio-based exome sequencing analysis from the Deciphering Developmental Disorders (DDD) cohort. This mechanism of disruption is an avenue for pathogenesis which has been largely unexplored in neurodevelopmental disorders to date [25].

5.2. Gene dosage

Gene dosage refers to the number of copies of a given gene that are present in the genome of an individual. Dosage has been found to play a substantial role in ASD pathogenesis, as demonstrated through CNV analysis, *i.e.* analysis of duplication or deletion variants of > 1Kb [26]. In 2004, two groups independently identified that large scale CNVs were often overlapping with genic regions [27,28]. The influence of these CNVs means either an increase or depletion in activity of the contained genes with potential for damaging functional consequences. A comprehensive analysis identified clinically relevant CNVs in 10.5% of neurodevelopmental disorder cases investigated, with 11.4% in ASD cases. Importantly many of the CNVs identified were found to occur across multiple neurodevelopmental disorders [29].

5.3. Gene regulation

As a complex trait, non-coding variants, particularly variants affecting gene regulation are likely to influence ASD [30]. Advances in whole genome sequencing and bioinformatic tools are enabling studies of non-coding regions of the genome. Yuen et al. estimated that non-coding and genic non-coding *de novo* variants account for 15.6% and 22.5% respectively, of predicted damaging *de novo* variants in ASD cases. Non-coding elements, *e.g.* untranslated regions, regulatory sequences involved in exon skipping and DNAse hypersensitivity regions were most enriched for *de novo* variants [31]. The first study significantly associating genome-wide non-coding variants with ASD shows convergence in the pathways and processes disrupted by both coding and non-coding variants in ASD, specifically in synaptic transmission and neuronal development [32]. Ruzzo et al. also provided evidence that non-coding variants impact neurobiology in ASD, reporting a recurrent 2.5 KB deletion within the promoter of *DLG2*, a gene associated with cognition and learning in mice and human [33].

Preferential transmission of structural non-coding variants has been

X

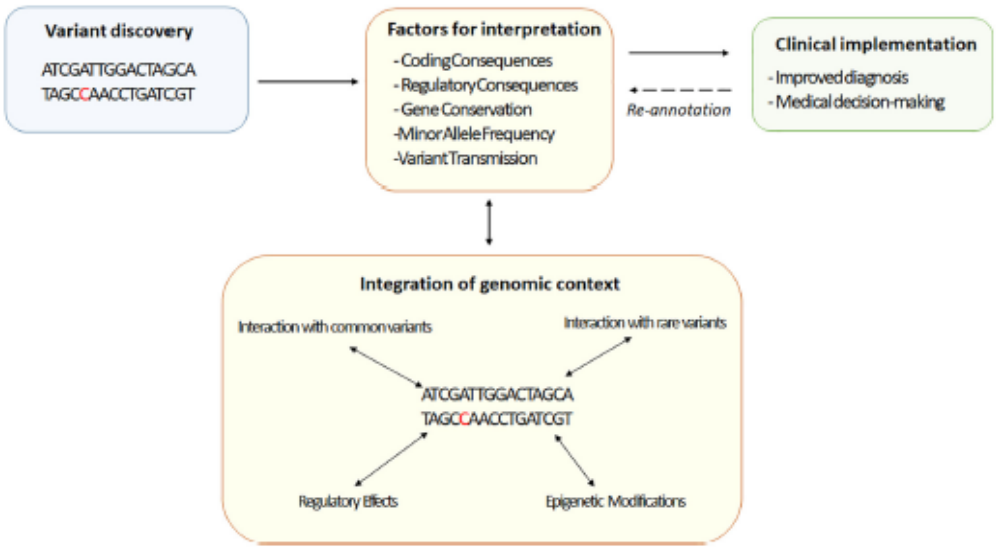


Fig. 1. Pathway from sequencing to clinical implementation. Outlined are the main stages of ASD gene discovery; from variant discovery (blue), through genomic data analysis (yellow), to accurate translation for meaningful diagnosis (green). *Re*-annotation refers to regular re-analysis of genetic diagnosis, as additional variants reach significant association with ASD. The variant highlighted in red, here a single nucleotide variant, represents any variant type detectable through application of genomic technologies (Table 1). Epigenetic modifications include methylation changes, histone modification or microRNA dysregulations (reviewed Eshraghi et al. 2018) [83]. Research is ongoing to integrate genomic variants with other variation within an individual's genome, as described by McGuire et al. (2020) [84]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reported in ASD, specifically the transmission of cis-regulatory elements from father to affected rather than to unaffected offspring [34]. These findings are suggestive that not only are rare inherited non-coding variants increasing risk to ASD, but also indicate a parent-of-origin effect from this non-coding variant class, highlighting a key benefit to the use of a family-based study design in studies of ASD.

6. Family-based studies are key to rare variant analysis in ASD

Family-based studies, previously the foundation of disease gene discovery, are re-emerging as an effective tool to identify potentially pathogenic variants in neuropsychiatric disorders, including ASD [35]. Family-based designs facilitate the analysis of parent to offspring variant transmission. These study designs take the form of i) simplex families (trios); parents and their affected child, ii) multiplex families; parents with more than one affected child, and iii) more complex extended pedigrees with multiple affected individuals. By design, trio studies such as those investigating the MSSNG cohort (Table 2), have been particularly key to uncovering the enrichment of *de novo* variants in cases by comparing rates of *de novo* variants in affected offspring with their unaffected respective siblings [31].

Family-based study designs also enable analyses of parent-of-origin effects that are not possible in case-control design. Furthermore, the presence of matched unaffected siblings in these studies, gives a background level of genetic variation that can be used to distinguish between disease relevant variants and those that are unrelated, such as population-specific background variation or biases introduced in sequencing. A number of large-scale genomic investigations of ASD apply a family-based approach, including the Simons Simplex Collection (Simplex), Autism Genetic Research Exchange (Simplex and Multiplex) and The Autism Genome Project (Simplex and Multiplex) (Table 2).

7. Multiplex and simplex cases of ASD show different genetic architectures

Family structure plays a major role in the types of putative variants expected to be causative of a given ASD proband. Earlier CNV studies in ASD provided some evidence of differences in genetic architecture between simplex and multiplex families [36]. These differences are centred on the contribution of *de novo* and inherited variants to ASD susceptibility.

7.1. De novo variants

A lower rate of *de novo* variation is seen in multiplex families compared to simplex families, as expected by study design. Sebat et al. reported *de novo* copy number variants in 10% of simplex cases and 3% of cases from multiplex families in their cohort [36]. Similarly Ruzzo et al. give evidence for depletion of rare *de novo* ASD risk variants in multiplex families [33]. While, this is observed across multiple studies, the difference between multiplex and simplex family structures is not consistently evident. In their CNV analyses, Pinto et al. did not report such differences [37]. A limitation to these analyses, such as analyses involving the Autism Genome Project cohort (Table 2), arises from challenges in reporting of simplex/multiplex status, *i.e.* identifying a family as a true simplex, or as a family for which just one offspring was investigated.

7.2. Inherited variants

Consistent with the enrichment of *de novo* variants in simplex cases of ASD, there is a depletion of inherited variants associated with ASD in these spontaneous cases [36,38]. Klei et al. estimate narrow sense heritability to exceed 60% for ASD cases in multiplex families but estimate just 40% of narrow sense heritability for simplex families [39]. This means that 60% of phenotypic variance may be attributed to

additive genetic variance in individuals of multiplex families. As in comparison of *de novo* variant enrichment of simplex and multiplex families, this effect is not reported consistently across analyses.

Interestingly, the same putative variant may not be found in all affected individuals within a multiplex family as highlighted recently [40]. This study reports a maternally inherited 15q11.2 deletion in an affected male child and no paternally inherited putative variants from an affected father. Other studies have identified non-sharing of CNVs [41] and single nucleotide variants (SNVs) in members of multiply affected families. In the latter study the two affected siblings did not harbour the same rare risk variant in more than half of the multiplex families studied [42]. Similarly, pathogenically significant CNVs have been identified that are transmitted to an ASD proband from an unaffected parent, and shared with a unaffected sibling [43], adding to evidence for asymptomatic carriers of neurodevelopmental disorder CNVs.

Family studies in epidemiological cohorts from isolated populations have also confirmed that both rare and common genetic variants contribute to the susceptibility to ASD. A study on the Faro Island genetic isolate, affirms the importance of both common and rare variants in ASD susceptibility [44]. This study identifies in a subset of individuals in the cohort carrying rare deleterious variants in genes known already associated with ASD and in this same cohort, common genetic variants were also associated.

Given these two mechanisms of genetic variation, *de novo* and inherited in ASD, genomic sequencing studies in families with multiple affected individuals offers greater opportunity to understand the relative contribution of inherited and *de novo* variation in the genetic architecture of ASD.

8. Establishing putative ASD variants faces many challenges

Heterogeneity in ASD diagnoses is a major challenge facing genomic sequencing studies in ASD. In particular, diagnosis of ASD in the presence of intellectual disability. Diagnostic procedures are found to differ between that used in a clinical and research setting. For a comprehensive discussion on these challenges refer to Schaaf et al. [45].

The greatest challenge in analysis of large-scale genomic data is in the establishment of pipelines for data interpretation. Interpretation of putative variants is complicated by a wide variety of technical factors, such as sequence coverage, variant validation, consistency in sequencing platforms and variant calling and filtering techniques. Robust clinical diagnoses and rich phenotyping increase confidence in variant association [46]. A variant that has been associated with ASD and has substantial evidence supporting its validity will be interrogated for its biological role (Fig. 1).

Variants associated with ASD disrupt a wide variety of pathways and biological processes [7]. Identifying pathways and processes showing an increased mutational burden enables the isolation of cellular processes and pathways disrupted in ASD. Gene-lists are often compiled listing genes involved in a given process [42]. These lists are useful in establishing the process which a putative variant may be disrupting, and such gene lists are often consulted for membership when investigating the impact of a variant [40].

The establishment and maintenance of collective databases, such as SFARI gene [23], Developmental Disorders Genotype-to-Phenotype database (DDG2P) [47] and ClinVar [48], that are openly shared among researchers give hope for the development of variant specific disease models which will expectedly lead to a greater understanding of ASD pathology. Consistent re-analysis of pathogenicity is key to gaining maximum insight from available genomic data, as proven fruitful in the re-annotation of developmental and epileptic encephalopathies genes [49] (Fig. 1). A key stride in the development of an ASD gene list comes from Schaaf et al. in their proposal to adapt the ClinGen curation framework to ASD [45]. Development of a high-confidence gene list for ASD would have great use in genomic investigation, specifically in the development of targeted gene panels and a 'clinical exome'. Without a consensus gene list in ASD, attempts to develop such genome analysis strategies have limited application (Table 1).

Advances in long-read sequencing technologies hold the potential for sequencing of "dark gene regions", genomic regions inaccessible through next-generation sequencing. With high coverage and *de novo* assembly, Nanopore technologies have potential to sequence up to 100% of the genome (Table 1), with the greatest level of 'recovered' genes when compared with other genomic technologies, including the recovery of genes associated with ASD [50]. This technology, to our knowledge, has yet to be applied to ASD cohorts, aside from use in variant validation [34]. Long-read sequencing will enable discovery of genetic variants which have thus far been largely under-explored in ASD, such as repeat expansions, haplotype phased variants and methylation changes. Repeat expansion variants have already been associated with ASD, most notably the *FMR1* repeat expansion associated with Fragile X syndrome (MIM: 30024). As shown in an early haplotype mapping study, identification of haplotypes can succeed in identifying loci involved in ASD susceptibility [51]. Even more relevant perhaps, long-read sequencing enables the detection of CNVs and rearrangement events without the need for bioinformatic re-assembly and alignment of short reads.

9. Putting ASD in the context of other neuropsychiatric disorders

Whole genome sequencing has potential to investigate some of the major questions remaining unanswered in ASD genomics, including investigation of the overlap of ASD with other neurodevelopmental and neuropsychiatric disorders, both clinically and genetically. As highlighted in a review from Lord et al., elucidation of the genetic overlap of ASD with other neuropsychiatric disorders is needed [52]. Clinically, ASD frequently occurs co-morbidly with other neuropsychiatric disorders, in particular attention-deficit hyperactivity disorder (28%), anxiety disorders (13%) and mood disorders (11%) [53].

At the systems-level there is substantial evidence of genetic overlap between ASD and neurodevelopmental and neuropsychiatric disorders [54]. There is overlap in the genes associated with ASD and those associated with other neuropsychiatric disorders, such as schizophrenia and bipolar disorder [55–57]. This has been demonstrated strongly in a large-scale meta-analysis of eight European psychiatric cohorts identifying 109 pleiotropic loci [58]. The genetic overlap of ASD with other disorders is also evident at the variant level with *de novo* variation in ASD shared with intellectual disabilities [8] and shared with epilepsy [59].

10. Next-generation sequencing technologies improve diagnostic yield

There is a demand for clinical genetic testing in ASD [60]. Clinical CNV detection has already been translated widely, advancing the clinical genetics understanding of the condition. This translation crystallised some of the issues that will emerge with widespread translation of genomic technologies; namely clinical interpretation, relative contribution of inherited variants and particularly variant specificity to ASD. Currently no gene, which when disrupted by a pathogenic variant, has been found to confer risk to ASD without conferring risk to intellectual disability or other neurodevelopmental disorders. In the absence of appropriate study design and explicit, robust diagnoses, there is insufficient evidence to assign meaningful specificity of gene involvement in ASD [61].

Genomic technologies, given the greater proportion of the genome covered, have the potential to transform the clinical genetic understanding of the condition. This is illustrated by the increase in diagnostic yield with genomic technologies. Diagnostic yield refers to the number of cases where a putative genetic variant associated with the condition is identified in a cohort. This can be interpreted as a measure

of the utility of the technique and analysis strategy for the condition.

A recent meta-analysis scoping review states that exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders, defined in this study as developmental delay, intellectual disability and/or ASD [62]. The diagnostic yield for whole exome sequencing overall from these meta-analyses is 36%, surpassing the estimated 15–20% diagnostic yield of candidate gene arrays.

Using whole exome sequencing technologies, Feliciano et al. in the SPARK pilot, report a returnable genetic result in 10.4% of their cohorts affected offspring [40]. Importantly, in individuals with more complex phenotypes, such as ASD with seizures or co-morbid intellectual disability, they report a higher diagnostic yield than overall (27% and 20% respectively). This finding is consistent with other studies [62,63]. The SPARK study also reports a higher diagnostic yield in cases from multiplex families (15.2%) than simplex families (10.1%) [40].

Yuen et al. find a diagnostic yield of ASD relevant variants using whole genome sequencing to be 42.4% in their cohort of 85 multiplex families of ASD. This mirrors the diagnostic yield estimated in intellectual disability using the same sequencing platform [42,64]. The increased diagnostic yield using whole genome sequencing highlights the great potential for use of the technology in families with ASD. This estimate can be expected to increase further with developments in variant interpretation strategies and increases in sample sizes, giving more power to investigations of common variants and variants in the non-coding regions of the genome.

The clinical utility of whole genome sequencing holds great promise; however, this sequencing approach also faces major challenges. These include the need for large-cohort analyses and the failure to replicate genomic findings. One example is the report of the enrichment of *de novo* and private disruptive mutations within fetal CNS DNase I hypersensitive sites within 50 kb of genes that have been previously associated with autism risk [65] that later did not replicate despite a larger sample size [66]. Furthermore, we face limitations to the current capacity to interpret variants in the non-coding genome, as discussed by Lee & Gleeson [67]. Notwithstanding these challenges, the decrease in sequencing costs (Table 1) and the increase in sample sizes under investigation, together with the greater understanding of family inheritance will continue to give a more precise estimate of the diagnostic yield in ASD. The return of genetic results, alongside current behavioural diagnoses, may be used to improve therapeutic avenues in the future. Genetic diagnoses may also be used to inform family planning on a family-by-family basis as illustrated by a recent family study showing the CNV findings, which would have been pre-symptomatically predictive of ASD or atypical development in 7% (11 of 157) of families analysed [68].

11. Conclusion

Whole genome sequencing is the most effective technology to improve our biological understanding of neurodevelopmental disorders. With near full coverage of the human genome, coupled with the increase in sample sizes and the development of cutting-edge analytical methods, we now have the potential to identify more variants across the genome, in particular more rare pathogenic genetic variants. The detection of rare variants by genomic technologies will improve our understanding of the genetic architecture of ASD and other neurodevelopmental and neuropsychiatric disorders. With advances in biological interpretation enabling delivery of genetic discovery into clinical translation, genomic technologies will become an achievable step towards personalised family medicine, ultimately aiding ASD diagnosis and informing medical decision-making.

Declaration of Competing Interest

None.

Acknowledgements

Thank you very much to Dr. Elaine Kenny, TrinSeq–Genomics Core Facility, Dublin, Ireland for estimating Illumina sequencing costs. This publication has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant No. 15/SIRG/3324.

References

[1] G. Baird, et al., Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the special needs and autism project (SNAP), Lancet 368 (2006) 210–215.
[2] American Psychiatric Association, Diagnostic and statistical manual of mental disorders, 5th ed., (2013).
[3] S. Sandin, et al., The heritability of autism spectrum disorder, JAMA 318 (2017) 1182.
[4] B. Tick, P. Bolton, F. Happé, M. Rutter, F. Rijsdijk, Heritability of autism spectrum disorders: a meta-analysis of twin studies, J. Child Psychol. Psychiatry 57 (2016) 585–595.
[5] T. Gaugler, et al., Most genetic risk for autism resides with common variation, Nat. Genet. 46 (2014) 881–885.
[6] K.H. Miga, et al., Telomere-to-telomere assembly of a complete human X chromosome, Nature (2020) 1–9, https://doi.org/10.1038/s41586-020-2547-7.
[7] S. De Rubeis, et al., Synaptic, transcriptional and chromatin genes disrupted in autism, Nature 515 (2014) 209–215.
[8] F.K. Satterstrom, et al., Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism, Cell 180 (2020) 568-584.e23.
[9] Schizophrenia Working Group of the Psychiatric Genomics Consortium, S. W. G. of the P. G, et al., Biological insights from 108 schizophrenia-associated genetic loci, Nature 511 (2014) 421–427.
[10] E.A. Stahl, et al., Genome-wide association study identifies 30 loci associated with bipolar disorder, Nat. Genet. 51 (2019) 793–803.
[11] B. Creese, et al., Examining the association between genetic liability for schizophrenia and psychotic symptoms in Alzheimer's disease, Transl. Psychiatry 9 (2019).
[12] Psychiatric GWAS Consortium Bipolar Disorder Working Group, P, et al., Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4, Nat. Genet. 43 (2011) 977–983.
[13] J. Grove, et al., Identification of common genetic risk variants for autism spectrum disorder, Nat. Genet. 51 (2019) 431–444.
[14] J.A.S. Vorstman, et al., No evidence that common genetic risk variation is shared between schizophrenia and autism, Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 162 (2013) 55–60.
[15] B.S. Abrahams, D.H. Geschwind, Advances in autism genetics: on the threshold of a new neurobiology, Nat. Rev. Genet. 9 (2008) 341–355.
[16] MENDELIAN.CO, Autism Spectrum Disorders: Definition and Top 150+ Rare Diseases Related to Them, MENDELIAN.CO., 2019 Available at: https://www.mendelian.co/autism-spectrum-disorders-150-rare-diseases-related (Accessed 3rd April 2019).
[17] C. Betancur, Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting, Brain Res. 1380 (2011) 42–77.
[18] K.E. Tansey, et al., Common alleles contribute to schizophrenia in CNV carriers, Mol. Psychiatry 21 (2016) 1085–1089.
[19] E.A. Boyle, Y.I. Li, J.K. Pritchard, Leading Edge Perspective An Expanded View of Complex Traits: From Polygenic to Omnigenic, (2017), https://doi.org/10.1016/j.cell.2017.05.038.
[20] X. Liu, Y.I. Li, J.K. Pritchard, Trans effects on gene expression can drive omnigenic inheritance, Cell 177 (2019) 1022-1034.e6.
[21] E. Fombonne, Epidemiological surveys of autism and other pervasive developmental disorders: An update, J. Autism Dev. Disord. 33 (2003) 365–382.
[22] Turner, T. N. et al. Sex-based analysis of De novo variants in neurodevelopmental disorders. Am. J. Hum. Genet. 0, (2019).
[23] B.S. Abrahams, et al., SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs), Mol. Autism 4 (2013) 36.
[24] J. Veenstra-VanderWeele, S.L. Christian, E.H. Cook Jr., Autism as a paradigmatic complex genetic disorder, Annu. Rev. Genomics Hum. Genet. 5 (379–405) (2004).
[25] E.J. Gardner, et al., Contribution of retrotransposition to developmental disorders, Nat. Commun. 10 (2019) 1–10.
[26] S.J. Sanders, et al., Insights into autism Spectrum disorder genomic architecture and biology from 71 risk loci, Neuron 87 (2015) 1215–1233.
[27] J. Sebat, et al., Large-scale copy number polymorphism in the human genome, Science 305 (2004) 525–528.
[28] A.J. Iafrate, et al., Detection of large-scale variation in the human genome, Nat. Genet. 36 (2004) 949–951.
[29] M. Zarrei, et al., A large data resource of genomic copy number variation across neurodevelopmental disorders, NPJ Genomic Med. 4 (2019) 1–13.
[30] D. Botstein, N. Risch, Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, Nat. Genet. 33 (2003) 228–237.
[31] R.K. Yuen, et al., Genome-wide characteristics of de novo mutations in autism, NPJ Genomic Med. 1 (2016), https://doi.org/10.1038/npjgenmed.2016.27.
[32] J. Zhou, et al., Whole-genome deep-learning analysis identifies contribution of

noncoding mutations to autism risk, Nat. Genet. 51 (2019) 973–980.
[33] E.K. Ruzzo, et al., Inherited and De novo genetic risk for autism impacts shared networks, Cell 178 (2019) 850-866.e26.
[34] W.M. Brandler, et al., Paternally inherited cis-regulatory structural variants are associated with autism, Science 360 (2018) 327–331.
[35] D.C. Glahn, et al., Rediscovering the value of families for psychiatric genetics research, Mol. Psychiatry 24 (2019) 523–535.
[36] J. Sebat, et al., Strong association of De novo copy number mutations with autism, Science 316 (2007) 445–449.
[37] D. Pinto, et al., Functional impact of global rare copy number variation in autism spectrum disorders, Nature 466 (2010) 368–372.
[38] M. Ronemus, I. Iossifov, D. Levy, M. Wigler, The Role of De Novo Mutations in the Genetics of Autism Spectrum Disorders, (2014), https://doi.org/10.1038/nrg3585.
[39] L. Klei, et al., Common genetic variants, acting additively, are a major source of risk for autism, Mol. Autism 3 (2012) 9.
[40] P. Feliciano, et al., Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes, NPJ Genomic Med. 4 (2019) 1–14.
[41] V.M. Leppa, et al., Rare Inherited and De novo CNVs reveal complex contributions to ASD risk in multiplex families, Am. J. Hum. Genet. 99 (2016) 540–554.
[42] R.K.C. Yuen, et al., Whole-genome sequencing of quartet families with autism spectrum disorder, Nat. Med. 21 (2015) 185–191.
[43] M. Woodbury-Smith, et al., Variable phenotype expression in a family segregating microdeletions of the NRXN1 and MBD5 autism spectrum disorder susceptibility genes, NPJ Genomic Med. 2 (2017) 1–8.
[44] C.S. Leblond, et al., Both rare and common genetic variants contribute to autism in the Faroe Islands, NPJ Genomic Med. 4 (2019) 1.
[45] C.P. Schaaf, et al., A framework for an evidence-based gene list relevant to autism spectrum disorder, Nat. Rev. Genet. (2020) 1–10, https://doi.org/10.1038/s41576-020-0231-2.
[46] D.B. Callaghan, et al., Whole genome sequencing and variant discovery in the ASPIRE autism spectrum disorder cohort, Clin. Genet. (2019), https://doi.org/10.1111/cge.13556 cge.13556.
[47] C.F. Wright, et al., Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data, Lancet (London, England) 385 (2015) 1305–1314.
[48] M.J. Landrum, et al., ClinVar: public archive of relationships among sequence variation and human phenotype, Nucleic Acids Res. 42 (2014).
[49] C.A. Steward, et al., Re-annotation of 191 developmental and epileptic encephalopathy-associated genes unmasks de novo variants in SCN1A, NPJ Genomic Med. 4 (2019).
[50] M.T.W. Ebbert, et al., Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight, Genome Biol. 20 (2019) 1–23.
[51] J.P. Casey, et al., A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder, Hum. Genet. 131 (2012) 565–579.
[52] C. Lord, et al., Autism spectrum disorder, Nat. Rev. Dis. Prim. 6 (2020) 1–23.
[53] M.-C. Lai, et al., Prevalence of co-occurring mental health diagnoses in the autism population: a systematic review and meta-analysis, Lancet Psychiatry 6 (2019) 819–829.
[54] J.Y. An, C. Claudianos, Genetic heterogeneity in autism: from single gene to a pathway perspective, Neurosci. Biobehav. Rev. 68 (2016) 442–453.
[55] L.S. Carroll, M.J. Owen, Genetic overlap between autism, schizophrenia and bipolar disorder, Genome Med. 1 (2009) 102.
[56] D.H. Geschwind, J. Flint, Genetics and genomics of psychiatric disease, Science 349 (2015) 1489–1494.
[57] P.H. Lee, et al., Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders, Cell 179 (2019) 1469-1482.e11.
[58] S.H. Lee, et al., Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs, Nat. Genet. 45 (2013) 984–994.
[59] H.O. Heyne, et al., De novo variants in neurodevelopmental disorders with epilepsy, Nat. Genet. 50 (2018) 1048–1053.
[60] K.S. Barton, et al., Pathways from autism spectrum disorder diagnosis to genetic testing, Genet. Med. 20 (2018) 737–744.
[61] S.M. Myers, et al., Insufficient evidence for "autism-specific" genes, Am. J. Hum. Genet. 106 (2020) 587–595.
[62] S. Srivastava, et al., Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders, Genet. Med. 1 (2019), https://doi.org/10.1038/s41436-019-0554-6.
[63] K. Tammimies, et al., Molecular diagnostic yield of chromosomal microarray analysis and whole-exome sequencing in children with autism Spectrum disorder, JAMA 314 (2015) 895.
[64] C. Gilissen, et al., Genome sequencing identifies major causes of severe intellectual disability, Nature 511 (2014) 344–347.
[65] T.N. Turner, et al., Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA, Am. J. Hum. Genet. 98 (2016) 58–74.
[66] T.N. Turner, et al., Genomic patterns of De novo mutation in simplex autism, Cell 171 (2017) 710-722.e12.
[67] S. Lee, J.G. Gleeson, Closing in on mechanisms of open neural tube defects, Trends Neurosci. 43 (2020) 519–532.
[68] L. D'Abate, et al., Predictive impact of rare genomic copy number variations in siblings of individuals with autism spectrum disorders, Nat. Commun. 10 (2019) 5519.
[69] I. Mitra, et al., The contribution of de novo tandem repeat mutations to autism spectrum disorders, bioRxiv (2020), https://doi.org/10.1101/2020.03.04.974170.
[70] N. Mousavi, S. Shleizer-Burko, R. Yanicky, M. Gymrek, Profiling the genome-wide landscape of tandem repeat expansions, Nucleic Acids Res. 47 (2019) 90.
[71] J.Y. An, et al., Towards a molecular characterization of autism spectrum disorders: An exome sequencing and systems approach, Transl. Psychiatry 4 (2014).
[72] J.D. Buxbaum, et al., The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders, Neuron 76 (2012) 1052–1056.
[74] C.B. Pedersen, et al., The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders, Mol. Psychiatry 23 (2018) 6–14.
[75] F.K. Satterstrom, et al., ASD and ADHD have a similar burden of rare protein-truncating variants, bioRxiv (2018), https://doi.org/10.1101/277707.
[76] R.K.C. Yuen, et al., Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder, Nat. Neurosci. 20 (2017) 602–611.
[77] D. Levy, et al., Rare De novo and transmitted copy-number variation in autistic Spectrum disorders, Neuron 70 (2011) 886–897.
[78] S.J. Sanders, et al., Multiple recurrent De novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism, Neuron 70 (2011) 863–885.
[79] J.Y. An, et al., Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder, Science 362 (2018).
[80] D.M. Werling, et al., An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder, Nat. Genet. 50 (2018) 727–736.
[81] T.A.G.P. Consortium, et al., Mapping autism risk loci using genetic linkage and chromosomal rearrangements, Nat. Genet. 39 (2007) 319–328.
[82] J.D. Buxbaum, et al., The autism simplex collection: an international, expertly phenotyped autism sample for genetic and phenotypic analyses, Mol. Autism 5 (2014) 34.
[83] A.A. Eshraghi, et al., Epigenetics and autism spectrum disorder: is there a correlation? Front. Cell. Neurosci. 12 (2018) 78.
[84] A.L. McGuire, et al., The road ahead in genetics and genomics, Nat. Rev. Genet. (2020) 1–16, https://doi.org/10.1038/s41576-020-0272-6.

Appendix IV-I: Ní Ghrálaigh, F. *et al.* (2022) 'Brief Report: Evaluating the Diagnostic Yield of Commercial Gene Panels in Autism', Journal of Autism and Developmental Disorders 2022. Springer, pp. 1–5. doi: 10.1007/S10803-021-05417-7.

Journal of Autism and Developmental Disorders
https://doi.org/10.1007/s10803-021-05417-7

BRIEF REPORT

Brief Report: Evaluating the Diagnostic Yield of Commercial Gene Panels in Autism

Fiana Ní Ghrálaigh[1,2] · Ellen McCarthy[1] · Daniel N. Murphy[1] · Louise Gallagher[2] · Lorna M. Lopez[1]

Accepted: 20 December 2021

Abstract
Autism is a prevalent neurodevelopmental condition, highly heterogenous in both genotype and phenotype. This communication adds to existing discussion of the heterogeneity of clinical sequencing tests, "gene panels", marketed for application in autism. We evaluate the clinical utility of available gene panels based on existing genetic evidence. We determine that diagnostic yields of these gene panels range from 0.22% to 10.02% and gene selection for the panels is variable in relevance, here measured as percentage overlap with SFARI Gene and ranging from 15.15% to 100%. We conclude that gene panels marketed for use in autism are currently of limited clinical utility, and that sequencing with greater coverage may be more appropriate.

Keywords Autism · Panel · Sequencing · Genomics · Utility

Introduction

The benefits of a genetic diagnosis of autism are extensive ("Genetic Testing Statement | ISPG—International Society of Psychiatric Genetics"). The International Society of Psychiatric Genetics propose in their consensus statement on genetic testing that the "*identification of known pathogenic variants may help diagnose rare conditions that have important medical and psychiatric implications for individual patients and may inform family counselling*". A genetic diagnosis of autism may allow for the prospect of genetic counselling for affected individuals and their families; it may also provide an affected individual with opportunity to take part in targeted research or receive anticipatory medical advice.

Genetic diagnosis in autism is limited by the ability to robustly determine the relevance of putatively pathogenic genetic variation. Genomic research in autism is progressing quickly, enabled by advancements in next-generation sequencing technologies and the subsequent establishment of large-scale sequencing cohorts and pedigree-based sequencing cohorts (Glahn et al., 2019; Ní Ghrálaigh et al., 2020). To date, many genes have been identified as having some link to autism (Satterstrom et al., 2020). The Simons Foundation Autism Research Initiative (SFARI) Gene database (Abrahams et al., 2013), collates more than 990 genes for which there is evidence of association with autism, however the clinical utility of this database is limited by the absence of systematic curation of gene-condition relationships (Schaaf et al., 2020). Despite this progress in autism genomic research, major challenges remain in the development of targeted gene panels with substantial clinical utility in autism.

At case-level, gene discovery is complicated by the nature of autism as a complex condition with a large degree of phenotypic heterogeneity. A candidate pathogenic variant may be evaluated, in the majority of autism cases, as being contributory to the genetic risk rather than being wholly causative of an individual's condition. At cohort-level, studies

✉ Fiana Ní Ghrálaigh
fiana.nighralaigh.2020@mumail.ie

Ellen McCarthy
ellen.mccarthy.2018@mumail.ie

Daniel N. Murphy
Daniel.N.Murphy@mu.ie

Louise Gallagher
LGALLAGH@tcd.ie

Lorna M. Lopez
lorna.lopez@mu.ie

[1] Department of Biology, Maynooth University, Maynooth, Co Kildare, Ireland

[2] Department of Psychiatry, Trinity College Dublin, Dublin, Ireland

Published online: 07 January 2022

discovering "autism genes" are compounded by an apparent lack of specificity to autism. For example, in individuals affected by both autism and intellectual disability, genes identified show relevance to both autism and other neurodevelopmental disorders (Myers, Challman, Bernier, et al., 2020). For these reasons, the development of effective gene panels to aid autism diagnosis is extremely complicated.

Despite these limitations, commercial gene panels are available and marketed for use in autism diagnosis. Hoang et al. (2018) evaluated many of these gene panels, clearly demonstrating their heterogeneity (Hoang et al., 2018). Their survey shows large variability in the number of genes being tested by panels, lack of consensus in the genes selected for inclusion, as well as variability in the reporting of laboratory qualification and reporting protocols.

Methods

Identifying Autism Gene Panels

Gene panels marketed for use in autism were identified and collated through the following approaches: web browser search (search terms *"autism gene panel"*, *"ASD gene panel"*, *"sequencing tests for autism spectrum disorder"*, *"gene panels for autism testing"* and *"autism genetic testing"*), gene panels analysed by Hoang et al. (2018) (Hoang et al., 2018) and Genomics England PanelApp (search terms *"Autism"*, *"ASD"*) (Martin et al., 2019). Panels identified for which gene lists were not provided were excluded from analyses (CGC genetics "Autism" panel & Michigan Medicine "Autism/ Intellectual Disability Panels"). Gene list sources are outlined in Supplemental Table 1 (collated October 2020-January 2021).

Refining Gene Lists

Each gene panel identified provides a list of genes targeted by the probes. By nature, these gene lists arise from a variety of sources and were compiled at varying times. For this reason, gene lists were run through HUGO Gene Nomenclature Committee (HGNC) Multi-Symbol Tool (Version: 2021–01-06 update). Where the gene symbol reported by the provider is an approved gene symbol in HGNC, it is used in analyses. Where the gene symbol is no longer approved by HGNC, it was updated to the approved gene symbol given by HGNC. A small number of deviations occurred that could not be resolved, which resulted in the removal of genes from the analyses. The resulting refined gene lists are provided in Supplemental Table 2. Gene counts reported in Table 1 also reflect these updates.

Estimating Percentage Overlap with SFARI Gene

To determine the relevance of genes targeted in autism, each panel was assessed for the proportion of genes covered that are included in the SFARI Gene database (all gene scores and genetic categories) of genes implicated in autism susceptibility (Version: 2021–01-13 release). Where necessary, the SFARI gene list (n = 1,003) was updated to HGNC approved gene symbols (n = 5) and genes with symbol mismatch (n = 3) were removed. The number of genes targeted by each panel that are included in SFARI gene are presented in Table 1 as a percentage of the total genes in the panel. SFARI Gene was subset to high-confidence autism associated genes, assigned as such based on SFARI gene scoring of 1 or 2. Percentage overlap was again calculated on this subset and presented in Table 1.

Selection of Clinically Relevant Variants

Clinically relevant variants, as identified and characterised by whole exome sequencing in the Simon's Powering Autism Research Knowledge cohort, were used to determine the clinical utility of each panel. Variants included in our analyses are those reported in Feliciano et al. (2019) data set 10 (Feliciano et al., 2019), comprising inherited and de novo single nucleotide variants (SNVs), insertion deletion variants (indels) and copy number variants (CNVs). Reported chromosomal abnormalities were not included. Gene lists were assembled to include those for which clinically relevant SNVs and indels could be defined and those that fall within the boundaries of clinically relevant CNVs. While targeted gene panels lack the ability to define copy number variant boundaries, genes within these variants will appear as deleted or duplicated, thus variation will be detected.

Determining and Reporting Diagnostic Yield

Diagnostic yield was determined by cross-referencing the gene list of each gene panel with the lists of implicated genes in the SPARK cohort. Diagnostic yield was calculated as the proportion of individuals sequenced, for which a relevant genetic variant was identified, corresponding with the genes contained on each gene panel. The number of individuals in the cohort was taken as 472 affected individuals (465 offspring and 7 parents) as detailed by Feliciano et al. (2019). In keeping with this study, 13 individuals, those in families self-reporting a genetic diagnosis were not included in the estimates of diagnostic yield. With this justification, diagnostic yield was calculated as the number of individuals with a relevant variant, as a percentage of the total cohort of 459 affected individuals without a genetic diagnosis.

The number of individuals for which a clinically relevant finding would have been identified by using each targeted gene panel is reported for both pathogenic and probable pathogenic variants, as assigned by Feliciano et al. (2019) (Table 1).

Determining and Reporting Correlation

Pearson's product-moment correlation was computed with n = 16 degrees freedom for diagnostic yield and number of genes targeted and for diagnostic yield against percentage overlap with SFARI gene (all genes).

Results

Here we estimate the clinical utility of commercial gene panels marketed for use in autism. Diagnostic yield, which is the proportion of cases interrogated for which a genetic cause can be determined, is a strong measure of the clinical utility of a sequencing technology. Feliciano

Table 1 Diagnostic yield of gene panels marketed for use in autism

Service provider	Panel name	Number of genes targeted	Percentage overlap with SFARI gene		Diagnostic yield in SPARK
			SFARI gene All genes	SFARI high confidence Genes (Scores 1 and 2)	
Ambry Genetics	AutismNext Panel	72	87.5%	76.39%	2.61%
Asper Neurogenetics	Autism Spectrum Disorders NGS Panel	76	88.16%	71.05%	2.83% (0.22%)
Blueprint Genetics	Autism Spectrum Disorders Panel	75	45.33%	36%	1.53% (0.44%)
Center for Human Genetics	Autism Spectrum Disorder 53-Gene Panel	53	84.91%	45.28%	1.96% (0.22%)
Centogene	Syndromic Autism Gene Panel	50	88%	76%	2.4% (0.22%)
Centogene	Intellectual Disability Panel	599	43.41%	24.54%	5.23% (1.31%)
EGL Genetics	Autism Spectrum Disorders Tier 2 Panel	62	74.19%	66.13%	2.18%
Fulgent Genetics	Autism NGS Panel	121	76.86%	55.37%	4.36% (0.44%)
GeneDx	Autism/ID Xpanded Panel	2641	20.64%	10.98%	10.02% (3.49%)
GENETAQ	Autism	27	92.59%	66.67%	1.53%
Genomics England PanelApp	Autism (Version 0.20)	733	100%	42.7%	7.63% (1.96%)
Greenwood Genetic Centre	Syndromic Autism Sequencing Panel	83	80.72%	69.88%	3.05%
GX Sciences	Developmental Nutrigenomic Panel	33	15.15%	0%	0.22%
MNG Laboratories	Comprehensive Disability/Autism Panel	1345	19.85%	12.04%	6.1% (1.3%)
Munroe-Meyer Institute	Autism/Intellectual Disability/Multiple Anomalies Panel	117	55.56%	41.88%	2.4% (0.22%)
Prevention Genetics	Autism Spectrum Disorders Panel	170	95.29%	90.59%	6.32% (0.44%)
Reference Laboratory Genomics	Autism Spectrum Disorders (Expanded Panel)	77	77.92%	64.94%	3.05% (0.44%)
Sema4	Comprehensive Autism Spectrum Disorders Panel (228)	228	57.46%	43.42%	4.79% (0.87%)

Presented are gene panels relevant to autism. The number of genes present in each gene panel are correct as of January 2021. Gene lists provided at the sources listed in Supplemental Table 1 were updated to HGNC approved gene symbols where necessary. Percentage overlap with SFARI is estimated as the proportion of genes within each respective gene list appearing in SFARI Gene (01–13-2021 release). This overlap is presented for both the complete SFARI Gene gene lists and the High Confidence SFARI genes only (Scores 1 and 2). Diagnostic yield is estimated as the number of individuals for which a genetic cause of autism was identified as a proportion of those investigated (459 affected individuals for which no genetic diagnosis was previously reported). Pathogenic variation is considered as variants listed in Feliciano et al. (2019). Variants considered are de novo and inherited single nucleotide variants, insertion-deletion variants and copy number variants. Diagnostic yield of pathogenic variation is listed, with the additional diagnostic yield achieved by inclusion of probable pathogenic variants listed in brackets alongside

et al. (2019) estimate the diagnostic yield of whole exome sequencing to be 10.4% in the initial 457 families enrolled in the Simons Powering Autism Research (SPARK) cohort (Feliciano et al., 2019). A 'likely pathogenic' variant was identified in a further 3.4% of families studied. This estimate comes from the identification of a variant that fulfils either the 'likely pathogenic' or 'pathogenic' criteria, according to American College of Medical Genetics and Genomics (ACMG) standards (Richards et al., 2015).

Gene panels relevant to autism are presented in Table 1. To determine the clinical utility of each autism gene panel, variants meeting 'likely pathogenic' or 'pathogenic' criteria in the SPARK cohort can be limited to those within the gene set of each panel, respectively. In doing so, we ask how many of the pathogenic variants identified by Feliciano et al. would have been identified in the SPARK cohort with application of an autism gene panel, instead of application of whole exome sequencing. The diagnostic yield of each gene panel, estimated with respect to Feliciano et al. analyses, is presented in Table 1. The diagnostic yields range from 0.22% to 10.02%, with most gene panels achieving a diagnostic yield below 3%.

Gene discovery in autism is ongoing. Most genes included in the commercial gene panels are autism relevant. This is illustrated by the inclusion of a large proportion of the panel-specific genes in the SFARI Gene database (Abrahams et al., 2013)(Table 1). Gene selection for inclusion in autism panels is variable in relevance, ranging in overlap with SFARI Gene from 15.15% to 100%. Diagnostic yield of the gene panels and size of the panel were found to be positively correlated, ($r = 0.82$, $p = 3.033e\text{-}05$), indicating an increased number of genes per gene panel enables detection of a clinically relevant variant in a greater number of individuals. No significant correlation between percentage overlap with SFARI Gene and diagnostic yield was detected.

Discussion

Considering the low diagnostic yield of the gene panels that were investigated, we can infer that, while the gene selection for inclusion in autism gene panels is evidence-based, these gene lists are not extensive enough to justify use in autism diagnosis, a complex trait for which hundreds of genes have been associated. Critically, if the application of a targeted gene panel to an affected individual's genome returns negative for pathogenic variation, one cannot conclude that a causative variant is not present. Rather, it is more likely that genetic causes have been missed due to the limited application of the gene panel.

The GeneDx "Autism/ID Xpanded Panel" represents the autism gene panel with the highest number of individuals for which a genetic diagnosis would have been obtained with its application (10.02%). This diagnostic yield is comparable to that of whole exome sequencing, 10.4% (Feliciano et al., 2019) and that of chromosomal microarray sequencing with a median diagnostic yield of 8.1% (Savatt & Myers, 2021). However, important to note is that this gene panel targets many more genes ($n = 2{,}641$) than some of the smaller gene panels, for example GENETAQ "Autism" panel ($n = 27$), with a diagnostic yield of just 1.53%. The positive correlation of diagnostic yield associated with inclusion of a larger number of genes, reflects well the complex genetic architecture of autism and the number of loci expected to be associated. This raises the question whether autism is an appropriate candidate for the development of commercial gene panels, that are reliant and limited due to the size of the gene panel, the cost and current knowledge of the genetic basis of autism, and questions whether developments should focus on application of sequencing technologies with a broader coverage, such as whole genome sequencing. Expanding beyond targeted autism genes, whole genome sequencing presents the opportunity to explore more of the human genome and, ultimately, to further increase the diagnostic yield in autism (Yuen et al., 2015). Progress in non-coding variant annotation and interpretation, accompanied by a decrease in sequencing costs, may further popularize the clinical use of whole genome sequencing. Currently, whole exome sequencing is proposed as the first-tier diagnostic test for neurodevelopmental disorders (Srivastava et al., 2019). The diagnostic yield in autism using clinical exome sequencing has been estimated at 6.1% in autism (20% overall yield in neurodevelopmental disorders) (Martinez-Granero et al., 2021). Genotyping chips have limited clinical utility for rare genetic variation of SNVs and should not be used to guide health decisions without validation (Mn et al., 2021). Autism genetic testing as minimal as cytogenetic microarray and Fragile X testing alone may be all that is feasible in a clinical setting, which is currently the situation in Ireland.

Provided the relevant expertise and infrastructure for variant interpretation are available and cost effective, gene panels have potential for clinical utility. However, current evidence does not support their applicability in autism (Buxbaum et al., 2020; Myers, Challman, Martin, et al. 2020). Achieving the ultimate goal of a comprehensive autism gene panel will require uniform robust phenotyping to account for the heterogeneity in autism presentation. Application of a formal evidence-based gene curation framework, such as that proposed by Schaaf et al. (Schaaf et al., 2020), would account for the degree of certainty in autism diagnoses in studies reporting association and account for co-morbid diagnoses, providing consistency throughout gene discovery. To conclude, evaluation of the diagnostic

yield of commercial gene panels marketed for autism determines that they are currently of very limited clinical utility.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10803-021-05417-7.

Acknowledgements This publication has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant No. 15/SIRG/3324.

Data Availability All data generated or analysed during this study are included or referenced in this published article and its supplementary information files.

Code Availability https://github.com/FianaNG/autism_gene_panels.

Declarations

Conflict of Interests The authors declare that they have no competing interests.

Ethical Approval Not applicable.

Consent for Publication Not applicable.

Consent to Participate Not applicable.

References

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., et al. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular Autism*. https://doi.org/10.1186/2040-2392-4-36

Buxbaum, J. D., Cutler, D. J., Daly, M. J., Devlin, B., Roeder, K., Sanders, S. J., et al. (2020). Not all autism genes are created equal: a response to myers. *American Journal of Human Genetics Cell Press*. https://doi.org/10.1016/j.ajhg.2020.09.013

Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T. N., Wang, T., Brueggeman, L., et al. (2019) Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *npj Genomic Medicine*, *4*(1), 1–14 Doi: https://doi.org/10.1038/s41525-019-0093-8

Genetic Testing Statement | ISPG - International Society of Psychiatric Genetics. (n.d.). *2013*. https://ispg.net/genetic-testing-statement/. Accessed 3 March 2021

Glahn, D. C., Nimgaonkar, V. L., Raventós, H., Contreras, J., McIntosh, A. M., Thomson, P. A., et al. (2019). Rediscovering the value of families for psychiatric genetics research. *Molecular Psychiatry. Nature Publishing Group*. https://doi.org/10.1038/s41380-018-0073-x

Hoang, N., Buchanan, J. A., & Scherer, S. W. (2018). Heterogeneity in clinical sequencing tests marketed for autism spectrum disorders. *Genomic Medicine*. https://doi.org/10.1038/s41525-018-0066-3

Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., et al. (2019). November 1) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics. Nature Publishing Group*. https://doi.org/10.1038/s41588-019-0528-2

Martinez-Granero, F., Blanco-Kelly, F., Sanchez-Jimeno, C., Avila-Fernandez, A., Arteche, A., Bustamante-Aragones, A., et al. (2021). Comparison of the diagnostic yield of aCGH and genome-wide sequencing across different neurodevelopmental disorders. *Genomic Medicine*, *6*(1), 1–12. https://doi.org/10.1038/s41525-021-00188-7

Mn, W., & L, J., Jw, H., Ks, R., J, T., At, H., & Cf, W. (2021). Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation. *BMJ (clinical Research Ed.)*, *372*, n214. https://doi.org/10.1136/bmj.n214

Myers, S. M., Challman, T. D., Martin, C. L., & Ledbetter, D. H. (2020) Response to Buxbaum et al. *American Journal of Human Genetics*. Cell Press Doi: https://doi.org/10.1016/j.ajhg.2020.09.012

Myers, S. M., Challman, T. D., Bernier, R., Bourgeron, T., Chung, W. K., Constantino, J. N., et al. (2020). Insufficient evidence for "autism-specific" genes. *American Journal of Human Genetics*, *106*(5), 587–595. https://doi.org/10.1016/j.ajhg.2020.04.004

Ní Ghrálaigh, F., Gallagher, L., & Lopez, L. M. (2020). Autism spectrum disorder genomics: the progress and potential of genomic technologies. *Genomics*. https://doi.org/10.1016/j.ygeno.2020.09.022

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J. Y., et al. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, *180*(3), 568-584.e23. https://doi.org/10.1016/j.cell.2019.12.036

Savatt, J. M., & Myers, S. M. (2021). Genetic testing in neurodevelopmental disorders. *Frontiers in Pediatrics*, *9*, 526779. https://doi.org/10.3389/FPED.2021.526779

Schaaf, C. P., Betancur, C., Yuen, R. K. C., Parr, J. R., Skuse, D. H., Gallagher, L., et al. (2020). A framework for an evidence-based gene list relevant to autism spectrum disorder. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-0231-2

Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., et al. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genetics in Medicine*. https://doi.org/10.1038/s41436-019-0554-6

Yuen, R. K. C., Thiruvahindrapuram, B., Merico, D., Walker, S., Tammimies, K., Hoang, N., et al. (2015). Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature Medicine*, *21*(2), 185–191. https://doi.org/10.1038/nm.3792

Appendix V-I: Determining the Clinical Utility of Autism Gene Panels

Available at https://github.com/FianaNG/autism_gene_panels.