Improving The Diagnostic Yield of Whole Genome Sequencing: Neurodevelopmental and Epilepsy Disorders



Thesis submitted in fulfilment for the degree of

Doctor of Philosophy

Trinity Term 2023

Mona Hashim

Linacre College



Declaration

I declare that this thesis is entirely my own work, except where otherwise stated, and describes my own research.

Mona Hashim

Linacre College

Publications

Pagnamenta, A. T., Yu, Y., Walker, S., Noble, A.J., Lord, J., Dutta, P., **Hashim, M**, Camps, C., … & Taylor, J. C. (2023). The impact of inversions across 33,924 families with rare disease from a national genome sequencing project. In press

Pagnamenta, A. T., Camps, C., Giacopuzzi, E., Taylor, J. M., **Hashim, M**., Calpena, E., ... & Taylor, J. C. (2023). Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases. Genome medicine, 15(1), 94.

Oppermann, H., Marcos-Grañeda, E., Weiss, L.A., Gurnett, C.A., Jelsig, A.M., Vineke, S.H., Isidor, B., Mercier, S., Magnussen, K., Zacher, P., **Hashim, M**.... & Platzer, K. (2023). CUX1-related neurodevelopmental disorder: deep insights into phenotype-genotype spectrum and underlying pathology. European Journal of Human Genetics, 31(11), 1251-1260.

Hashim, M., Stewart, H., Yu, J., Banos-Pinero, B., Pagnamenta, A. T., & Taylor, J. C. (2023). Genome sequencing identifies KMT2E-disrupting cryptic structural variant in a female with O'Donnell-Luria-Rodan syndrome. Clinical Genetics.

O'Donnell-Luria AH, Pais LS, Faundes V, Wood JC, Sveden A, Luria V, Abou Jamra R, Accogli A, Amburgey K, Anderlid BM, Azzarello-Burri S, Basinger AA, Bianchini C, Bird LM, Buchert R, Carre W, Ceulemans S, Charles P, Cox H, Culliton L, Currò A; Deciphering Developmental Disorders (DDD) Study; Demurger F, Dowling JJ, Duban-Bedu B, Dubourg C, Eiset SE, Escobar LF, Ferrarini A, Haack TB, **Hashim M**., ... & Islam, L. (2019). Heterozygous variants in KMT2E cause a spectrum of neurodevelopmental disorders and epilepsy. The American Journal of Human Genetics, 104(6), 1210-1222.

إهـداء

الي من كانت الداعم الاول لتحقيق طموحي.. ويدي اليمني في هذه الرحلة...

الى من أبصرت بها الطريق... حياتي واعتزازي...

الي القلب الحنون.. الي من كانت دعواتها تحيطني...

إليك أمي الحبيبة ... لتقر عينك...

Abstract

Background: Whole genome sequencing (WGS) has significantly enhanced the diagnostic yield for rare genetic disorders, providing molecular diagnoses and opening avenues for therapeutic interventions, cascade testing, and reproductive planning. However, the diagnostic yield remains at 25-40%, presenting challenges in deciphering the underlying causes of incompletely characterised conditions. This limitation primarily arises from gaps in knowledge about causative genes and the restricted use of gene panels in genome sequencing analysis. Panels, by design, may overlook diagnoses beyond their predefined scope.

Objective: The primary objective of this work was to identify novel disease-causing genes to enable new diagnoses and to characterise and validate the detection of complex and elusive structural variants responsible for neurodevelopmental disorders. Given the clinical and genetic diversity among affected individuals, a range of analytical approaches was employed, extending beyond single nucleotide variants to include structural variants. Leveraging large cohorts and gene-based approaches expedited the discovery of novel genes and prioritised those requiring functional validation.

Methods: The study systematically analysed WGS data from patients with undiagnosed neurodevelopmental disorders and epilepsy phenotypes, recruited through the Health Innovation Challenge Fund 2 (HICF2) project and the 100,000 Genomes project (100KGP). These patients included a cohort with known karyotypic abnormalities suspected as the main cause of the phenotype. Furthermore, the effect of variants in a potential novel disease gene (*PLPPR5*) were investigated functionally.

Results:

In patients where no karyotypic abnormalities were detected, the study led to the identification of two novel disease genes: *KMT2E* and *PLPPR5*. Additionally, strong genetic evidence from patient variants and the presence of KO mouse models exhibiting similar phenotypes propose *MGRN1* as a further potential disease-causing gene, pending functional validation.

KMT2E was validated through the identification of additional patients with similar phenotypes through international collaborations. Following the initial report that described *KMT2E* as a

neurodevelopmental disorder-causing gene, a new syndrome called O'Donnell-Luria-Rodan-Rodan syndrome (ODLURO) has been recognised, and the gene has received approval for panel testing. Notably, this recognition has already facilitated the diagnosis of over 70 new patients, including eight from the 100KGP.

On the other hand, functional validation of compound heterozygous variants in *PLPPR5* revealed an upregulated expression of key autophagy genes, an increase in lysosomal number, and significant mitochondrial morphological and functional impairment. Additionally, there was a significant reduction in the mutant's capacity to generate filopodia-like extensions, possibly leading to phenotypes observed in patients with Jeavons syndrome. These findings expand our knowledge of the genetic basis of these disorders.

In patients with karyotypic abnormalities, short-read WGS data analysis demonstrated an efficacy of ~91% in identifying the breakpoints, leading to the reassignment of breakpoints to neighbouring cytogenetic bands in ~88% of cases. Moreover, the base resolution of WGS revealed complexities cryptic to karyotyping in 7 cases of the apparently balanced translocations. At least 30% of the apparently balanced translocations (ABTs) resulted in gene disruption that is likely to cause the phenotype, 15% disrupted topologically associated domains (TADs) for known disease genes, and 5% were associated with pathogenic genomic imbalances.

Amongst the disrupted a novel gene was *PPP2R2A*, which encodes for the regulatory subunit of Protein phosphatase 2 complex enzyme. It is part of the Phospho Protein Phosphatase 2 (PPP2) gene family of which six members are known to cause PP2A-related neurodevelopmental disorders. The balanced translocation disrupting *PPP2R2A* co-segregates with the phenotype in a large multi-generational kindred, and other variants in the gene were identified through GeneMatcher in seven additional independent families with similar phenotypes. Furthermore, loss-of-function variants in *ECPAS* are proposed for functional validation based on strong genetic evidence and the presence of KO mouse models exhibiting similar phenotypes.

Conclusion: The utility of novel analytical in-house pipelines, such as Variant Explorer and SVRare, in this thesis work facilitated the identification of complex structural rearrangements in known genes that were previously missed. These included copy-neutral structural variants and translocations, which are challenging to detect and characterise using conventional genetic-testing methods.

Moreover, using a combinatorial genomic approach led to the discovery of various novel disease-causing genes and potential associations to disease (*KMT2E, PPP2R2A, PLPPR5, MGRN1, ECPAS*) thus providing molecular diagnosis to patients, deeper understanding of the underlying mechanism of the observed phenotypes, and directions for future functional studies. These findings underscore the continued potential of WGS in advancing our understanding of the genetic underpinnings of neurodevelopmental disorders and epilepsy.

Acknowledgments

Completing this DPhil was not a solitary journey; it is a culmination of countless moments of support, encouragement, and guidance from individuals who accompanied me in this amazing experience, and I extend my heartfelt gratitude to every one of you who enriched my academic and personal growth during this remarkable journey.

First and foremost, I would like to express my deepest appreciation to my supervisor, Prof. Jenny Taylor. Your unwavering support has been a beacon of light throughout. You have been a nurturing presence, a facilitator of my academic pursuits, a support during times of distress and a constant source of encouragement. I am ever so grateful for the opportunity you provided me with to do this DPhil and for your being there every step of the way.

I owe a debt of gratitude to Prof. Andrew Wilkie for being an incredible supervisor. Your rigorous questioning of my scientific thinking has pushed me to expand my boundaries and explore the limits of my capabilities. Your guidance and direction have been instrumental in shaping my research.

Prof. Ira Milosevic, you introduced me to a fascinating world through the lens of microscopes. Your words of encouragement have consistently kept me motivated, even in the face of challenges. Your mentorship has been invaluable, and our discussions have been a source of inspiration and a newly found passion for cell biology. This would not have been complete except when sharing the excitement of looking at mesmerising live cells with James Bancroft and Edward Drydale so thank you for making that learning experience so much fun!

Dr. Alistair Pagnamenta, I am fortunate to have had you as a teacher and mentor. The excitement and passion you bring to work are contagious especially when it comes to solving SV puzzles! You have guided me through it all with unfaltering patience. You have been a constant pillar of support.

To Dr. Phalguni Rath, you have been an amazing friend. Your patience, unwavering dedication, chats, and discussions have been a source of inspiration, and your positivity has been a shining light.

Dr. Carme Camps, your presence has been a source of solace, whether in scientific or non-scientific conversations. Your support, empathy, and shared laughter have been invaluable.

At times during this journey, I had to do some statistical analyses (as one would expect from a DPhil) but that was when my brain did two things, go blank and think where is Dimitris?!! I cannot thank

you enough (Dr. Dimitris Vavoulis) for your patience in explaining statistical approaches and for tolerating my myriad of questions. Your ability to simplify complex concepts is a rare gift.

To all past and current members of the Taylor group, it has been a pleasure working alongside and learning from each one of you, solving countless crosswords over the years together, supporting each other, creating a stimulating and a positive environment. Also, how would this be complete without the continuous heavenly baked cakes and delicious cookies made by a smiling happy amazing Donna! No matter what I was feeling, I knew who to go to and I will have a smile by the end so thank you!

To my dear friends who embarked on this DPhil journey with me, your camaraderie has made the experience an unforgettable one. Aljawhara- gives you the motivation to work no matter what is going on keeping you focused on the prize, Hira- where do I start? Truly you are a treasure! you bring so much energy, joy, and happiness to all those around you, Faseeha- your friendship, kindness, and support can never be replaced, Tayyaba – you truly represent the meaning of your name "kindness", I am glad our paths have crossed. The list of dear friends extends a mile but together, we weathered the ups and downs of thesis writing, and I hope our friendship continues to thrive, filled with joy and achievements.

My deepest gratitude goes to my mother, whose love, support, and prayers have been a continuous source of strength. I am ever grateful for you being my mother.

To my siblings, especially Hiba, your support and care have meant the world to me. You were with me every step of the way; no words or thanks can convey the love I hold dear in my heart for you.

To my daughters, Mariam, Maha, Noor, and Fatima, and to all my nieces and nephews, you are my biggest fans and cheerleaders. Your love has fuelled my determination.

Thank you from the bottom of my heart.

Mona

Table of Contents

Declaration....2

Publications....3

Abstract....5

Acknowledgments....8

Table of Contents....10

Abbreviation list....15

Web Resources....25

List of Figures....26

List of Tables....28

1. Introduction....29

1.1 The burden of rare diseases....29

1.2 Neurodevelopmental disorders and the significance of receiving a molecular diagnosis....30

1.2.1 Global developmental delay (GDD)....31

1.2.2 Intellectual disability (ID)....32

1.2.3 Autism spectrum disorder (ASD)....33

1.3 Epilepsy....34

1.4 Types of genetic tests available....35

1.4.4 Genetic and Genomic Sequencing....39

1.4.5 From Sanger to NGS: The Evolution of Genomic Testing....39

1.5 Filtering, prioritisation, and classification in WGS Data: Methods and Approaches....48

1.6 Ethical Issues Associated with Whole Genome Sequencing....52

1.7 Thesis objectives....54

2. Materials & Method....56
2.1 Study recruitment and participation....56
2.2 Ethics and patients consent....57
2.3 Accessing whole-genome sequencing data for the 100KGP....57
2.4 Whole-genome sequencing for HICF2 and the 100KGP cases....58
2.4.1 DNA extraction and library preparation for the HICF2 samples....58
2.4.2 Whole-genome sequencing....58
2.4.3 Read alignment, mapping, and quality check for HICF2 samples....58
2.5 Genetic analysis....62
2.5.1 Copy number and zygosity analyses....62
2.5.2 Variant identification....63
2.5.3 Variant annotation....63
2.5.4 Variant prioritisation....64
2.6 Validation of identified causative variants....69
2.6.1 PCR and dideoxy sequencing....69
2.6.2 BioNano optical mapping....73
2.6.3 Minigene assay....74
2.7 Whole transcriptome sequencing....83
2.7.1 RNA processing and quantitation....83
2.7.2 Library preparation, sequencing, and data processing....84
2.8 Molecular mechanics....84
3. Combinatorial genomic approaches for novel gene discovery in neurodevelopmental disorders and epilepsy....86
3.1 Introduction to Neurodevelopmental Disorders....86
3.2 Neurodevelopmental Disorders: Findings from Whole Genome Sequencing....88
3.2.1 Aicardi syndrome....88

3.2.2 Complex structural variants leading to X-linked neurodevelopmental disorder 99

3.2.3 Bi-allelic variants in DOCK7 in a patient with Developmental and Epileptic Encephalopathy-23 (DEE23) 105

3.2.4 KMT2E is a novel gene associated with a neurodevelopmental disorder, now known as ODLURO syndrome. 113

3.3 Epilepsy: the example of Jeavons Syndrome 121

3.3.1 PLPPR5 as a novel disease-causing gene for epilepsy 123

3.3.2 A second case of Jeavons syndrome points towards another potential novel epilepsy gene: MGRN1 128

4. Functional validation of Jeavons Syndrome patient-specific mutations in PLPPR5 145

4.1 Introduction 145

4.2 Materials and Methods 148

4.2.1 Materials 148

4.2.2 Method 154

4.3 Results 175

4.3.1 PLPPR5 is trafficked to the plasma membrane via galT-positive vesicles in HEK293T and COS7 cells 175

4.3.2 The patient-specific mutations impeded the trafficking of PLPPR5 to the plasma membrane, and affected the formation of filopodia-like extensions induced by PLPPR5 177

4.3.3 PLPPR5 mutations p.L13F and p.R117C lead to an increased number LAMP1-positive vesicles 180

4.3.4 Impaired mitochondrial dynamics in cells overexpressing either PLPPR5 mutations (p.L13F or p.R117C) 183

4.3.5 Mitochondrial bioenergetic profile in cells overexpressing either of the PLPPR5 mutations (p.L13F or p.R117C) indicates a lowered mitochondrial fitness under baseline conditions when compared to the WT 186

4.4 Discussion 191

4.5 Limitations of the current work....196

5. Structural Variants: Rapidly increasing the diagnostic yield of Whole-Genome Sequencing ...198

5.1: Introduction....198

5.1.1 Challenges to SV detection.... 200

5.1.2 Structural variants in disease.... 202

5.2: Method....205

5.2.1 CNV/SV pipeline of the 100KGP.... 205

5.2.2 Structural variant analysis using SVRare.... 205

5.4 Results....213

5.4.1 Discovery of complex structural variants using a gene-oriented approach rapidly increases the diagnostic yield of WGS.... 213

5.4.2 Clinically relevant structural variants hiding in plain sight.... 215

5.4.3 Potential clinically relevant complex structural variants (cxSVs) in patients with neurodevelopmental disorder/ phenotype:.... 217

5.4.4 Systematic analysis of cases with karyotypic abnormalities including apparently balanced translocations from the 100KGP.... 234

5.4.5 PPP2R2A and ECPAS as potential novel disease genes directly disrupted by translocations in patients with neurodevelopmental disorders.... 239

5.4.6 Known neurodevelopmental disorder-associated genes disrupted by de novo translocations.... 245

5.4.7 Positional effect through the disruption of chromatin architecture and lnc-RNAs..... 251

5.4.8 Germline chromothripsis and chromoanasynthesis: Complex and Challenging Diagnosis 256

5.5 Discussion....262

6. Discussion and conclusion....265

6.1 Summary of results....265

6.2 Research impact....270

6.3 Limitations and barriers 273

6.4 Future work 274

References 276

Appendix 344

Abbreviation list

Abbreviation	Abbreviation Full Term
1000G	1000 Genomes Project
100KGP	100,000 Genomes Project
a-CGH	Array-Comparative Genomic Hybridisation
ABT	Apparently Balanced Translocation
ACMG	American College of Medical Genetics and Genomics
AD	Autosomal Dominant
ADHD	Attention Deficit Hyperactivity Disorder
AF	Allele Frequency
ANKRD11	Ankyrin Repeat Domain Containing 11
AP-1	Activator Protein-1
AR	Autosomal Recessive
ARHGEF3	Rho Guanine Nucleotide Exchange Factor 3
ARID1B	At-Rich Interaction Domain 1b
ARX	Aristaless Related Homeobox
AS	Aicardi Syndrome
ASD	Autism Spectrum Disorder
ATF7IP	Activating Transcription Factor 7 Interacting Protein
ATG	Autophagy-Related Proteins
ATXN1	Ataxin 1
AUH	AU RNA Binding Methylglutaconyl-CoA Hydratase
BCORL1	BCL6 Corepressor Like 1
BND	Breakends
BRF	Base Recall Fraction
BWA	Burrows-Wheeler Aligner
CACNA1E	Calcium Voltage-Gated Channel Subunit Alpha1 E

Abbreviation	Abbreviation Full Term
CADD	Combined Annotation-Dependent Depletion.
CAPN2	Calpain 2
CAPN8	Calpain 8
CASK	Calcium/Calmodulin Dependent Serine Protein Kinase
CCR	Complex Chromosomal Rearrangement
CDKL5	Cyclin Dependent Kinase Like 5
CHD1	Chromodomain Helicase DNA Binding Protein 1
CHD2	Chromodomain-Helicase-DNA-Binding Protein 2
CLN3	CLN3 Lysosomal/Endosomal Transmembrane Protein, Battenin
CMA	Chromosomal Microarray Analysis
CN	Copy Number
CNV	Copy Number Variant.
COQ10A	Coenzyme Q10A
COSMIC	Catalogue Of Somatic Mutations in Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTCF	Corrected Total Cell Fluorescence
CTR9	CTR9 Homolog, Paf1/RNA Polymerase II Complex Component
CUL4B	Cullin 4B
CUX1	Cut-Like Homeobox 1
cxSV	Complex Structural Variants
CYBC1	Cytochrome B-245 Chaperone 1
dbSNP	Single Nucleotide Polymorphism Database
DCC	DCC Netrin 1 Receptor
DDD	Deciphering Developmental Disorders Study
ddNTPs	Dideoxynucleotides
DEE 1/23	Developmental And Epileptic Encephalopathy 1/ 23
DEE1	Developmental And Epileptic Encephalopathy 1

Abbreviation	Abbreviation Full Term
DHR2	DOCK Homology Region 2
DHX30	DEXh-Box Helicase 30
DIRAS2	DIRAS Family GTPase 2
DLS	Direct Label and Stain
DN	De Novo
DNA	Deoxyribonucleic Acid
dNTPs	Deoxyribonucleotide Triphosphates
DOCK7	Dedicator Of Cytokinesis 7
DOCK8	Dedicator Of Cytokinesis 8
dsDNA	Double Stranded DNA
DTT	Dithiothreitol
DYRK1A	Dual Specificity Tyrosine Phosphorylation Regulated Kinase 1A
EBF3	Early B-Cell Factor 3
ECAR	Extracellular Acidification Rate
ECPAS	Ecm29 Proteasome Adaptor and Scaffold Protein
EEG	Electroencephalogram
EFTUD2	Elongation Factor Tu GTP Binding Domain Containing 2
ELFN2	Extracellular Leucine Rich Repeat and Fibronectin Type III Domain Containing 2
ERG	Electroretinographic
ES	Embryonic Stem
ESF1	ESF1 Nucleolar Pre-rRNA Processing Protein Homolog
ESP6500	Exome Sequencing Project V. 6500
EtBr	Ethidium Bromide
ETC	Electron Transport Chain
ExAC	Exome Aggregation Consortium
FBS	Foetal Bovine Serum

Abbreviation	Abbreviation Full Term
FCCP	Carbonyl Cyanide-P-Trifluoromethoxyphenylhydrazone
FCGR3A	Fc Gamma Receptor IIIa
FGF13	Fibroblast Growth Factor 13
FISH	Fluorescence In Situ Hybridisation
FLNA	Filamin A
FMR1	Fragile X Messenger Ribonucleoprotein 1
FoSTeS	Fork Stalling and Template Switching
FOXG1	Forkhead Box G1
FOXP1	Forkhead Box P1
FP	False-Positive
FXS	Fragile X Syndrome
GABRA2	Gamma-Aminobutyric Acid Type A Receptor Subunit Alpha2
GADD45G	Growth Arrest and DNA Damage-Inducible Gamma
GADO	Gene Network Assisted Diagnostic Optimization
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
GDD	Global Developmental Delay
gDNA	Genomic DNA
GeCIP	Genomics England Clinical Interpretation Partnership
GEFS+	Genetic Epilepsy with Febrile Seizures Plus
GEL	Genomics England Limited
GEMIN8	Gem Nuclear Organelle Associated Protein 8
GGE	Genetic Generalised Epilepsies
GJB2	Gap Junction Protein Beta 2
GM-MDT	Genomic Medicine Multi-Disciplinary Team
GMC	Genomic Medicine Centre
GMS	Genomic Medicine Service

Abbreviation	Abbreviation Full Term
gnomAD	Genome Aggregation Database
GOF	Gain-Of-Function
GRIN2B	Glutamate Ionotropic Receptor NMDA Type Subunit 2B
GTEx	Genotype-Tissue Expression Project
GUCY2C	Guanylate Cyclase 2C
GWAS	Genome-Wide Association Studies
H2O2	Hydrogen Peroxide
HEK293-T	Human Embryonic Kidney 293 Cells
HGMD	Human Gene Mutation Database
HGVS	Human Genome Variation Society
HI	Haploinsufficiency
Hi-C	High-Throughput Chromosome Conforma4on Capture
HICF2	Health Innovation Challenge Fund 2
HPO	Human Phenotype Ontology
IBD	Identical-By-Descent
ID	Intellectual Disability
IGE	Idiopathic Generalized Epilepsy
IGV	Integrative Genomics Viewer
INV	Inversion
iPSCs	Induced Pluripotent Stem Cells
IQ	Intelligence Quotient
IVA	Ingenuity Variant Analysis
JME	Juvenile Myoclonic Epilepsy
JS	Jeavons Syndrome
KANK1	KN Motif and Ankyrin Repeat Domains 1
KCNMA1	Potassium Calcium-Activated Channel Subfamily M Alpha 1

Abbreviation	Abbreviation Full Term
KCNQ2/3/5	Potassium Voltage-Gated Channel Subfamily Q Member 2/3/5
KCNQ5	Potassium Voltage-Gated Channel Subfamily Q Member 5
KCNT1	Potassium Sodium-Activated Channel Subfamily T Member 1
KCP	Kielin /Chordin-Like Protein
KIF1A	Kinesin Family Member 1A
KMO	Kynurenine 3-Monooxygenase
KMT2E	Lysine Methyltransferase 2E
KO	Knockout
LAMP1	Lysosome Associated Membrane Protein-1
LCFAs	Long Chain Fatty Acids
LCR	Low Copy Repeat
LINE-1/L1	Long-Interspersed Element 1
lncNR2F1	NR2F1 Antisense RNA 1
lncNR2F2	NR2F2 Antisense RNA 1
lncRNAs	Long Non-Coding RNAs
LOEUF	Loss-Of-Function Observed/Expected Upper-Bound Fraction
LoF	Loss Of Function
LPA	Lipoprotein(A)
MAGED4	Melanoma Antigen, Family D, 4
MBD	Methyl-CpG Binding Domain
MCF2L	Mcf.2 Cell Line Derived Transforming Sequence Like
MECP2	Methyl-CpG Binding Protein-2
MEF2C	Myocyte Enhancer Factor 2C
mEPSC	Miniature Excitatory Synaptic Current
MGMT	Methylguanine Methyltransferase
MGRN1	Mahogunin Ring Finger 1
MLPA	Multiplex Ligation-Dependent Probe Amplification

Abbreviation	Abbreviation Full Term
MMBIR	Microhomology-Mediated Breakage Induced Repair
mTOR	Mammalian Target of Rapamycin
MYO5A	Myosin VA
NAC	N-Acetylcysteine
NAHR	Non-Allelic Homologous Recombination
NALCN	Sodium Leak Channel Non-Selective Protein
NBEA	Neurobeachin
NCBI	National Centre For Biotechnology Information
NDDs	Neurodevelopmental Disorders
NGS	Next-Generation Sequencing
NHEJ	Non-Homologous End Joining
NHS	National Health Service
NR2F1	Nuclear Receptor Subfamily 2 Group F Member 1
NR2F2	Nuclear Receptor Subfamily 2 Group F Member 2
NRF-1	Nuclear Respiratory Factor 1
NRXN1	Neurexin 1
NUS1	NUS1 Dehydrodolichyl Diphosphate Synthase Subunit
OCEL1	Occludin/Ell Domain Containing 1
OCR	Oxygen Consumption Rate
ODLURO	O'Donnell-Luria-Rodan-Rodan Syndrome
OMIM	Online Mendelian Inheritance In Man
OXPHOS	Oxidative Phosphorylation
PAFAH1B1	Platelet Activating Factor Acetylhydrolase 1b Regulatory Subunit 1
PAK5	P21 (RAC1) Activated Kinase 5
PBS	Phosphate Buffered Saline
PCDH7	Protocadherin 7
PCR	Polymerase Chain Reaction

Abbreviation	Abbreviation Full Term
PDD-NOS	Pervasive Developmental Disorder Not Otherwise Specified
PDHA1	Pyruvate Dehydrogenase E1 Subunit Alpha 1
PE	Paired-End
PIKFYVE	Phosphoinositide Kinase, FYVE-Type Zinc Finger Containing
PLBD1	Phospholipase B Domain Containing 1
PLCB4	Phospholipase C Beta 4
pLI	Probability Of Being Loss of Function Intolerant
PLPPR5	Phospholipid Phosphatase Related 5
PolyPhen-2	Polymorphism Phenotyping V2 Algorithm
PPP1R15B	Protein Phosphatase 1 Regulatory Subunit 15B
PPP2	Phospho Protein Phosphatase 2
PPP2R2A	Protein Phosphatase 2 Regulatory Subunit B alpha
pREC	Probability Of Intolerance to Recessive Loss Of Function
PRGs	Plasticity-Related Genes
PTEN	Phosphatase And Tensin Homolog
pTriplo	Probability Of Triplosensitivity
PTVs	Protein Truncating Variants
RIN	RNA Integrity Number
RNA	Ribonucleic Acid
ROH	Regions Of Homozygosity
ROIs	Regions Of Interest
ROS	Reactive Oxygen Species
RPE	Retinal Pigment Epithelium
RPS14	Ribosomal Protein S14
RT-PCR	Reverse Transcription Polymerase Chain Reaction
RVIS	Residual Variation Intolerance Score
SBS	Sequencing By Synthesis

Abbreviation	Abbreviation Full Term
SCN1A	Sodium Voltage-Gated Channel Alpha Subunit 1
SCN1B	Sodium Voltage-Gated Channel Beta Subunit 1
SHANK2	SH3 And Multiple Ankyrin Repeat Domains 2
SIFT	Sorting Intolerant From Tolerant
SINE	Short-Interspersed Element
SLC17A5	Solute Carrier Family 17 Member 5
SLC2A1	Solute Carrier Family 2 Member 1
SLC9A6	Solute Carrier Family 9 Member A6
SNAP25	Synaptosome Associated Protein 25
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SNX30	Sorting Nexin Family Member 30
SOCS6	Suppressor Of Cytokine Signalling 6
SOD	Superoxide Dismutase
SOX3	SRY-Box 3
SV	Structural Variant
SYK	Spleen Associated Tyrosine Kinase
SYN1	Synapsin I
T2T	Telomere-To-Telomere
TAD	Topologically Associated Domains
TBCD	Tubulin Folding Cofactor D
TBCK	Tbc1 Domain Containing Kinase
TDRD15	Tudor Domain Containing 15
TEAD1	Tea Domain Transcription Factor 1
TENM1	Teneurin Transmembrane Protein 1
TF	Transcription Factor
TFEB	Transcription Factor EB

Abbreviation	Abbreviation Full Term
TIRF	Total Internal Reflection Fluorescence
TMRM	Tetramethyl-Rhodamine Methyl Ester
TP53BP2	Tumor Protein P53-Binding Protein 2
TRPA1	Transient Receptor Potential Cation Channel Subfamily A Member 1
TS	Triplosensitivity
TSC1 / 2	TSC Complex Subunit 1 / 2
UHMW	Ultra-High Molecular Weight
UMI-4C	Unique Molecular Iden4fier - Circular Chromosome Conforma4on Capture
UPLC	Ultra-Performance Liquid Chromatography
UROS	Uroporphyrinogen III Synthase
UTR	Untranslated Region
VCF	Variant Call Format
VEP	Variant Effect Predictor
VLCFAs	Very Long Chain Fatty Acid
VUS	Variant Of Uncertain Significance
WDR45B	WD Repeat Domain 45B
WDR45B	WD Repeat Domain 45B
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WGS500	Whole Genome Sequencing 500 Study
WT	Wild Type
ZNF750	Zinc Finger Protein 750
ZNF750	Zinc Finger Protein 750
ZNF778	Zinc Finger Protein 778

Web Resources

CADD, https://cadd.gs.washington.edu/

ClinVar, https://www.ncbi.nlm.nih.gov/clinvar

DECIPHER, https://decipher.sanger.ac.uk/

GenBank, https://www.ncbi.nlm.nih.gov/genbank/

GeneMatcher, https://genematcher.org/

Genome Aggregation Database (gnomAD), https://gnomad.broadinstitute.org

GTEx Portal, https://gtexportal.org/home/

Michela**ngl**o-VENUS, https://michelanglo.sgc.ox.ac.uk/venus

Mutalyzer, https://mutalyzer.nl/normalizer

MutationTaster, https://www.mutationtaster.org/

NMDEscPredictor, https://nmdprediction.shinyapps.io/nmdescpredictor/

Online Mendelian Inheritance in Man (OMIM), https://omim.org/index

SpliceAI-lookup, https://spliceailookup.broadinstitute.org/

UCSC Human Genome Browser, http://www.genome.ucsc.edu

VuTR: Visualising the consequences of 5' UTR variants, https://vutr.rarediseasegenomics.org/

List of Figures

FIGURE 1.1: THE TIERING SYSTEM DESIGNED BY GENOMICS ENGLAND 51

FIGURE 2.1: VARIANT EXPLORER PIPELINE……...61

FIGURE 2.2: DYNAMIC VARIANT FILTERING AND PRIORITISATION PROCESSES 66

FIGURE 2.3: A SCHEMATIC OF THE DE NOVO ARX DELETION.....72

FIGURE 2.4: FIGURE 2.4: DIAGRAM OF RHCGLO MINIGENE REPORTER SYSTEM. 75

FIGURE 2.5: : CLONING STRATEGY TO GENERATE RHCGLO-MGRN1-EX14-15 PLASMID. 78

FIGURE 2.6: SUCCESSFUL CLONING OF MGRN1 EXONS 14-15 INTO RHCGLO. 80

FIGURE 2.7: VALIDATION OF SUCCESSFUL SDM. 82

FIGURE 3.1: OVERVIEW OF THE AS COHORT…….94

FIGURE 3.2: DETECTION AND VALIDATION OF A DE NOVO DELETION IN ARX 98

FIGURE 3.3: PEDIGREE OF PATIENTS WITH X-LINKED NEURODEVELOPMENTAL DISORDER. 100

FIGURE 3.4: X CHROMOSOME COMPLEX STRUCTURAL VARIANT.....102

FIGURE 3.5: HI-C PROFILE AND CONTACT DOMAINS AT THE XQ27.1 LOCUS. 103

FIGURE 3.6: BI-ALLELIC VARIANTS IN DOCK7 IN A PATIENT WITH DEE23.....112

FIGURE 3.7: VALIDATION OF A DE NOVO 2 BP DELETION IN KMT2E. 114

FIGURE 3.8: DISTRIBUTION OF IDENTIFIED & REPORTED MISSENSE VARIANTS IN KMT2E. .119

FIGURE 3.9: SEGREGATION OF A COMPOUND HETEROZYGOUS VARIANTS IN PLPPR5. 126

FIGURE 3.10: VALIDATION OF A RARE HOMOZYGOUS C.1482+1 G>A VARIANT IN MGRN1.....134

FIGURE 3.11: VALIDATION OF C.1482+1 G>A IN MGRN1 USING MINIGENE SYSTEM 137

FIGURE 3.12: HOMOZYGOUS C.1482+1G>A VARIANT IN MGRN1 138

FIGURE 3.13: RNA-SEQ DATA FROM PROBAND, MOTHER, AND CONTROL SAMPLES. 141

FIGURE 4.1: PCDNA3.1 VECTOR WITH PLPPR5-3XFLAG. 154

FIGURE 4.2: VERIFICATION OF SUCCESSFUL SDM TO INTRODUCE PLPPR5 VARIANTS.....156

FIGURE 4.3: STRATEGY TO INTRODUCE FLUORESCENT TAGS INTO PCDNA3.1-PLPPR5.....159

FIGURE 4.4: CLONING VERIFICATION STRATEGY USING RESTRICTION DIGESTION. 161

FIGURE 4.5: GATING STRATEGY TO DETERMINE TRUE TMRM SIGNAL.....169

FIGURE 4.6: ILLUSTRATION OF THE MITO-STRESS TEST ASSAY AND OCR PROFILE. 171

FIGURE 4.7: CELL NUMBER OPTIMISATION FOR SEAHORSE ASSAY. 172

FIGURE 4.8: CHARACTERISATION OF PLPPR5 WT IN COS7 CELLS.....176
FIGURE 4.9: EFFECT OF JS PATIENT-SPECIFIC MUTATIONS ON PLPPR5 LOCALISATION 178
FIGURE 4.10: TOTAL INTERNAL REFLECTION FLUORESCENCE (TIRF) ANALYSIS. 179
FIGURE 4.11: LYSOSOMAL CHARACTERISATION IN COS7 CELLS.....181
FIGURE 4.12: LYSOSOMAL AND AUTOPHAGOSOMAL MARKERS ANALYSIS 182
FIGURE 4.13: MITOCHONDRIAL ANALYSIS IN COS7 CELLS.....185
FIGURE 4.14: OXYGEN CONSUMPTION RATE (OCR) PROFILE IN HEK293T CELLS.....190
FIGURE 5.1: CONVENTIONAL STRUCTURAL VARIANT TYPES……199
FIGURE 5.2: IDENTIFICATION OF RARE STRUCTURAL VARIANTS WITH SVRARE PIPELINE. .212
FIGURE 5.3: OVERALL SUMMARY OF SV FINDINGS.....214
FIGURE 5.4: A MISSED DE NOVO SMALL DELETION IN ARID1B.....216
FIGURE 5.5: REPRESENTATION OF COMPLEX STRUCTURAL VARIANT CLASSES.....219
FIGURE 5.6: CHALLENGES TO VALIDATION OF CXSV IN MECP2 227
FIGURE 5.7: PACBIO VALIDATION OF CXSV IN MECP2.....228
FIGURE 5.8: CONFIRMATION OF THE FIRST KMT2E CRYPTIC STRUCTURAL VARIANT 233
FIGURE 5.9: CLINICAL PHENOTYPES AS REPORTED IN THE 100KGP DATABASE.....236
FIGURE 5.10: OVERVIEW SUMMARY RESULTS KARYOTYPIC ABNORMALITY COHORT 238
FIGURE 5.11: A SEGREGATING BALANCED TRANSLOCATION DISRUPTING PPP2R2A.....242
FIGURE 5.12: COMPLEX REARRANGEMENT IN APPARENTLY BALANCED TRANSLOCATION244
FIGURE 5.13: WGS DATA REVEALS CRYPTIC COMPLEXITIES IN CUX1 250
FIGURE 5.14: CHARACTERISATION OF A LARGE INVERSION DOWNSTREAM OF FOXG1.....255
FIGURE 5.15: LINEAR SCHEMATIC OF POSSIBLE CHROMOANASYNTHESIS EVENT.....258
FIGURE 5.16: LINEAR SCHEMATIC OF THE REARRANGEMENTS IN CHROMOTHRIPSIS 261

List of Tables

TABLE 2.1: A LIST OF HPO TERMS PROVIDED BY THE RECRUITING CLINICIANS 67

TABLE 2.2: VALIDATION PRIMER LIST. 71

TABLE 2.3:PRIMERS USED FOR CLONING MGRN1 EXONS 14-15 INTO RHCGLO. 76

TABLE 3.1: CLINICAL FEATURES AND EXAMINATION OF PROBANDS IN AS COHORT……...94

TABLE 3.2: CANDIDATE VARIANTS CONSIDERED FOR FURTHER ANALYSIS IN DEE23 CASE.108

TABLE 3.3: VERIFIED CANDIDATE GENES & VARIANTS IN THE PROBAND SEC 3.2.3.....114

TABLE 3.4: DIAGNOSTIC LOF VARIANTS IDENTIFIED IN KMT2E 118

TABLE 3.5: LIST OF RARE DELETERIOUS MISSENSE VARIANTS IN THE 100KGP.....120

TABLE 3.6: PRIORITISED INHERITED HETEROZYGOUS VARIANTS IN AFFECTED SIBLINGS..124

TABLE 3.7: A LIST OF VARIANTS IDENTIFIED IN FIVE FAMILIES WITH JS 129

TABLE 3.8: IN SILICO PANELS APPLIED TO THE WGS DATA OF FAMILIES 1-4 130

TABLE 3.9: LIST OF HPO TERMS REPORTED FOR THE PROBAND IN FAMILY 5 131

TABLE 3.10: LIST OF VARIANTS IN ROH IN THE PROBAND.....133

TABLE 4.1: LIST OF CHEMICALS AND THEIR SUPPLIERS……...148

TABLE 4.2: LIST OF KITS AND CONSUMABLE PRODUCTS 149

TABLE 4.3: CELL CULTURE CONSUMABLES.....150

TABLE 4.4: SOLUTIONS 151

TABLE 4.5: PRIMARY AND SECONDARY ANTIBODIES USED FOR WESTERN BLOT 152

TABLE 4.6: PLASMIDS.....152

TABLE 4.7: LIST OF EQUIPMENT 153

TABLE 4.8: LIST OF SOFTWARE PROGRAM 153

TABLE 4.9: CLONING PRIMER SEQUENCES 157

TABLE 4.10: FUNCTIONS OF DRUGS ON INNER MITOCHONDRIAL MEMBRANE.....172

TABLE 4.11: OCR CALCULATIONS. 173

TABLE 4.12: Z-SCORE TABLE OF SIGNIFICANTLY DIFFERENT BIOCHEMICALS.....189

TABLE 5.1: EXPERT CURATED AND APPROVED AUTOSOMAL DOMINANT GENES……207

TABLE 5.2: FAMILY STRUCTURE AND PARTICIPANT COUNTS WITHIN FAMILIES IN GEL 210

TABLE 5.3: KMT2E SVS IDENTIFIED IN P5. 231

TABLE 5.4: LIST OF REGIONS WITH LOSS OR GAIN AS DETECTED BY WGS IN CASE 6.....253

Chapter 1. Introduction

1.1 The burden of rare diseases

Rare diseases individually affect fewer than 5 in every 10,000 people. However, about 6,000- 8,000 different conditions are recognised, amounting to approximately 3.5 million affected people in the UK alone, and about 30 million in Europe. Despite the scale of the issue, a significant proportion of these patients do not have access to the proper care and treatment (APPG, 2016).

Rare diseases in general, and neurodevelopmental disorders in particular, pose several challenges for patients attempting to access the appropriate medical care. Some of these challenges are: i) the lack of understanding of each particular condition due to scarcity of patients living with any single condition, ii) unidentified causative genes, iii) incorrect diagnosis due to the heterogeneity of phenotypes, iv) limitations of genetic testing modalities available in clinics and v) lack of targeted treatments and therapies. Therefore, finding a genetic diagnosis will substantially impact the different groups involved in the diagnostic odyssey ranging from the patients and their families to clinicians, researchers, and the healthcare system.

From the clinical perspective, finding the underlying genetic cause of the observed phenotypes would help provide a more accurate genetic diagnosis. This could lead to changes in the patient care pathway, such as initiating changes to treatment and management, prompting participation in screening programmes or advice on access to alternative educational and social provisions previously unavailable to them. Moreover, the clinician would be able to provide the family with informed counselling on the general risk of having more children and the reproductive options available based on the genetic findings (Splinter et al., 2018). From the research aspect, delineating the genetic factors contributing to rare diseases helps identify further patients with similar conditions (O'Donnell-Luria et al., 2019). This can lead to a better insight into the primary biological mechanism of the disease as well as create an opportunity for developing new therapeutic targets (Zweier et al., 2019). Finally, rare diseases are usually accompanied by a financial burden on the

family and the healthcare system. Having a genetic diagnosis was shown to reduce the cost to 6-7% of the original estimate (Splinter et al., 2018).

This thesis is focused on neurodevelopmental disorders, which range from readily identifiable conditions to rare, more heterogeneous, hard-to-diagnose diseases. The main reason for using this group of diseases to explore the utility of WGS entirely is the range of challenges it presents to diagnosis as well as being the largest group of conditions recruited to the 100KGP, reflecting the increased demand for molecular diagnosis for these disorders.

1.2 Neurodevelopmental disorders and the significance of receiving a molecular diagnosis

Neurodevelopmental disorders (NDDs) are a diverse group of conditions that stem from abnormal brain development. They can impair cognitive, communication, adaptive, and psychomotor abilities (Moreno-De-Luca et al., 2013). Examples of NDDs include autism spectrum disorder, social communication disorders, intellectual disability, developmental delay, and attention deficit hyperactivity disorder (ADHD) (Moreno-De-Luca et al., 2013). These disorders pose a significant societal health challenge, affecting over 3% of children worldwide (Gilissen et al., 2014). NDDs may result from prenatal, perinatal, or postnatal injuries but an estimated 20% to 30% of cases are considered to be disorders with a genetic origin (Michelson & Clark, 2020). Nevertheless, diagnosing NDDs can be difficult due to their significant clinical and genetic heterogeneity (Stevenson et al., 2003). Large-scale studies have shown a high degree of phenotypic heterogeneity and overlap between NDDs, making an accurate diagnosis based on clinical presentation challenging (Deng & Pan, 2018; Posey et al., 2019). Currently, no universal biomarkers are available to diagnose NDDs, and it is mainly based on clinical presentation. However, many symptoms are not unique to single NDDs, and several may present with similar symptoms, leading to overlapping phenotypes (Deng & Pan, 2018; Posey et al., 2019). Different variants in the same gene can also result in different disease phenotypes, further complicating diagnoses (Deng & Pan, 2018). Traditional genetic investigations and thorough consultations can be costly and may not result in a definitive diagnosis when no recognisable syndrome is suspected.

Understanding the specific cause of NDDs confirming or establishing a new clinical diagnosis provides numerous benefits to clinicians, patients, and their families. These include improved early intervention through referral to therapy programmes, support groups, advocacy organisations, and research projects. Early identification of NDDs through monitoring child development enables the initiation of supportive therapies at an early stage when they are most effective in improving outcomes. In addition, by providing specific treatments based on the underlying pathophysiology of the condition, a better understanding of co-morbidities such as vision loss, hearing loss, and epilepsy can be achieved, leading to improved predictions and outcomes. Finally, it helps provide informed genetic counselling to the family regarding recurrence risk and fertility, life expectancy, and long-term prognosis (Michelson & Clark, 2020).

The following is a short overview of the main NDDs in the cases discussed in this thesis such as global developmental delay, intellectual disability, and autism spectrum disorder.

1.2.1 Global developmental delay (GDD)

Global developmental delay (GDD) pertains to children below the age of five who display significant delays of at least two standard deviations below the mean in two or more areas of development. These areas may include gross and fine motor skills, communication and language, social skills, cognitive abilities, and activities of daily living (Moeschler, 2008).

GDD can be associated with physical feature, such as dysmorphology or internal organ malformations, which could suggest the presence of a syndromic genetic disorder in about 15% of cases. However, most patients with GDD/ID do not present any, or only minor signs, of syndromic disorder (Moeschler et al., 2014). In such cases, the intellectual disability is referred to as non-syndromic. Nevertheless, clinical signs in syndromic forms may not always be obvious enough to facilitate a straightforward diagnosis (Bélanger & Caron, 2018). Usually, the presence of syndromic form indicates a larger genetic defect than in isolated developmental disabilities making the genetic aetiology of GDD/ID highly diverse and complex (Bélanger & Caron, 2018).

1.2.2 Intellectual disability (ID)

Intellectual disability (ID) is a condition characterised by significant limitations in intellectual functioning and adaptive behaviour, which affects approximately 1-3% of the global population (Karam et al., 2015).

ID can be classified into one of four groups: mild, moderate, severe, and profound. While mild or moderate ID may only manifest as academic struggles in school-age children, those with moderate to severe ID typically experience additional medical issues and are identified within the first two years of their life (Boat and Wu, 2015).

Individuals with profound intellectual disability account for 1.5% of all cases of ID and have Intelligence Quotient (IQ) scores below 20. They display severely limited capacity to acquire self-care skills and require continuous support and assistance with day-to-day activities (Leonard & Wen, 2002).

Studies have shown that genetics plays a significant role in the aetiology of ID, with a high level of genetic heterogeneity (Bélanger & Caron, 2018). ID can be caused by chromosomal abnormalities, single gene mutations, and copy number variations (CNVs) (Vissers et al., 2016). Chromosomal abnormalities, such as Down syndrome (trisomy 21) and Turner Syndrome (45,X), are well-known genetic causes of ID. Single gene mutations associated with ID have been identified in genes such as *MECP2, FMR1*, and *ARX* (Bélanger & Caron, 2018). CNVs, which are deletions or duplications of DNA segments larger than 1 kilobase, have also been associated with ID (Vissers et al., 2016). Studies have shown that CNVs account for approximately 15% of cases of unexplained ID (Mefford et al., 2012). Additionally, research has shown that *de novo* CNVs, which arise spontaneously and are not inherited from parents, are more common in individuals with ID compared to the general population (Mefford et al., 2012).

The genetic causes of ID are complex and can involve interactions between multiple genes and environmental factors. In some cases, the genetic cause of ID may not be identified. However, advances in genetic testing technology have allowed for more accurate diagnosis and identification of the genetic basis of ID, which can aid in developing personalised treatments and support for affected individuals and their families (Vissers et al., 2016).

1.2.3 Autism spectrum disorder (ASD)

Autism spectrum disorder (ASD) is a neurodevelopmental disorder that affects communication, social interaction, and behaviour. ASD has a complex intricate nature characterised by a spectrum of symptoms and levels of severity. It is commonly referred to as a "spectrum" disorder because of the broad range of symptoms, abilities, and degrees of impairment that individuals with ASD can display.

ASD is typically diagnosed in early childhood, although some individuals may not receive a diagnosis until later in life. It has been estimated to have a high occurrence rate of about 1 in 36 children (Sharma et al., 2018) with a 3-fold higher prevalence in males than in females (Loomes et al., 2017). The symptoms of ASD can vary widely from person to person, but some common symptoms include difficulty with social interactions, communication, and behaviour.

There are several different subtypes of ASD, including Asperger syndrome, childhood disintegrative disorder, and pervasive developmental disorder not otherwise specified (PDD-NOS). These subtypes were previously used to describe different types of autism, but they are no longer used in the most recent diagnostic manual (DSM-5) and have been replaced by the umbrella term “autism spectrum disorder” (Volkmar & McPartland, 2014).

Although the precise aetiology of ASD remains uncertain, studies indicate that a combination of genetic factors, including single nucleotide variants (SNVs) and large genomic abnormalities (Lee et al., 2015), metabolic conditions such as phenylketonuria (Baieli et al., 2003), as well as environmental factors - whether prenatal, such as congenital rubella (Chess, 1977) or postnatal - may play a role. Furthermore, all of these factors have been shown to be associated with epilepsy (Betancur, 2011). Due to being less commonly affected by autism, women may have a higher threshold for receiving a diagnosis and are more likely to have an identifiable genetic cause for their condition.

In general, individuals with ASDs have a higher likelihood of experiencing other neuropsychological disabilities, such as receptive and expressive language impairments, specific learning difficulties, and ID, as well as other neurological comorbidities, including psychiatric disorders and epilepsy, compared to the general population (Lukmanji et al., 2019).

1.3 Epilepsy

Epilepsy is a neurological disorder that is characterized by an enduring predisposition to seizures, with associated cognitive, psychological and social consequences (Devinsky et al., 2018a). It occurs as the main manifestation or as part of a syndrome.

There is a growing list of epilepsies that are considered of unknown origin but are thought to have a genetic component. The lack of understanding of the underlying pathogenic cause of epilepsy or syndromes with epilepsy hinders the treatment and management of the patient's condition.

One of the most common genetic causes of epilepsy is mutations in ion channel genes, which affect the electrical activity of neurons in the brain. For example, mutations in the *SCN1A* gene, which encodes the alpha subunit of the voltage-gated sodium channel Nav1.1, have been linked to several forms of epilepsy, including Dravet syndrome, a severe childhood epilepsy with multiple seizure types (Bartolini, 2021). Other ion channel genes, such as *KCNQ2* and *KCNQ3*, which encode subunits of the potassium channel Kv7.2 and Kv7.3, respectively, regulate the frequency and duration of action potential have also been associated with epilepsy (Kuang et al., 2015).

In addition to ion channel genes, other genetic factors have been implicated in epilepsy. For example, CNVs have been long established to be linked to several forms of epilepsy, including West syndrome and idiopathic generalized epilepsy (IGE) (Striano et al., 2011). Furthermore, rare genetic variants in GABAA receptor subunit genes, which encode proteins that mediate inhibitory neurotransmission in the brain, have been associated with epilepsy and related disorders (Macdonald & Kang, 2009).

Advances in genomic technologies have facilitated the identification of additional genetic variants associated with epilepsy. Genome-wide association studies (GWAS) have identified common variants that increase the risk of epilepsy, including variants in the *PCDH7* gene, which encodes a protein involved in cell-cell adhesion (ILAECCE, 2014). Exome sequencing, which examines the protein-coding regions of the genome, has identified rare genetic variants in numerous genes that increase the risk of epilepsy, including *CACNA1E*, which encodes a subunit of the voltage-gated calcium channel Cav2.3 (Epi4K and EPGP, 2013).

Genetic factors are crucial in the development of epilepsy, and identifying genetic variants associated with the disorder has advanced our comprehension of its underlying mechanisms. With

the increasing use of genetic testing, it is expected that identifying genetic variants will contribute to more accurate clinical diagnosis of epilepsy, enabling the detection of associated health conditions, their monitoring, and sometimes impacting treatment decisions. One example in which the genetic finding impacts the treatment is *KCNT1*. Pathogenic variants specifically gain-of-function (GOF) ones in *KCNT1* lead to multiple epilepsy disorders that are refractory to anti-epileptic medications. Patients with GOF variants were found to benefit from quinidine administration, an anti-arrhythmic drug, reducing the seizure burden (Barcia et al., 2012; Bearden et al., 2014; Milligan et al., 2014; R. Liu et al., 2023) Therefore, different types of testing should be considered at the onset of clinical diagnosis to achieve an early accurate diagnosis.

1.4 Types of genetic tests available

Until recently, clinical genetic testing to identify genetic causes of various disorders involved procedures that interrogate DNA, RNA, and chromosomes using cytogenetic, molecular techniques or a combination. DNA-based methods, such as Sanger and targeted gene panel sequencing, are laborious and time-consuming. Although they provide high-quality molecular diagnoses where genes have already been identified, the continuous discovery of novel genes associated with the conditions limits the panel's sensitivity (Katsanis & Katsanis, 2013). This is due to concurrently targeting one or a few loci/genes. Whilst oligonucleotide arrays enable the simultaneous examination of the patient's entire genome, their resolution is limited to detecting medium to large genomic changes between 50-200 kb (McCarroll et al., 2006). The interrogation of the whole genome at a single nucleotide resolution has only become possible with the development of faster and cheaper whole-exome (WES) and whole-genome (WGS) sequencing (Pasche & Absher, 2011).

The current strategy for genetic testing in clinical laboratories varies significantly depending on the test types used for variant identification/analysis. The tests offered range in scope from single gene approach to genome-wide tests. Moreover, they also range in their resolution from SNVs to CNVs that affect millions of bases encompassing hundreds of individual genes.

WGS can replace all other methods, including chromosomal analyses, to detect CNVs and structural variants (SVs) previously detected using cytogenetic techniques. This gives WGS the potential to become a first-tier diagnostic test (Lindstrand et al., 2019). Nonetheless, it is essential to gain a clear understanding of the specific capabilities of each test in detecting various types of variants before we can fully grasp the potential of WGS.

1.4.1 Chromosomal Analysis: Low-Resolution G-banded karyotype Approach

A G-banded karyotype is a visual representation of all 46 chromosomes captured during cell division in metaphase. It is used to examine the whole genome for large structural and numerical abnormalities such as deletions, duplications, and translocations of 3-5 Mb size or greater (Mandahl , 1992). The Giemsa-stained chromosomes are captured in an image after undergoing trypsin digestion, resulting in this characteristic representation.

The light and dark bands that appear on the chromosomes serve to differentiate between different regions within them. The euchromatic regions, which are transcriptionally active, have a less condensed configuration with proteins that are more accessible to trypsin. On the other hand, heterochromatic regions, which are less active, have a more condensed configuration and tend to bind more strongly to the Giemsa stain (Flore & Milunsky, 2012). Chromosome analysis is typically performed using blood samples, in which T-cell lymphocytes are stimulated to divide *ex vivo* using phytohemagglutinin. If the analysis fails, it may be due to insufficient or abnormal T-cells, as seen in cases of 22q11.2 deletion syndrome. In such cases, alternative tissues, such as fibroblast culture from skin biopsies, can be used for chromosome analysis.

Despite providing a resolution of 750-800 bands, karyotyping is not uniform in its distribution, making the detection of changes in some regions challenging. This leads to missed detection of copy number changes, even those as large as 10Mb. Additionally, the subjective visual inspection involved in karyotype analysis is prone to analyst variability and inconsistencies in abnormality detection.

In cases of developmental disability of unknown cause, chromosome analysis has a yield of only about 3% (Shevell et al., 2003). The American College of Medical Genetics and Genomics recommends its use only when a large chromosomal rearrangement is suspected or when family

history suggests a parent may carry a balanced translocation indicated by several miscarriages or infertility (Miller et al., 2010; Schaaf et al., 2011). However, with fewer karyotypes being performed in clinical cytogenetics laboratories post-2010, expertise in interpretation has declined over time (Hochstenbach et al., 2017).

1.4.2 High-Resolution chromosomal microarray analysis (CMA)

Karyotyping was replaced by CMA, as it has a higher resolution with a minimum of ~100 kb genome-wide (Wain & Thorland, 2011). Thus, CMA provides enhanced detection of disease-causing microdeletions and microduplications and the ability to reveal cryptic genomic imbalances related to apparently balanced translocation. Additionally, it offers more accurate characterisation of unbalanced chromosome rearrangements detected through G-banded karyotyping (McGillivray et al., 2012). The major CMA types include comparative genomic hybridisation arrays (a-CGH) and Single Nucleotide Polymorphism (SNP) arrays. Both types are widely used as first-line genetic testing as they allow for the detection of disease-causing CNVs, which were found to be the cause of 5%-10% cases of autism and between 20-25% in patients with more severe neurodevelopmental syndromes (Beaudet, 2013). In addition to detecting copy number changes, SNP arrays can also identify copy-neutral variations, such as regions of homozygosity (ROH), which is an indication of uniparental disomy or parental consanguinity.

Furthermore, SNP arrays are more sensitive to low-level mosaicism, down to 5%, than a-CGH, with the lowest level reported at 7% (Menten et al., 2006). However, it is essential to note that probes in SNP arrays are not uniformly distributed across the genome, which can result in missing copy number changes in regions with a lower number of probes. Furthermore, CMA fails to detect copy-neutral structural variants' orientation and size, such as balanced translocations, rings, and inversions. Detection of CNVs using arrays is challenged in regions with low copy repeat (LCR) sequences, sequences with high levels of repetition, and highly homologous sequences such as pseudogenes. Therefore, clinical genetics laboratories often validate findings from arrays using a different technique, such as fluorescence in situ hybridisation (FISH) or multiplex ligation-dependent probe amplification (MLPA) analysis of the suspected region.

1.4.3 Fluorescence in situ hybridisation (FISH)

FISH is a highly sensitive test using fluorescently labelled probes on cells during metaphase (dividing cells) or interphase (fixed cells). FISH is valuable when a particular deletion is suspected, parental testing is needed following a child's diagnosis with a deletion, or when the presence of mosaicism is being evaluated. However, it is less effective for detecting duplications because two positive signals can appear as one when they are close together, which is why interphase preparations are often preferred (Jensen, 2014). FISH is a cost-effective and relatively fast option when a specific, known CNV is suspected and a single gene deletion or duplication is a significant contributor to the condition, such as in the case of 22q11.2 deletion; also known as DiGeorge syndrome (Ning et al., 1996). On the other hand, it is not the best choice when the phenotype is non-specific for which a microarray test would be a better option.

1.4.4 Genetic and Genomic Sequencing

1.4.4.1 The Trade-Offs of the gold standard Sanger Sequencing: High Precision, Low Volume

Sanger Sequencing, also known as dideoxy sequencing or chain termination sequencing, is a widely used method for determining the sequence of DNA molecules. It was developed by Frederick Sanger and colleagues in 1977 and was the first successful method for sequencing DNA. The method works by using a DNA polymerase enzyme to synthesise a complementary strand of DNA using a single-stranded template and a mixture of four fluorescently labelled dideoxynucleotides (ddNTPs). These labelled ddNTPs can be incorporated into the growing strand, however, they cause premature termination of the reaction resulting in a mixture of extended fragments with different lengths. These are separated by size using high-resolution electrophoresis in a capillary based polymer gel. Using laser excitation of the fluorescently labelled ddNTPs provides a readout that corresponds to the identity of the nucleotide in the terminal position in what is known as sanger sequence (Shendure & Ji, 2008). This process generates a series of overlapping DNA sequences that can be combined to produce the complete DNA sequence (Sanger et al., 1977). Sanger sequencing is considered the gold standard for accurate genetic testing and has been instrumental in many scientific breakthroughs, including the initial sequencing of the human genome (Lander et al., 2001).

1.4.5 From Sanger to NGS: The Evolution of Genomic Testing

Next-generation sequencing (NGS) or second-generation sequencing is an innovative technology that has revolutionised the field of genomics. It has enabled the simultaneous sequencing of billions of DNA fragments in a single run, resulting in high-volume and low-cost genomic testing (Shendure & Ji, 2008). This technology has transformed the field of genomics by making large-scale genome sequencing projects feasible and has opened up new avenues for medical and scientific research (Metzker, 2010). It is characterised by its ability to produce vast amounts of genomic data without any prior hypotheses or preconceived ideas, making it a powerful tool for gene-agnostic testing (Metzker, 2010).

More than 90% of NGS sequencing data has been generated using a technique called sequencing by synthesis (SBS) chemistry (Quail et al., 2012). This method involves four main steps: library preparation, cluster generation, sequencing, and data analysis.

During the library preparation step, the genomic DNA (gDNA) or coding DNA (cDNA) is fragmented using mechanical shearing via ultrasonication or nebulisation, or enzymatical digestion (Knierim et al., 2011). The blunt ends of the sheared or digested fragments are ligated to unique adapters at both ends serving as sample identifiers. After adaptor ligation, the DNA fragments are either size-selected using agarose gel electrophoresis or paramagnetic beads, which remove the ligation duplicates. The next step is cluster generation, in which the DNA fragments are melted, and the resulting single-stranded DNAs are immobilised. The immobilisation can occur on planar solid surfaces of a flow cell, as is the case with Illumina sequencers, or on the surface of micron-scale beads, such as with 454-Roche/SOLiD sequencers. Alternatively, the single-stranded DNAs can be immobilised on ionised spheres, which is the case with Ion Torrent sequencers (Loman et al., 2012). The immobilisation step is followed by template amplification to generate clusters, which is accomplished through polymerase chain reaction (PCR), either on a solid surface or through emulsion PCR in individual microreactors, beads, or spheres within the sequencers. The next step involves sequencing by synthesis, which is achieved by incorporating modified nucleotides in the case of illumina (with reversible terminators and fluorescent dyes), and 454-Roche/SOLiD (with pyrophosphate group) or unmodified nucleotides in the ion torrent. In the case of illumina, the fluorescently labelled deoxyribonucleotide triphosphates (dNTPs) with reversible terminators are incorporated into the newly synthesised complementary strand to the clustered DNA fragments on the flow cell. These dNTPs are added sequentially, halting the synthesis upon incorporation into any single position based on the sequence of the template. A light source then excites the synthesised complementary strand, followed by a cleavage of the terminator bound to the dNTP to allow the progression of the synthesis and minimise the bias of incorporation as the four modified bases have an equal chance of being incorporated. In 454-Roche/SOLiD, the modified nucleotides with a pyrophosphate group release light upon incorporation, revealing the number of incorporated bases at each position. On the other hand, the unmodified nucleotides, utilised in the ion torrent, are

directly detected through pH changes caused by nucleotide incorporation, without labels or terminators (Ross et al., 2013). The emission of light or hydrogen ions from each dot on the solid surface or sphere during the enzymatic attachment of complementary nucleotides to clusters of identical single-stranded DNA fragments is captured by a camera and recorded as a fragment sequence (Metzker, 2010). The record of the multiple copies representing the DNA fragment sequences generated is saved in a FASTQ file.

To achieve the necessary level of resolution for an NGS experiment, adequate coverage is essential, which refers to the average number of reads that align to each base within the sample DNA. Furthermore, to ensure that every DNA region is represented multiple times in different read frames, the sequences of the fragments (reads) must overlap. In modern NGS protocols, purified DNA is obtained from multiple cells, and DNA shearing is performed on several genome copies to provide a sufficient number of overlapping fragments. In single-cell sequencing, DNA is PCR-amplified before fragmentation to generate the required number of overlapping reads (Nawy, 2013). Regardless of the approach used, the overall length of the sequenced reads must exceed the genome size by a significant factor to ensure the unambiguous determination of the whole genome sequence (Shendure & Ji, 2008).

Paired-end (PE) sequencing represents a significant improvement in NGS technology. This technique involves sequencing both ends of DNA fragments in a library and then aligning the forward and reverse reads as read pairs. Compared to single-read sequencing, PE sequencing produces twice as many reads in the same amount of time and effort, which increases the accuracy of read alignment and allows for the detection of indels (Nakazato et al., 2013). Moreover, the analysis of differential read-pair spacing enables the removal of PCR duplicates that may arise during library preparation. Although single-read sequencing is suitable for some applications, such as small RNA sequencing, the majority of researchers currently prefer the paired-end approach due to its ability to produce a higher number of SNV calls following read-pair alignment (Nakazato et al., 2013).

Finally, the reads are aligned to a reference genome ready to be analysed. Algorithms dedicated to the alignment of sequencing data have shown rapid growth in number and development to

accommodate the fast-evolving sequencing technologies. The developments have targeted read length, throughput, and the error rate of the specific technology. One of the first algorithms and still used to this date, is BLAT (Kent, 2002), which was initially designed to align sequences that are <1000bp unidirectional nucleotide sequence reads known as expressed sequence tag (EST) sequences and are derived from cDNA (Boguski et al., 1993). However, such algorithms are not well suited to cope with high throughput technologies generating millions of short-read sequencing data to be aligned to the reference genome. For aligning WGS short-reads, established options like Burrows-Wheeler Aligner (BWA, specifically BWA-MEM) remain popular for their speed and accuracy, particularly for mapping reads to a single reference genome (Li & Durbin, 2009a). Minimap2 offers versatility, handling both short and long reads with good performance (Li, 2018), while Stampy provides high sensitivity for indel discovery, making it valuable for variant calling studies (Lunter & Goodson, 2011). Several other tools have been developed and commonly used such as Bowtie2 (Langmead & Salzberg, 2012), and Isaac (Raczy et al., 2013), as well as other specialised tools like NovoAlign (http://novocraft.com/) for aligning to closely related reference genomes.

Once aligned, a variety of analyses can be conducted, including identifying SNPs or insertion-deletions (indels), counting reads for RNA methods, conducting phylogenetic or metagenomic analyses, and more.

NGS has brought about a paradigm shift in genomics by enabling high-volume, low-cost, and hypothesis-free genomic testing. Various companies have developed multiple techniques over the years; however, the Illumina platform is currently the dominant player in the market. Their technique relies on massively parallel sequencing of short DNA fragments ranging from 300-500 bp in length to generate millions of short reads (50-300 bp). The generation of such high coverage for each genomic position is to circumvent the inherently higher sequencing error rate, compared to the Sanger technique, by reaching a consensus sequence. However, higher coverage does not necessarily improve error rates when they are driven by genomic context. These errors occur in low-complexity regions of the genome, such as those with stretches of the same base, known as homopolymer sequences (Slatko et al., 2018).

NGS techniques make it feasible to detect novel genes and gene-disease associations. However, the indiscriminate detection of variants can lead to the identification of variants of unknown significance (VUS), which are difficult to interpret due to a lack of scientific evidence regarding their role. Consequently, clinical interpretation and translation of such data for the patient can be complicated. The utility of this technique extends beyond providing insights into disease-related sequence information. It can also reveal methylation patterns, which can offer an additional layer of genetic information linked to disease (Barros-Silva et al., 2018).

The technology is widely used for various applications, including panel testing, whole-genome sequencing, exome sequencing, transcriptome sequencing, and epigenetic analysis (Shendure & Ji, 2008).

1.4.5.1 Targeted sequencing and gene panels

NGS has the capability to create targeted multigene panels for any group of genes exhibiting a shared phenotype or regions linked to the disease. This targeted approach offers advantages over genome-wide testing, as the average coverage obtained from WGS ranges between 30-50x per genome, while targeted sequencing could reach 500-1000x coverage (Illumina). Such a level of resolution would enable precise identification of variants, which are limited only to genes or regions linked with the patient's particular phenotype. However, there is no scope for novel gene discovery using this targeted approach.

For conditions with genetic heterogeneity, such as non-syndromic ID or ASD, panels can be designed to enable concentrated sampling of some or all of the more than one thousand genes previously associated with these conditions (Aspromonte et al., 2019). While this approach offers flexibility for the customisation of panels, it also carries the potential for inconsistencies in the number of genes and delivery time among laboratories that provide genetic testing for similar phenotypes. Therefore, the selection of the most suitable panel for testing different patients can be a trial-and-error process and might not always yield positive results. In addition, the expanding number of genes associated with neurodevelopmental disorders makes the curation and updating of panels a laborious task, and the panels rapidly become obsolete. Therefore, a genomic approach

would be more appropriate for both novel disease gene discovery as well as for patients where the targeted panels yield negative results.

1.4.5.2 Whole Exome sequencing

Whole exome sequencing (WES) is an NGS technique that targets the protein-coding regions of the genome, known as the exome, which account for only about 1-2% of the total genome (Bamshad et al., 2011).

This method is based on capture enrichment of the exons, usually including ±25 bp flanking intronic regions of the genes involving the splicing regions. This enrichment of specific regions of the genome leads to non-uniform coverage peaking around the exons and tailing off on the flanking regions resulting in a lower coverage and only a few reads being captured in these splicing regions. WES allows for the identification of coding genetic variations, such as SNVs and indels, that might contribute to Mendelian diseases. Historically, approximately 85% of mutations with significant impacts on disease-related traits were believed to reside within the exome (Majewski et al., 2011) and thought to cause the majority of monogenic mendelian diseases (Kuhlenbäumer et al., 2011). However, these estimates are based on WES data, which exclude non-coding regions from capture or eliminated during the bioinformatic processing of data. As a result, variations within non-coding regions that regulate well-established disease genes are frequently underreported. This is reflected in databases such as ClinVar where less than 0.5% of high confidence pathogenic variants are in untranslated regions (UTRs) and have an essential regulatory role (Sangeeta & Pal, 2009). Nonetheless, the interpretation of non-coding variants can be complex, and our understanding of their impact on disease development may lag behind that of the exome, which could introduce bias. This is also clear within the classification of regulatory variants in ClinVar as ~63% are classified as VUSs compared to ~44% of coding variants (Ellingford et al., 2022). Yet, the use of WES in genetic studies has led to a rapid increase in the identification of disease-causing mutations, as reflected in the addition of approximately 2,000 new entries in OMIM describing the molecular basis of specific phenotypes (Petersen et al., 2017).

WES has emerged as a valuable tool for identifying disease-causing genetic variants in both rare and common complex diseases. It has been used to identify genetic variants associated with various

neurodevelopmental and epilepsy disorders, including ID, ASD, and epilepsy. For example, a recent study used WES to identify new candidate genes for ASD and found that variants in the *NALCN* gene, previously implicated in developmental disorders, were significantly associated with ASD (Wang et al., 2022). Another study applied WES to 1,244 individuals with ID and found that rare *de novo* variants in genes related to synaptic function were significantly enriched in ID cases compared to controls (De Rubeis et al., 2014).

One of the major projects to scale up the use of WES is The Deciphering Developmental Disorders (DDD) project, which has WES data for ~13,000 cases from the UK, with 88% having neurodevelopmental disorders. The diagnostic rate is 34%, with parent-child trios contributing more towards the yield than singleton cases (Wright et al., 2022).

However, WES has its limitations as it only captures a portion of the non-coding genome, which includes vital regulatory elements that impact gene expression and play significant roles that are becoming increasingly recognised (Warr et al., 2015). Furthermore, the non-uniformity of coverage, particularly at the ends of exons, including splicing regions, increases the risk of missing up to 15% of disease-causing variants (Taylor et al., 2015). Additionally, it is not effective in identifying structural variants >50bp or copy number variations, which can be detected by other techniques, such as chromosomal microarray analysis (Tankard et al., 2018). Several key challenges contribute to the low diagnostic yield of WES, such as in the DDD study, including inefficiencies in the capture process, lack of coverage of regions outside the exome, the possibility that pathogenic variants lie in yet unrecognised novel disease-causing genes, size of cohorts which limit possibility of identifying recurrent variants, and reduced power for identifying pathogenic variants in cases of pre/perinatal death. Additionally, the estimation that roughly 1,000 possibly disease-causing genes associated with developmental disorders are yet to be discovered further complicates matters, along with the challenge of incomplete penetrance (Kaplanis et al., 2020). Although routine WES has improved diagnostic capabilities in clinical diagnostic laboratories, only approximately one-third of patients with presumed genetic neurodevelopmental disorders receive a definite diagnosis (Srivastava et al., 2019; Vissers et al., 2017). This diagnostic uncertainty can significantly impact patients, parents, and families and providing more conclusive molecular diagnoses could alleviate some of this burden (Graungaard & Skov, 2007).

1.4.5.3 Whole-genome sequencing

Whole-genome sequencing (WGS) offers a more comprehensive approach than WES as it does not limit sequencing to a subset of the genome, and it overcomes many of the limitations of WES. Some of the benefits of WGS include providing uniform coverage across the genome, allowing for the detection of various types of variants, encompassing SNVs, CNVs, translocations, inversions, as well as variants in the non-coding regions of the genome (Lionel et al., 2018).

An overview of the types of variants detected using the various tests discussed so far is shown in Table 1.1, while Table 1.2 summarises the attributes of currently available tests (JAX, 2022).

Table 1.1: Comparison of current tests available and the variants detected by each test. Adapted from (JAX, 2022)

Test type	SNV	Indel	CNV	TNR	Methyl	mtDNA
Targeted mutation sanger sequencing	x	#	#	#	#	#
Gene panels	x	x	-	-	-	-
Chromosomal analysis (CA)	-	x	x	-	-	-
Whole Exome sequencing	x	x	-	-	-	-
Whole genome sequencing	x	x	x	-	x	-

SNV=Single Nucleotide Variants; Indel=duplications and deletions; CNV= Copy Number Variants; TNR=Trinucleotide repeat expansions; Methyl=Methylation; mtDNA=Mitochondrial DNA variants; X denotes the type of variant detected by the test; # denotes variants that could be assessed as appropriate for the condition tested; - denotes variants that are not detected by the test. (Gross et al., 2019)

Table 1.2: Comparison of attributes for tests used for variant detection. Adapted from (JAX, 2022)

Test type	Scope	Sensitivity	Uncertain findings	Secondary findings	Unexpected findings	Cost	Turnaround time
Targeted mutation Sanger sequencing	+	+++++	+	N/A	+	+	+
Gene panels	+++	+++	+++	++	+++	++	+
Chromosomal analysis (CMA)	++++	++	++++	+++	++++	+++	+++
Whole Exome sequencing	+++++	+	+++++	+++	+++++	++++	++++
Whole Genome sequencing	++++++	++	++++++	++++	++++++	+++++	+++++

Scope= the specific types of variants that the test assesses; Sensitivity= How well the variants are detected; Uncertain findings=Variants detected with unknown impact or a specific function; Secondary findings= Variants that are detected but unrelated to the phenotypes/condition tested; Unexpected findings= unexpected familial relationships that relate to the findings; Turnaround time= Time taken to return test results to the patient or the clinician

The potential of WGS in clinical diagnostics has been demonstrated by the WGS500 project, which evaluated its application for clinical use. The study recruited 217 individuals, amongst whom 157 were patients with varying conditions left undiagnosed after investigations using traditional genetic methods (Taylor et al., 2015). This study uncovered two key results: i) a higher diagnostic yield was achieved compared to testing of known genes using conventional methods, with a range of 21-57% depending on the inheritance pattern of the disease, and ii) WGS was able to identify pathogenic causative variants within genes that had previously been studied using targeted methods. This demonstrates that a broad spectrum of conditions can be diagnosed using WGS. The diagnostic yield for WGS from the WGS500 project is consistent with results from other studies with stringent criteria that report similar percentages (Lionel et al., 2018). However, some studies report a higher diagnostic outcome of 73% harder-to-diagnose conditions such as neurodevelopmental disorders (Soden et al., 2014). This could be attributed to the smaller sample size recruited for the study, the lack of pre-screening inclusion criteria implemented in the studies mentioned above, the focus on one specific disorder rather than broad spectrum and the predominance of *de novo* variants in NDD. The application of WGS over the years has led to a continuing expansion in neurogenetic research revolutionising the clinical diagnostics of rare neurological disorders (Pittman & Hardy, 2013). This could help alleviate the "diagnostic odyssey" that many patients with rare neurodevelopmental diseases experience as they wait 5-30 years for a diagnosis (EURORDIS, 2007).

With the increasing potential of WGS in the genetic diagnostic field, there is a high degree of risks presented by the analysis. These disadvantages include unsolicited secondary findings beyond the American College of Medical Genetics and Genomics (ACMG) approved list of 61 actionable genes (Miller et al., 2023), unexpected familial relationships, and scientific reporting of common VUSs. Together these could potentially impact the patient's ethical rights and affect family relations where a finding might implicate other untested members. The reporting of VUSs as causative variants increase the burden on the diagnostic process (van et al., 2013). This issue is reflected by an earlier study conducted on 104 sequenced individuals with severe childhood diseases, which reported 406 variants as causative pathogenic variants using <5% frequency to filter for recessive variants (Bell et al., 2011). A later examination of those reported variants found that 27% (122/406) were either

common polymorphisms or did not have enough supporting evidence for pathogenicity (Bell et al., 2011).

The need for stringent criteria when performing WGS analysis is becoming increasingly recognised in efforts to avoid reporting on false positives, incidental and unexpected findings. The ACMG guidelines have been adopted to mitigate these instances in reporting of genetic findings. This is particularly important where WGS is encouraged to be the first line for diagnostic testing in developmental disorders.

1.5 Filtering, prioritisation, and classification in WGS Data: Methods and Approaches

The emergence of WGS and its integration into clinical settings has driven the need for efficient and precise bioinformatics tools to analyse WGS data. The primary objective is to detect disease-causing variants from the vast number of approximately 5 million candidates identified in human genomes. Employing filtering strategies can decrease the number of variants that necessitate manual scrutiny by utilising diverse types of information annotated to variants stored in the Variant Call Format (VCF) file following variant calling (Biesecker, 2010; Carson et al., 2014).

Variant filtering and prioritisation are performed using various approaches, including those that focus on genotypic information only and tools that utilise both genotypic and phenotypic information. For genotype-based analyses, several functional biological annotations have been developed and used to assess various aspects and rank the identified variants. These include annotations for protein function such as SIFT and PolyPhen-2 (Adzhubei et al., 2013a; Ng & Henikoff, 2003), ancestry-specific allele frequency from large population databases such as Genome Aggregation Database (gnomAD; Karczewski, et al., 2020), conservation using constraint scores (Cooper et al., 2010), epigenetics through the utilisation of databases such as ENCODE (Davis et al., 2018), spatial genome organisation from chromatin interaction analysis studies (Tang et al., 2015), protein networks which explore the gene expressions through highly connected hubs (Yu et al., 2007), mappability, which indicates the reliability of sequencing reads (Karimzadeh et al., 2018), local nucleotide diversity and heritability (Gazal et al., 2017), genomic sequence (O'Leary et al., 2016), as well as integrative annotations indicating variant deleteriousness. Some of these integrative prediction tools include CADD (Rentzsch et al., 2019), SpliceAI (Agathe et al., 2023), FATHMM-

XF (Rogers et al., 2018), and VEP (McLaren et al., 2016). However, several limitations face the genotype-based analysis, including the fact that most of the prediction tools target point mutations and small indels within coding regions of the genome. Therefore, there is an increased demand for calling and annotation tools, both for structural variants that are >50 bp and for non-coding variants. Moreover, relying solely on genotypic information can be challenging, as human genomes typically contain around 100 loss-of-function (LoF) variants, which can potentially inactivate approximately 20 genes (MacArthur et al., 2012).

The limitation of relying solely on genotypic information was recognised early on, leading to the development of genotypic-phenotypic ranking algorithms such as Exomiser (Smedley et al., 2015), which focuses on exome data and Genomiser that utilises WGS data (Smedley et al., 2016). These tools take advantage of frameworks such as the Human Phenotype Ontology (HPO), which standardises clinical terms used for describing abnormalities linked to human disease (Robinson et al., 2008). Moreover, they exploit mouse, rat, and zebrafish phenotype databases to draw connections between genotypic variants and patients' specific phenotype information to rank identified variants (Smedley et al., 2015).

Clinical and research laboratories employ various filtering and prioritisation approaches that focus on specific variant types based on their expertise, interests, and the nature of the disease being studied. This limitation can hinder the analysis of comprehensive datasets such as those produced by WGS. However, SNVs and CNVs are only meaningful for clinical interpretation if they can be linked to a specific disease. To ensure a consistent approach to variant interpretation for clinical use, the ACMG has established policies and standards. Variants are classified into five categories: pathogenic, likely pathogenic, uncertain, likely benign, and benign (Richards et al., 2015). Variants that are classified as pathogenic mutations or benign polymorphisms allow for straightforward interpretation and management: pathogenic variants trigger further assessment and patient communication, while benign variants provide reassurance and minimal follow-up, unless specific circumstances arise. Variants are considered likely benign or likely pathogenic when there is a 90% or greater certainty about their pathogenic or benign nature. During variant classification, the inheritance pattern is considered, and if a variant is inherited from an unaffected parent, it supports the classification of a likely benign or benign category. A study of nearly 600,000 variants reported

in the ClinVar database over three years found that those categorised as "likely pathogenic" were most frequently reclassified as pathogenic, including 70% of those anticipated to cause a missense mutation and 90% of those expected to cause a loss-of-function mutation (Harrison & Rehm, 2019). About 40% of total variants end up with a classification of VUS, which presents a challenge in the clinical context and medical reporting (Federici & Soddu, 2020). Usually, these variants are omitted from reports relayed to patients due to insufficient clinical and functional information available, as they could lead to confusion and anxiety for patients. Therefore, they require further functional validation or replication in independent families, to establish their causality before reporting back to patients (Makhnoon et al., 2019).

Incidental findings are another set of variants that are typically reported and follow ACMG reporting standards and policy. These positive findings are unexpected and not related to the primary analysis or findings that led to the referral phenotype, but they still have medical significance for the patients. The ACMG has identified a set of genes that have been well documented in relation to the disease and would impact the treatment or management of the patient (Green et al., 2013) which they recommend should be screened for incidental findings.

Large sequencing projects such as the 100,000 Genome Project (100KGP) follow a similar approach for variant prioritisation, and classification for reporting.

The 100KGP was introduced in 2013 to diagnose patients with rare diseases and some types of cancers. One of the main aims is to embed the utility of WGS within the National Health System (NHS) in the United Kingdom (Turnbull et al., 2018). Therefore, Genomic Medicine Service (GMS) was created in 2018 to make genetic testing part of the routine healthcare. The GMS collaboration involves seven regional genetic laboratory hubs, various NHS clinical genetics services, as well as primary, secondary and tertiary care in the NHS, and Genomics England. This collaborative work facilitates for equitable access to genetic medicine to more people across the UK (Pichini & Bishop, 2022).

The clinical diagnostic pipeline of the 100KGP for filtering and analysing WGS data uses a tiering system designed by Genomics England (GEL; Figure 1.1). Identified variants are assigned into one of three categories (tier 1-3) based on the mode of inheritance, consequence, and the classification of gene disrupted according to PanelApp (https://panelapp.genomicsengland.co.uk). It is a

knowledge base of *in silico* gene panels for a broad range of human diseases applied to the WGS data of sequenced participants according to their phenotypes and assumed clinical diagnosis. In PanelApp, the genes are classified into three classes: approved genes for clinical diagnosis and have a high level of evidence for gene-disease association (Green), genes with moderate evidence but not enough for diagnostic grade use (Amber), or genes with insufficient evidence/ conflicting evidence of causality, therefore, should not be used for interpretation (Red). The data from the 100KGP is held in the National Genomic Research Library (Caulfield et al., 2017) with secure access only to approved clinicians and academic researchers to enable scientific discovery and timely translation into patient care.

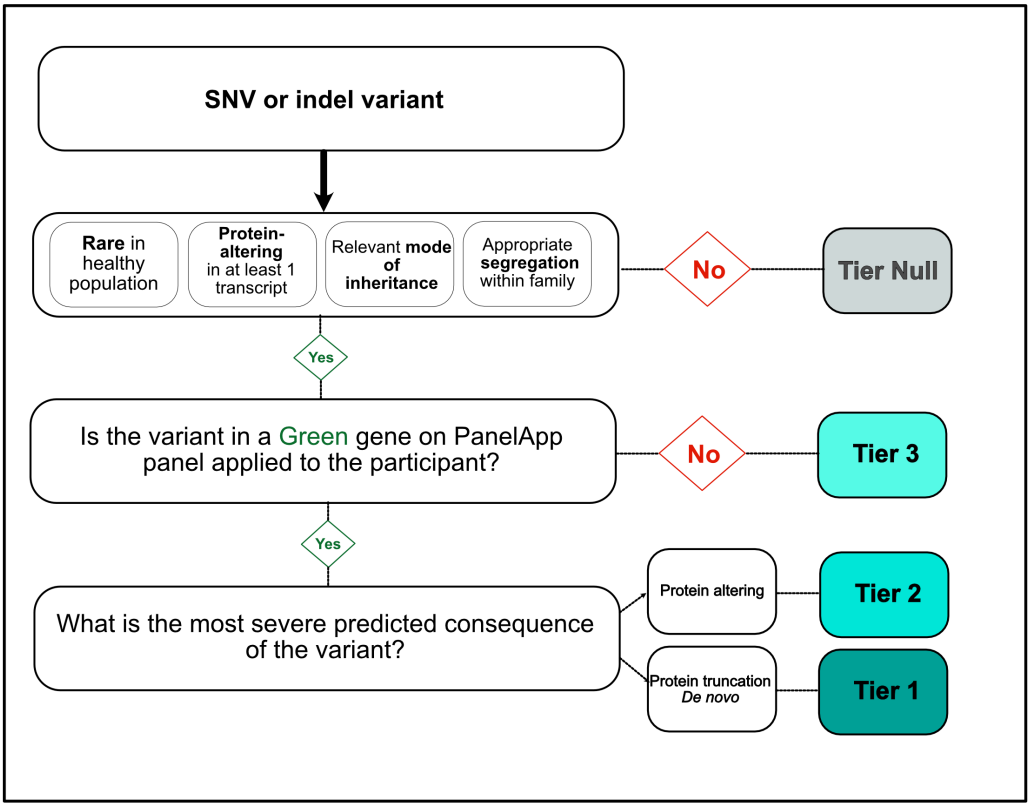


Figure 1.1: The tiering system designed by GEL for single nucleotide variants and indels. Adapted from the tiering process flowchart from GEL "Rare Disease Results Guide"

Regardless of the outcome of the genetic analysis of WGS data, the results are typically delivered to patients and their families by the clinician providing genetic counselling on the outcome.

1.6 Ethical Issues Associated with Whole Genome Sequencing

Thorough pre-test counselling and informed consent can help address numerous ethical considerations associated with genetic testing. However, several potential pitfalls exist, including interpreting a positive result as a reason for feeling guilty or placing blame, patient privacy loss due to the posting of clinical information in public databases, identifying abnormal results in family members tested to classify a variant of uncertain significance, and parental misidentification (due to adoption or gamete donation) and unexpected consanguinity/incest.

To mitigate these issues, it is crucial to ensure that patients are fully informed about the potential risks and benefits of genetic testing before undergoing the procedure. Healthcare providers should provide clear explanations of the results and the implications for patients and their families. Additionally, privacy laws and policies must be in place to protect patient's personal information from unauthorised access or disclosure. By addressing these concerns through effective communication and privacy protections, genetic testing can be a valuable tool for improving patient health and well-being.

Even when provided with appropriate counselling, patients and their families may still misinterpret the results of a genetic test. This misinterpretation is particularly true for non-physicians who may view genetics as rigidly deterministic and need to become more familiar with scientific concepts of uncertainty. Additionally, patients, their families, and their healthcare providers may be so invested in finding a positive test result after a long and challenging diagnostic journey that they view VUS as pathologic, despite receiving counselling (Reiff et al., 2012).

Prior to genetic testing for neurodevelopmental disorders, it is essential to obtain informed consent. The informed consent process should include a thorough review of the potential implications of test results and how the family may perceive them, whether positive, negative, or inconclusive. The family needs to properly understand the risks and benefits associated with the testing before deciding how much-unsolicited information about pathologic variants they wish to receive. By ensuring that the family has adequate knowledge and understanding of the testing process and potential outcomes, informed consent can help to promote more effective communication and decision-making.

Positive results from genetic testing can have a significant impact on various aspects of a person's life, including changes in therapy service eligibility, insurance discrimination, and increased fear

and stigma. Therefore, obtaining consent before conducting genetic testing is important, and this should include adolescents with the capacity to participate in the decision-making process. It is essential to seek out and respect their feelings and opinions about testing to ensure they are fully informed and empowered to make the best decision for themselves. By involving adolescents in the decision-making process, their rights and autonomy can be respected, and the potential impacts of genetic testing on their lives can be minimised (Bush et al., 2018; Rosenfeld & Patel, 2017).

1.7 Thesis objectives

As the list of rare genetic diseases with molecular diagnoses continues to expand, it has the potential to reduce the diagnostic odyssey for a growing number of patients. However, the diagnostic yield from WGS data remains between 25-40%, largely because the structural, splice site and non-coding variants have been inadequately explored. This poses a challenge for deciphering the underlying causes of the remaining, incompletely characterised conditions.

In general, the main aim of this thesis is to analyse WGS data using a combination of approaches with a focus on under-explored variant types such as structural variants and validate the functional effects of rare genetic disease variants that stem from WGS data of patients with neurodevelopmental disorders.

The primary objectives of this thesis are as follows:

1. To conduct a comprehensive analysis of WGS data from unsolved neurodevelopmental and epilepsy cases recruited to the WGS project funded by the Health Innovation Challenge Fund 2 (HICF2) and to the 100KGP, beyond the SNVs present in the coding sequence. This will involve exploring copy number variations, structural variants, and the non-coding regions of DNA in Chapter 3.
2. To identify novel disease-causing genes to facilitate new diagnoses in Chapters 3 and 5.
3. To utilise various techniques to functionally assess the genetic findings from Chapters 3 (*PLPPR5*). This will enhance our understanding of the cellular biology underlying epilepsy, leading to a deeper comprehension of the underlying biological causes of the disorder in chapter 4.
4. To characterise and validate the detection of complex and difficult-to-detect structural variants in Chapter 5.

Data disclaimer(s): the sequencing data included in this thesis are part of collaborative work and access was generously granted by the study collaborators. I would like to extend sincere appreciation to all mentioned in this disclaimer for their invaluable contributions, expertise, and collaborative efforts, which significantly enriched the research presented in this thesis.

- **Chapter 3:**
    - Whole genome sequencing for samples recruited under the HICF2 project was performed by the accredited Oxford Molecular Diagnostics Laboratory at the John Radcliffe Hospital
    - VARAN in-house annotation tool developed as well as read alignment, mapping and quality check steps by Dr. Niko Popitsch (Taylor group, University of Oxford, UK)
    - Variant Explorer pipeline developed by Dr. Edoardo Giacopuzzi (Taylor group, University of Oxford, UK)
    - The molecular mechanics analysis for DOCK7 protein discussed in chapter 3 section 3.2.3 was performed by Dr. Matteo Ferla (Taylor group, University of Oxford, UK)
    - MLPA validation of variant described in chapter 3 section 3.2.1.1 was performed by Dr. Davide Mei (neurogenetics laboratory, Meyer University Hospital)
    - Optical mapping was performed at the Bionano Services Laboratory in San Diego, USA
    - Transcriptomics sequencing was performed by the Oxford Genomics Centre
- **Chapter 5:**
    - Sequencing and SV calling were performed by GEL.
    - SVRare pipeline used for analysis was developed by Dr. Jing Yu (Taylor group, University of Oxford, UK)
    - The breakpoints of the translocation identified in I14_5 have been validated by dideoxy sequencing by the Wessex clinical genetics laboratory.

Chapter 2. Materials & Method

2.1 Study recruitment and participation

The study involved neurodevelopmental disorders and epilepsy cases recruited as a part of the HICF2 project or the 100KG project under the rare disease domain.

The majority of patients for the HICF2 project were recruited through the Genomic Medicine Multi-Disciplinary Team (GM-MDT) which verified patient eligibility for whole genome sequencing, approved participation in the project, and reviewed case findings. The GM-MDT consisted of diverse specialists, including clinical scientists, genetic counsellors, non-clinical researchers, physician-scientists, and physicians. The referring clinicians wishing to enrol a patient in the HICF2 study were required to submit a detailed application. The form detailed the patient's clinical information, disease diagnosis, phenotype, family pedigree, prior genetic testing performed, and the availability of samples from the proband and relevant family members (parents, siblings, and affected extended family members).

To be recruited for WGS, patients had to have had extensive pre-screening with high-resolution a-CGH and gene panel-based testing of known genes for their condition. This pre-screening was conducted at various time points of the patients' diagnostic odysseys using NGS techniques and arrays that were evolving. However, some diagnoses were missed due to factors such as the use of older versions of a-CGH with lower resolution, sequencing of hotspot variants only, use of older gene panels, or human error and variations between different clinical laboratories. This made it challenging to identify the genetic causes for the patients' conditions. Information on the submission form was used to guide downstream analysis and extract HPO expressions. Additional HPO terms were obtained directly from clinicians if the initial form did not have sufficient detail. Patients were referred from Oxford mainly but also from other UK centres, including Great Ormond Street, Birmingham Women's and Children's NHS Trust, University Hospital Coventry, Northampton General Hospital, and overseas centres such as University Hospital of Ulm, Germany; University Medical Centre, Utrecht, Netherlands; Odense University Hospital Denmark; Meyer

Children's Hospital IRCCS, Florence, Italy.

2.2 Ethics and patients consent

This study involved participants from the HICF2 and the 100KGP. For participants recruited under the HICF2 project, a genetic counsellor or medical specialist obtained fully informed consent from all participants. Most participants consented under the Molecular Genetic Analysis and Clinical Studies of Individuals and Families at Risk of Genetic Disease study (MGAC) protocol approved by the West Midlands Research Ethics Committee (WMREC), reference 13/WM/0466. Additional ethics approvals were obtained from Oxford Research Ethics Committee (OXREC) B: 04.OXB.017; OXREC C 09/H0606/74; OXREC A (05/Q1605/66); WMREC 09/H1204/30; London Riverside REC 09/H0706/20; REC: 12/SC/0044; HRA Committee East of England, Cambridge South (REC: 14/EE/1112); MREC 12/WA/0001; REC 10/H0604/85; Wales Research Ethics Committee (REC5) (13/WA/0371) and Paediatric Ethics Committee of the Tuscany Region/2014/000559-15.

For cases from the 100KGP, ethics approval was obtained from Cambridge South REC (14/EE/1112), and participants gave informed consent to partake in the study (as described in Smedley et al 2021). The recruitment processes for both projects were conducted over several years during which the pipelines evolved.

2.3 Accessing whole-genome sequencing data for the 100KGP

For cases recruited under the 100KGP, the WGS data are held securely in the National Genomic Research Library (https://doi.org/10.6084/m9.figshare.4530893.v6). The data access was granted via application to join one of the Genomics England Clinical Interpretation Partnership (GeCIP) domain at www.genomicsengland. co.uk/join-a-gecip-domain. For the purpose of analysis in this thesis, patients' data were reviewed periodically until data release (main-programme_v16; 2022-10-13).

2.4 Whole-genome sequencing for HICF2 and the 100KGP cases

2.4.1 DNA extraction and library preparation for the HICF2 samples

DNA was extracted from peripheral blood samples using the Gentra Puregene kit (Qiagen), yielding 3-3.5 µg each, at a minimum concentration of 30ng/µl, which was quantified using the High-Sensitivity Picogreen or Qubit system (Invitrogen). The DNA purity was verified to be between 1.8 and 2.0 at A260/280, and the sample integrity was assessed using 1% E-Gel EX (Invitrogen), showing no band below 23Kb. The sequencing library for each sample was prepared using Illumina TruSeq PCR-Free according to the manufacturer's guidelines.

2.4.2 Whole-genome sequencing

Whole genome sequencing for samples recruited under the HICF2 project was performed by the accredited Oxford Molecular Diagnostics Laboratory (OMDC) at the John Radcliffe Hospital using either HiSeq2500 or HiSeq4000 platforms (Illumina) in high throughput mode. The sequencing generated 150-bp reads with targeted mean coverage was 30× and obtained a minimum coverage of 22.5×.

For the 100KGP cases, the high throughput library preparation was done using TruSeq PCR-free and then sequenced using high throughput mode generating 150 bp paired-read on the HiSeqX (Illumina).

2.4.3 Read alignment, mapping, and quality check for HICF2 samples

2.4.3.1 HICF2 bioinformatics research pipeline

Initially, all sequence reads were generated using the Illumina offline base-caller (OLB v1.9.3) by the OMDC. The data from samples in the initial stages of the HICF2 project were mapped to the GRCh37d5 human reference sequence, which contains a 35.48-Mb decoy chromosome. The decoy chromosome reduces the misalignment of repetitive sequences, particularly around the centromeres, improves the accuracy of SNP discovery, and reduces the rate of false positive calls (Genovese et

al., 2013). On the other hand, the later samples were mapped to build GRCh38 and both genome references were obtained from the 1000 Genomes Project.

Samples data processing was performed using an in-house bioinformatics research pipeline developed for the HICF2 project. The read alignment, mapping and quality check steps for all samples was led by Dr. Niko Popitsch, Burrows-Wheeler Aligner (BWA) v.0.7.10 (Li & Durbin, 2009) was used to align the FASTQ files per sample per lane. The aligned files were then refined per lane using Stampy v.1.0.23 (Lunter & Goodson, 2011). The BAM files were sorted per lane. Subsequently, all bam files for each sample were merged into a single BAM file using Samtools (v.1.3). Duplicates were removed from the BAM files using Picard markduplicates (v1.111). During the later stages of the project, an in-house streamlined pipeline was developed by Dr. Edoardo Giacopuzzi (Variant Explorer; Figure 2.1).

2.4.3.2 Variant Explorer pipeline

The Variant Explorer pipeline was used to systematically re-analyse all samples and construct a genetic variant cohort representation. It is an automated pipeline that was built on Nextflow (Di Tommaso et al., 2017). It conducts read alignment to GRCh38, perform quality control on the alignments which includes mapping stats, coverage, sample relatedness, and ancestry, as well as variant calling, and annotations including ranking based on HPO terms. For the HICF2 samples, sequence reads were aligned to GRCh38 using BWA v0.7.15 (Li & Durbin, 2009b), and duplicated reads marked using Samblaster v0.1.24 (Faust & Hall, 2014). Low mappability regions were computed for GRCh38 reference genome using GenMap (Pockrandt et al., 2020) and a 150 bp k-mer size, corresponding to the read-length used in WGS sequencing. Subsequently, small variants were called using DeepVariant v1.0.0 (Poplin et al., 2018) and structural variants using a combination of Lumpy v.0.2.13 (Layer et al., 2014) and CNVnator v0.4.1 (Abyzov et al., 2011), stretches of homozygosity (ROH) were detected using bcftools roh command, and repeat expansions were identified using ExpansionHunter v3.2.2 (Dolzhenko et al., 2019). The pipeline is publicly accessible at https://github.com/edg1983/WGS_pipeline. The resulting VCF files, which encompass both small variants and structural variants, underwent a filtrating process based on

quality metrics. These included retaining variants with genotype quality above 20 that were present in at least one individual and removing missing genotypes to avoid variants falsely assigned reference genotype due vcf merging. Following this refinement process, the variants were subjected to annotations, including gene-level consequences utilising SnpEFF v4.3t (Cingolani et al., 2012), population allele frequencies (AF) from sources like gnomAD v2 and 1000 Genomes phase 3, predictive impact scores such as CADD (Rentzsch et al., 2019b), DANN (Quang et al., 2015), REVEL (Ioannidis et al., 2016), FATHMM (Shihab et al., 2015), ncER (Wells et al., 2019), ReMM (Smedley et al., 2016b), spliceAI (Jaganathan et al., 2019), and MaxEntScan (Yeo & Burge, 2004), as well as identification of known disease-causing variants from databases like ClinVar (Landrum et al., 2016), dbVAR (Lappalainen et al., 2013), and DECIPHER (Swaminathan et al., 2012). Gene-level ranking scores based on the HPO terms of the family were calculated using GADO v1.0.1 (Deelen et al., 2019) and variants were further annotated with probability of being loss-of-function intolerant (pLI) values from gnomAD v2.1.1, the GDI score (human damage index; (Itan et al., 2015), and the RVIS (Residual Variation Intolerance Score) value (Petrovski et al., 2015) based on ExAC v2.

Family-based analyses were performed using the Variant Explorer graphical user interface to interrogate the different variant types (coding and non-coding SNVs and SVs) that fit any of the inheritance patterns. The initial quality filter excluded low mappability, low complexity, and Segmental duplications regions to minimise false positives and cohort AF was set to 1%. The variant type filter was set to include: SNVs, INDELs, deletions, duplications, and inversions that had maximum population AF of 0.1%. For missense variant filter: minimum REVEL score= 0.3, CADD=20, and DANN score=0.3. For both INDELs and SNVs predicted to affect splicing, the SpliceAI minimum score was set to 0.2. Segregation filters were used to determine the inheritance pattern for which the variants can be further prioritised with a GADO score $\geq$ 0.6. Where rare suspicious SVs were identified in a patient, disruption of TADs and predicted loops were considered. Genomic coordinates of the SVs were visualised on the 3D Genome Browser against data from chromatin structures studies using human embryonic stem (ES) cells and human ES-cell-derived lineages (Dixon et al., 2015), as well as human lymphoblastoid cells (Rao et al., 2014). Genes

contained within a domain that had at least one insulating boundary disrupted by the SV were considered. Only genes that had been associated with dominant developmental disorders were considered for potential positional effects. A comparison of the phenotypes in the patient to phenotypes associated with disrupted genes was performed. An in-depth description and validation of the pipeline is available in (Pagnamenta et al., 2023, *in press*)

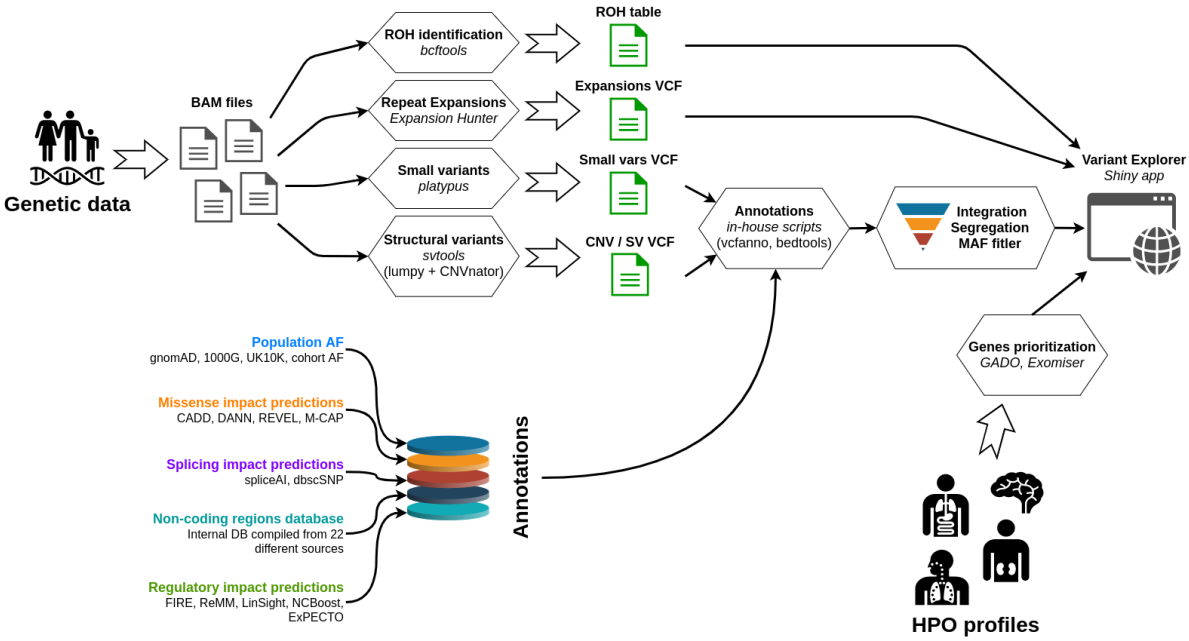


Figure 2.1: Variant Explorer pipeline- an in-house developed for analysing HICF2 WGS data. The pipeline automates the process of compiling and annotating WGS data. It provides a graphical user-friendly interface for filtering and prioritising variants. The pipeline was developed by Dr. Edoardo Giacopuzzi.

2.5 Genetic analysis

During the period of my DPhil, I used two distinct pipelines to conduct the genetic analyses aimed at investigating the WGS data. The initial analyses were performed using the HICF2 bioinformatics research pipeline described in sections 2.5.1-2.5.4. The same cases were later revisited using the Variant Explorer pipeline (Figure 2.1) described in section 2.4.3.2 and was used to either confirm the finding from the initial pipeline or identify new variants that were previously missed.

2.5.1 Copy number and zygosity analyses

CNVs calling, and annotation were initially conducted by Dr. Samantha Knight and for the HICF2 cases described in this thesis I analysed the generated annotated CNV lists.

The CNVs were identified using bam read account information employing the ngCGH tool (Petrovski et al., 2015) with a configured window size of 300 reads and the ngbin tool (Sun Microsystems Inc.) with a window size of 1000bp and a minimum read depth of 0 reads. B-allele frequency (BAF) information was obtained from vcf files generated by Platypus 0.5.2-py2.7 (Rimmer et al., 2014). Log2R and BAF outputs were imported, processed, calls flagged and visualized using Nexus Discovery Edition 8.0 (BioDiscovery, Inc., El Segundo, California, USA). The statistical approach was SNPRank Segmentation (based on Circular Binary Segmentation) and the settings used for flagging regions of interest (CN events and regions of homozygosity (ROH) that may be due to uniparental isodisomy or identity by descent) from imported ngCGH Log2R and BAF outputs were: Significance Threshold=1x10-6; Max Contiguous probe spacing (Kbp) = 5.0; Min number of probes per segment = 2; High Gain=1.5; Gain=0.13; Loss=-0.12; Big Loss=-1.5; homozygous Frequency threshold=0.88; Homozygous value threshold=0.8; Heterozygous Imbalance Threshold = 0.4; Minimum LOH length (KB) = 1000; Minimum SNP probe Density (Probes/MB)=0.0; Percent outliers to remove=3.0. The calls were curated to remove CNVs occurring commonly in the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home), regions of homozygosity <2Mb (i. e. within the normal population size range distribution) and frequently encountered signal 'artefacts' (eg due to segmental duplication and repeat sequence regions). The remaining calls were annotated comprehensively and used to assess pathogenicity

either individually or in combination with sequence level variants, aiding variant filtering as well as guiding the analyses.

2.5.2 Variant identification

For the cases described in this thesis, I performed the variant calling and annotation steps when using the HICF2 bioinformatics research pipeline and performed genetic analysis steps using the Variant Explorer pipeline.

When using the HICF2 bioinformatics research pipeline, variant calling was performed using Platypus V0.8.1 (Rimmer et al., 2014) and variants were stored in VCF files. This algorithm has the ability to detect SNPs and short indels (<50bp) as well as being sensitive to somatic mutations at low frequencies (Pagnamenta et al., 2012; Ruark et al., 2013). The variant calling was joint calling of related samples in a trio or a quartet so that the same position is examined simultaneously in all samples of the family. Joint calling of trios/ quartet allows the distinction of homozygous reference variant from no-calls due to the lack of supporting reads for that position and this is important especially for accurate detection of *de novo* calls.

2.5.3 Variant annotation

The multi-sample output VCF files from the previous variant calling step were annotated using VARAN (in-house annotation tool developed by Dr. Niko Popitsch). The algorithm utilises Variant Effect Predictor (VEP) from Ensembl to predict the effect of called variants on genes, protein structure and function, and regulatory regions. Other annotations from various external and in- house databases are then added to the files to include allele frequencies (includes: 1000G, ESP6500, gnomAD, UK10K, COSMIC, Human Gene Mutation Data (HGMD), ClinVar, GERP++, PhastCons, PhyloP, CADD, and WGS500).

2.5.4 Variant prioritisation

I analysed the annotated VCF files using two different variant prioritisation tools; Ingenuity Variant Analysis (IVA) and Exomiser.

2.5.4.1 Ingenuity Variant Analysis

The analysis of the various cases was performed while taking the different possible inheritance patterns into account. These modes included autosomal dominant including *de novo*, autosomal recessive (homozygous and compound heterozygous), and X-linked. The variants were retained if they conformed with the following confidence criteria:

a) had a call quality of at least 20.0 in cases or at least 20.0 in controls (unaffected parents of the proband and other healthy participants recruited to HICF2),
b) pass upstream pipeline filtering,
c) had read depth of at least 10.0 in cases or at least 6.0 in controls,
d) had an allele fraction of at least 5.0 in cases or at least 5.0 in controls. Furthermore, the top 5.0% of most exonically variable 100-base windows in healthy public genomes (1000 genomes), and top 1.0% most exonically variable genes in healthy public genomes (1000 genomes) were excluded.

The quality of the filtered variants was further ensured by excluding variants that had at least 20% of the surrounding reads fail the PASS filters (BRF). Various thresholds were tested and <20% threshold for the BRF filter was found to remove SNPs that are less likely to be in dbSNP suggesting they are likely to be false positive while maintaining high quality variants.

The variants that pass the confidence criteria were then further filtered for allele frequency. For *de novo* variant screening, variants with frequency lower than 0.1- 0.2% in various databases were kept (1000 genomes project, the NHLBI ESP exomes, gnomAD, and WGS500). For compound heterozygous and recessive inheritance patterns, the threshold of allele frequency was raised to 0.5% in the same databases.

Variants that have been predicted to be deleterious by various predictor scores such as SIFT (Kumar et al., 2009), Polyphen-2 (Adzhubei et al., 2013), CADD (Rentzsch et al., 2019), and SpliceAI (Jaganathan et al., 2019) were kept. Moreover, variants that were established as pathogenic, possibly pathogenic, disease-associated according to HGMD were also considered. Deleterious missense variants were determined using CADD score cut off >20 and those predicted to be deleterious by both SIFT and Polyphen-2. For protein truncating variants, three categories are expected to be observed with respect to tolerance loss of gene function: null (where loss of both copies of the gene is tolerated), recessive (where loss of a single copy of the gene is tolerated, but not loss of both copies), and haploinsufficient (where loss of a single copy of the gene is not tolerated). The scores range from 0-1 where 0 is tolerant to loss of function and 1 is intolerant to loss of function (Lek et al., 2016). Genes with pLI or pREC scores >0.6 were used as a guide for inclusion for further investigation. Variants with SpliceAI scores >0.5 for donor loss/gain and acceptor loss/gain variants were considered splice site candidates.

The variants that agree with those pathogenicity criteria were then subjected to genetic analysis to obtain the variants that are found in the proband only but not the parents (*de novo*), or that are found in the proband as homozygous variants but heterozygous in the parent (simple recessive), or genes where there are two heterozygous (compound heterozygous) alteration, one inherited from each parent or variants that satisfy X-linked inheritance pattern dominant or recessive.

Biologically relevant variants were prioritised according to association to patient's clinical phenotypes. The HPO terms used for identifying the biological links for cases discussed in chapter 3 are described in Table 2.1. In cases where no variants that have been biologically linked with any of the specified phenotypes were found, variants that matched the genetic analysis criteria were retained as potential candidate variants. These filters describe the more stringent criteria that were used for filtering, however, where no predicted deleterious variants were identified, the filters were relaxed. The dynamic nature of variant filtering and prioritisation is shown in Figure 2.2. Each of the final variants was manually checked in the Integrated Genomics Viewer (IGV).

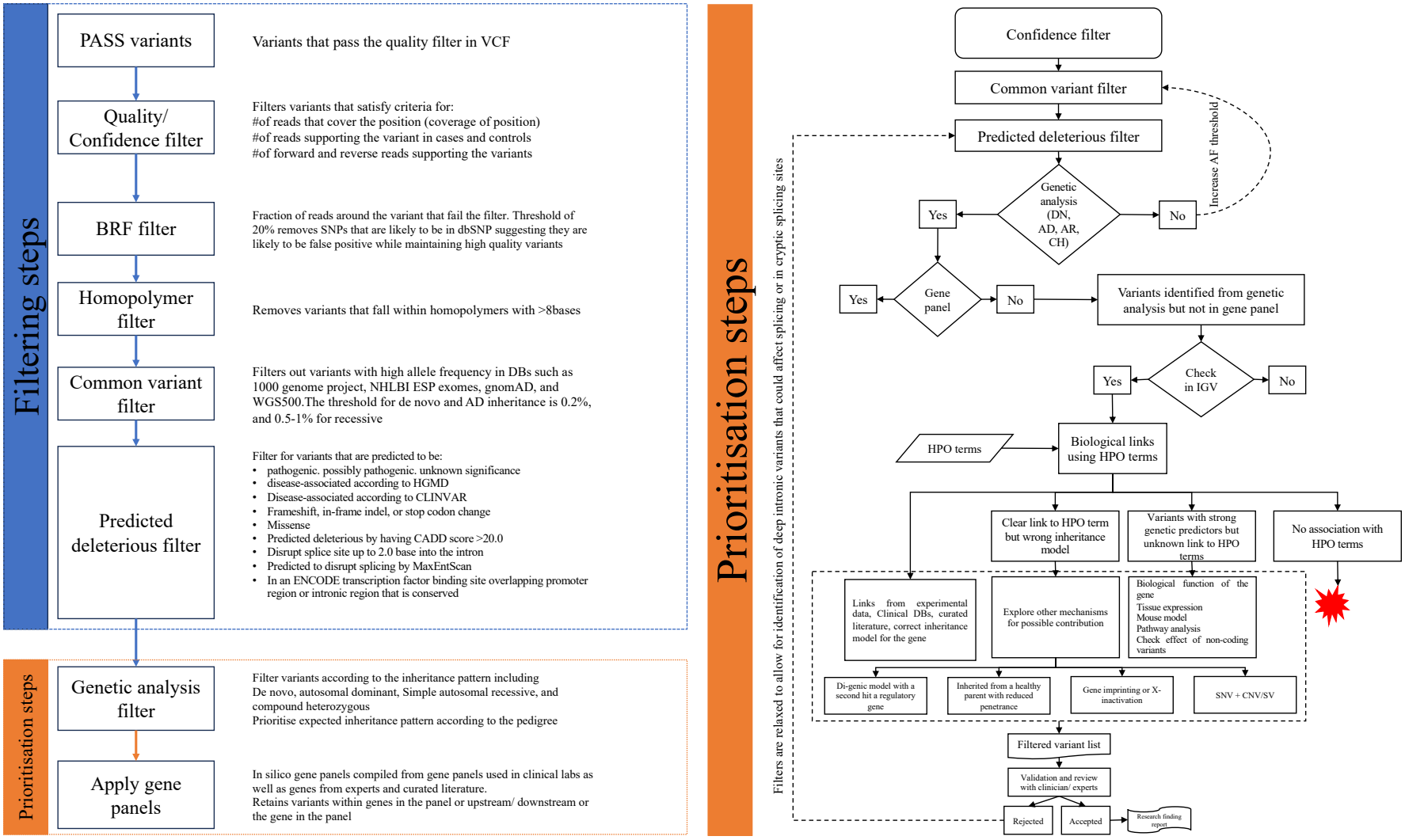


Figure 2.2: Dynamic variant filtering and prioritisation processes in the initial HICF2 pipeline.

Table 2.1: A list of HPO terms provided by the recruiting clinicians for the conditions discussed in chapter 3. These terms were utilised in the variant prioritization analysis to establish the biological connection between the variants detected and the patients' phenotypes. This was accomplished by searching databases for documented links between the genes containing deleterious variants and diseases that involve the expression of the observed phenotypes. The reported connections are curated by expert scientist who review the literature periodically and update the IVA database.

Condition	Consanguinity	Age of onset	Sex	Ethnicity	HPO Id	HPO terms
Aicardi syndrome (Sec 3.2.1.1)	No	20 days	Female	Caucasian	HP:0012469 HP:0002079 HP:0001263	Infantile spasms Hypoplasia of the corpus callosum Global developmental delay
X-linked neurodevelopmental disorder (Affected male proband- sec 3.2.2)	No	20 months	Male	Mixed: Asian/Caucasian	HP:0001252 HP:0001324 HP:0004305 HP:0011968 HP:0100022	Muscular hypotonia Muscle weakness Involuntary movements Feeding difficulties Abnormality of movement
X-linked neurodevelopmental disorder (Affected male sibling- sec 3.2.2)	No	8 months	Male	Mixed: Asian/Caucasian	HP:0001252 HP:0001324 HP:0004305 HP:0011968 HP:0100022	Muscular hypotonia Muscle weakness Involuntary movements Feeding difficulties Abnormality of movement
DEE23 (Sec 3.2.3)	Yes	7 yrs-old	Female	Lebanese	HP:0001250 HP:0010864 HP:0000252	Seizure Intellectual disability (severe) Microcephaly
ODLURO (3.2.4)	No	6 yrs-old	Male	Caucasian	HP:0010864 HP:0025160 HP:0000256 HP:0001883 HP:0000023	Intellectual disability, severe Abnormal temper tantrums Macrocephaly Talipes Inguinal hernia

Condition	Consanguinity	Age of onset	Sex	Ethnicity	HPO Id	HPO terms
Jeavons Syndrome (Male proband- sec 3.3.1)	No	12 yrs-old	Male	Caucasian	HP:0002069 HP:0012000 HP:0012001 HP:0010852 HP:0002392 HP:0025097	Generalised tonic-clonic seizures EEG with generalised spikes EEG with generalised poly-spikes EEG with photoparoxysmal response EEG with polyspike wave complexes Eyelid myoclonus
Jeavons Syndrome (Affected female sibling - sec 3.3.1)	No	8 yrs-old	Female	Caucasian	HP:0002069 HP:0012000 HP:0012001 HP:0010852 HP:0002392 HP:0011149	Generalised tonic-clonic seizures EEG with generalised spikes EEG with generalised poly-spikes EEG with photoparoxysmal response EEG with polyspike wave complexes Absence seizure with eyelid myoclonia
Jeavons Syndrome (sec 3.3.2)	Yes	18 yrs-old	Female	Asian	HP:0000347 HP:0000486 HP:0000664 HP:0000708 HP:0001249 HP:0001263 HP:0001382 HP:0002019 HP:0002069 HP:0002121 HP:0002197 HP:0002353 HP:0002360 HP:0002373 HP:0002384 HP:0007359 HP:0011149 HP:0011169 HP:0011968 HP:0100512	Micrognathia Strabismus Synophrys Behavioural abnormality Intellectual disability Global developmental delay Joint hypermobility Constipation Generalized tonic-clonic seizures Absence seizures Generalized seizures EEG abnormality Sleep disturbance Febrile seizures Focal seizures with impairment of consciousness or awareness Focal seizures Absence seizures with eyelid myoclonia Generalized conic seizures Feeding difficulties Vitamin D deficiency

2.5.4.2 Exomiser analysis

The multi-sample VCF for each family was also analysed separately by Exomiser (following the Exomiser protocol using the hiPHIVE algorithm) for phenotype comparison to confirm candidate variants identified by IVA and to ensure that there were no better candidates (Smedley et al., 2015b). Exomiser prioritises the variants in the multi-sample VCF files according to a variant's quality, inheritance pattern, and patient's clinical phenotype. The algorithm determines the candidate variants according to protein interaction networks (interactome), clinical relevance and cross- species phenotype comparisons in Mouse and Zebra fish models (Smedley et al., 2015). Read alignments supporting this final set of candidates suggested by Exomiser were viewed manually in IGV.

2.6 Validation of identified causative variants

The prioritised potentially causative variants in each proband/ affected sibling were validated using various techniques. I performed PCR or RT-PCR and dideoxy sequencing (unless otherwise stated) for SNVs and splice site variants, MLPA (kit P189-C2, MRC Holland, performed by Dr. Davide Mei in the neurogenetics laboratory, Meyer University Hospital) in the case of ARX discussed in chapter 3 section 3.2.1.1, or BioNano optical mapping for the complex structural variant discussed in chapter 3 section 3.2.2. The primers used for PCR amplification are listed in Table 2.2 and were designed using Primer3 to amplify the genomic region surrounding the putative pathogenic variant and ordered from Eurofins genomics.

2.6.1 PCR and dideoxy sequencing

A 50ng aliquot of the patient's DNA was used as a template for PCR amplification using FastStart Taq DNA polymerase kit (Roche, Cat. # 12 032 902 001) according to the following conditions: 2.5µl 10X Buffer with MgCl2, 0.25µl Taq polymerase (5U/µl), 0.5µl dNTP (10mM each), 2µl primer mix (10pM of each primer), 1µl DNA template (at 50 ng/µl), 5µl 5X CG rich solution (only when required), and water to make a final volume of 25µl. The GC rich solution was only added when the first reaction failed or when the region to be amplified was GC rich. The PCR amplification was performed using a 96X Universal peqSTAR thermal cycler (VWR, Cat. # 732-2887), and the cycling parameters were:

denaturation step at 95ºC for two minutes, followed by 35 amplification cycles of 30 seconds at 95ºC, 30 seconds at 55ºC, and 30 seconds at 72ºC, followed by six minutes at 72ºC and hold at 4ºC. The PCR product was purified by ExoSAP reaction using Exonuclease I (Exo I; New England Biolabs, Cat. # M0293S) and Shrimp Alkaline Phosphatase (rSAP; New England Biolabs, Cat. # M0371S) to digest primers and unused nucleotides. The ExoSAP reaction was set up by adding 1µl of 10x Buffer, 0.1µl of Exo I, 1µl of rSAP, and 0.9µl of water to 15µl of PCR product. Samples were incubated at 37 °C for 60 min, 85 °C for 15min and then held at 4°C in a thermal cycler. The purified PCR product was sent to Eurofins Genomics where the dideoxy sequencing was conducted. The sequence files (*.ab1) generated were analysed using SnapGene software (from Dotmatics, available at snapgene.com).

Table 2.2: Validation primer list. List of primers used for the validation of the putative pathogenic candidates for each case discussed in chapter 3 and 5

#	Individual/ section	Primer	Tm (°C)	Size (bp)	Cycling conditions
1	DOCK7 (section 3.2.3)	a: 5'-GTTTGTGGATGGGAATAATGTCA-3', b: 5'-TGTTCATGAAGTTCCCCATGG-3'	58	587	Standard cycling conditions as described in 2.6.1
2	*KMT2E* (section 3.2.4)	a: 5'-TGTATCACCCCATGCAAGTG-3', b: 5'-ATCAGCTTAGCCTTGGGAGA-3'	57	298	
3	*PLPPR5* (section 3.3.1)	Maternal variant: b: 5'-CCAGCGAGTAGAGGAGCAC-3', b: 5'-CTCTTCCTCCAGCGGCTC-3'	56	367	
		Paternal variant: a: 5'-TGCAAACAGTTAACATGTGGG-3', b: 5'-ACGGACAAGGCTATAATCAAAGG-3'	56	390	
4	*MGRN1* (section 3.3.2)	a: 5'-GTTGCCTGTTGGCCGTTT-3', b: 5'-GAAACAAAGGGGAAAGTAGCC-3'	57	578	
		RT-PCR: a1: 5'-TTCAGGCATCTCGGACGG-3', b1: 5'-GTGACGACGACTCATCAACCTCT-3' a2: 5'- TCTCCGAGGACGTGGACG-3', b2: 5'- GGTCTGATTCCAGCGACAGAAG-3'	60	-	
5	*MECP2*- I5_5 (section 5.4.3.1)	MECP2_Afii: 5'-TGCAAATAATTCTAAGCTGTCCC-3', MECP2_DF: 5'-GCCACCCACAAGTCTCCTA-3'	57		
6	*KMT2E*- I6_5 (section 5.4.3.1)	a: 5'-ACATTTACGCTTGGAAATTAAAA-3'; b: 5'-GCTCTCTGATAACTCCTTTCTCTGA-3'	57	649	
7	Translocation- I13_5 (section 5.4.7)	FOXG1_BND1: a: 5'-TGCCTCCTTTTACTTCCTACCTC-3' ; b: 5'-TTTCTCCTGCCCCTATAGCC-3'	57	498	
		FOXG1_BND2: a: 5'-CTCATGGCAGTTCTGGATAGC-3'; b: 5'-TGGGGGCCTTGTAACAGATA-3'	57	394	
8	Translocation- I14_5 (section 5.4.8)	I14_5_bp_1: a: 5'-GCCACTGCACTCGCACTC-3'; b: 5'-AAAAACACTGTATGGTCAGTCT-3'	54	389	
		I14_5_bp_2: a: 5'-GGAATCTCCTTGTAGATGTCT-3'; b: 5'-ACTAGAAGACGTCCAAGCAG-3'	53	460	
		I14_5_bp_3: a: 5'-ACACTGCAAAAAGGCAGCTG-3'; b: 5'-CATAAAGCAATAGAGTCCAGC-3'	54	427	
		I14_5_bp_4:a: 5'-TTTATCTAATTCTCTCCAAGCAT-3'; b: 5'-ATTCTGTGATGGTCCCACTG-3'	55	363	
		I14_5_bp_5: a: 5'-CGCTGTCTCCTAGGCTGG-3'; b: 5'-AATGACAAGAGACAGTGATGG-3'	54	390	
		I14_5_bp_6: a: 5'-TCGAACGCCTGACCTCAGA-3'; b: 5'-CCAAGATGACCTGGCTGCT-3'	59	400	
		I14_5_bp_7: a: 5'-AATTGTCAGTTCAACTCTTCTG-3'; b: 5'-GCATATATCTTTGTAATAGAATGA-3'	50	320	
		I14_5_bp_8: a: 5'-TGGACTGAGATCACACAGCT-3'; b: 5'-GCATGCAAATGGGACTGATG-3'	57	496	
		I14_5_bp_9: a: 5'-CACTGAACTTGCATTCTCAGT-3'; b: 5'-ACACATAGGGTCTCAAGTTCT-3'	55	311	
		I14_5_bp_10: a: 5'-CTTTTCTGGGCAAAGTTGAGA-3'; b: 5'-AACAGTTTGTACATTTTCATTCC-3'	53	377	
		I14_5_bp_11: a: 5'-CTAAATAGTGCTTGAATGAATGA-3'; b: 5'-TGGGTGAGTGTGCAAGATAG-3'	53	283	

For the validation of a *de novo* deletion affecting *ARX* gene, a set of primers were designed (Figure 2.3) and tested but failed to generate PCR products, which is discussed in chapter 3 section 3.2.1.1. The validation was performed using MLPA kit (P189-C2, MRC Holland) in the Neurogenetics laboratory - Meyer University Hospital, following the manufacturer's instructions.

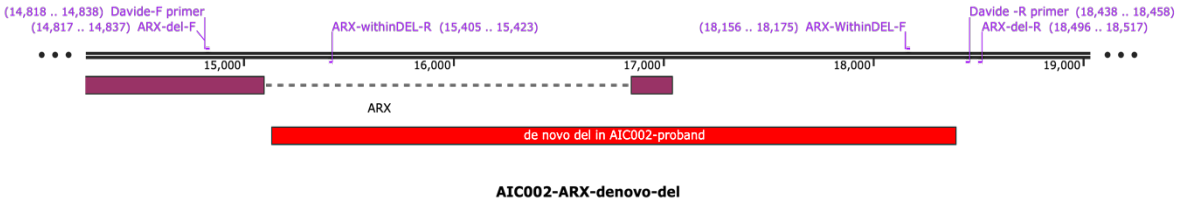


Figure 2.3: A schematic of the de novo deletion in ARX. The positions of the primers relative to the de novo deletion in the ARX gene identified in the proband of case 3.2.1.1 are shown in this schematic. Various combinations of these primers were tested, but no PCR product was observed for any pairs that spanned the deleted region in the proband. Due to the large size (~3.7Kb), the wild type allele was not expected to be amplified using these primer sets and PCR cycling conditions. These primers should theoretically amplify the mutant allele, which has a reduced size of 439 bp due to the deletion

2.6.2 BioNano optical mapping

This procedure was performed at the Bionano Services Laboratory in San Diego, USA.

2.6.2.1 Ultra-high molecular weight DNA extraction and labelling

For each sample, a minimum of 650µl of whole peripheral blood was used to purify ultra-high molecular weight (UHMW) DNA using the Bionano Prep SP Frozen Human Blood DNA Isolation Protocol v2, following manufacturer's instructions (Bionano Genomics, San Diego, USA). Briefly, white blood cells were counted, pelleted (2200g for 2min), and treated with LBB lysis buffer and proteinase K to release gDNA. After inactivation of proteinase K by PMSF treatment, gDNA was bound to a paramagnetic disk, washed, and eluted in an appropriate buffer. UHMW DNA was left to homogenise at room temperature overnight. The following day, DNA molecules were labelled using the Direct Label and Stain (DLS) Protocol (Bionano Genomics, San Diego, USA). Briefly, 750ng of gDNA were labelled in presence of Direct Label Enzyme (DLE-1) and DL-green fluorophores. After clean-up of the excess of DL-Green fluorophores and rapid digestion of the remaining DLE-1 enzyme by proteinase K, DNA backbone was counterstained overnight.

2.6.2.2 Data collection

Labelled UHMW gDNA solution (4-12ng/µl) was loaded on a Saphyr chip and imaged by the Saphyr instrument (Bionano Genomics, San Diego, USA). Saphyr chip was run at a target of 400Gbp to aim for 100X coverage.

2.6.2.3 Bionano Rare variant analysis and structural variant calling

The *de novo* assembly and Variant Annotation Pipeline were executed on Bionano Solve software v3.7. Reporting and direct visualisation of structural variants was done on Bionano Access v1.7. DNA molecules were first aligned against the human reference genome GRCh38, and SVs and CNVs were identified based on discrepancies in the alignment between the sample and the reference. For each SV and CNV call, confidence scores were calculated and provided by Bionano Genomics (Bionano Genomics, 2018). For data filtering, the following recommended confidence scores were used: Insertion/Deletion= 0; Inversion= 0.7; Duplication= -1; Intra-fusion/Inter-translocation= 0.05; Copy Number= 0.99; Aneuploidy= 0.95. Events detected by *de novo* assembly were subsequently

filtered against the OGM control database. Additional filtering by pairwise comparison (dual analysis) was performed to easily distinguish inherited and non-inherited variants.

2.6.3 Minigene assay

2.6.3.1 RHCglo plasmid DNA stock preparation

The RHCglo plasmid (~5.7 kb), was a gift from Thomas Cooper (Addgene plasmid #80169; http://n2t.net/addgene:80169; RRID: Addgene_80169; Figure 2.4: A). The plasmid incorporates several unique restriction sites, enabling easy "cut and paste" procedures.

In order to isolate the RHCglo plasmid, bacterial swabs were plated onto LB agar plates (Sigma-Aldrich) supplemented with ampicillin (Amp) and incubated overnight at 37°C. Single colonies were selected for inoculation of LB broth supplemented with Amp and grown overnight at 37°C (~16 hours). The plasmid extraction was performed using the NucleoBond® Xtra Midi Kit (Macherey-Nagel; Cat. No. 740410.10/.50/.100). A schematic of the RHCglo plasmid, depicting its structure and the included restriction sites for replacing an artificial exon with a gene segment is shown in Figure 2.4: B.

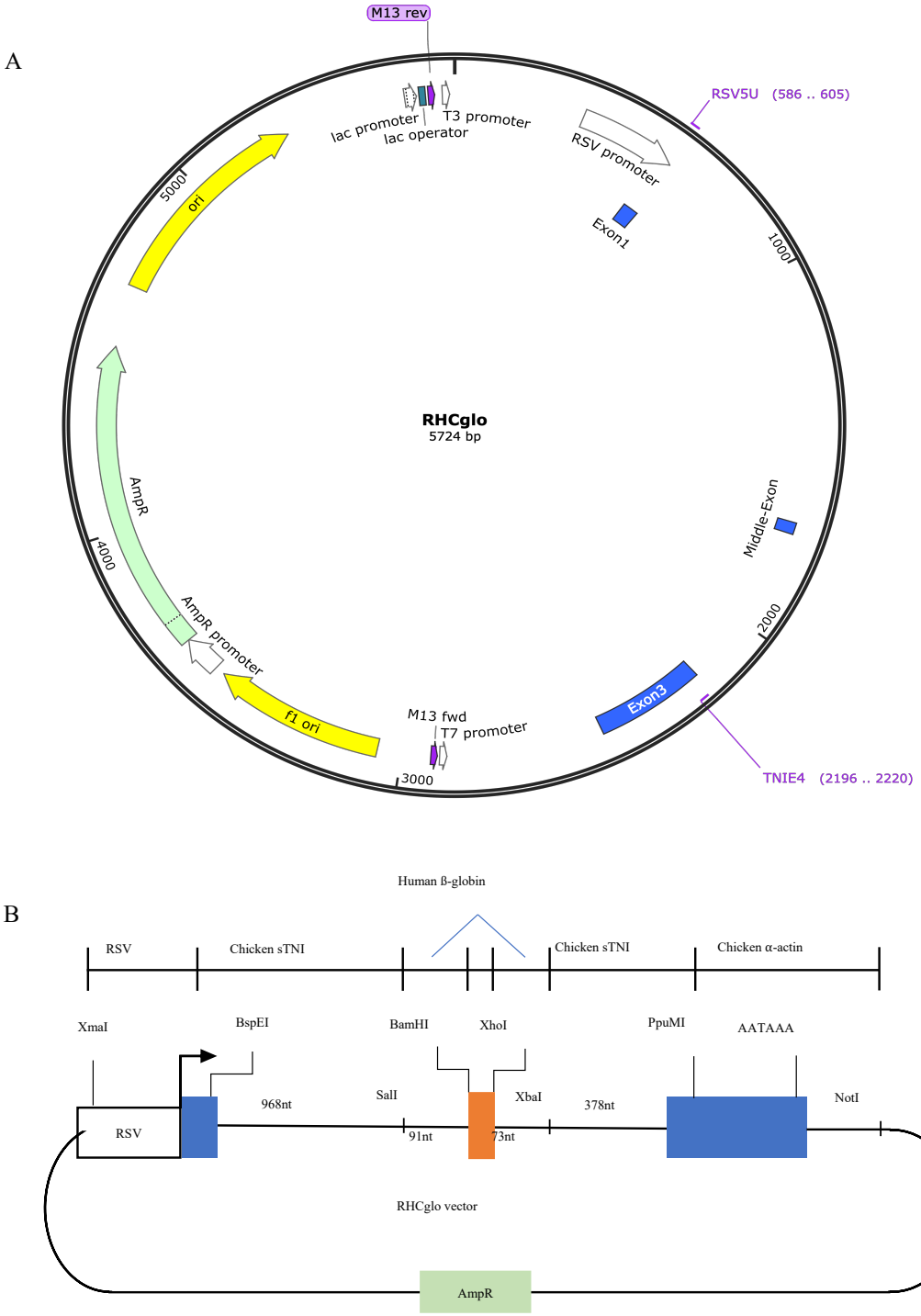


Figure 2.4: Figure 2.4: Diagram of RHCglo minigene reporter system. A) Map of the RHCglo showing the vector backbone and bacterial resistance to antibiotic (Ampicillin). B) a schematic showing the various segments with distinct restriction sites exclusive to the vector, adapted from Singh & Cooper, 2006. The middle exon is directly bordered by the last 91 nucleotides (nt) of human β-globin intron 1 on the upstream side and the first 73 nt of human β-globin intron 1 on the downstream side. Additionally, the distal upstream region of intron 1 encompasses introns 1 and 3 of chicken skeletal troponin I (sTNI), while the distal downstream region of intron 2 encompasses the final 364 nt of sTNI intron 3. It utilises the RSV, Rous sarcoma virus enhancer/promoter to drive expression of this artificial minigene.

2.6.3.2 Primer design

Control human genomic DNA was used as a template to generate the *MGRN1* (exon 14-15) insert with flanking intronic regions.

The primers (Table 2.3) were designed using SnapGene to generate XbaI and SalI restriction sites overhangs. The primers were synthesised by Eurofins Genomics and dissolved in deionised distilled H_2O.

Table 2.3:Primers used for cloning MGRN1 exons 14-15 into RHCglo.Primer sequences were modified to generate restriction sites for XbaI/SalI (red colour) on the 5' and 3'ends of the MGRN1 exons 14-15 with flanking intronic sequences.

Primer name	Primer sequence
MGRN1-F-Ex14-15-SalI	5'-ACGCGTCGACGTCGGCCATAGCGGCCGCGGAATCCCGGTGGCTGGTGATG-3'
MGRN1-R-Ex14-15-XbaI	5'-ATGGTACCCTGGAGGTGCTGTTCCAGG-3'

2.6.3.3 Amplification of MGRN1 exon14-15 insert using high fidelity Q5 polymerase

To perform the PCR, two reaction mixtures (50µl each) were prepared, each containing 10µl (5X) Q5 Reaction Buffer, 1µl of 10mM dNTPs, 10µM of both forward and reverse primers, 0.5µl of Q5 High-Fidelity DNA Polymerase, and water up to a final reaction volume of 50µl. The PCR reaction used 100ng control human genomic DNA as a template. The thermal cycle programme consisted of an initial denaturation step at 98°C for 30 seconds, followed by 35 cycles at 98°C for 10 seconds, 70.5°C for 30 seconds, and 72°C for 40 seconds, with a final extension step at 72°C for 2 minutes. PCR products were visualised using 1% agarose gel. Subsequently, the PCR products were purified using the Monarch PCR & DNA Clean-up Kit (New England BioLabs) following the manufacturer's instructions and 15µl eluates obtained. The concentration of PCR DNA was determined using a NanoDrop Spectrophotometer.

The amplified PCR product was digested with XbaI and SalI-HF (New England BioLabs; Cat. No. R0145S and R3138, respectively) to create sticky ends (Figure 2.5: A). The vector RHCglo was also digested either with XbaI and SalI-HF to create compatible sticky ends with which the insert sequence was to be ligated. The digestion reaction (20µl) was setup as follows: 1µg plasmid DNA or insert, 1µl of XbaI and 1µl of SalI, 2µl of 10X buffer, and water to make up 20µl. The reaction mixture was incubated for 1 hour at 37°C then 1µl of Shrimp Alkaline Phosphatase was added to the plasmid reaction only and incubated for further 30 minutes 37°C, and the digestion reactions terminated at 65°C for 20 minutes. The digested PCR product and vectors were electrophoresed on 1% agarose gel containing Ethidium Bromide (EtBr) for 60 minutes at 80 volts then extracted using the QIAquick Gel Extraction Kit (Qiagen; Cat. No./ID: 287049). The complete purification and correct sizes for the inserts and vectors were confirmed using DNA gel electrophoresis on 1% agarose gel containing EtBr (Figure 2.5: B) as previously described.

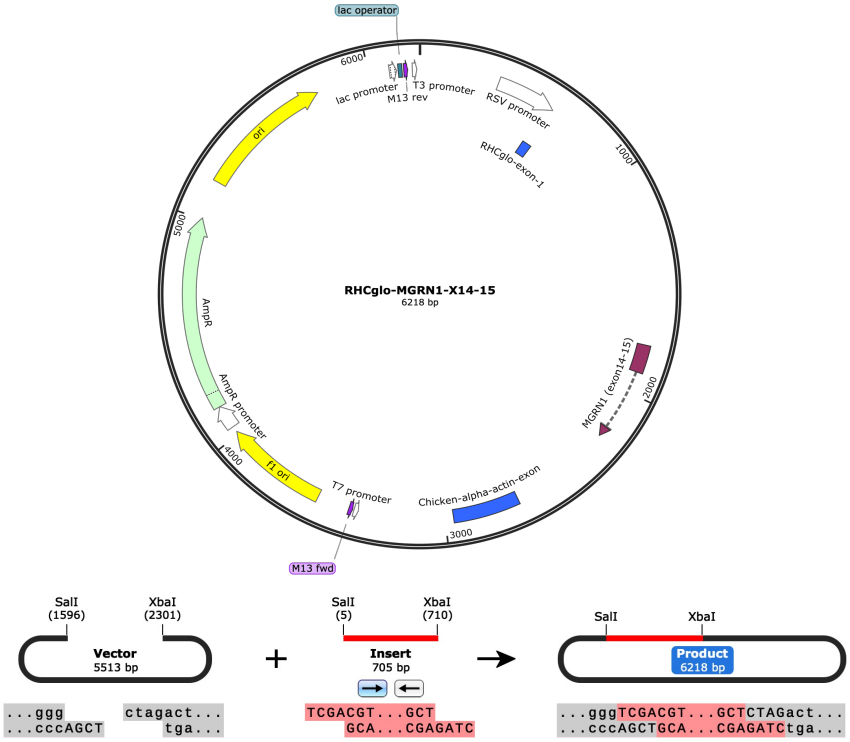
2.6.3.4 Ligation of RHCglo Vector and MGRN1 exons 14-15 Insert

The concentrations of the purified PCR product and vector were determined using Qubit 2.0 fluorometer with Qubit™ dsDNA HS Assay Kit according to the manufacturer's instructions. These values were used to calculate the amounts needed for 1:3 vector: insert ligation reaction using NEBioCalculator (https://nebiocalculator.neb.com/#!/ligation).

T4 DNA Ligase was used . 10µl reactions were set up by mixing 1µl of T4 DNA Ligase Buffer (10X), 75 ng vector DNA, 25ng insert DNA, 1µl of T4 DNA Ligase, and water to make up10µl then incubated at room temperature for 1 hour.

After the ligation reaction, 5µl of each reaction mixture was transformed into 50µl of DH5α *E. coli* competent cells by incubating on ice for 30 minutes. After 30 minutes, the reactions were heat shocked for 30 seconds at 42°C and immediately incubated on ice for 5 minutes. Each transformed mixture was spread evenly on prewarmed Ampicillin-enriched LB-agar plates and incubated overnight.

A



B

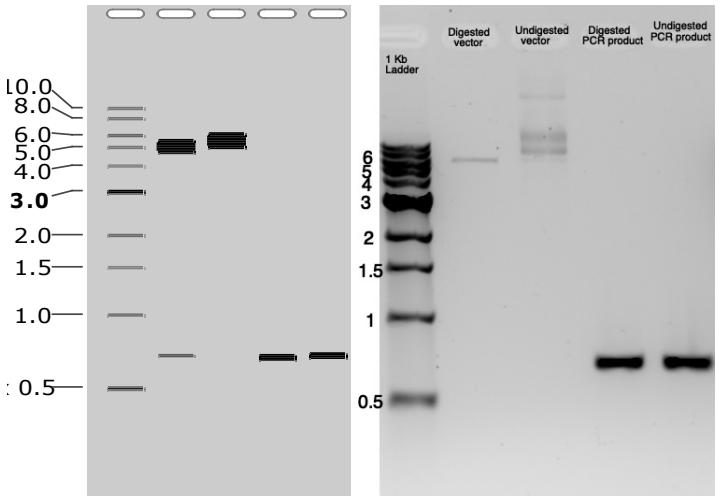


Figure 2.5: : Cloning strategy to generate RHCglo-MGRN1-Ex14-15 plasmid. A) Cloning strategy of exons 14-15 with flanking intronic sequence including 183bp upstream of exon 14 and 144bp downstream of exon 15 to ensure capture of regulatory regions required by the splicing machinery in the cell. B) Agarose gel (1%) simulation of the expected digest product sizes (left) and actual obtained digested/ undigested vector and PCR products.

2.6.3.5 Analysing Transformants

Four single colonies were selected from the plate to inoculate 3ml of Ampicillin containing LB medium. Each tube was incubated for 16 hours before making 50% glycerol stocks of the selected transformed colonies (stored at -80°C) and harvesting the cells for DNA isolation by Miniprep with the QIAprep Spin Miniprep Kit by following the manufacturer's instructions.

In order to confirm that these clones contained the correct inserts of expected size, the selected clones were digested using restriction enzymes that cut within the insert towards one of the ends, which confirms both the size and orientation of the insert. For the RHCglo-MGRN1-Ex14-15 colonies, 1µl of EcoRI (dual cuts in plasmid) and 1µl of HindIII (cuts into the insert) were used for double digestion as previously described in section 2.6.3.3. The expected sizes from the digestion of the RHCglo-MGRN1-Ex14-15 are 4.308 kb, 1.4. kb, and 0.5 kb (Figure 2.6). These clones were also sent for dideoxy sequencing using a primer positioned in the 5-prime UTR of RSV promotor (RSVU5_primer: 5'-CATTCACCACATTGGTGTGC-3') by Eurofins Genomics to verify the presence of the *MGRN1* exon14-15 insert and correct orientation. Results from sequencing were analysed with SnapGene to confirm ligation of *MGRN1* exon14-15 insert in RHCglo vector; all restriction sites were identified. Correct sequence of the insert was confirmed, and no mutations were introduced by the amplification step of the insert. To prepare for subsequent experiments, bacterial overnight cultures were generated for Midiprep using the NucleoBond Xtra Midi kit and for glycerol stock of the RHCglo-MGRN1-Ex14-15 plasmid.

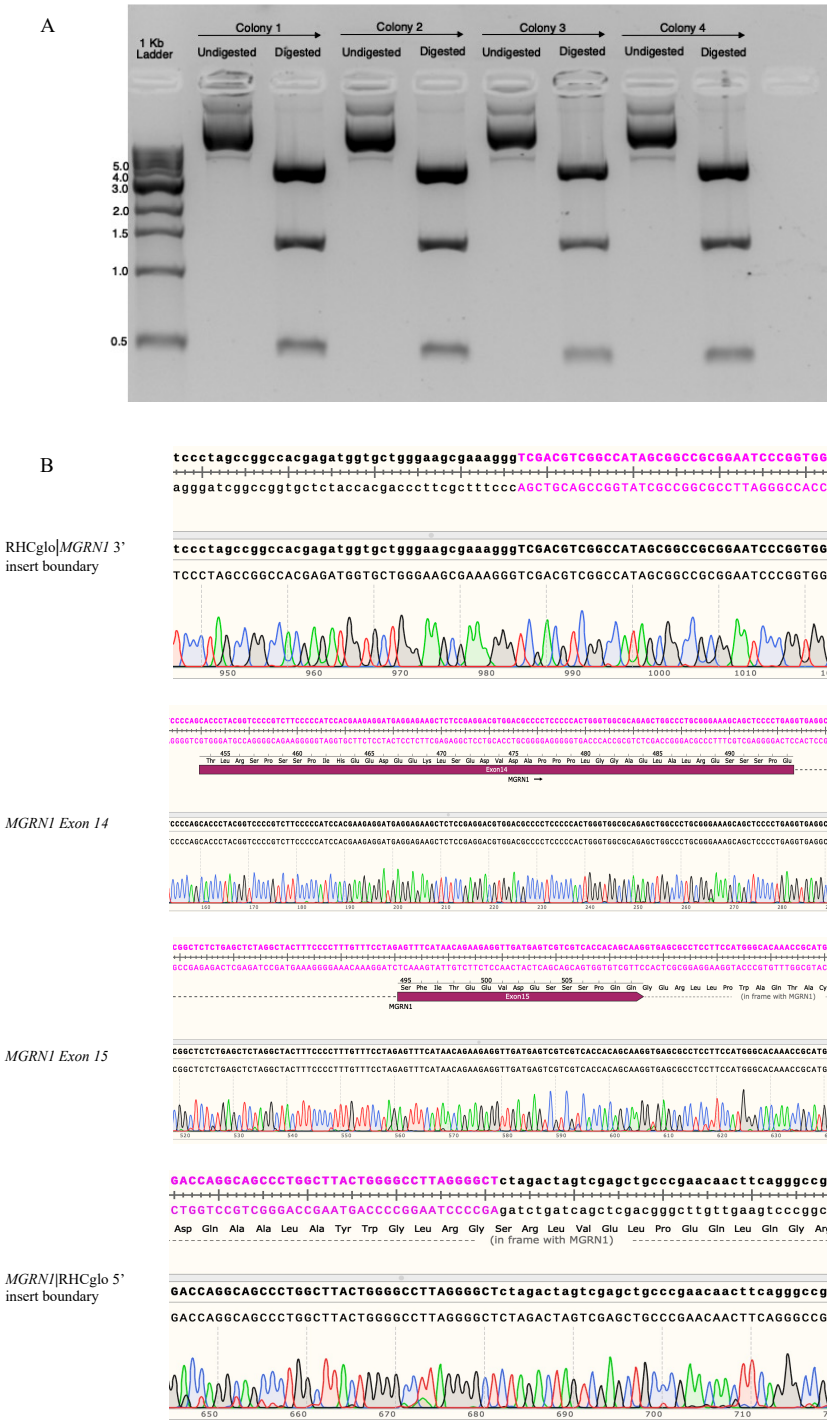


Figure 2.6: Successful cloning of MGRN1 exons 14-15 into RHCglo. A) Double digest strategy to validate the presence of the MGRN1 exon14-15 insert and correct orientation. B) dideoxy sequencing of selected colonies to verify the correct ligation of the insert into the vector using RSVU5 primer, which allowed confirmation of the boundaries for insertion as well as the sequence of the insert, ensuring that no mutations had been introduced during the amplification step.

2.6.3.6 Site-directed mutagenesis

Site-directed mutagenesis based on PCR was used to generate the patient's identified donor splice site variant in *MGRN1* (NM_015246.4(*MGRN1*); c.1482+1G>A). An oligo containing the specific mutation was designed using SnapGene (5'-CAGCTCCCCTGAGATGAGGCCCCCCCG-3') then ordered from Eurofins Genomics. The oligo was resuspended using double-distilled water (ddH_2O) to obtain 100ng/µl of the mutagenesis primer. The mutant strand was synthesised by PCR using QuikChange Lightning Site-Directed Mutagenesis Kit. The PCR was setup in a 25µl reaction mixture containing 2.5µl of 10×QuikChange Lightning Reaction Buffer, 0.75µl of QuikSolution reagent, 30ng of dsDNA template (plasmid), 1µl of the mutagenesis primer (100ng/µl), 1µl of dNTP mix, 5µl 200µM dNTPs, 1µl QuikChange Lightning Enzyme Mix, and PCR-grade water to a final volume of 25µl. The cycling parameters were set to 1 cycle at 95°C for 2 minutes, followed by 30 cycles of 95°C for 20 seconds, 55°C for 30 seconds, and 65°C for 3 minutes 30 seconds (30 seconds/kb of the ~7Kb plasmid), and x1 cycle at 65°C for 5 minutes. Once the PCR reaction was finished, 1µl of the *DpnI* provided in the kit was added to the PCR product and incubated at 37°C for 5 minutes. *DpnI* will digest the parental methylated and hemi-methylated DNA, leaving only the mutated DNA intact.

To increase the transformation efficiency, 2µl of β-mercaptoethanol were added to XL10-Gold ultracompetent cells prior to the addition of 2µl of the *DpnI*-treated DNA. The treated cells were incubated on ice for 30 minutes then Heat-pulsed in a 42°C water bath for 30 seconds then incubated on ice for 2 minutes. The cells were incubated in at 37°C for 1 hour with shaking at 225x RPM after the addition of 0.5ml of preheated LB broth to 42°C in a water bath. The cell suspension was briefly centrifuged at 2000xg for 1 minute to remove the supernatant, leaving 100µl for cell resuspension to be plated on Ampicillin-enriched LB-agar plates, and incubated overnight at 37°C. Single colonies were used for overnight culture and plasmid was purified using the GeneJET Plasmid Miniprep Kit according to the manufacturer's protocol. The presence of the mutation was verified by sequencing the purified plasmid DNA using the Eurofins Genomics dideoxy sequencing service and visualised using SnapGene (Figure 2.7).

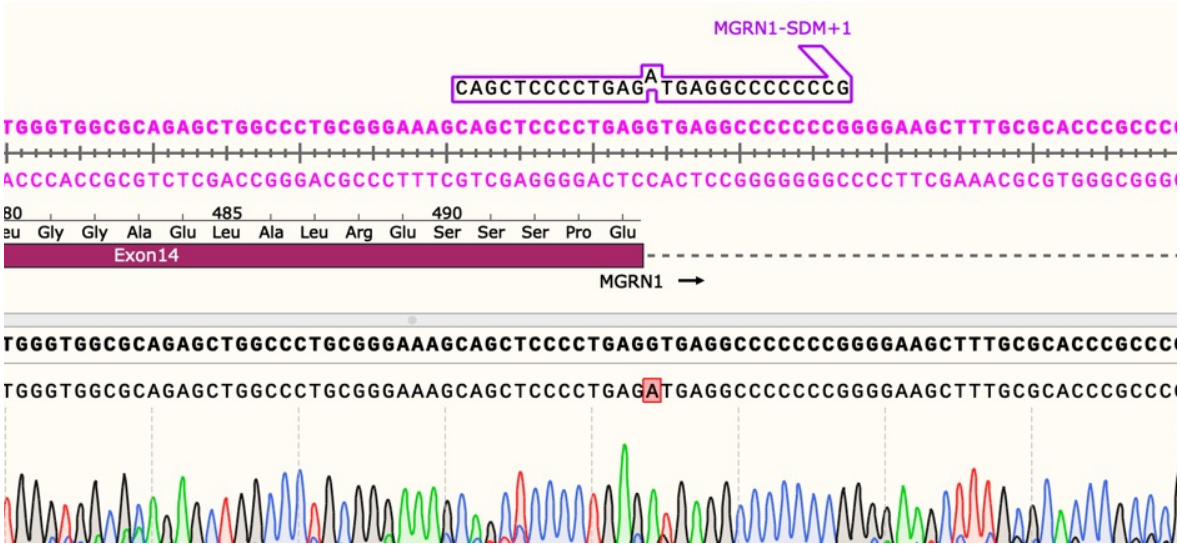


Figure 2.7: Validation of successful SDM. Sanger sequence trace of one of the selected colonies showing the presence of the patient specific mutation generated from the SDM step and the sequence of the mutagenic primers used for generating the mutation.

2.6.3.7 Cell culture

Human embryonic kidney 293 cells (HEK293-T) were maintained in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 2mM L-Glutamine, and 1% penicillin/streptomycin. The cell line was verified to be free of mycoplasma contamination.

2.6.3.8 Transfection, Sample preparation, RNA extraction, and RT-PCR

For the transfection, cells were dissociated using 0.25% trypsin-EDTA, counted and seeded at 5×10^5 cells/well using 6-well format and incubated overnight in a humidified incubator at 37°C, 5% CO_2. The 50-60% confluent cells were subsequently transfected with 1µg of RHCglo-MGRN1-Ex14-15 plasmid DNA (wild type or mutant) using Lipofectamine 3000 following the manufacturer's protocol. The cells were incubated with the medium containing the transfection mix for 24 hours then the cells were washed with ice-cold PBS, scraped with cell scrapers, and collected on 1.5 Eppendorf microtubes and total RNA was extracted with commercially available RNeasy Mini Kit. RNA pellets were dissolved in RNAse-free water and the concentration was determined using a NanoDrop 2000 spectrophotometer.

One microgram of total RNA was reverse–transcribed to cDNA using QuantiTect Reverse Transcription Kit according to the manufacturer's instructions. A modification to the manufacturer's protocol was to use 0.5µl of the kit primer Mix and 0.5µl of the downstream primer TNIE4RT (10 uM; 5'-AGGTGCTGCCGCCGGGCGGTGGCTG-3') to prime reverse transcription.

The PCR was performed using 2µl of the resulting cDNA as a template with the primer pair RSV5U(5'-CATTCACCACATTGGTGTGC-3') and TNIE4RT(5'-AGGTGCTGCCGCCGGGCGGTGGCTG-3') as described in section 2.6.3.3. The PCR products were visualised using a 2% agarose gel.

2.7 Whole transcriptome sequencing

RNA-seq was used to assess the effect of homozygous splice donor site variants identified in the probands and unaffected parents. This is relevant to cases discussed in sections 3.2.3 and 3.3.2.

2.7.1 RNA processing and quantitation

Whole peripheral blood (2.5 mL/tube) was collected into PAXgene tubes (Qiagen/Preanalytix, Hombrechtikon, Switzerland) from the proband, unaffected parents, age and gender matched controls. I extracted total RNA using the PAXgene Blood RNA kit (Qiagen Cat. No. 762174) according to the manufacturer's instructions. RNA purity was assessed using the 260/280 ratio detected on a NanoDrop 2000 Spectrophotometer (ThermoFisher Scientific Cat. # ND-2000). All samples had a ratio of 2.0-2.12 showing that the RNA was pure. The RNA quantity was assessed using Qubit 2.0 and ranged between 33-100 ng/ul. RNA quality was then assessed on an Agilent Bioanalyser 2100 using RNA 6000 Pico Total RNA Kit (Agilent Technologies). The RNA integrity number (RIN) score ranged between 7.5-9 indicating good integrity of RNA.

2.7.2 Library preparation, sequencing, and data processing

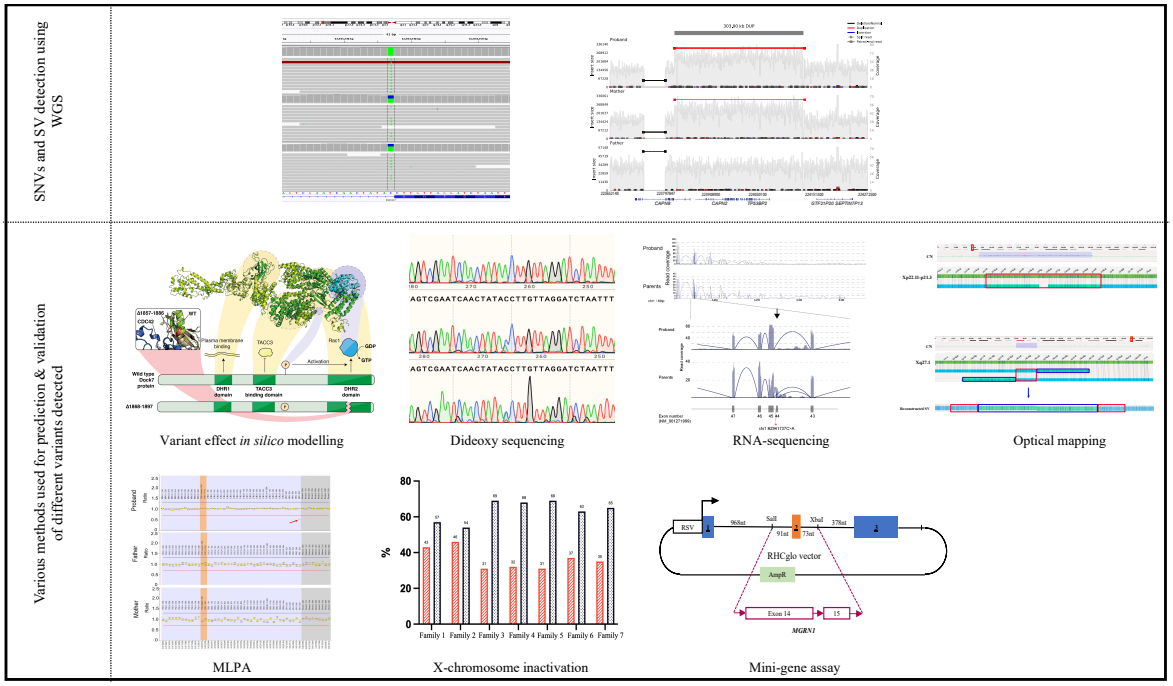
The library preparation was performed using TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero (Illumina) and haemoglobin depletion according to the manufacturer's instructions and sequencing was performed on the HiSeq-4000 by the Oxford Genomics Centre. The FASTQ files were aligned to the reference genome GRCh37.EBVB95 using HISAT2. The BAM files were sorted and indexed using Samtools and duplicates were removed using Picard markduplicates. Quality control checks were done using GATK/RNA-SeQC v1.1.8.1 module. Count tables for reads aligning to gene features from BAM alignment were generated using featurecounts. Visualisation of transcripts and splice junctions was performed either using IGV v2.11.2 or SplAdder v2.2.1 (Kahles et al., 2016).

2.8 Molecular mechanics

The molecular mechanics analysis for DOCK7 protein discussed in chapter 3 section 3.2.3 was performed by Dr. Matteo Ferla, a postdoc in the Taylor group. The change in thermostability was predicted *in silico*. The Dock7 model was initially made with I-Tasser with alterations based on the crystal structures PDB:6AJ4, a structure of a complex of the DHR-2 domain of DOCK7 and CDC42 (Kukimoto-Niino et al., 2019a) and PDB: 2WMO (J. Yang et al., 2009), but subsequently a ColabFold (Mirdita et al., 2022) model of Dock7-Cdc7 was used. To predict the effect of the in-frame slippage of exon 44, PyRosetta Remodel was used to remove the residues 1868-1897 and reposition residues 1865-1903 (Huang et al., 2011), this was followed by 15 cycles of FastRelax with the scorefunction (Nivón et al., 2013). Twenty replicates were performed to find the lowest energy state. To verify that the *in-silico* experiment converged on the solution, the same number of replicates was performed without the deletion. Interactive visualizations were made in MichelaNGLo (Ferla et al., 2020) available at https://michelanglo.sgc.ox.ac.uk/r/dock7.

Graphical abstract chapter 3

Combinatorial genomic approach was employed to detect, characterise, and validate variants, which facilitated identification of new disease-causing genetic associations and hard-to-detect variants in individuals with neurodevelopmental disorders and epilepsy. These included RNA-seq, Optical mapping sequencing, and minigene assay in addition to routinely used methods such as dideoxy sequencing, MLPA, and X-chromosome inactivation. This multifaceted approach not only revealed novel genetic associations but also highlighted the challenges of integrating these advanced methods into routine clinical practice. Despite these challenges, such an approach holds promise for a deeper understanding and providing more precise diagnoses.



Chapter 3. Combinatorial genomic approaches for novel gene discovery in neurodevelopmental disorders and epilepsy

This chapter presents the genetic analysis results for undiagnosed patients with neurodevelopmental disorder and epilepsy phenotypes recruited to the Health Innovation Challenge Fund 2 (HICF2) project or the 100,000 Genomes project (100KGP) with Jeavons syndrome. Some of the findings in this chapter are published in (Pagnamenta et al., 2023; and O'Donnell-Luria et al., 2019).

3.1 Introduction to Neurodevelopmental Disorders

WGS has revolutionised genetics, allowing more patients with rare diseases to receive a molecular diagnosis than ever as more genes are tested in clinics. Furthermore, it has opened up new opportunities for discovering novel disease-causing genes and providing new insights into the molecular mechanisms underpinning rare and complex disorders. However, despite the vast amount of data generated by WGS, the diagnostic yield for patients remains relatively low, with up to 43-70% of patients with rare diseases lacking a definitive molecular diagnosis (Taylor et al., 2015; Lionel et al., 2018; Seaby et al., 2022).

One challenge that may limit the diagnostic yield of WGS is the presence of variants in disease-causing genes within "dark and camouflaged" regions that are not captured or represented well in sequencing data (Ebbert et al., 2019). These regions include those difficult to map or align to the reference genome using standard short-read sequences. The difficulties arise either from being inherently problematic to sequence due to high GC content leading to few mapped reads, or easy to sequence but difficult to map, presenting a bioinformatic challenge. The latter difficulty is usually associated with regions of low complexity and was described in a recent study in which a systematic analysis of short-read sequencing highlighted the impact of such regions in missing the detection of mutations across 76 "dark" genes associated with 326 human disease phenotypes in OMIM (Ebbert et al., 2019). Therefore, standardised pipelines that utilise a combination of tools that does not only rely on coverage for structural variants identification should be used. This is especially vital for large

sequencing projects for patients with rare diseases, to mitigate challenges faced with dark and camouflaged regions.

The low diagnostic yield of patients with rare diseases also stems from insufficient knowledge of disease-causing genes. Therefore, there is a continuous need for novel gene discovery for which WGS provides a suitable platform.

One significant barrier to novel gene discovery is using *in silico* gene panels for clinical genetic reporting. This strategy mimics experimental gene panel sequencing for clinical genetic testing. These *in silico* panels, typically composed of a pre-selected set of genes associated with specific disorders or phenotypes, are often used as the primary test or as the basis for analysis strategy applied to exome and genome sequencing (Green et al., 2013). While this approach holds immediate clinical relevance, it imposes limitations on the analysis, resulting in a reduced diagnostic yield. Additionally, it heightens the risk of failing to identify rare disease-causing variants in newly discovered genes. This is due to the exclusion of the vast majority of the genome from the analysis, which contradicts the purpose of conducting WGS.

Moreover, panels for neurodevelopmental disorders vary significantly between the different clinical laboratories. Examples are the panels offered for targeted testing for autism spectrum disorder in 21 clinical settings ranging from panels containing 11 to those containing 2,562 genes. The only gene present across all ASD-related panels compared was *MECP2* (Hoang et al., 2018). The low diagnostic yield of static gene panels has been quickly realised, and WES became the first-line testing for various neurodevelopmental disorders. Nonetheless, it does not come without limitations, which include issues with structural rearrangements and non-coding variants. Therefore, WGS is considered a more reliable method for identifying genetic variants in these underrepresented regions, ultimately leading to a more conclusive molecular diagnosis. However, most analytical pipelines for WGS data lack a comprehensive analysis of such variants, which have been shown to play a significant role in the pathogenesis of rare diseases (Vaz-Drago et al., 2017). This restriction to the coding region of the genome can lead to omitting important variant types that may hold the key to novel gene discoveries.

Adopting a holistic and experimental approach to analysing WGS data is essential to overcome these barriers and improve diagnostic yield. This includes using complementary technologies such as RNA sequencing and protein informatics to determine the pathogenicity of identified variants and the systematic analysis of structural, non-coding and splice site variants. Such an approach will be critical for identifying new disease-causing genes and developing novel diagnostic and therapeutic options for patients with neurodevelopmental disorders.

The low frequency of disease-causing variants in any individual rare disease poses a significant challenge to novel disease-gene discovery. This is primarily due to the small sample sizes available for a particular condition. Nevertheless, this hurdle can be surmounted by leveraging large databases, international collaborations, and data sharing initiatives. Large-scale databases like the 100KGP and the DDD, along with online services and testing companies that facilitate collaboration among researchers and clinicians, such as GeneMatcher and GeneDx, offer an extensive wealth of information. This vast repository of genetic data plays a pivotal role in identifying novel genetic causes, developing new diagnostic tools, and improving existing treatments. Therefore, multi-disciplinary collaborations can help overcome some of the challenges associated with the rarity of the diseases, such as the difficulty in identifying patients with similar phenotypes or genotypes.

3.2 Neurodevelopmental Disorders: Findings from Whole Genome Sequencing

3.2.1 Aicardi syndrome

One of the disorders that continue to defy the advances in scientific knowledge and technologies is Aicardi syndrome, which was initially described briefly in 1965 and more fully in 1969. Aicardi syndrome (AS) is a rare, dominant disorder postulated to be X-linked since it mainly affects females and XXY Klinefelter males (Hopkins et al., 1979). It is characterised by a triad of chorioretinal lacunae, agenesis of the corpus callosum, and infantile spasms (Aicardi et al., 1965). These three features are the primary hallmarks of the syndrome. Nevertheless, a majority of individuals diagnosed with Aicardi syndrome commonly exhibit developmental delay and intellectual disability. However, there have been reports of cases where individuals with mild or no intellectual disability or developmental delay have been observed (Wong & Sutton, 2018). Moreover, the clinical phenotypic spectrum and severity are continuously expanding to include other features such as cystic

formations around third cerebral ventricle and/or choroid plexus, choroid plexus papillomas, coloboma, as well as lacunae, and focal seizures rather than spasms (Aicardi, 1999). Given that AS diagnosis is based on clinical presentation, which is continuously expanding, revised clinical diagnostic criteria were suggested. Patients with the triad are diagnostic of AS while the presence of two of the classic triad in addition to at least two other major or supporting features is strongly suggestive of AS diagnosis. The major features associated with AS include brain malformations (mostly polymicrogyria), periventricular and subcortical heterotopia, cysts around the third cerebral ventricle and/or choroid plexus, optic disc/nerve coloboma or hypoplasia. The supporting features include vertebral and rib abnormalities, microphthalmia, "Split-brain" electroencephalogram (EEG), gross cerebral hemispheric asymmetry, vascular malformations, or vascular malignancy. It has an estimated occurrence of 4000 cases worldwide, but the exact prevalence is unknown (Kroner et al., 2008).

The implied hypothesis is that a *de novo* mutation on the X chromosome causes the development of Aicardi syndrome in females but is lethal to hemizygous males (Van-Den-Vyver, 2002). The assumption that the genetic cause of AS is a *de novo* mutation stems from the fact that all AS cases are sporadic, including two affected sisters and one monozygotic twin pair, and no familial cases have been reported (Molina et al., 1989; Pons & Garcia, 2008). Moreover, females affected by AS have not been reported to be reproductively active; hence cannot pass on the pathogenic variants to their offspring; therefore, every case is expected be caused by a *de novo* variant (Donnenfeld et al., 1989). Apart from the inheritance pattern suggesting that AS is an X-linked condition, a few studies of AS cases have shown the involvement of deletions and translocations of the short arm of the X chromosome in the development of AS clinical phenotypes. These studies have narrowed the involved region within the X chromosome to Xp22.2-Xp22.3 using clinical and cytogenetic data from several previous studies on the distal short arm of the X chromosome (Ballabio & Andria, 1992). The notion of AS having an X-linked association gained further support through X-inactivation studies. These studies revealed that females afflicted with AS exhibit a skewed, non-random X-inactivation pattern in genomic DNA isolated from peripheral blood (Neidich et al., 1990). Notably, this skewed X-inactivation significantly differs from the random pattern seen in the healthy population; $p < 0.0001$ (Eble et al., 2009). However, diverse findings in the literature,

including discordant monozygotic twins, challenge the concept that AS development is linked to skewed X-inactivation (Costa et al., 1997; Hoag et al., 1997). These reports also derive from X-inactivation patterns detected in genomic DNA samples from either EBV-transformed lymphoblastoid cell lines or cultured skin fibroblasts. While these conflicting results regarding X-inactivation patterns in AS patients neither confirm nor dismiss the role of X-inactivation in AS development, they could provide insight into the observed variability in clinical severity among AS patients. (Neidich et al., 1990).

To date, no genes have been confirmed as contributing to the onset of AS even though a few have been proposed. These suggested genes included ones located on the X chromosome such as Filamin A (*FLNA*), and Cyclin Dependent Kinase Like 5 (*CDKL5*), as well as autosomal genes such as TEA Domain Transcription Factor 1 (*TEAD1*), and Occludin/ELL Domain Containing 1 (*OCEL1*). Both *FLNA* (Van-Den-Vyver et al., 2004) and *CDKL5* were hypothesised as causes for Aicardi syndrome as they are associated with diseases that have some overlapping phenotypes with AS such as Periventricular Nodular Heterotopia 1 (OMIM # 300049) and Developmental and epileptic encephalopathy 2 (OMIM # 300672), respectively. Moreover, immunostaining studies on post-mortem brains from AS patients showed strong immunolabelling of filamin, however, no variants were identified in these patients (Van-Den-Vyver et al., 2004). This hypothesis was further investigated in a later study on a male patient with normal karyotype but no variants in *FLNA* were identified (Anderson et al., 2009). As *CDKL5* has been associated with early-onset intractable seizures and infantile spasms, a study by Nemos et al., 2009 conducted a screening on a cohort of patients with early-onset seizures, including ten females with Aicardi syndrome. However, this yielded negative results for the AS patients.

Another study in 2015 suggested that *TEAD1* and *OCEL1* might be associated with non-X-linked AS in single cases of the condition (Schrauwen et al., 2015). The study was based on WES/WGS analysis of ten AS patients. However, this finding has not been replicated in subsequent larger studies that analysed 38 girls affected by AS (Wong et al., 2017). The failure to identify disease-causing variants or genes involved in the development of AS has been primarily due to unsuitable detection techniques and the lack of sufficient samples to perform analysis with significant statistical power. The advances in WGS and other genomics techniques have shown the capacity to robustly identify

genetic causes for individual cases, which could be employed to detect the pathogenic variants in the case of AS. Therefore, a gene-agnostic approach would be more suitable for a cohort of AS patients.

In this study, a cohort of 12 affected probands from 12 independent families, all reported as non-consanguineous (Figure 3.1: A), were recruited to the HICF2 programme in the hope of establishing a genetic cause for Aicardi syndrome. The recruitment of this cohort was an international collaborative effort, as only a small number of individuals are diagnosed with AS worldwide.

Prior to conducting WGS, the patients were screened using array-CGH. Moreover, an X-inactivation assay was performed as it would be interesting to see if non-random X-inactivation is playing a role in our cohort. It was only possible to perform the assay on 7/12 AS probands due to the lack of availability of DNA samples from the remaining five. All tested probands showed a random X-inactivation pattern in genomic DNA extracted from peripheral blood (Figure 3.1: B), which is not completely surprising as random X-inactivation patterns have been noted in previous reports (Hoag et al., 1997). However, these results do not reflect the X-inactivation pattern in the brain tissue of these patients, which was not performed in this cohort. The presence of random X-inactivation patterns in AS patients implies that the syndrome may not be solely caused by the inactivation of one X chromosome in females. This finding could lead to several possible conclusions or interpretations including that AS has a complex genetic mechanism and may involve intricate genetic interactions beyond X-inactivation. Other genetic factors, mutations, or epigenetic changes could be contributing to the syndrome. The possibility of a mutation on an autosome with sex-limited expression rather than an X-linked mutation becomes of an interest. For instance, Forkhead Box P1 (*FOXP1*) on chromosome 13 (Fröhlich et al., 2017) and DCC Netrin 1 Receptor (*DCC*) in chromosome 18 (Marsh et al., 2017) showcase male-biased phenotypic manifestations, possibly influenced by androgens.

It's conceivable that a comparable mechanism operates in the development of AS, where autosomal genes contribute to the disorder's emergence, while X-inactivation plays a role in modifying phenotypes. The surge of testosterone levels in males during foetal development (between weeks 8-24 of gestation; Collaer & Hines, 1995; Hines, 2005) could potentially offer a protective effect

during crucial developmental stages of the corpus callosum, Retinal Pigment Epithelium (RPE), and choroid, which occur between the 12th and 16th weeks of gestation (Kornienko & Pronin, 2009). However, it's important to acknowledge that this perspective does not rule out the possibility that yet-to-be-identified genes associated with AS in females could result in distinct male phenotypes, potentially leading to different diagnoses. This could potentially contribute to the rarity of AS diagnoses in karyotypically normal males.

Another interpretation of these results is that X-inactivation contributes to the clinical heterogeneity of AS, meaning that different individuals with AS can exhibit varying levels of severity and symptoms. Therefore, a random X-inactivation pattern could contribute to this variability, but it might not be the primary determinant. Furthermore, while random X-inactivation patterns are observed, it's still possible that X-inactivation plays a secondary or modifying role in the syndrome. It might interact with other genetic or environmental factors to influence the presentation of AS.

Another note is that X-inactivation can vary between different tissues in the body. While random X-inactivation might be observed in peripheral blood leukocytes, it could differ in other tissues that are relevant to the development of AS. One major obstacle to decisively determine the role of X-inactivation in the development of AS is sample limitations. The observed random X-inactivation patterns might be influenced by the specific samples studied (e.g., peripheral blood leukocytes, cell lines). Other tissues or cell types could have different X-inactivation patterns. In essence, while random X-inactivation patterns suggest that X-inactivation might not be the sole driver of AS, they don't completely rule out the possibility of its involvement. The genetic and molecular mechanisms underlying AS are likely complex and multifaceted, involving interactions between various factors beyond X-inactivation.

The probands were selected upon clinical diagnosis of Aicardi Syndrome by a clinician and a specialised ophthalmologist. The clinical features of all probands are shown in Table 3.1. The proband and the unaffected sister in Family 5 were confirmed to be monozygotic twins by assessment of genome-wide regions in which both sample share both chromosomes identical-by-descent (IBD). For all chromosomes analysed, paternal and maternal copies were shown to be shared Identical-by-Descent (i.e., IBD2) between the proband and the unaffected sister in Family 5 which is expected for monozygotic twins. In contrast, a comparison of sibling pairs using this method typically shows alternating between different IBD states with approximately 25% of genome IBD2, 50% IBD1 and 25% IBD0. VCF-intersect gave a genotype concordance of 99.94%, confirming that these sisters were indeed monozygotic twins.

The joint WGS analysis of all cases did not identify a recurrent variant shared between all probands indicating that the condition is not genetically homogeneous. Moreover, it highlights that oother factors such as environmental ones might have a bigger role than originally anticipated and might explain the discordance between the monozygotic twins. Therefore, a separate interrogation of the WGS data of each family was performed to identify genetic variants that might be causal in the individual cases. X-chromosome targeted analysis identified no variants, which was the initial hypothesis of AS being an X-linked syndrome; therefore, the analysis was extended to the autosomal chromosomes. Potential candidates for each family were prioritised identifying a rare structural variant in family 2. However, in the other families lead candidates were identified but none were strong enough to pursue further functional studies.

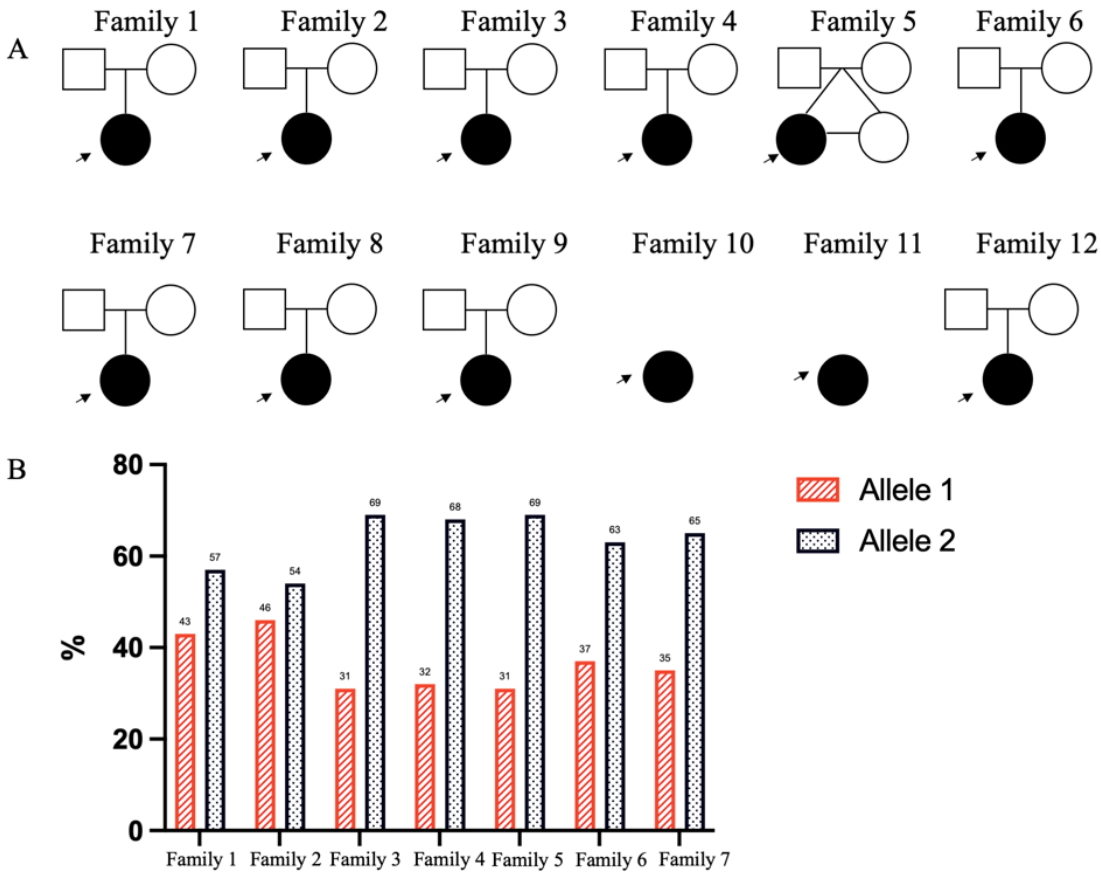


Figure 3.1: Overview of the AS cohort. A) Family pedigrees for probands diagnosed with Aicardi syndrome based on clinical phenotypes but no molecular diagnosis was established. In family 5, the proband was confirmed to be a monozygotic twin to the unaffected sister. All members of the pedigrees shown here were recruited and whole genome sequenced as part of the HICF2 project. B) X-inactivation assay using the highly polymorphic, differentially methylated androgen receptor gene shows random X-inactivation in probands from families 1-7.

Table 3.1: Clinical features and examination of probands in AS cohort. A (-) is placed where no information is available so the absence or presence of the missing feature cannot be confirmed.

Family ID	WGS	Sex	Infantile spasms	Corpus callosum	chorioretinal lacunae	Cysts	Coloboma of optic nerve	periventricular heterotopia	Developmental delay	Other brain malformations/ neurological findings
Family 1	P, F, M	Female	+	U/N	Bilateral	Interhemispheric arachnoid	Right	U/N	Mild psychomotor	U/N
Family 2	P, F, M	Female	+	Hypoplastic	U/N	U/N	U/N	U/N	Mild psychomotor	Hippocampal malrotation
Family 3	P, F, M	Female	+	Hypoplastic	Bilateral	Arachnoid and Choroid plexus	U/N	Bilateral	Global	U/N
Family 4	P, F, M	Female	+	Agenesis	Bilateral	Arachnoid and Choroid plexus	U/N	Bilateral	U/N	Hydrocephalus Enlarged ventricles Polymicrogyric cortex severe hypotonia
Family 5	P, F, M, MZ twin	Female	+	U/N	Left eye	U/N	Left	U/N	U/N	Complex cerebral malformation
Family 6	P, F, M	Female	+	U/N	Bilateral	Arachnoid cyct	U/N	Bilateral	U/N	
Family 7	P, F, M	Female	+	Partial agenesis	Bilateral	Multiple	U/N	Bilateral	U/N	Focal polymicrogyria
Family 8	P, F, M	Female	+	Agenesis	Bilateral	U/N	U/N	U/N	U/N	U/N
Family 9	P, F, M	Female	+	Agenesis	Bilateral	U/N	U/N	U/N	U/N	U/N
Family 10	P	Female	+	Agenesis	Bilateral	U/N	U/N	U/N	U/N	U/N
Family 11	P	Female	+	Agenesis	Bilateral	U/N	U/N	U/N	U/N	U/N
Family 12	P, F, M	Female	+	Agenesis	Bilateral	U/N	U/N	U/N	U/N	U/N

P= Proband, F= Father, M= Mother, MZ= Monozygotic, U/N= Unknown, +=present

One of the main results in this cohort was the significant genetic finding that led to clinical reassessment of the patient in Family 2, which is described in the following section.

3.2.1.1 A structural variant in Aristaless Related Homeobox (ARX) leads to a change in diagnosis from Aicardi syndrome to Developmental and Epileptic Encephalopathy 1 (DEE1)

This section describes a 9-year-old girl born to non-consanguineous parents. She had onset of spasms at 20 days old, mild psychomotor delay, hypoplastic corpus callosum and bilateral hippocampal malrotation, which was more severe on the left side. Ophthalmological examination at around one year old showed normal eye fundus and no chorioretinal lacunae. Array CGH testing was negative. The patient was initially diagnosed with Aicardi syndrome. Although the diagnostic chorioretinal lacunae were not observed during the ophthalmological consultation, the diagnosis was still made based on the presence of two features from the diagnostic triad plus at least two other major or supporting features, according to the expansion of diagnostic criteria (Sutton & Van-Den-Vyver, 2020).

The initial analysis of WGS data of this case was performed using the HICF2 bioinformatics research pipeline. However, this approach did not identify any convincing CNVs and SNVs. The WGS data was then re-analysed using the Variant Explorer pipeline, which revealed a *de novo* 3.3 kb deletion (chrX:g.25013828_25017089del; build 38) in the Aristaless-Related Homeobox (*ARX*) gene, which encodes Aristaless-related homeobox protein that is part of the homeodomain transcription factor family (Figure 3.2: A).

This finding is an example of missed variants due a combination of genomic sequence context and a suboptimal initial pipeline for such regions. *ARX* falls within a region where the GC content is high and reaches above 80% in some of the exonic regions (Figure 3.2: B). This presents a challenge for sequencing resulting in lower coverage compared to other regions in the genome. This is further confounded by the presence of multiple repeat regions in *ARX* leading to lower mappability, hence making the detection of SVs based on coverage alone challenging.

In the initial analysis using the HICF2 bioinformatic research pipeline (Method section 2.5.1), the CNVs were detected by extracting SNPs genotyping data called by Platypus, which are used as pseudo-CGH probes. These SNPs are used to simply computing the coverage to detect CNVs

through B-allele frequencies and Log2R imported from ngCGH, which simulates methods used for array data. These methods usually have less resolution as they struggle to detect small events and they are more affected by uneven distribution of SNVs that are used as probes.

In the re-analysis using Variant Explorer pipeline (Method section 2.4.3.2), the SV call set was directly generated from the WGS alignments using Lumpy, which considers both split reads and discordant read pairs information. This was complemented with CNV estimation by CNVnator. This method gives a more systematic evaluation of SV events across the genome and in general it is more sensitive for small events. There are still regions where variants cannot be detected with confidence due to low mappability or difficult-to-sequence regions, but still one can obtain a more comprehensive view of SVs across the genome. This comes with the cost of an increased proportion of false positives in WGS-based events compared to array-based methods. However, as the whole HICF2 cohort was processed simultaneously for cohort-wise variant calling, multiple pieces of evidence from all the samples were collected enabling a more reliable list of SV events to be compiled. These identified SV events were subsequently subjected to re-genotyping across all samples, guided by the supporting evidence obtained from the alignment data. This in general increases sensitivity in SV detection and allows a better estimate of true positives and false positives using the overall distribution of technical parameters. Therefore, using methods that do not rely purely on coverage for SV detection would be better suited for dark/camouflaged regions and in this case the refined pipeline allowed for SV detection in *ARX*.

ARX is required for brain development, specifically the cerebral hemispheres, and patterning (Bienvenu et al., 2002) and has been linked to developmental and epileptic encephalopathy 1 (DEE1; OMIM #308350). The condition follows a recessive X-linked pattern affecting males harbouring mutations which either occur *de novo* or which are inherited from asymptomatic carrier mothers. However, more commonly, heterozygous affected females have been identified with milder forms of the disease (Bettella et al., 2013; Ekşioğlu et al., 2011; Kato et al., 2004; Kwong et al., 2015; Marsh et al., 2009; Wallerstein et al., 2008). It is noteworthy that the patient in this cohort exhibits a phenotype similar to that commonly seen in males and more severe than expected in females. Compared with affected females, the severity in this patient is more comparable to the two cases reported with truncating variants, except for the presence of structural brain abnormalities in our

patient that are absent in the two reported cases (Bettella et al., 2013; Wallerstein et al., 2008). In these females, the variants (c.1459delA and c.1465delG) are predicted to lead to a truncated protein with a premature stop in amino acid 491. The stability of the truncated proteins has not been ascertained in these patients but if we assume that the truncated form of the protein is expressed preserving some functional aspects of ARX such as DNA binding through the Homeodomain, which is positioned before the predicted truncation, a highly conserved domain in the C-terminal would be disrupted (aristaless domain 526-544 amino acid) impacting the function of the protein. These patients also had skewed X inactivation, which possibly contributes to their phenotypes. In our patient, the deletion removes ~1.2 kb intergenic region upstream of *ARX*, the 5'UTR, the first coding exon, as well as the majority of intron 1. These deleted regions compose the regulatory elements as well as the transcriptional start site of the gene, therefore, it is predicted that this deletion would prevent the transcription of *ARX* resulting in a complete loss of a copy of this haploinsufficient gene. Furthermore, we only ascertained the X inactivation pattern in peripheral blood in our patient (Figure 3.1: B), which does not necessarily reflect the patten in the brain tissue further complicating the interpretation of heterozygous variants in *ARX* in female patients.

Validation of the 3.3 kb deletion in the proband using PCR and Sanger sequencing (various primer pairs tested) failed to yield either a specific PCR product or any product. The design of the primers was complicated by the presence of short tandem repeats at both ends of the deletion and further confounded by the high GC content reaching >80% at the distal breakpoint (Figure 3.2: B). The referring clinical laboratory validated the deletion using MLPA (Method section 2.6), and the signal for exon 1 probes showed a drop only in the proband that was not observed in the healthy parents, confirming that the deletion is *de novo* (Figure 3.2: C). In light of this finding, the clinical phenotype of the proband in Family 2 was re-evaluated, and the diagnosis was considered to better align with DEE1, especially considering the lack of chorioretinal lacunae, a key feature of AS.

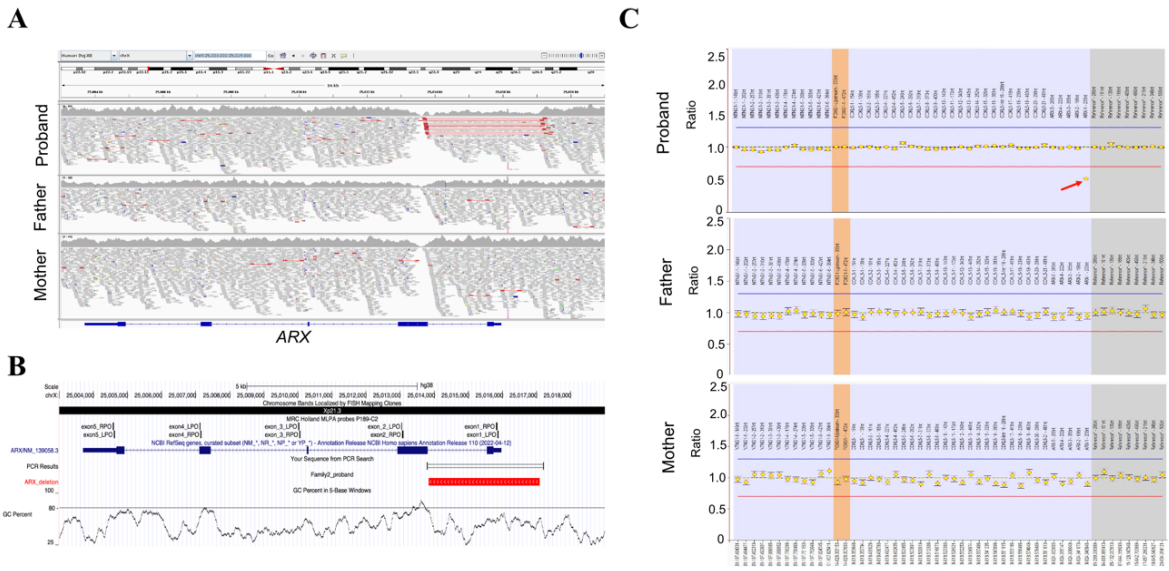


Figure 3.2: Detection and validation of a de novo deletion in ARX revised the diagnosis to DEE1 in AS Family 2. A) Screenshot showing 125bp read alignments supporting a de novo deletion involving ARX exon 1. Region shown is chrX:25,003,000-25,019,000 (GRCh38) encompassing the full ARX gene. Visualisation is using IGV v2.11.2, with squished and "view as pairs" options. Alignments are coloured by insert size and transcript shown is NM_139058.3. B) UCSC genome browser session showing the position of the deletion in relation to the PCR primers and MLPA probes that were used for validation. Also shown is the GC content which rises to >80% near the distal breakpoint where the coverage drops also in parental genomes indicating "dark by depth" region. An interactive version is available at https://genome.ucsc.edu/s/mhashim/ARX . C) The ~3 kb deletion was confirmed by the MLPA validation data, which shows a drop in the signal only for the proband for the exon 1 probe (red arrow). The MLPA figure was provided by the NHS lab

3.2.2 Complex structural variants leading to X-linked neurodevelopmental disorder

The second neurodevelopmental case involved two brothers born to a non-consanguineous couple. They were referred to the Clinical Genetics department for evaluation of global developmental delays. The first child was born following a pregnancy complicated by gestational diabetes, for which the mother took metformin during her pregnancy. He was referred at 14 months old due to severe muscle weakness and developmental delays. His birth weight was 3.03 kg and he developed hypoglycaemia, which was treated with glucose and nasogastric feeds. He developed jerky movements from 3 days of age. His developmental milestones were severely delayed, and he was unable to hold up his neck at the age of 14 months; he had no words in his speech and with limited understanding of language. At the age of 3 years, he was still not walking and used a personalised wheelchair, was non-verbal and had increased tone in all limbs with jerky, abnormal movements. The younger brother was seen at the genetics clinic at the age of 8 months with similar but slightly milder features. He was floppy with dystonic movements. Both parents showed no symptoms, but a history of a neuromuscular disorder among the mother's male cousins and the mother's maternal uncles was reported suggesting X-linked mode of inheritance and that the likely cause of the disorder is on the X chromosome.

Genetic testing performed by the clinical laboratory using microarray CGH identified a 590 kb duplication of Xp22.1. Further molecular cytogenetic investigations by the clinical laboratory using FISH identified a derivative X chromosome where one copy of Xp22.1 was inserted into the long arm of the X chromosome with a possible breakpoint at Xq27.1. Both brothers had the same X chromosome rearrangement, which had been inherited from their mother. Further cascade testing in the family performed by the clinical laboratories showed that the maternal grandmother, her sister and the maternal great-grandmother carried the X chromosome anomaly. This indicates that the neuromuscular disorder in their affected sons is consistent with X linked recessive inheritance and that the likely causative variant is located on the X chromosome (Figure 3.3).

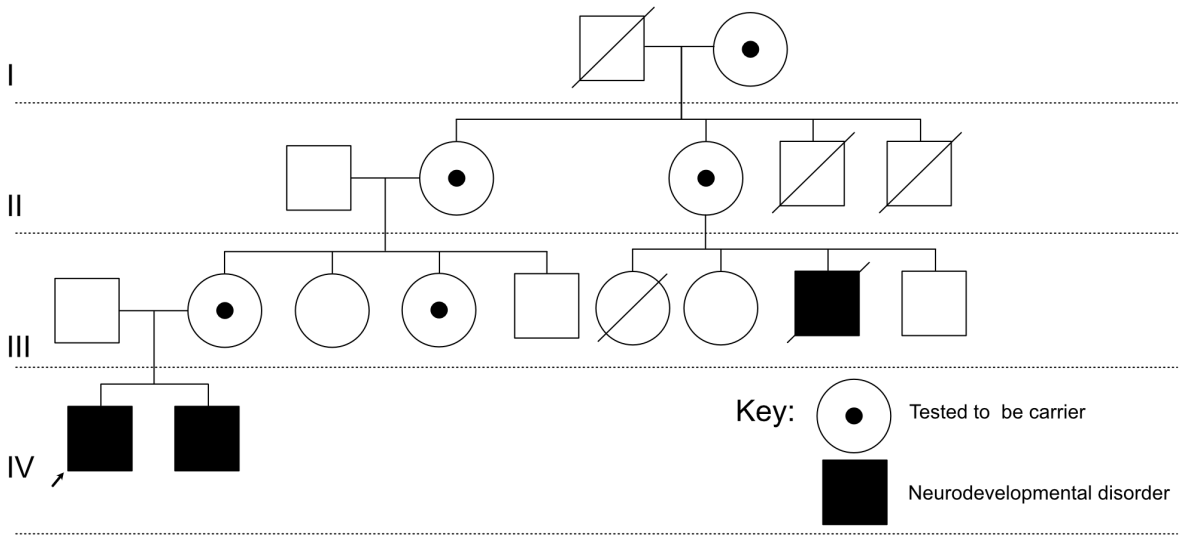


Figure 3.3: Pedigree of patients with X-linked neurodevelopmental disorder. The pedigree shows the segregation of phenotype with structural variants detected in affected males inherited from asymptomatic female carriers.

To confirm the possible mechanism of this rearrangement, it was essential to characterise the breakpoints of this non-tandem duplication precisely at a base level resolution, therefore, a sample from one of the affected siblings was whole genome sequenced. I analysed the WGS data, which robustly detected the breakpoints of the previously identified duplication and is now identified as ~633 kb of Xp22.11p21.3 (GRCh38(chrX):g.24321648_24954644dup) inserted into Xq27.1. However, WGS also identified another duplication ~102 kb at the site of insertion in Xq27.1 (GRCh38(chrX):g.140297721_140400415dup), suggesting an added complexity to this structural variant (SV), which was not detected previously with the other used methods (Figure 3.4: A).

While short-read Illumina data revealed two duplications, it did not resolve the two possible configurations that could be consistent with short-read data. Conformation 1 suggests that the larger duplication of ~633 kb has been inserted between the two smaller duplicated segments of ~102 kb in Xq27.1. Conformation 2 implies that the smaller duplicated segment has been inserted in between the duplicated larger segment in Xp22.11p21.3 (Figure 3.4: B). From the previous FISH results, conformation 1 is more likely; however, there is ambiguity surrounding the location of the smaller duplicated segment. Therefore, long read sequencing such as whole-genome optical mapping was considered ideal for resolving the uncertainty concerning which conformation was correct. This was essential for the interpretation of the pathogenicity of this variant.

Conformation 1 was validated using optical mapping (Figure 3.4: C-E), performed at the Bionano Services Laboratory (San Diego, USA).

In conformation 1, the insertion of a large segment would disrupt a region that is highly regulated through long-range regulatory elements and 3D chromatin configurations used to control the expression of multiple genes (Rao et al., 2014). Of the several disease-causing genes in the area, SRY-BOX 3 (*SOX3*) and Fibroblast Growth Factor 13 (*FGF13*) are OMIM disease genes involved in early neural development and are linked to intellectual developmental disorder with isolated growth hormone deficiency and Intellectual developmental disorder 110, respectively. The large, duplicated segment is inserted into Xq27.1 at GRCh38(chrX):g.140400415, which is ~1.2 Mb upstream of *FGF13* and ~102 kb downstream of *SOX3*. The insertion of both larger and smaller duplication adds ~735 kb within a region that is predicted to be a contact domain for *SOX3* and adjacent to a contact domain for *FGF13*. This addition would possibly push the contact domain corners for *SOX3* further apart hence disrupting the formation of the chromatin loop, which are usually marked by the peak in the contact maps (Figure 3.5). These loops bring genomic sites that are located far apart in a linear genome into proximity allowing for interaction of distal and proximal regulatory elements such as enhancers, and insulators affect promoters to regulate gene expression (Blackwood & Kadonaga, 1998; Gaszner & Felsenfeld, 2006). While the insertion does not seem to directly affect the contact domain for *FGF13*, the insertion would affect the predicted long-range interactions between the TADs boundaries (Figure 3.5). Therefore, it might affect their expression.

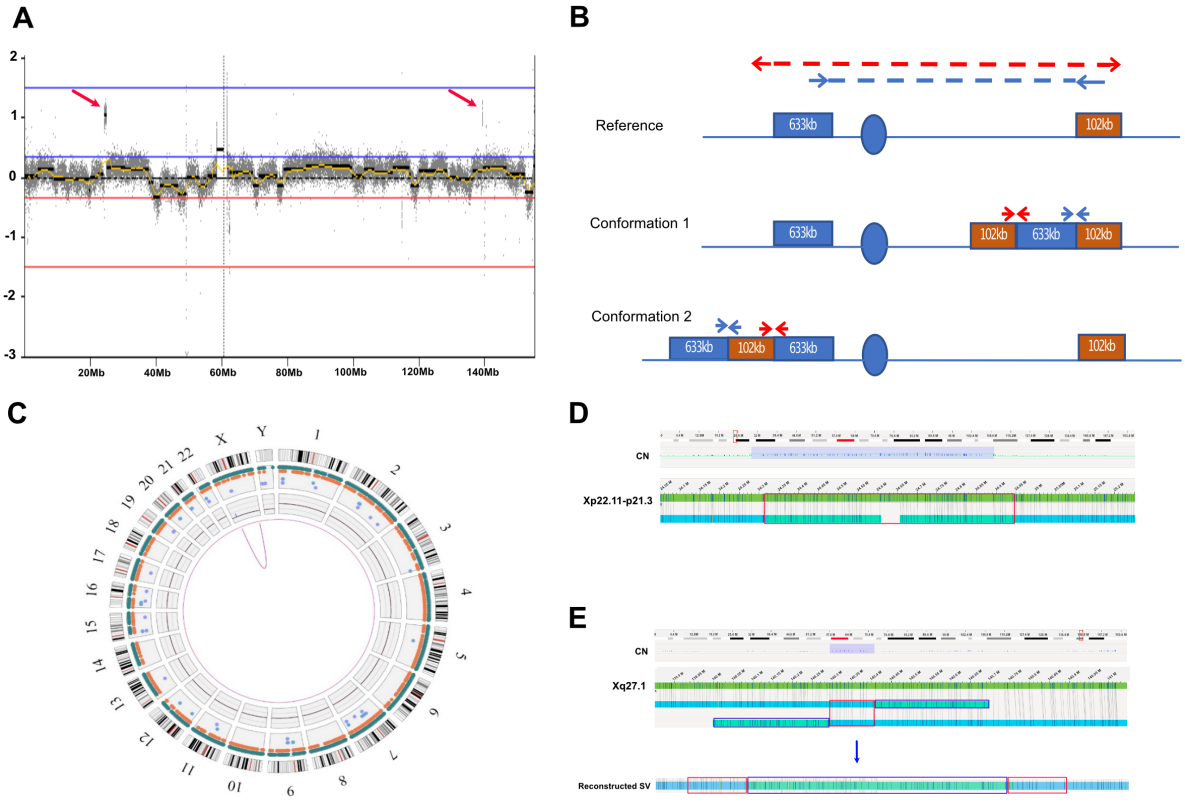


Figure 3.4: X chromosome complex structural variant. A) Read count information from short read sequencing normalised by ngCGH, a CNV calling tool, showing two X chromosome duplications (red arrows). B) Split Illumina read-pairs suggest the two duplications are inter-linked. However, two possible configurations can explain the split read pattern. Centromeres represented by blue circles. C) Circos plot highlights the only SV identified by the Bionano pipeline above the threshold for SV detection. D) Genome browser view of the optical maps robustly detects the ~600 kb duplication of Xp22.11p21.3 being inserted into Xq27.1, which is present in carrier mother, and both affected male siblings using the Bionano pipeline excluding Complex Multi-Path Regions (CMPR). The red boxes highlight the left and right segments of the duplication inserted into Xq27.1 (blue boxes). E) The Bionano pipeline without masking CMPR detects ~85 kb tandem duplication (red box) flanking either side of the 600 kb insertion from Xp22.11p21.3

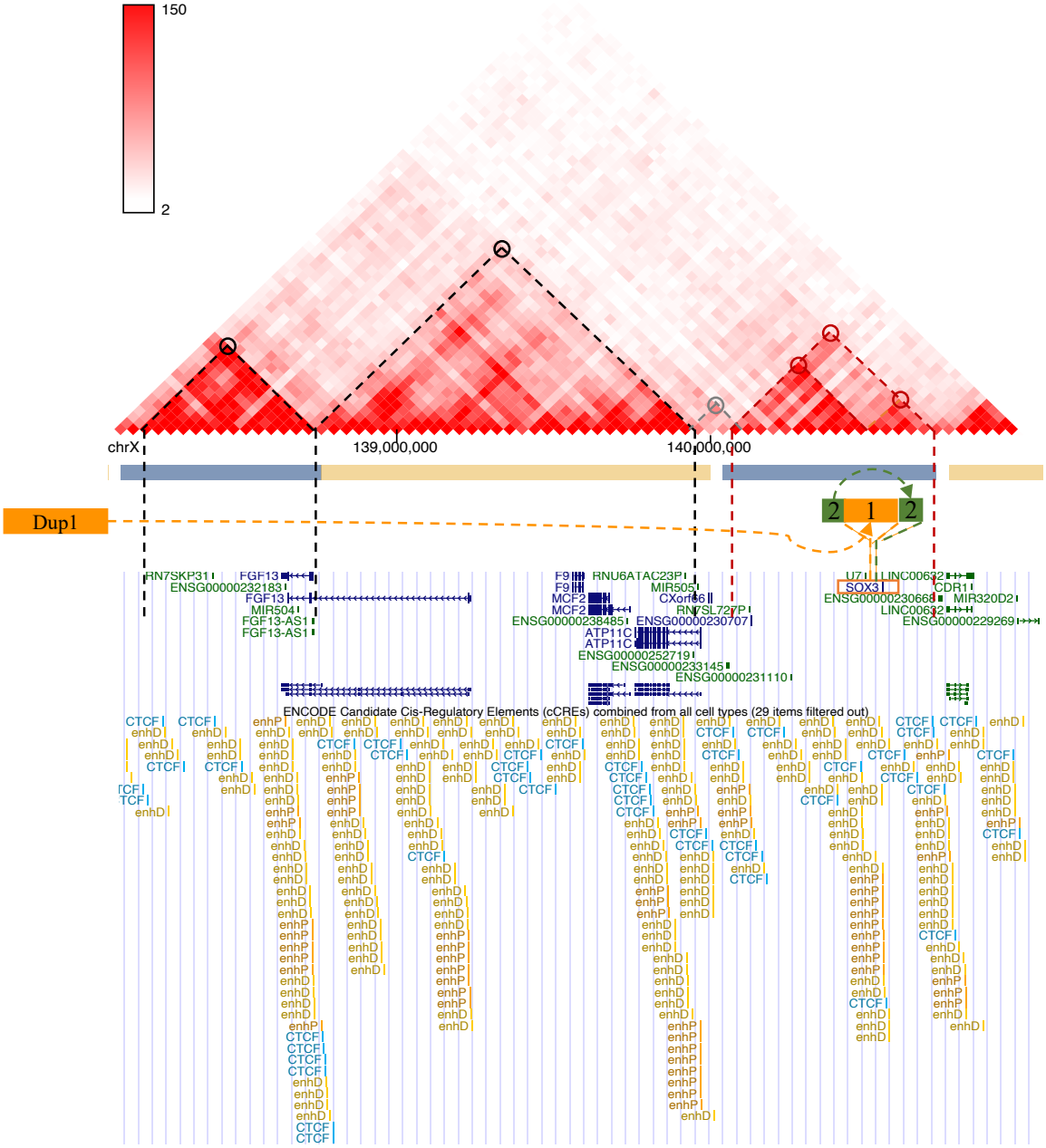


Figure 3.5: Hi-C profile and contact domains at the Xq27.1 locus.The displayed profile is based on data derived from human lymphoblastoid cell lines. A ~633 kb duplication (Dup1; orange box) from Xp22.11-p21.3 is inserted into Xq27.1 at GRCh38(chrX):g.140400415 flanked by two smaller duplicated segments (green box) spanning GRCh38(chrX):g.140297721_140400415dup. Overlapping Hi-C data suggest that the topology of the SOX3 contact domain might be altered. Enhancers located in the region based on ENCODE, loops involving SOX3 (red circles) or FGF13 (black circles), and CTCF-binding sites (blue lines) are indicated. Dashed lines indicate reported TADs

It would be interesting to measure the expression levels of *FGF13* and *SOX3* in the affected siblings; however, *SOX3* is not expressed in whole blood, and *FGF13* is expressed at low levels, with significant variability between samples, as shown in the genes' entry in https://gtexportal.org/. Thus, using whole blood for RNA expression might not yield meaningful data for these two genes of interest. To overcome this challenge, generating a patient's induced pluripotent stem (iPS) cell line and differentiating it into a neuronal lineage for accurate characterisation of the disease mechanism could be considered part of the future work plan.

The discovery of pathogenic SVs in the individuals mentioned in sections 3.2.1 and 3.2.2 highlights the significance of this often-overlooked type of disease-causing variations. Moreover, while it highlights the ability of short-read WGS data to detect these variants, it also demonstrates its inability to resolve the correct configuration of complex structural variants such as the one described in section 3.2.2 and the need for other methods such as long-read sequencing for resolving the structure that is important for the clinical interpretation of the variant. Thus, the impact of SVs in neurodevelopmental disorders and the effectiveness of WGS in SV identification and characterisation will be further examined in chapter 5.

3.2.3 Bi-allelic variants in DOCK7 in a patient with Developmental and Epileptic Encephalopathy-23 (DEE23)

This section describes a 22-year-old female born to Lebanese consanguineous healthy parents (Figure 3.6:A-B). The proband presented with seizures from 9 months of age, has short stature (-3 standard deviation; SD), and microcephaly (-2 SD). Her developmental milestones were delayed: sitting independently at ten months, walking at two years with poor balance, and being able to run at almost 7 years. Speech development was also delayed, with her having only a few words at 1.5 years, a situation which did not improve much, and some autistic features: clapping hands and not being interested in other children, self-mutilating with headbanging when angry. She exhibits dysmorphic features, including bitemporal narrowing, low anterior hairline, periorbital fullness, and prominent earlobes.

The proband presented in status epilepticus. Generalised seizures were preceded by lip smacking and chewing noises followed by a focal seizure of the right leg. An EEG showed paroxysmal changes in the left temporooccipital region and over the right occipital region. The seizures persisted despite being treated with valproate and levetiracetam.

Ophthalmological examination revealed normal ocular pressure, unremarkable ophthalmoscopic appearance and normal electroretinographic (ERG) traces. However, she has impaired vision with reported amblyopia and mild hypermetropia, indicating likely cortical causes of her vision impairment. Brain MRI detected prominence of frontal extra-axial spaces along with effacement of the superolateral aspects of the lumens of lateral ventricles. However, the corpus callosum appeared normal.

Genetic testing using array-CGH (Agilent 400 K) in 2011 revealed an ~255 kb maternally inherited copy number (CN) gain involving chromosome 1 at 1q41-q42.11(chr1:g.223820857_224076362dup; build 37). The duplication encompasses full copies of Calpain 2 (*CAPN2*) and Tumor Protein p53-Binding Protein 2 (*TP53BP2*) as well as a partial copy of Calpain 8 (*CAPN8*). However, none of these genes are linked to human disease. Nonetheless, copy number changes in the long arm of chromosome 1 have been associated with partial trisomy 1q41 syndrome and 1q41-q42 microdeletion syndrome featuring ID and brain abnormalities

(Emberger et al., 2001; Rosenfeld et al., 2011; Zak et al., 2016). Moreover, *TP53BP2* has a predicted probability of Triplosensitivity (pTriplo) of 0.94, indicating intolerance to an increase in dosage. The pTriplo scores are from the recent Collins *et al*., study, which calculated the scores using an ensemble machine-learning model based on large, rare copy-number variant analysis (Collins, et al., 2022). Therefore, the maternally inherited duplication could contribute to the phenotype but is not considered causative of the proband's phenotype as the carrier mother is regarded to be healthy.

The proband was recruited to the HICF2 project for whole genome sequencing as part of a parent-child trio in 2016. WGS data quality and sample relatedness were assessed as previously described (Methods section 2.4.3). The WGS data confirmed the maternally inherited CN gain, refining the breakpoints to (GRCh37(chr1):g.223818141_224123722dup; Figure 3.6: C). It revealed that the proximal breakpoint lies within a tandem repeat region between exons 2 and 3 of Caplin 8 (*CAPN8*) and the distal breakpoint within an intergenic segmental duplication, thereby excluding the possibility of a fusion gene. Complete copies of Caplin 2 (*CAPN2*), Tumor Protein P53-Binding Protein 2 (*TP53BP2*), and HM358976 genes are duplicated within the region of CN gain. No other CNVs of clinical interest were identified. Importantly, zygosity analyses identified large regions of homozygosity (ROH) of ~ 36Mb in total. This analysis guided the SNV analyses towards homozygous SNVs mapping within the ROH, consistent with an autosomal recessive inheritance model.

To investigate potential clinically relevant candidate variants, I focused on SNVs with modes of inheritance characterised by homozygous or compound heterozygous recessive patterns, as well as *de novo* dominant variants. The trio analyses identified a total of 670,545 high-confidence variants. Within this set, 181,044 were classified as rare, with an allele frequency of less than 0.2% in the highest allele frequency subpopulation of gnomAD, which was adjusted for recessive patterns of inheritance to 0.5%. Among these rare variants, 4,622 were predicted to potentially exert a deleterious impact, as determined by one of the following factors: CADD score, splice site alterations, Indels, disruption of known ENCODE transcription factor (TF) binding sites, or association with diseases based on databases such as HGMD or ClinVar. This acts as a soft filter to remove common variants reducing the number of variants but allows for more variants to be retained for further analysis.

Of the 4622 rare variants with a potential deleterious impact, 21 variants were identified as *de novo*, 13 variants satisfied the autosomal recessive model, and variants in 96 genes fitted the compound heterozygous mode of inheritance. Closer inspection of the read alignments in Integrative Genomics Viewer (IGV), revealed that 4/21 were true *de novo* variants, 10/13 autosomal recessive variants (6/10 within ROH), and compound heterozygous variants in 5/95 genes as true positives (Appendix Table S3.1).

The identified variants were then closely examined for pathogenicity predictors, and association to phenotypes of the patient. Only six variants had a CADD score > 20, 1 *de novo*, 4 homozygous all within ROH, and 1 compound heterozygous (Table 3.2)

The *de novo* heterozygous variant in Fc Gamma Receptor IIIa (*FCGR3A*) is linked to Immunodeficiency 20; autosomal recessive (OMIM #615707). While the variant has a high CADD score of 35, it introduces a stop codon only 20 amino acids before the natural stop codon for the protein (total length 253 amino acid). Moreover, it is predicted to escape nonsense-mediated decay by NMDEscPredictor tool (Coban-Akdemir et al., 2018). Therefore, it can be excluded. The homozygous variant in Gamma-Aminobutyric Acid Type A Receptor Subunit Alpha2 (*GABRA2*) affects the 3'UTR, however, it is not predicted to change splicing according to SpliceAI, nor it is predicted to affect the poly(A) signal (by MutationTaster tool). The gene is linked to Developmental and epileptic encephalopathy 78 (OMIM #618557), which is relevant to the phenotype of the patient, therefore, it cannot be fully excluded yet, but the variant is weak. The next homozygous variant is in Ribosomal Protein S14 (*RPS14*), which is linked to Refractory Macrocytic anaemia (OMIIM #153550). The phenotype is not relevant to the patient and is predicted benign by PolyPhen-2, which reduces the strength of the variant. The third homozygous variant is in MCF.2 Cell Line Derived Transforming Sequence Like (*MCF2L*), which is not linked to disease in OMIM. However, it was found to be activated during neurogenesis as part of a ChIP-seq analysis (Raposo et al., 2015) but the level of evidence for causality of this variant is weak. Finally, the homozygous variant in Dedicator of Cytokinesis 7 (*DOCK7*) is in a gene which has been linked to Developmental and epileptic encephalopathy 23 (OMIIM #615859) and has reported biological links to microcephaly (Yang et al., 2012), ID or epilepsy (Perrault et al., 2014). Moreover, *DOCK7* is highly expressed in the brain from early stages (Watabe-Uchida et al., 2006) and has a role in various processes that are

relevant to the phenotypes observed in the patient. These including actin cytoskeleton remodelling, cell migration, neuronal polarity, axonal development, and development of GABAergic interneurons (Watabe-Uchida et al., 2006; Majewski et al., 2012).

Table 3.2: Candidate variants considered for further analysis in DEE23 case. Highlight in red indicates a reason for downgrading the variant.

MOI	Gene	Variant	Pathogenicity predictors	gnomAD
AD	*FCGR3A*	NC_000001.10(NM_000569.8): c.700G>A; p.Arg233*	CADD = 35 pLoF= Low-confidence (END_TRUNC)	0.00005
AR	*GABRA2*	NC_000004.11(NM_000807.4): c.*1880G>A	CADD = 22.4	-
	RPS14	NC_000005.9(NM_005617.4): c.382G>A; p.Arg128Trp	CADD = 23.2 SIFT = deleterious PolyPhen-2 = benign	-
	MCF2L	NC_000013.10(NM_001112732.3): c.40G>A; p.(Glu14Lys)	CADD = 24.2	0.00004
	DOCK7	NC_000001.10(NM_001271999.1): c.5691+1G>T	CADD = 35 SpliceAI _DS_DL= 1	-
CH	*COQ10A*	NC_000012.11:g.56660759A>G	CADD = 17.1	0.00003
		NC_000012.11(NM_144576.4): c.233C>T; p.Ala78Val	CADD = 22.1 SIFT = tolerated Polyphen-2 = benign	0.0003

SpliceAI _DS_DL= SpliceAI delta score for donor loss; END_TRUNC = the LoF variant falls in the last of the transcript

The homozygous variant (NM_001271999.1(*DOCK7*): c.5691+1G>T) lies within a ROH and is absent from publicly available genomic databases, including gnomAD. It disrupts the canonical splice-donor site and is predicted by MaxEntScan score to decrease splicing efficiency by 100% (from 7.98 to -0.52) and has a SpliceAI score of 1 for donor loss. Therefore, it can be considered as the strongest candidate for this patient. The nearest strong alternative splice site is 91 bases from the exon/intron boundary on the 3' side, predicted to result in an insertion leading to a frameshift and thus likely to result in protein truncation and loss of function, however, this needs to be validated. The presence of the +1 splicing donor site variant was confirmed by Sanger sequencing (Figure 3.6: D).

At the time of sequencing, only three patients with *DOCK7* variants were reported in the literature, all compound heterozygous for protein-truncating variants (Perrault et al., 2014). Thus, the exact effect of the homozygous *DOCK7* splice site variant and its involvement in disease development was uncertain. Therefore, further investigations were undertaken. Moreover, additional cases were sought using the DDD database and the GeneMatcher online network service. A continuous review of the literature for other cases was also conducted.

3.2.3.1 A canonical splice donor site variant in DOCK7 causes in-frame deletion

To confirm that the homozygous splice variant was indeed null, RNA-sequencing was performed (Oxford Genomics Centre). Paired-end library preparation of blood-derived RNA followed by alignment to the genome allowed the detection of reads spanning multiple exons per gene, thus facilitating the analysis of alternative splicing. A total of 266-644 paired-end reads per sample aligned to at least one exon in the *DOCK7* locus, and the expression of *DOCK7* in the proband's sample was not reduced compared with the parents. Interestingly, the proband and her parents showed patterns of correctly spliced exons in the extreme 3' end of the transcript (exons 46-49 of NM_001271999.1) and a low number of intronic reads, which suggests that there is no intron retention or accumulation of unspliced intermediates.

Strikingly, all samples from the trio showed clear evidence of exon 44 skipping (Figure 3.6: E). The proband's RNA sample showed the splicing of exon 43 to 45 exclusively, whereas the parents showed both exon 44-skipped and exon 44-spliced reads (Figure 3.6: E). This partial skipping is consistent with the expression of the heterozygous c.5691+1G>T allele in each parent. Exon 44 comprises 90 nucleotides corresponding to 30 complete codons; therefore, exon skipping is not expected to result in a frameshift. Thus, the blood-derived splicing pattern of *DOCK7* in the proband suggests an in-frame deletion of 30 residues in the protein. Rarer, naturally-occurring isoforms of *DOCK7* that lack exon 44 also lack the highly conserved DOCK homology region 2 (DHR2) domain (gtexportal.org). A molecular mechanics analysis for DOCK7 lacking exon 44 was performed by Dr. Matteo Ferla (Method section 2.8). The exon-deletion is in a dynamic region of the DHR2 domain, known as lobe B, which confers Rac1/Cdc42 dual specificity and GEF activity (Kukimoto-Niino et al., 2019b). Therefore, the thermodynamic stability of the protein was predicted *in silico* (Method section 2.8). Whereas the deletion is not predicted to destabilise the protein ($\Delta\Delta G_{folding}$ -26 kcal/mol relative to wild type, for shared residues only), nor significantly weaken the interaction with Cdc42 ($\Delta G_{interface}$ wild type: -156 kcal/mol, deletion: -151 kcal/mol), it shortens a loop that controls this specificity making it virtually impossible for the α-6 helix to adopt the open conformation described in (Kukimoto-Niino et al., 2019b). It is expected to affect binding to the GTPase at residues affected by the deletion (p. Δ1868-1897), including those in direct contact with Cdc42 in the crystal structure (Figure 3.6: F); therefore, it most likely abrogates the guanine nucleotide exchange factor (GEF) activity. These data suggest that splice site variant-associated skipping of exon 44 may cause loss of function in DOCK7.

Pathogenic splice site variants resulting in in-frame deletions have been reported in other conditions (Aoyagi et al., 1999; Klima et al., 1993). These variants can be associated with atypical disease and variable expressivity (Kuivaniemi et al., 1988). A similar implication of the pathogenic splice site variant disrupting *DOCK7* can be observed on the DHR2 domain and DOCK7's binding to Rac1/Cdc42 needed during the developmental process. The DHR2 domain is required for neuroblast migration in mice (Nakamuta et al., 2017). Interestingly, more cases of patients with bi-allelic variants in *DOCK7* have been published since the case analysis. A comparison of the clinical phenotypes of all cases with this proband showed that they share core features of *DOCK7*-associated

syndromic disease, suggesting that the DHR2 domain may mediate essential functions lost in patients with truncating variants (Appendix Table S3.2).

While *DOCK7* is now a well-known gene for DEE23, at the time of analysis, it was not included in any gene panels due to a low number of reported patients, three from two unrelated families. Therefore, WGS has enabled the detection of this novel splice site variant in a gene that was not well-known at the time. While the prediction tools are a good indicator of the expected effect, they must be validated. For this variant, RNAseq confirmed that it does not lead to complete protein loss through truncation as predicted; instead, it leads to an in-frame deletion in a highly conserved domain that could severely affect the function of the protein. The causality of the variant was further substantiated through structure-based analysis, which suggested it disrupted the GTPase pathway necessary for brain remodelling during development, thus providing evidence for its involvement in disease development.

Despite the clinical overlap between this proband and the reported patients (Appendix section S3.2.3 and Table S3.2), there is a tremendous clinical heterogeneity. Therefore, it is difficult to discern the core features related to *DOCK7* from a small number. Hence, collecting more patients and reviewing the clinical phenotypes across all patients with *DOCK7* variants can help to delineate genotype-phenotype correlation. The case presented here serves as a reminder of the need to utilise a multi-faceted, dynamic approach when assessing the pathogenicity and causality of genetic variants.

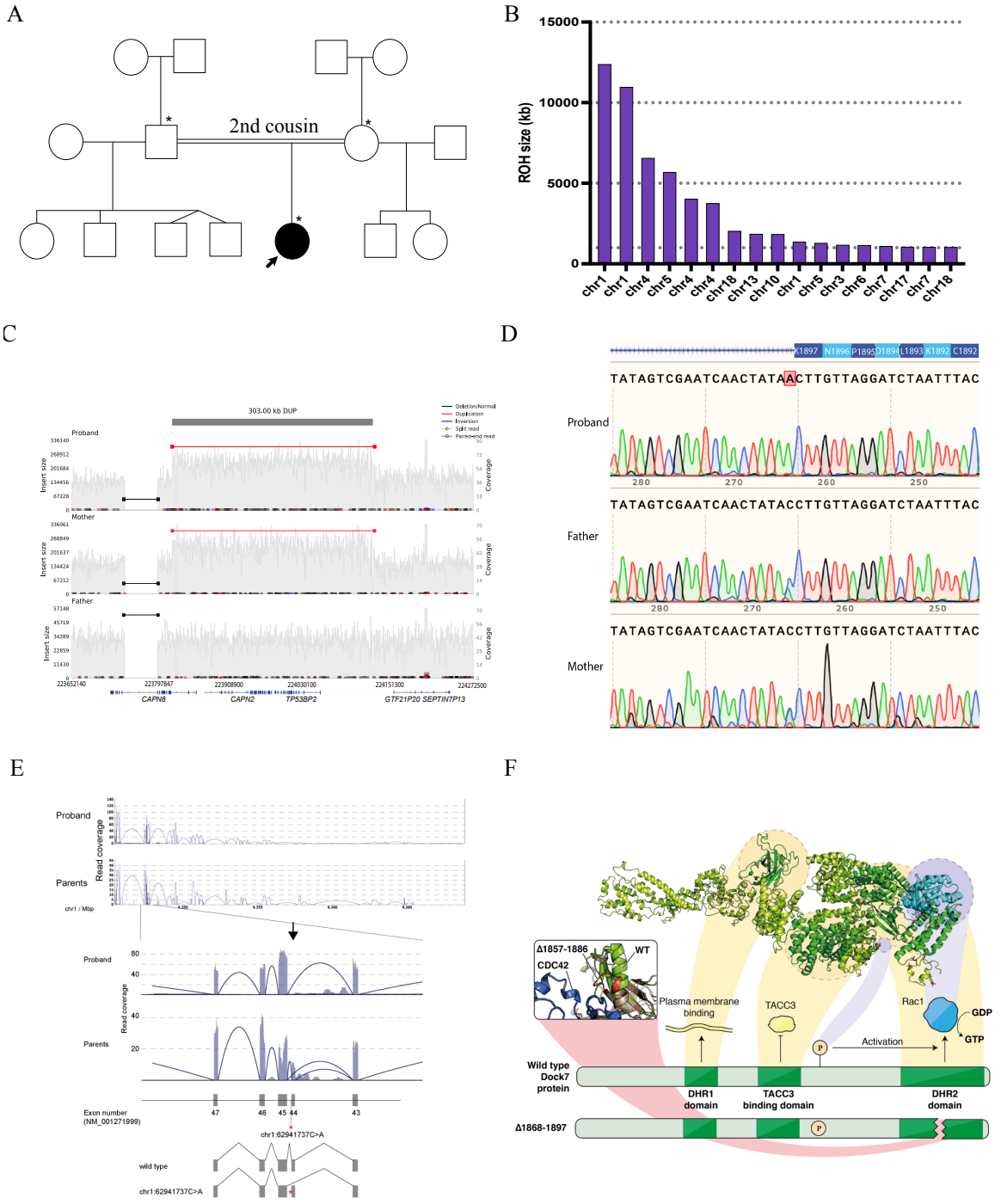


*Figure 3.6: Bi-allelic variants in DOCK7 in a patient with DEE23. A) Family pedigree showing consanguineous parents, which indicates AR as the possible inheritance mechanism for disease-causing variants. The consanguinity was not reported initially but identified through the analysis of WGS data and later confirmed by the clinician. * indicates samples sequenced. B) Maternally inherited copy-number gain that is absent from the father and encompasses a duplication of full copy of CAPN2, TP53BP2 and a partial copy of CAPN8. C) Sanger validation of DOCK7 (reverse strand). A homozygous variant G>T in the proband, and heterozygous variants in both mother and father are observed. D) RNA Sequencing analysis showing exclusive skipping of exon 44 of DOCK7 in proband and both skipped and spliced in exon 44 in parents. E) Diagram to show the domain structure of DOCK7 and the associated 3D structures based on crystal structure PDB:6AJL. Deletion of exon 44 of DOCK7 protein is predicted to cause a shift in the helical structure (shown in box with enlarged region) preventing the binding to Cdc42 and Rac1/ Rac3.*

3.2.4 KMT2E is a novel gene associated with a neurodevelopmental disorder, now known as ODLURO syndrome.

This section describes a 6-yr old, affected male born to non-consanguineous European healthy parents. The prenatal and neonatal courses were uncomplicated. The proband had delays in his developmental milestones, such as walking at 17 months. His speech development was delayed with his first word at two years, but he did not improve much and had persisting articulation problems. Mild intellectual disability and behavioural difficulties such as anger and crying were also noted. Other prominent abnormalities observed during examination included macrocephaly (+2.8 SD), facial leanness, mild pectus excavatum, clubfoot, and distinct facial features such as large ears, large nasal bridge, and thin upper lip.

Prior to WGS, a rare small maternally inherited ~306.6 kb tandem duplication of 1q41(GRCh38(chr1):g.223546528_223838782dup) was detected using array-CGH. The duplication encompasses partial copies of Tumor Protein P53-Binding Protein 2 (*TP53BP2*), and Caplin 8 (*CAPN8),* and a complete copy of Caplin 2 (*CAPN2*), similar to that observed in the patient discussed in section 3.2.3. However, as it is inherited from an unaffected mother, the involvement of the duplication in the manifestation of the proband's phenotype was uncertain; therefore, WGS was performed on a parent-proband trio.

WGS data quality and sample relatedness were assessed as described in the methods section 2.4.3, and ROH regions across the genome were <5Mb, consistent with non-consanguinity.

The WGS data confirmed the maternally inherited CN gain, refining the breakpoints to (GRCh38(chr1):g.223478557_223859081dup; Figure 3.7: A). It revealed that the duplication encompasses complete copies of *CAPN2*, *TP53BP2* and *CAPN8*. No other CNVs of clinical interest were identified. The SNV analysis interrogated variants that follow de novo autosomal dominant, autosomal recessive (simple and compound heterozygous), or X-linked, prioritised 132 deleterious variants, which were subsequently inspected using IGV.

The WGS alignment review revealed that out of the 132 variants, only two *de novo* heterozygous, three hemizygous, and four compound-heterozygous variants were true positives (Table 3.3). However, only the *de novo* frameshift variant in lysine methyltransferase 2E; *KMT2E*

(NM_182931.2(*KMT2E*): c.3672_3673delTA; p.TYR1224*) satisfied the selection criteria for further assessment. These criteria include rare (absent or AF <0.2 % with no homozygous/hemizygous present in gnomAD), expressed in brain tissue, and CADD >20 for missense variants. The *de novo* variant in *KMT2E* was Sanger validated (Figure 3.7: B).

Table 3.3: Verified candidate genes & variants in the proband sec 3.2.3. Only the de novo frameshift variant in KMT2E passed all selection criteria for consideration as a final candidate. In contrast, AR and CH variants were rejected due to a low CADD score of variants or the presence of homozygous/hemizygous variants in gnomAD. The coordinates shown are in build 38. Entries in red colour highlight one of the main reasons for exclusion.

MOI	CHR	Position	REF	ALT	Region	Gene	Protein Variant	Translational Impact	CADD	gnomAD AF
DN	7	104749590	TA	-	Exonic	***KMT2E***	p.Y1224*	frameshift	-	-
XR	X	14027247	C	T	Exonic	*GEMIN8*	p.V172M	missense	22.5	0.00078 (64 Hemi)
XR	X	47441631	C	T	Intronic	*SYN1*	-	-	7.463	0.000516 (5 Hemi)
XR	X	129149660	C	G	Exonic	*BCORL1*	p.A971G	missense	16.51	0.0001808 (12 Hemi)
CH	1	204374586	T	C	3UTR	*PPP1R15B*	-	-	3.569	0.0004
CH	1	204379133	T	A	Exonic	*PPP1R15B*	p.E469D	missense	13.93	0.0008
CH	3	56762347	G	A	3UTR	*ARHGEF3*	-	-	9.029	-
CH	3	56950890	A	G	Intronic		-	-	9.691	-

MOI= Mode of Inheritance; DN= de novo; XLR= Autosomal recessive (simple); CH=recessive compound heterozygous

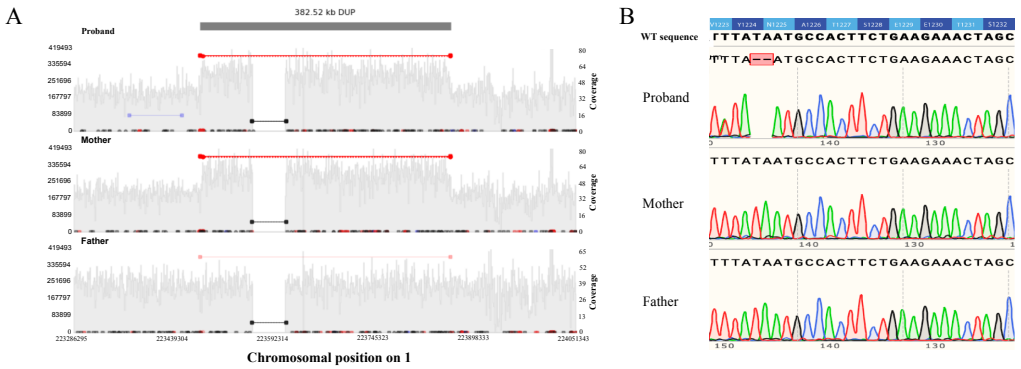


Figure 3.7: Validation of a de novo 2 bp deletion in KMT2E. A) Maternally inherited ~400 kb copy-number gain that is absent in the father and it encompasses a duplication of full copy of CAPN2, TP53BP2 and CAPN8. B) Sanger validation of KMT2E variant in proband leading to a frameshift that is absent in father and mother

KMT2E is also known as the Mixed Lineage Leukaemia member 5 (MLL5) and is a member of the KMT2 gene family. It encodes a histone methyltransferase, an epigenetic protein that functions as a transcriptional regulator and has a vital role in various biological processes such as cell cycle progression, maintaining genetic stability, adult blood cell formation, and sperm production. It is highly expressed in the brain, especially during foetal development. (Dong et al., 2014). KMT2 family proteins have a SET domain that is enzymatically active and responsible for its methyltransferase function (Faundes et al., 2018; Rao & Dou, 2015). KMT2E stands apart from other members of the family due to its unique sequence and location of the SET domain within the protein. Despite suggestions that KMT2E may lack intrinsic methyltransferase activity (Mas-Y-Mas et al., 2016), it is noteworthy that this protein remains highly conserved. It is suggested that KMT2E may impact H3K4 methylation indirectly through the regulation of other histone-modifying enzymes. Other KMT2 proteins have multiple PHD finger domains that function as readers of H3K4 methylation, but KMT2E has only one PHD finger domain. These domains usually bind to specific histone markers to bring transcription factors and nucleosome-associated complexes to chromatin. Unlike most KMT2 proteins, which act as global activators of open chromatin, KMT2E is thought to be a repressor. However, the exact mechanisms behind its regulation of gene transcription are still not apparent (Zhang et al., 2017).

The constraint data from gnomAD show that *KMT2E* is a candidate haploinsufficient gene. *KMT2E* is significantly depleted for protein-truncating variants, presumably due to negative selection, with a LEOUF score of 0.06 and a pLI score of 1.0.

Several *KMT2E* (*MLL5*) deficient mouse models have been created and characterised (Heuser et al., 2009; Madan et al., 2009; Yap et al., 2011; Zhang et al., 2009). *KMT2E* KO mouse models exhibit stunted growth, higher mortality, and compromised haematopoiesis. However, no neurological phenotypes were noted in these mice (Tasdogan et al., 2016).

The data presented collectively supported *KMT2E* as a strong candidate for causing disease and provided a solid basis for further investigation of patients with heterozygous variants in the gene through collaboration and large databases.

Through international collaboration, this case contributed to a case series of patients with heterozygous *KMT2E* variants and neurodevelopmental phenotypes (O'Donnell-Luria et al., 2019). This study described a new neurodevelopmental disorder related to *KMT2E* and added to the expanding list of diseases caused by variants in the KMT2 gene family. The majority of patients described harboured protein truncating variants (PTVs) but four cases had missense variants. Individuals with PTVs experienced mild developmental delays and intellectual disability, as well as a high occurrence of autism. Other common symptoms included large head size, low muscle tone, and digestive problems. The facial appearance was also distinctive. Epilepsy was uncommon in these cases, and haploinsufficiency was suspected as the cause of the disorder. The similar symptoms in individuals with small deletions in this region supported this hypothesis. Furthermore, a second cohort was subsequently published, confirming haploinsufficiency as the disease mechanism. The syndrome was named ODLURO (Velmans et al., 2022) and led to the inclusion of *KMT2E* in the Genetic epilepsy syndromes, Intellectual disability and Severe Paediatric Disorders *in silico* panels in GEL's PanelApp.

On the other hand, individuals with missense variants experienced severe developmental delays, microcephaly instead of macrocephaly and epilepsy, including a type of epilepsy that begins in infancy (O'Donnell-Luria et al., 2019). This difference in symptoms may be due to changes in the function of the *KMT2E* gene, and further research is needed to understand this connection, and more cases with missense variants would be required. Therefore, I have looked at WGS data of participants with rare, deleterious missense variants within *KMT2E* in the rare disease cohort of the 100KGP, blinded to their reported phenotype. This strategy would allow for an unbiased analysis of the involvement of missense variants in *KMT2E* in disease development since the initial study (O'Donnell-Luria et al., 2019) only had four individuals with missense variants. In the meantime, another two case reports with three individuals harbouring missense variants were published, bringing the total to seven (Li et al., 2021; Sharawat et al., 2021).

KMT2E targeted analysis in the 100KGP database interrogated SNVs and indels (described in this chapter) and structural variants (described in chapter 5).

Variants in *KMT2E* were extracted from the aggregate files and annotated as described in section 2.3.2 then prioritised as described in 2.5.4.1.

The analysis identified eight LoF variants in individuals that were not solved including a complex SV (Table 3.4). The individual with complex SV is described in section 5.4.2.1 (I6_5). For seven individuals, the variant explains the phenotypes, so they were submitted as diagnostic variants that were later accepted by the clinician. However, for the remaining one, which harboured a +2-splice acceptor variant (NM_182931.3:c.3844+2T>C), a "contact the clinician" form was submitted to obtain further clinical phenotypes for the participant as only two terms were available in the research environment (RE) including polycythaemia, and hypertrophic cardiomyopathy. Despite several attempts, the contact was not established, therefore, it remains as "VUS".

In two individuals a recurrent variant NM_182931.3(*KMT2E*):c.4829dup; p.Leu1610Phefs*259 was identified, which was also reported in O'Donnell-Luria et al., 2019. This position has a T homopolymer, which would explain being a hotspot. The data for these individuals were checked to ensure these findings correspond to different individuals and are different to the patient described in the initial report. The two individuals reported here are males while the one in the paper is female. The two males have different year of birth and from different ethnicities (one is 13 yrs-old who is from mixed ethnicity: White and Black Caribbean while the other is 12 yrs-old who is White Irish).

Table 3.4: Diagnostic LoF variants identified in KMT2E in participants within the 100KGP and their reported phenotypes

Individual	Variant	HPO terms
1_3	NC_000007.14(NM_182931.3): c.4829dup; p.(Leu1610Phefs*259)	Abnormality of the gastrointestinal tract Intellectual disability Delayed gross motor development Tall stature Global developmental delay Autistic behaviour Large for gestational age Delayed fine motor development Macrocephaly Abnormality of earlobe Broad hallux phalanx
2_3	NC_000007.14(NM_182931.3): c.4829dup; p.(Leu1610Phefs*259)	Generalised hypotonia Scoliosis Abnormal muscle tone Abnormality of brain morphology Premature birth Nystagmus
3_3	NC_000007.14(NM_182931.3): c.2346_2356del; p.(Val783Serfs*2)	Epileptic spasms EEG with generalised epileptiform discharges Seizure Infantile spasms Multifocal epileptiform discharges Infantile encephalopathy Generalised seizures Intellectual disability - moderate EEG abnormality hypsarrhythmia
4_3	NC_000007.14(NM_182931.3): c.1944_1948del; p.(Lys649Glufs*8)	Focal seizures Severe visual Impairment Uplifted earlobe Severe global developmental delay
5_3	NC_000007.14(NM_182931.3): c.1213C>T; p.Arg405*	Seizures Intellectual disability Limbic malformations
6_3	NC_000007.14(NM_182931.3): c.-189_-189+1insAA	Premature birth Wide nasal bridge Prominent forehead Macroglossia Delayed speech and language development Hypertelorism Global developmental delay Communicating hydrocephalus Muscular hypotonia Inability to walk Delayed fine motor development Macrocephaly Delayed gross motor development
7*_3	NM_182931.3:c.2197-1181_2524del NM_182931.3:c.2596+137_2596+223del NM_182931.3:c.2597-127_3756-74del NM_182931.3:c.2596+223_2597-128inv	Global developmental delay Macrocephaly Delayed fine motor development Delayed gross motor development Intellectual disability, moderate Delayed speech and language development

For missense and in-frame indel variants, the analysis found 63 instances predicted as deleterious by either SIFT or PolyPhen-2, and with a CADD score greater than 20. Given the uncertainty surrounding the functional impact of missense variants, particularly in newly described genes without replication in published research, a stringent filtering approach was applied. This involved retaining variants where both prediction tools agreed on variant being deleterious or damaging. This prioritised eight deleterious missense variants with CADD scores ranging from 24.5-32 (Table 3.5). Only one is confirmed *de novo*, one segregating, and six are singleton with no parental data to confirm origin (Figure 3.8). While the patients with LoF variants exhibit similar phenotypes to that seen in ODLURO syndrome, the phenotypes of patients with missense variants are very heterogeneous.

In individual_9 with the *de novo* variant (NM_182931.3(KMT2E):c.1225C>T; p.Arg409Trp), only one HPO term (proportionate short stature) is similar to the reported clinical phenotypes in patients with ODLURO (Li et al., 2021; O'Donnell-Luria et al., 2019; Sharawat et al., 2021). On the other hand, individual_12 presents a close phenotypic similarity to ODLURO, but the origin of the variant cannot be ascertained. Therefore, 'contact the clinician' forms were submitted; however, contact was not established and both variants remain as a "VUS".

Nonetheless, with no apparent clustering of the missense variants across *KMT2E* and the clinical phenotypic heterogeneity in patients harbouring deleterious rare variants therefore, the implication of missense variants in the development of ODLURO calls for further investigations.

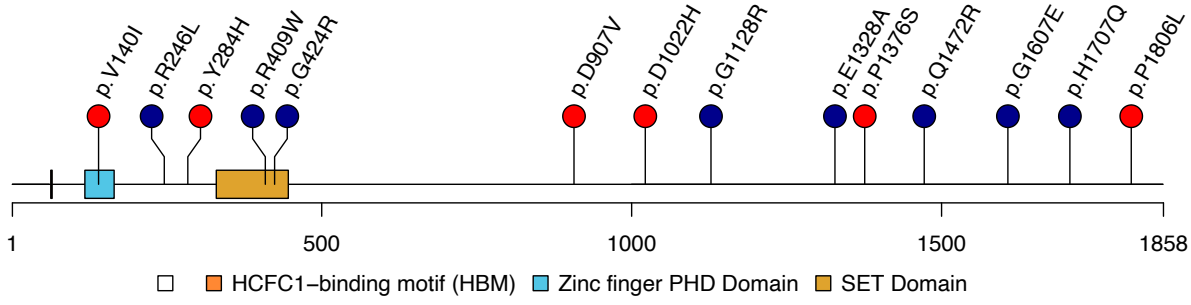


Figure 3.8: Distribution of identified & reported missense variants in KMT2E. Previously reported in the literature and those in patients recruited to the 100KGP are shown in this graph. The previously reported missense variants are in red colour, while the 100KGP cases are in blue

Table 3.5: List of rare deleterious missense variants in the 100KGP. HPO terms in bold indicates overlap with ODLURO clinical phenotypes.

Case	Gender	Normalised disease	HPO	Variant	Existing_ variation	SIFT	Polyphen-2	AF	CADD	Inheritance
8	Male	Cystic Kidney disease	Hypertension Multiple renal cysts	NM_182931.3(KMT2E): c.737G>T; p.Arg246Leu	rs768601012	deleterious	probably damaging	-	24.6	Singleton
9	Male	IUGR and IGF abnormalities	Unilateral conductive hearing impairment Anterior pituitary hypoplasia Preauricular skin tag Ectopic posterior pituitary **Proportionate short stature** Growth hormone deficiency	NM_182931.3(KMT2E): c.1225C>T; p.Arg409Trp	rs1421636564	Deleterious	probably damaging	-	32	***De novo*** (Trio with mother and father)
10	Female	Early onset and familial Parkinson's disease	Resting tremor Spasticity Dysphagia Abnormality of the basal ganglia Parkinsonism Bradykinesia	NM_182931.3(KMT2E): c.1270G>A; p.Gly424Arg	rs1473615786	Deleterious	probably damaging	1 in gnomAD (0.000005)	32	Singleton
11	Male	Congenital myopathy	Central hypoventilation **Feeding difficulties in infancy** **Abnormality of the nervous system** **Central hypotonia** Abnormality of musculature Generalised limb weakness	NM_182931.3(KMT2E): c.3382G>C; p.Gly1128Arg	rs1343024595	Deleterious	probably damaging	1 in gnomAD (0.000005)	27	Singleton
12	Female	**Intellectual disability**	Short ribs **Autistic behaviour** **Sleep disturbance** **Generalised hypotonia** Coxa valga **Severe global developmental delay** **Microcephaly** **Abnormality of the clavicle** **Delayed fine motor development.** **Delayed speech and language development** Slender long bones with narrow diaphysis **Delayed ross motor development** **Proportionate short stature** **Anxiety** **Intellectual disability** Delayed CNS myelination	NM_182931.3(KMT2E): c.3983A>C; p.Glu1328Ala	-	Deleterious	probably damaging	-	28	Singleton
13	Male	Charcot Marie Tooth disease	Impaired pain sensation Impaired vibratory sensation Peripheral axonal neuropathy Distal lower limb muscle weakness Pes cavus	NM_182931.3(KMT2E): c.4415A>G; p.Gln1472Arg	rs1418100350	Deleterious	probably damaging	-	25.4	Singleton
14	Female	Developmental macular and foveal dystrophy	Abnormality of the fundus Mildly reduced visual acuity	NM_182931.3(KMT2E): c.4820G>A; p.Gly1607Glu	rs201171142	Deleterious	probably damaging	-	24.7	Duo with affected relative (other)
15	Male	Early onset and familial Parkinson's disease	Resting tremor Parkinsonism Bradykinesia	NM_182931.3(KMT2E): c.5121C>A; p.His1707Gln	rs769644635	Deleterious	probably damaging	-	24.5	Singleton

Currently, there is no cure for *KMT2E*-related disorders, but animal studies suggest that treatments with N-acetylcysteine (NAC) or other antioxidants may be effective. This suggestion is based on mice lacking both copies or only one copy of the mll5 gene showing DNA damage and elevated levels of reactive oxygen species (ROS). The cellular effects were reversed when the mice were given NAC, a precursor of the antioxidant glutathione (Tasdogan et al., 2016). This finding holds potential therapeutic implications for humans, as NAC supplementation has previously been used to treat glutathione depletion in acetaminophen overdose and inborn errors of metabolism that result in increased free radical damage. Further research is needed to determine whether individuals with *KMT2E* haploinsufficiency are also susceptible to increased ROS and if treatment with NAC or other antioxidants may be beneficial.

3.3 Epilepsy: the example of Jeavons Syndrome

The second group of diseases investigated in this thesis is epilepsy. It is a neurological disorder characterised by an enduring predisposition to seizures, with associated cognitive, psychological and social consequences (Devinsky et al., 2018). It occurs as the primary manifestation or as part of a syndrome. A growing list of epilepsies, considered of unknown origin, is thought to have a genetic component. The lack of understanding of the underlying pathogenic cause of epilepsy or syndromes with epilepsy hinders the treatment and management of the patient's condition. Here I focus on a specific type of epilepsy known as Jeavons syndrome (JS).

JS (also known as Eyelid myoclonia-EMA) is an early childhood generalised form of reflex epilepsy with an onset between 4-12 years old. It is characterised by (1) eyelid myoclonia with and without absences, eye closure-induced seizures and EEG paroxysms, and photosensitivity. The syndrome is slightly more frequent and has higher severity in females than males (ORPHA#: 139431). The definitive prevalence of JS is unknown but appears to represent around 7-8% of all unknown generalised epilepsies.

The main focus of earlier studies in the literature was the differential clinical diagnosis of the syndrome as there are overlaps with other genetic generalised epilepsies (GGE) presenting with eyelid myoclonia. This phenotypic heterogeneity makes it difficult for clinicians to arrive at the correct diagnosis; thus, a great emphasis was placed on clinical aspects. It is becoming increasingly

apparent that a strong genetic component plays a significant role in the onset of JS, supported by family and twin studies. In one report, >50% of patients with eyelid-myoclonia had a known family history, with affected individuals mostly being first- or second-degree relatives with various epilepsy phenotypes (S. Striano et al., 2002). These studies suggest a complex inheritance for EMA with shared genetic determinants overlapping with GGE and Genetic epilepsy with febrile seizures plus; GEFS+ (Sadleir et al., 2012).

Most reported cases suggest that the inheritance of the condition follows an autosomal dominant pattern with reduced penetrance. However, a consanguineous Palestinian family with three affected siblings suggests that a recessive inheritance may also be possible (Helbig, 2012). Although most JS/EMA cases become seizure-free with polytherapy, some patients remain drug-resistant. Hence, determining the exact genetic cause of the condition could help guide treatment and management.

3.3.1 PLPPR5 as a novel disease-causing gene for epilepsy

A quartet with two affected siblings (male and female) born to non-consanguineous parents was recruited to the HICF2 project for WGS (Figure 3.9: A). Both siblings showed typical JS symptoms. The sister's photosensitive epilepsy was more severe than the brother's, and she had convulsions, absences, and eyelid myoclonia. Her first seizure occurred at age 8 and was triggered by light emitted from the TV, with likely prior eyelid myoclonia. An EEG investigation showed a tendency to generalised epilepsy, photosensitivity, and rhythmic sharp waves with eyelid myoclonia. The brother also had a tonic-clonic seizure at a later age, at 12 years, followed by absence attacks with eyelid myoclonia. EEG showed susceptibility to generalised epilepsy and paroxysms of generalised spike-wave discharge on eye closure. He is now seizure-free on valproate and lamotrigine.

WGS data quality and sample relatedness were assessed (Methods section 2.4.3). The pedigree suggests a recessive inheritance pattern for variant prioritisation, although other models were also considered. Under the recessive model, compound-heterozygous missense variants in Phospholipid phosphatase-related gene 5 (*PLPPR5*; NC_000001.10 (NM_001037317.2):c.349C>T and (NM_001037317.2):c.37C>T) were identified through IVA (Figure 3.9: B) and Exomiser ranked *PLPPR5* the top candidate. Furthermore, it has a high score for the probability of intolerance to recessive loss of function (pREC) of 0.94.

Prioritising variants under a dominant model with reduced penetrance identified eight variants from either parent. They were prioritised as they have CADD score >20; but only three were predicted damaging by SIFT and PolyPhen-2, including one of the heterozygous *PLPPR5* variants. However, the variants in *PLPPR5* are the only variants shared between the affected siblings and show biparental inheritance (Table 3.6).

Table 3.6: Prioritised inherited heterozygous variants in affected siblings from either parent

Chr	POS	REF	ALT	Gene	AA change	Inherited	SIFT	PolyPhen-2	CADD	gnomAD AF
11	10779088	T	C	***CTR9***	p.S1145P	M	Damaging	Probably Damaging	28.1	-
1	98956630	G	A	***PLPPR5***	p.R117C	F	Damaging	Probably Damaging	27.4	0.00019
2	240742970	G	A	*KIF1A*	p.S1200L	M	Tolerated	Benign	23.3	0.00006
22	37374406	A	G	*ELFN2*	p.F377L	F	Tolerated	Benign	22.9	0.00001
1	99004635	G	A	*PLPPR5*	p.L13F	M	Tolerated	Benign	22.8	0.00003
15	52384258	G	A	*MYO5A*	p.S606F	M	Tolerated	Benign	22.6	0
18	70326192	A	G	***SOCS6***	p.I508M	F	Damaging	Probably Damaging	22.2	0.00001
2	208346075	A	G	*PIKFYVE*	p.S1713G	F	Tolerated	Benign	21.4	0.00004

AA: Amino acid; M= mother; F= father

The variants were Sanger validated and showed segregation with the phenotypes when the unaffected sibling DNA was checked for the presence of the variants (Figure 3.9: C).

The paternally inherited (p.R117C) variant is predicted to have a more deleterious effect than the maternally inherited p.L13F. However, the latter falls within the first transmembrane domain, and a leucine-to-phenylalanine substitution would not be expected to change the hydrophobicity of the membrane. Yet, the aromatic ring in phenylalanine may alter the membrane curvature, as it is likely to interact more favourably with lipid headgroups compared to leucine. In contrast, p.R117C is part of the di-arginine retention/retrieval motif, which is expected to be disrupted by the substitution. This motif functions as a quality control mechanism for the correct folding and assembly of the complex governing endoplasmic reticulum exit (Gouw et al., 2018). Disruption of this motif could prevent the protein from reaching the plasma membrane and failing to form the neurites/ filopodia. In these individuals, the p.R117C mutation is predicted to be the primary driver of the phenotype, with the p.L13F mutation partially contributing through altering the membrane curvature but not

fully behaving as the WT of PLPPR5. Yet, it might be partially mitigating, but not fully restoring, the normal function.

To date, *PLPPR5* has not been linked to any disease. However, a recent large WGS study on 17,600 patients with confirmed epilepsies identified *PLPPR5* in the top 200 identified epilepsy-related candidate genes (Epi25 Collaborative, 2019). While the different pieces of evidence point towards *PLPPR5* to have a potential disease association, it is vital to confirm the causality of *PLPPR5* in this family. Therefore, it is essential to find other cases with similar genotypes and phenotypes, which could be done by interrogating larger databases such as the 100KGP, and variant/gene matching services such as GeneMatcher. The search followed two routes genotype-to-phenotype or vice versa. The genotype approach investigated biallelic variants in *PLPPR5* in the 100KGP participants as well as submitting the gene for GeneMatcher. The 100KGP investigation involved investigating biallelic variants that are rare and considered deleterious by one of the pathogenicity predictor tools such as CADD/Polyphen/SpliceAI and are not found in a homozygous status in any unaffected participants. This analysis was inconclusive as no patients were identified to harbour rare deleterious recessive variants in *PLPPR5*. The GeneMatcher submission did not identify any matches by the time of thesis submission. The second approach was a targeted analysis of WGS data of patients with JS in the 100KGP looking at PLPPR5 specifically as well as trying to identify possible causes where no recessive variants in *PLPPR5* was identified. For this group of patients, no biallelic variants in PLPPR5 were prioritised but other candidate genes were identified, , discussed in section 3.3.2.

Phospholipid phosphatase-related gene 5 (*PLPPR5;* aliases *PRG5, LPPR5*) encodes a transmembrane protein related to the lipid phosphate phosphatase (LPP) superfamily and is mainly expressed in brain tissue. It plays a role in neuronal regeneration and plasticity (Bräuer & Nitsch, 2008; Coiro et al., 2014). *PLPPR4* (*PRG1*) regulates spinal density and memory formation (Liu et al., 2016; Schneider et al., 2018), *PLPPR3* (*PRG2*) promotes axonal branch growth (Brosig et al., 2019), *PLPPR1* (*PRG3*) helps restore function after spinal cord injury by promoting neurite growth (Broggini et al., 2010), and *PLPPR2* (*PRG4*), expressed during synaptic development and regeneration, regulates synaptic LPA levels and is linked to epilepsy and brain damage. *PLPPR5* (*PRG5*), a new member of the plasticity-related protein family, is expressed in the brain and spinal cord, promotes filopodia formation and neurite growth and is highly expressed in high plasticity

areas like the hippocampus. *PLPPR5* expression is lower in the cortex, cerebellum, stratum radiatum, and striatum and is expressed in the spinal cord during development (Broggini et al., 2010; Coiro et al., 2014).

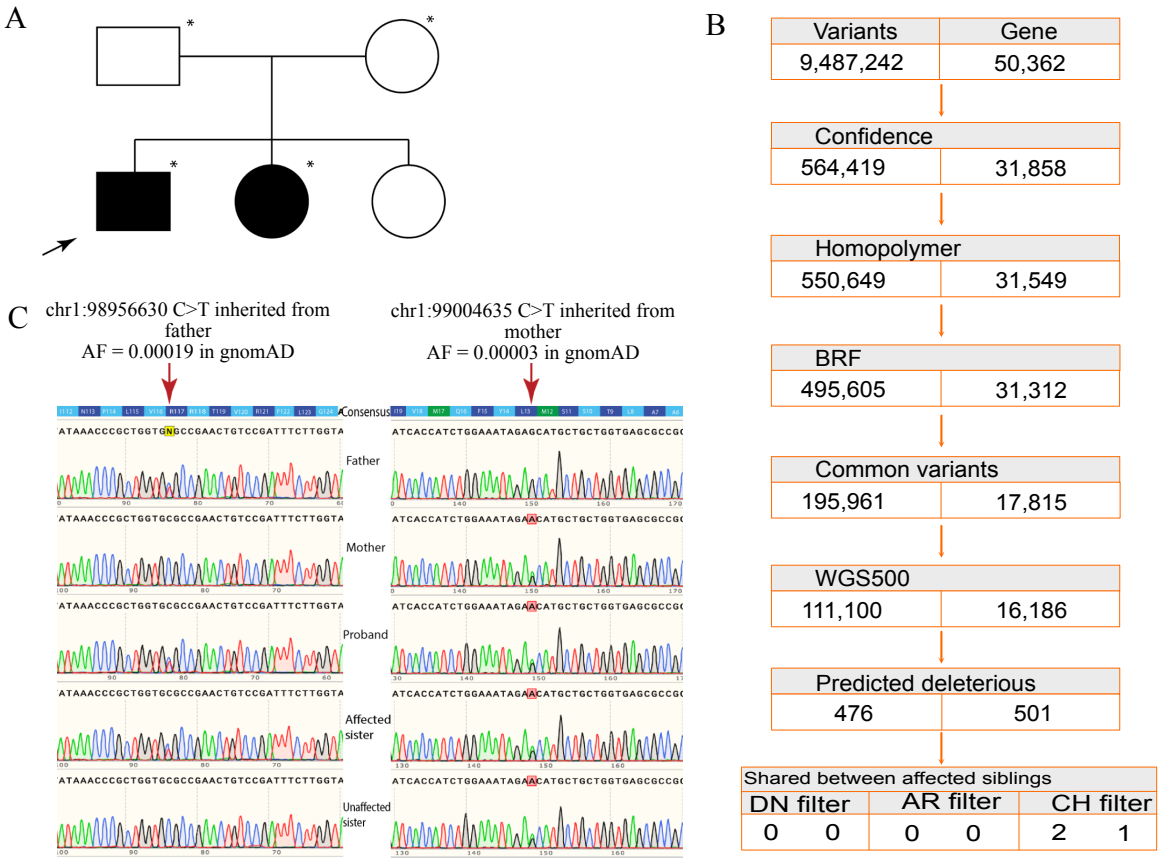


*Figure 3.9: Segregation of a compound heterozygous variants in PLPPR5. A) Pedigree of the JS case recruited to the HICF2 programme comprising two affected siblings with JS, unaffected parents and an unaffected sister. * denotes individuals sequenced and arrow denotes proband. B) PLPPR5 compound heterozygous variants were the only variants shared between the two affected siblings. C) Sanger validation of the paternally and maternally inherited compound heterozygous variants in PLPPR5 in both affected siblings and confirmed segregation with phenotype as only one of the variants is in the unaffected sister.*

Functional studies showed that PLPPR5 localises to the plasma membrane of neuronal cells leading to an increase in filopodia formation when overexpressed in N1E-115 neuronal cells (Broggini et al., 2010a). This effect was further supported by a study on primary murine hippocampal neurons showing that upregulation or downregulation of *PLPPR5* affected the number of dendritic spines (Coiro et al., 2014). Electrophysiological recordings in *PLPPR5*-knockdown neurons showed reduced spontaneous miniature excitatory synaptic current (mEPSC) frequencies. This reduction resembles what is observed in Stargazer- mice characterised by absence seizures (Menuz & Nicoll, 2008).

Only recently, a KO mouse model was generated using CRISPR/Cas9, showing that *Plppr5*$^{-/-}$ mice have lower hippocampal mossy fibre sprouting and a lower seizure threshold compared to wild type mice (Sun et al., 2021). Mossy fibre sprouting is a dynamic process that may involve replacing or restoring lost synaptic connections rather than forming new excitatory circuits in dentate granule cells (Bittencourt et al., 2015; Buckmaster, 2014; Ni et al., 2019; Yamawaki et al., 2015). However, its role in epilepsy is debated. Curiously, optogenetic activation of sprouted mossy fibres led to a reliable firing of action potentials in postsynaptic dentate granule cells following single light pulses, contributing to network hyperexcitability (Beenhakker & Ritger, 2019).

Furthermore, it was reported that *PLPPR5* has a protective effect against injury linked to melatonin on certain neurobehavior and cognition through an unknown mechanism (Sun et al., 2021). The same group showed that *Plppr5* KO directly impacts neuronal zinc homeostasis and worsens hypoxia-induced mitochondrial oxidative stress damage (Wang et al., 2022). Seizures and acquired epilepsy are influenced by changes in neuronal metabolism, mitochondrial homeostasis, and increased reactive oxygen species (Kovac et al., 2017). Additionally, *Plppr5* KO was found to worsen oxidative stress-induced neuronal damage, lower the number of surviving mitochondria, and increase mitophagy activity (Wang et al., 2022).

Taken together, the deleterious variations in a gene highly expressed in the brain and belonging to a family involved in various stages of neuronal development and maintenance, with some members implicated in seizure development, make *PLPPR5* a novel candidate gene for epilepsy. This is supported by mouse models, and the functional studies undertaken to validate this conclusion as part of this thesis are discussed in Chapter 4.

3.3.2 A second case of Jeavons syndrome points towards another potential novel epilepsy gene: MGRN1

The aetiology of JS is not well understood, and the mechanisms underlying the eye closure-induced seizures remain unclear. Therefore, identifying patients with JS is essential for several reasons. First, it allows for the accurate diagnosis of JS and differentiates it from other forms of epilepsy. Second, it enhances our understanding of the clinical features, diagnosis, management, and outcome of this rare form of epilepsy. Third, identifying patients with JS allows for investigating the underlying mechanisms and potential genetic factors involved in developing JS. Therefore, it is crucial to continue the efforts to identify more patients and to increase our understanding of this rare form of epilepsy.

The 100KGP database was explored to find other JS cases, which led to the identification of 17 families with either one or multiple affected members exhibiting phenotypes indicative of JS diagnosis (release main_programme_v16_2022_10_13). The main phenotype for prioritisation used was eyelid myoclonia, with or without absence seizure as it is one of the distinctive diagnostic features of JS (Smith et al., 2018).

Only one of the 17 families was solved by the 100KGP pipeline due the presence of a *de novo* missense variant (NC_000015.9(NM_001271.4):c.3937; p.R1313*; CADD = 37) in Chromodomain-helicase-DNA-binding protein 2 (*CHD2*), a known gene for JS, in an affected male proband.

For the remaining families, I performed SNV and SV analyses, which identified five candidates, of which four were in known epilepsy genes (Table 3.7). These potential candidates were then reported back to the clinician through a “contact the clinician” form, which is then reviewed by the clinical collaborator team before forwarding it on to the recruiting clinician.

Table 3.7: A list of variants identified in five families with JS in the 100KGP

Family	Pedigree	Gene	Variant	Inheritance	gnomAD	SIFT/ PolyPhen-2	CADD
1	Trio- with affected father and grand-father-father	*KCNMA1*	*NM_001161352.2(KCNMA1):c.2109C>A (p.TYR703Ter)*	Present in the three affected family members	-	-	**43**
2	Trio- with affected father and unaffected mother	*SCN1A*	*NM_001165963.4(SCN1A):c.1327G>A (p.Glu443Lys)*	Inherited from affected father (VCV001309815.7)	0.003%	PoyPhen-2: probably damaging	24.8
3	Quartet- with affected mother, grand-mother, and maternal aunt	*SCN1B*	*NM_001037.5(SCN1B):c.363C>G (p.Cys121Trp)*	Present in the four affected family members VCV000009252.28	0.0028%	PoyPhen-2: Deleterious/ probably SIFT: damaging	25.7
4	Duo- with affected mother	*KCNQ5*	*NM_019842.4(KCNQ5):c.863A>G (p. Glu288Gly)*	Missense inherited from affected mother	-	Damaging	32
5	Duo- with unaffected mother	*MGRN1*	*NM_015246.4(MGRN1):c.1482+1G>A*	In ROH (consanguineous family) absent from gnomAD and no homozygous LoF variants in gnomAD	-	Damaging	33

The variants in *KCNMA1, SCN1A, SCN1B,* and *KCNQ5* were considered as causative for the respective families. These were all in families with multiple affected members with at least five years diagnostic odyssey. The reporting of variants identified within the GEL RE follows a tiering system (please refer to Figure 1.1), which classifies the variants identified for families 1-4 as tier 3 variants. These are variants that are not expected to be reviewed or reported by the NHS genomic medicine centres (GMCs). However, given that the variant in *KCNMA1* is a stop gain in a known gene, it was puzzling as to why it has not been classified as tier 1. One possibility is the classification of the gene as approved "green" gene in the *in-silico* panels applied to the WGS data of the patient for analysis, therefore, I have inspected the panels applied to the patients (Table 3.8). The panels applied for this family was an earlier version in which the gene (*KCNMA1*) was not approved as a green gene, hence it was a tier 3 variant as it is in all affected members.

Table 3.8: In silico panels applied to the WGS data of families 1-4

Family	Panel applied	Version with gene listed as green
1	Brain channelopathy **v1.52**	**v1.55**
	Genetic epilepsy syndromes **v1.13**	**v1.266**
	Mitochondrial disorders v1.9	-
2	Familial Focal Epilepsies v1.9	v0
	Familial Genetic Generalised Epilepsies v1.22	v0
3	Genetic epilepsy syndromes v1.365	v0
4	Epileptic encephalopathy v1.143	v0.335
	Intellectual disability v2.447	v0

For family 5, a rare deleterious homozygous LoF variant in a gene that is not linked to any human disease in OMIM, or the literature was identified and a contact the clinician form was initiated to explore its potential effect further. However, the remaining 11 families no candidate variants were identified.

Family 5: An 18-year-old female patient diagnosed with JS was enrolled in the 100KGP for WGS along with her mother, for whom no clinical information was recorded. The information in the RE also indicated the presence of consanguinity. Furthermore, there is a detailed list of phenotypic terms (Table 3.9). Structural variants analysis (Chapter 5 Section 5.2) did not identify any SVs of clinical relevance. ROH across the genome in the proband covered ~158 Mb (5.2%), consistent with second cousins and with the 100KGP record, however, the degree of consanguinity was not further elaborated by the clinician or the family.

Table 3.9: List of HPO terms reported for the proband in family 5 in the 100KGP RE

HEAD & NECK	***Face***	Micrognathia
	Eyes	Strabismus Synophrys
ABDOMEN	***Gastrointestinal***	Feeding difficulties Constipation
SKELETAL		Joint hypermobility
NEUROLOGIC	***Central Nervous System***	Absence seizures with eyelid myoclonia EEG abnormality Febrile seizures Focal seizures Generalized clonic seizures Generalized seizures Generalized tonic- clonic seizures Global developmental delay Intellectual disability Sleep disturbance
	Behavioural Psychiatric Manifestations	Behavioural abnormality

SNV filtering prioritised six variants, for which the alignments were inspected in IGV to confirm that they are not false positive.

Three variants of the prioritised ones had low CADD scores and predicted as benign or tolerated variants (Table 3.10). Therefore, they were not considered for further analyses or literature curation. The other three candidates were in Tudor Domain Containing 15 (*TDRD15*), Transient Receptor Potential Cation Channel Subfamily A Member 1 (*TRPA1*), and Mahogunin Ring Finger 1 (*MGRN1*). *TDRD15* has not been linked to disease in OMIM and it is expressed mainly in the testis, as well as in the liver (GTEx Portal, Sjöstedt et al., 2020). The gene expression pattern is similar to that in swine (Gòdia et al., 2020) and teleost fish (Liu et al., 2022) models being linked to spermatogenesis. In a recent study on patients with Atopic Dermatitis, skin swap samples analysed with mass spectrometry were found to have an upregulated *TDRD15* (Morelli et al., 2021). Therefore, in our JS patient, this gene is not a prime candidate for the phenotypes observed. On the other hand, *TRPA1*

is linked to autosomal dominant, Episodic pain syndrome, familial, 1 (OMIM # 615040), which is characterised by episodic pain that affects the upper body, heightened pain response on mustard oil application, as well as breathing difficulties and tachycardia during the pain episodes. While the variant cannot be completely excluded as it is predicted deleterious, these associated phenotypes are not recorded in our patient, hence the relevance to our patient was not clear and remained as a "VUS". However, at a later stage a contact with the clinician was established and the variant was disregarded.

The last candidate is a canonical splice donor site variant in *MGRN1*, falling within ~10.4 Mb ROH. The variant is absent from gnomAD and in a gene with a high pREC score of 0.89. In gnomAD, *MGRN1* is depleted for heterozygous LoF variants (LOEUF = 0.47) and no homozygous LoF. The homozygous variant c.1482+1G>A is predicted to decrease splicing efficiency by 97%, as predicted by the MaxEntScan score (from 8.41 to 0.23) and has a score of 1 for donor loss from SpliceAI. There is no nearby alternative splice site; therefore, the variant may result in exon 14 skipping leading to a frameshift and early termination.

Based on WGS data, the mother, for whom no abnormal phenotype was recorded, carried the variant in a heterozygous state. However, due to the known consanguinity, it can be inferred that the father, whose genomic data were unavailable, is also a carrier of the variant. (Figure 3.10: A).

Approval for release of stored DNA samples for the proband and the mother from GEL was delayed eight months, after which I validated the variant using sanger sequencing (Figure 3.10: B). In the interim, it was also vital to validate the effect of the variant on splicing to confirm its deleteriousness, and its effect on mRNA synthesised and in turn protein before drawing any conclusions as splicing regulatory elements are known for their degeneracy (Cartegni et al., 2002).

The most effective approach for assessing the impact of this suspected variant on splicing would be to directly analyse RNA samples obtained from individuals harbouring the variant (Acedo et al., 2012). However, no suitable stored RNA samples for the proband and the mother were available in GEL sample repository and establishing a contact with the clinician took 16 months after initial submission of the "contact the clinician" form. Therefore, using alternative approaches for validation such as minigene assay was imperative.

Table 3.10: List of variants in ROH in the proband. An ROH with varying sizes has been identified in every chromosome except chr15, chr17, chr20, chr21, and chr22. Only ROHs with a candidate variant are shown in the table below. Entries in red indicate a reason for downgrading the variant/gene as a candidate.

ROH in	Gene	Variant	gnomAD	Predictors of gene intolerance	Pathogenicity predictions	OMIM morbid gene
Chr1	Kynurenine 3-Monooxygenase (*KMO*)	NC_000001.11(NM_003679.5): c.361+8T>C	0.03%	LOEUF = 0.70 pLI = 0 pRec = 0.99 RVIS = 0.66 (84.55%*)	CADD: 1.47 SpliceAI: 0.0300	No
Chr2	Tudor Domain Containing 15 (*TDRD15*)	NC_000002.12(NM_001306137.2): c.2396C>A; p.Pro799His	-	LOEUF = 0.53 pLI = 0 pRec = - RVIS = -	CADD: 23.3 Polyphen-2: probably damaging SIFT: deleterious	No
Chr7	Kielin /Chordin-Like Protein (*KCP*)	NC_000007.14(NM_001366122.1): c.572-176C>T; (p.=) NC_000007.14(NM_001135914.1): c.553C>T; p.Pro185Ser	0.005%	LOEUF = - pLI = - pRec = - RVIS = -	CADD: 13.0 Polyphen-2:benign SIFT: deleterious	No
Chr8	Transient Receptor Potential Cation Channel, Subfamily A, Member 1 (*TRPA1*)	NC_000008.11(NM_007332.3): c.352C>G; p.Leu118Val	0.02%	LOEUF = 1.16 pLI = 0 pRec = 0.007 RVIS = 0.79 (87.29%*)	CADD: 25.0 Polyphen-2: probably damaging SIFT: deleterious	#615040 (AD)
Chr16	Mahogunin Ring Finger 1 (*MGRN1*)	NC_000016.10(NM_015246.4): c.1482+1G>A	-	LOEUF = 0.47 pLI = 0.1 pREC = 0.89 RVIS = -1.77 (2.3%*)	CADD: 34 SpliceAI_DS_DL: 1	No
Chr19	Dedicator of Cytokinesis 6 (*DOCK6*)	NC_000019.10(NM_020812.4): c.760C>T; p.Arg254Cys	0.06% (1 hmz)	LOEUF = 0.72 pLI = 0 pRec = 0.98 RVIS = -0.94 (9.42%*)	CADD: 20.9 Polyphen-2: benign SIFT: tolerated	#614219 (AR)

*Percentage in RVIS score represents the percentile of the gene amongst the most intolerant of human genes.

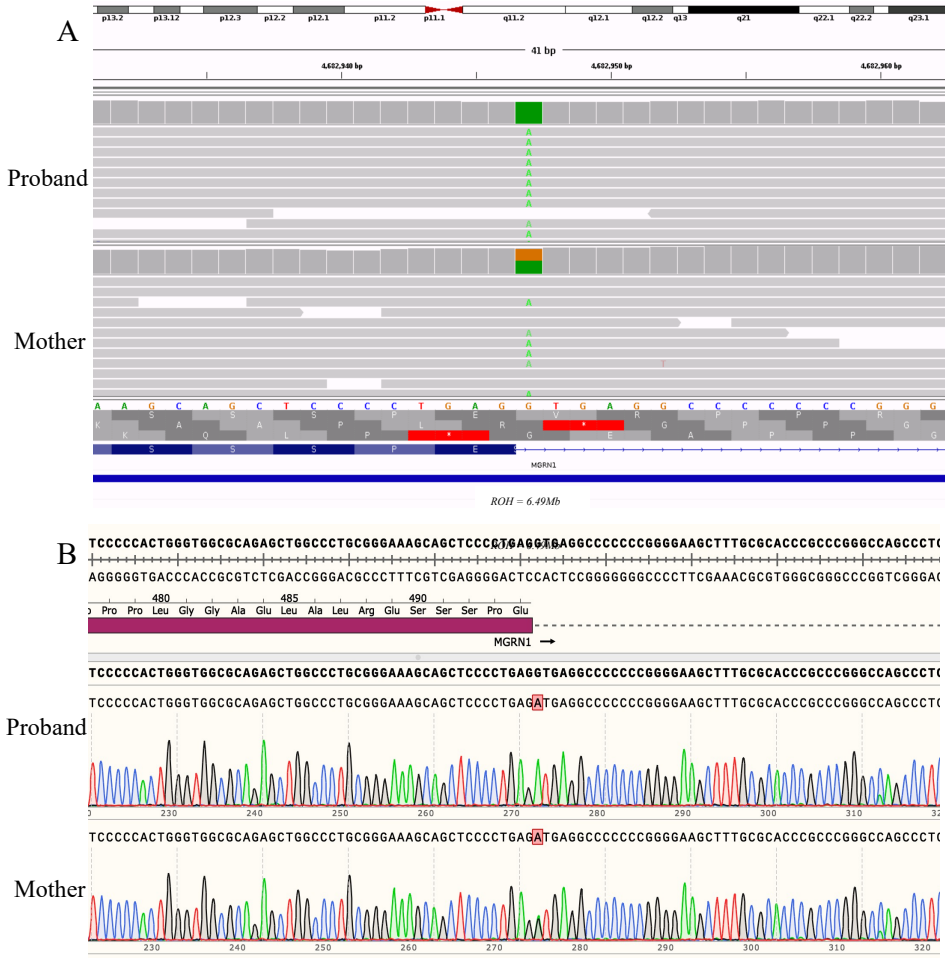


Figure 3.10: Validation of a rare homozygous c.1482+1 G>A variant in MGRN1. A) Short-read illumina alignments show the homozygous variant in the proband present in all reads and as heterozygous in the mother. The gene falls in a large region of homozygosity of 6.49 Mb. Father's genome is not available, but consanguinity is reported for this family supporting the inheritance of the second allele from a father carrying variant. B) Sanger sequencing confirms the homozygosity in the proband and heterozygous status in the mother.

3.3.2.1 The ultra-rare homozygous splice donor site variant (c.1482+1G>A) in MGRN1 leads to a frameshift and early termination using minigene system and RNA-seq

In this section, I present the work conducted to investigate the influence of the c.1482+1G>A variant in *MGRN1* on alternative splicing by analysing the mRNA isoforms synthesised using the minigene assay. The utilisation of the minigene system offers a straightforward and robust approach to identify the effect of variants on splicing, thereby giving us an insight into the pathogenicity of the splicing variant.

3.3.2.1.1 Minigene analysis confirms exon skipping due to c.1482+1G>A variant in MGRN1

The minigene reporter assay described by Singh & Cooper, 2006 was followed for this experiment. HEK293T cells were transfected with RHCglo-MGRN1-Ex14-15 (WT or mutant) or with RHCglo vector only (Figure 3.11: A-B). At 24-hour post transfection, RNA was extracted for each construct (section 4.2.2.6.1) and RT-PCR was performed. To ensure the specificity of the PCR targeting transcripts generated from exogenous expression driven by the introduced *MGRN1* exons in the minigene system, the reverse primer was designed in exon 15 of *MGRN1*, while the forward primer was positioned in the first artificial exon of the RHCglo vector (Figure 3.11: C). This primer configuration allowed for selective amplification of the desired transcripts synthesised from the minigene system.

The RT-PCR analysis of RNA extracted from cells transfected with the wild type variant revealed a prominent band representing normal splicing of exon 14 and 15 (202 bp). In contrast, cells transfected with the plasmid carrying the variant exhibited a band indicative of exon skipping. An additional weaker band, observed in the wild type RHCglo-MGRN1-Ex14-15 plasmid sample, corresponded to exon skipping at a size of 78 bp. Although unexpected, this phenomenon could be attributed to the different genomic context adjacent to the exons, as the minigene system may not encompass the entire length of intronic regions flanking the exons and might be required for splicing efficiency (Buratti et al., 2006). In this instance, 183 bp upstream of exon 14, and 144 bp downstream of exon 15 of *MGRN1*. Nevertheless, the complete absence of the larger isoform when the mutant construct is used, strongly suggests an evident exon skipping effect induced by the variant on

splicing. Notably, no bands were detected in the samples from un-transfected cells or those transfected with the RHCglo plasmid lacking the *MGRN1* exons, confirming the specificity of the screening approach (Figure 3.11: D). Therefore, we can support the deleterious effect of the variant on splicing of *MGRN1* leading to exon 14 skipping causing frameshift and early termination (p.S345Rfs*4). Supported by this finding, it was possible to obtain fresh blood samples from both the proband and the carrier mother for a more definitive analysis by RT-PCR and sanger sequencing of the product (Method section 2.6.3.8). The results of the RT-PCR and sanger sequencing confirmed exclusive exon skipping in the proband and both WT and exon skipping in the mother (Figure 3.12: A-C).

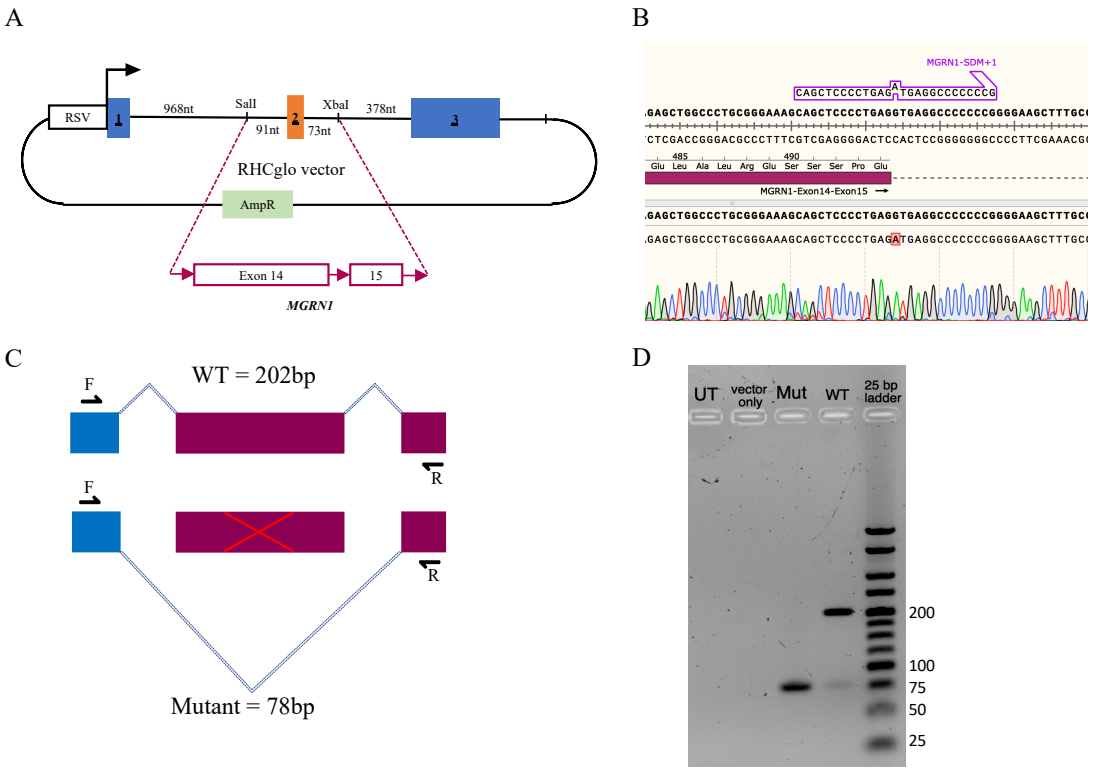


Figure 3.11: Validation of c.1482+1 G>A in MGRN1 using minigene system in HEK293T cells. A) Map of RHCglo minigene vector, which is comprised of 3 artificial exons represented as blue and orange boxes. The first artificial exon (blue), marked 1, is comprised of the last 96 nucleotides of human b-globin intron 1, the first nucleotide of globin exon 2, the 30-nucleotide skeletal troponin I (sTNI) exon 2, the last 3 nucleotides of globin exon 1, and the first 75 nucleotides of globin intron 1. The middle artificial exon, marked as 2, is flanked by 91 and 73 nucleotides of intron 1 of the human ß-globin gene. The third artificial exon, marked as 3, is the last exon of the chicken skeletal α-actin gene. This vector was used to clone exons 14-15 of MGRN1 flanked by 183 bp upstream of exon 14 and 144 bp downstream of exon 15, which are needed to facilitate the regulation and binding of the splicing machinery of the cell. The cloning utilised restriction enzymes SalI and XbaI for restriction fragment insertion. Exon-14 is predicted to be skipped due to the presence of the variant. B) primer sequence used for site-directed mutagenesis to introduce the c.1482+1G>A variant in MGRN1. The sanger sequence trace validates successful introduction of the splicing variant. C) schematic illustration of predicted mRNA size in WT (202bp) and only 78bp in the presence of variant (mutant) due exon skipping. D) RT-PCR products from RNA samples obtained from HEK293T cells transfected with RHCglo-MGRN1 plasmids. The samples include those carrying the wild type donor splice site (WT), the plasmid with the c.1482+1G>A donor splice site variant (Mut), the native RHCglo vector without MGRN1 exons cloned in (vector only), as well as un-transfected HEK293T cells (UT). The RT-PCR products are visualised in 2% agarose gel. RNA sample from cells transfected with the WT shows a dominant band at the correct size for mRNA with spliced in exon 14, while the mutant (Mut) sample shows a band at the size

corresponding to that with exon skipping.

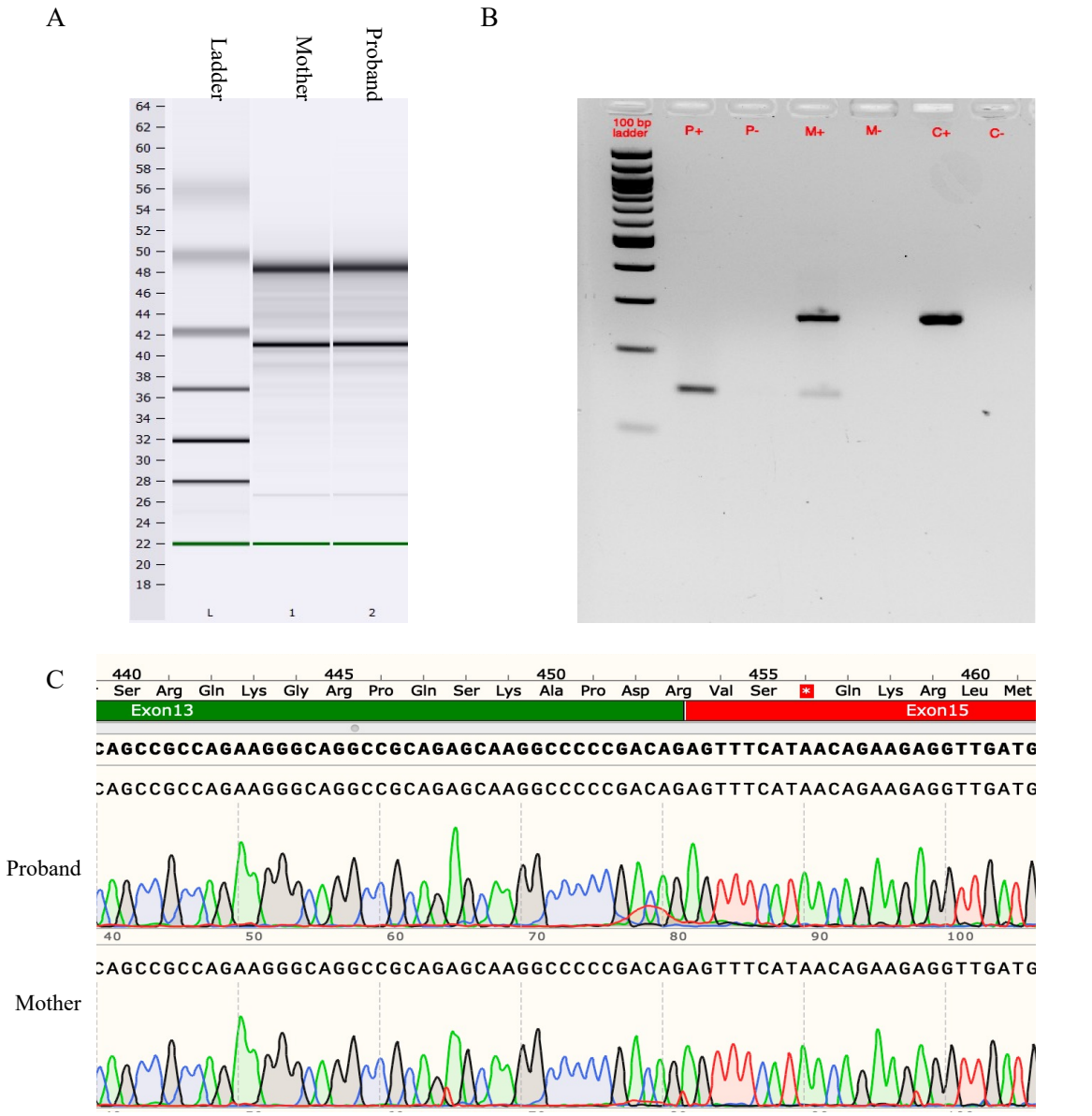


Figure 3.12: Homozygous c.1482+1G>A variant in MGRN1 shows abberant splicing in whole-blood samples from the proband and the heterozygous carrier mother. A) Quality check of RNA integrity using Bioanalyzer system giving RNA Inegrity Number (RIN) of 8.7 and 8.6 for the sample from the mother and the proband respectively. This indicates a high degree of RNA integrity. RNA extraction was performed as described in method section 2.7.1. B) visulaisation of RT-PCR products (primers in method table 2.2) shows two bands in the mother's sample (M+), one at the expected size for WT (272bp) and a second band at the expected size for exon skipping (148bp). The sample of the proband (P+) shows exclusively the band corresponding to exon skipping (148bp). Control (C+) sample from an unaffected participant shows the WT pattern. Samples with (-) sign are controls for the RT reaction, showing that bands are due to amplification of cDNA and not from genomic DNA. C) Sanger sequence of the lower band shown in the agarose gel in B.

3.3.2.1.2 RNA-seq of patient sample reveals a different splicing profile including intron retention induced in addition to exon skipping due to the presence of the c.1482+1G>A variant in MGRN1

RNA-seq provides a valuable tool for not only validating or disproving the impact of variants on splicing (Cummings et al., 2017), but also for identifying aberrant splicing events and the downstream effect of the identified variants on expression of other genes across the entire transcriptome (Kremer et al., 2017; Gonorazky et al., 2019).

RNA sequencing of whole-blood derived RNA samples from the proband, unaffected mother, and two age-matched control samples was performed (Oxford Genomics Centre). Alignment to the genome allowed the detection of reads spanning multiple exons per gene, thus facilitating the analysis of alternative splicing. The expression of *MGRN1* was not reduced in comparison to the mother and the two controls (Figure 3.13: A). All samples showed correct splicing of exons 1-13 and exons 15-17 (ENST00000262370.11). However, samples from the proband and mother showed clear evidence of exon 14 skipping as expected (Figure 3.13: B). Moreover, evidence of intron 14 retention that is linked to the presence of c.1482+1 G>A variant can be seen in both proband and mother but absent from both controls (Figure 3.13: A-B). As the minigene and RT-PCR validation did not reveal this observation (Figure 3.11: D and Figure 3.12: B), despite using primers designed to detect both the intron retained and exon skipped isoforms, an additional RT-PCR was conducted using a forward primer positioned within exon 14 and the reverse primer within intron 14. Consequently, in control cDNAs without the variant and exhibiting normal splicing, no amplification should be observed. The newly designed primers substantiated the RNA-seq findings regarding the presence of intron retained mRNAs (Figure 3.13: C). The heterozygous mother also showed correct splicing of exons 13 to 15 that is similar to the controls and consistent with the expression of the WT allele in the mother (Figure 3.13: B). Exon 14 comprises 124 nucleotides corresponding to 41.3 codons; therefore, exon skipping is expected to result in a frameshift. The intron 14 retention also introduces a premature termination codon predicted to elicit nonsense-mediated mRNA decay, which eventually results in loss of transcript and translated protein (Cartegni et al., 2002). Therefore, from the RNA-seq data we further validate the deleterious effect of the c.1482+1G>T variant on the splicing profile of *MGRN1*.

One important technical observation from this experiment is the presence of a bias in PCR-based technique validation, including RT-PCR, towards the preferential amplification of shorter DNA or cDNA fragments. This bias should be considered when validating splicing events that could result in either shorter (due to exon skipping) or longer fragments (due to intron retention). The bias arises from the inherent nature of the PCR process, which involves the annealing and extension of primers to amplify a specific DNA region. As the fragment length increases, the efficiency of primer binding and amplification decreases. Consequently, shorter fragments are amplified more efficiently, leading to a bias in favour of shorter products. In contrast, RNA-seq offers an advantage by enabling the unbiased detection and quantification of fragments across a wide range of lengths. This capability allows for the comprehensive analysis of both short and long RNA molecules, thereby minimising the length bias observed in PCR-based techniques.

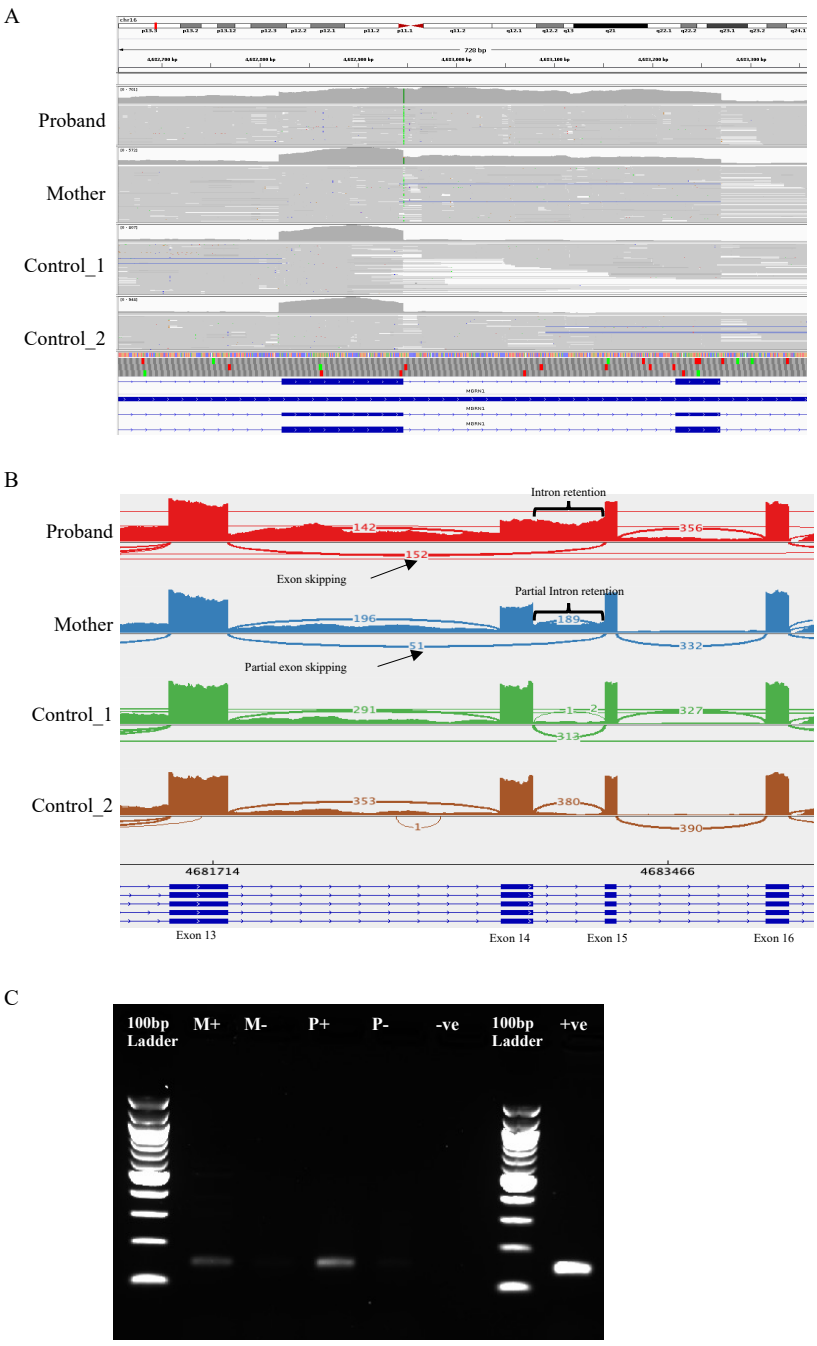


Figure 3.13: RNA-seq data from proband, mother, and control samples. The data reveal a different splicing profile including exon skipping and intron retention that was not observed from PCR and minigene assay. A) alignment of reads from RNA-seq data confirms the presence of c.1482+1 G>A variant in RNA transcripts with intron 14 retention that is absent in the controls. B) Sashimi plot shows two RNA transcripts for MGRN1 with one transcript exhibiting complete exon 14 skipping and the other showing complete retention of intron 14. In the mother an additional RNA transcript with wild type pattern splicing can be observed. Transcripts exhibiting exon 14 skipping or intron 14 retention are absent from controls who show exclusively the wild type pattern of splicing. D) Primer pair designed to capture specifically the intron retained allele show bands (150bp) in cDNA samples from proband (P+) and mother (M+) but not in RT negative samples (M- and P-) or in cDNA from a control sample (-ve) who does not carry the variant and is not expected to have a band at the expected size. gDNA sample was included as a positive control (+ve)

Mahogunin Ring Finger 1 (*MGRN1*) is a cytosolic E3 ligase that regulates various cellular functions, including intracellular trafficking, to alleviate cellular insults generated by various stress-inducing agents (Upadhyay et al., 2016; Soto & Pritzkow, 2018). It encodes a protein with a highly conserved C3HC4 RING domain, which led to the belief that it could have E3 ubiquitin ligase activity based on sequence similarities (Phan et al., 2002). A study confirmed this prediction when recombinant mahogunin partnered with Ubiquitin-conjugating enzyme E2 5 (UBC5), indicating that it has E3 ubiquitin ligase activity. This study also suggests that a lack of functional mahogunin could result in the build-up of its substrate (He et al., 2003). *MGRN1* has four isoforms resulting from the alternative splicing of exons 12 and 17, with isoforms I, II, and III highly expressed in the brain, kidney, and heart, with similar levels found in liver tissue (Bagher et al., 2006). The expression of *MGRN1* in brain tissue was also found to be high in mice and humans, with over 90% sequence similarity among orthologues and high conservation across both invertebrates and vertebrates' genomes (He et al., 2003). *MGRN2*, a paralogue of *MGRN1*, was found to play a crucial role in maintaining and regulating dendritic growth in cultured neurons and is essential for neuronal survival (Matz et al., 2015).

Several homozygous *MGRN1* KO mouse strains for the null allele ($Mgrn1^{md\text{-}nc}$) were generated using different mouse genetic backgrounds including C3H/HeJ, 129S1/SvImJ and 101/H * C3H/HeH. Embryonic lethality in C3H/HeJ and 129S1/SvImJ reached 100% with complete penetrance. Yet, mouse litters from 101/H * C3H/HeH background carrying the $Mgrn1^{md\text{-}nc}$ null allele with no detectable mRNA transcripts had a reduced viability of ~50%. The surviving litters exhibited late-onset spongiform neurodegeneration, with vacuolation and neuronal loss in various brain regions, such as the cerebral cortex, hippocampus, thalamus, brain stem, caudate putamen, and cerebellum granule layer (He et al., 2003). Animals with the same genetic background but carrying a hypomorphic allele ($Mgrn1^{md/md}$) with detectable but reduced mRNA transcripts levels did not exhibit any of these neuropathologies (Phan et al., 2002; Bagher et al., 2006). The high mortality rates and developmental defects, including congenital heart defects, abnormal craniofacial patterning, and incorrect patterning of the left-right body axis were also noted in Cota et al., 2006. While the phenotypes of the mouse models do not completely align with the main presentation of neurodevelopmental clinical features and epilepsy in the proband, the mouse model clearly does not

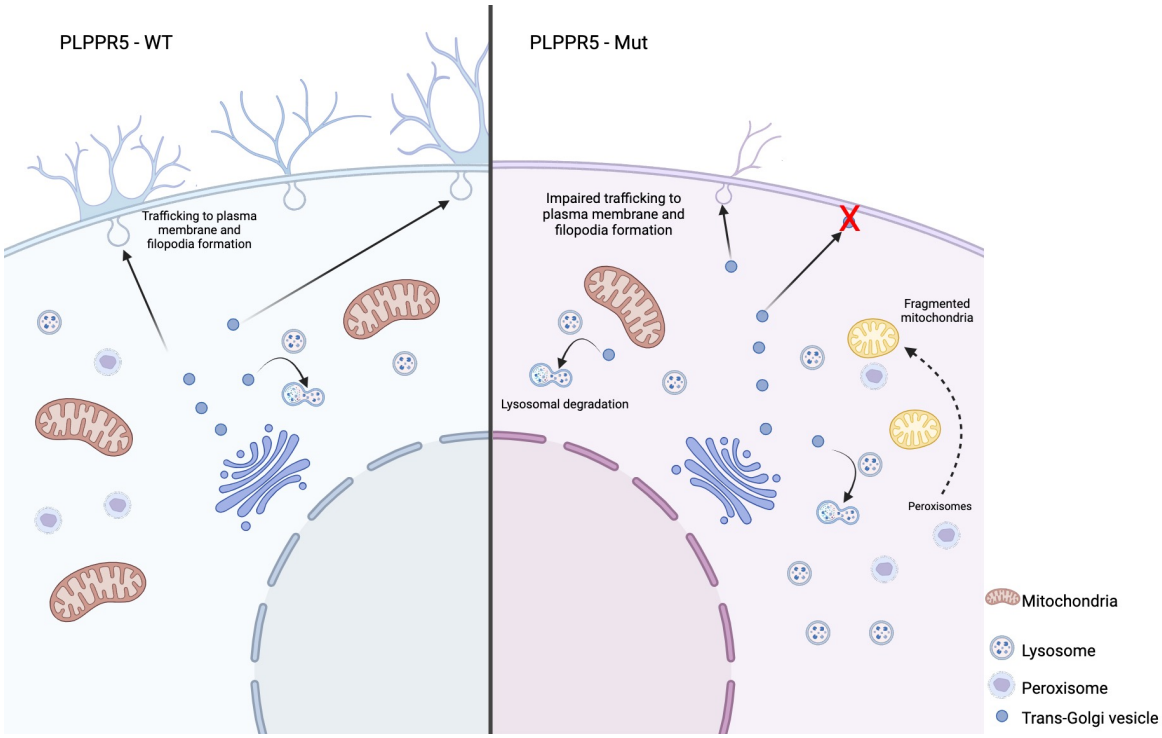
tolerate recessive loss of *MGRN1,* and the surviving litters present with variable phenotypes including neurological ones. The phenotypic differences could be attributed to species differences that are heightened when trying to capture the human brain physiology on mouse models.

Nonetheless, given the nature of identified variant in our patient and its deleterious effect on splicing, absence from population databases, high expression in brain, conserved function, and negative impact on viability and phenotype demonstrated in the KO mouse model, *MGRN1* is a strong candidate for being a novel human disease gene. Further investigations into the impact of the recessive variants in *MGRN1*, whether conducted using patient-derived cell lines or induced pluripotent stem cells (iPSCs), hold the potential to significantly enhance our understanding of its contribution to disease development. Moreover, these studies may also contribute to revealing the disrupted pathways underlying JS. Notably, it is possible that the endosomal-lysosomal and autophagy pathways in which *MGRN1* plays a main role, and which are typically associated with lysosomal storage disorders (Myerowitz et al., 2020), may play a role in the pathogenesis of JS, as recent evidence has demonstrated their impairment in other forms of absence epilepsies (Tanaka et al., 2018).

In conclusion, this chapter has discussed the genetic analysis of WGS data from patients with neurodevelopmental disorders and epilepsy. The findings highlight the importance of WGS in identifying previously missed structural variants in either newly described *DOCK7* or hard-to-detect *ARX* and complex rearrangements. Additionally, the discovery of three novel disease associations further expands our understanding of the genetic basis of these disorders. *KMT2E* has already led to the diagnosis of more than 70 new patients, including seven new patients in the 100KGP, while *PLPPR5* and *MGRN1* are promising candidates based on strong evidence and merit further functional studies. These results demonstrate the ongoing potential of WGS in improving our understanding of the genetic causes of neurodevelopmental disorders and epilepsy.

Graphical abstract chapter 4

Impact of deleterious variants in *PLPPR5* in filopodia formation, autophagy, and mitochondrial function, highlighting a potential novel disease-causing association for *PLPPR5* contributing to epilepsy development



Chapter 4. Functional validation of Jeavons Syndrome patient-specific mutations in *PLPPR5*

4.1 Introduction

Understanding the impact of JS patient-specific variants in a gene that has a potential novel disease association might shed light on the molecular mechanisms and pathways involved in disease development. Conducting *in vitro* functional studies can result in valuable insights into the intricate molecular mechanisms underlying the disorder and offer prospect of uncovering new therapeutic strategies.

In cases where a potentially novel disease-causing gene is involved, determining the effect of genetic variants on gene expression and function is crucial, and *in vitro* functional studies play a particularly important role in this regard. Validation of the effect of an identified variant serves as robust evidence of its pathogenicity, and this helps to classify it as a pathogenic or likely pathogenic variant in accordance with the ACMG guideline (Richards et al., 2015). In turn, this contributes significantly to confirming a genetic diagnosis for patients with rare diseases.

I previously discussed a case from the HICF2 project involving a quartet with two affected siblings (male and female) by JS born to non-consanguineous parents (section 3.3.1). JS is an early childhood generalised form of reflex epilepsy with an onset between 4-12 years old. It is characterised by eyelid myoclonia with and without absences, eye closure-induced seizures and EEG paroxysms, and photosensitivity. The syndrome is slightly more frequent and has higher severity in females than males (ORPHA#: 139431). Compound heterozygous variants in *PLPPR5* were identified as the top candidate for JS through whole genome sequencing, supported by several lines of evidence. These include the co-segregation of the variants in the affected family, a high pREC score for *PLPPR5*, the expression and function of the gene, and the suggestive literature implicating it in epilepsy development.

PLPPR5 belongs to phospholipid phosphatase-related proteins (PLPPRs), also known as PRGs (plasticity-related genes), which are key regulators of neuronal growth and synaptic transmission in the central nervous system. Through multiple studies, it has been revealed that the different members

of PLPPRs play important roles in processes such as filopodia formation (Broggini et al., 2010; Coiro et al., 2014; Savaskan et al., 2004; Velmans et al., 2013), axonal guidance (Cheng et al., 2016), branching during nervous system development and regeneration, and regulation of dendritic spine number and excitability (Liu et al., 2016b). Loss of PLPPRs expression in KO mice has been shown to increase susceptibility to seizures (Wang et al., 2021) and lead to defects in sensory information processing (Cheng et al., 2016), development of psychiatric disorders, stress-related behaviours, and abnormal social interaction (Vogt et al., 2016). However, despite these findings, the exact role of PLPPRs in the development of neurological diseases remains largely ambiguous.

PLPPR5, like other PLPPR family member proteins, is an integral transmembrane protein and has six membrane-spanning domains with both C- and N-termini being locked in the cytosol (Bräuer & Nitsch, 2008; Brindley, 2004; Strauss & Bräuer, 2013). It exhibits detectable expression during embryonic stages, particularly in the dentate gyrus of the hippocampus (Broggini et al., 2010; Coiro et al., 2014; Gross et al., 2022). In adult animals, *PLPPR5* is prominently expressed in the hippocampus, cerebellum, and striatum. Its levels are comparatively low in regions such as the thalamus, hypothalamus and pons (I. Gross, Brandt, et al., 2022). Both inhibitory and excitatory neurons express *PLPPR5* (Gross et al., 2022). Previous studies have shown that overexpression of *PLPPR5* induces the formation of membrane protrusions in non-neuronal cell lines (Broggini et al., 2010), and dendritic filopodia and spine formation in mature hippocampal and cortical neurons (Broggini et al., 2010; Coiro et al., 2014). On the other hand, the downregulation of *PLPPR5* was shown to reduce the number of neurites in immature cortical neurons (Broggini et al., 2010).

Filopodia and dendritic spines play a critical role in facilitating excitatory synaptic inputs in cortical neurons, thereby contributing significantly to excitatory neuronal signalling within the brain (Tønnesen & Nägerl, 2016). Given that epilepsy primarily arises from disruptions in electrical excitability, it is feasible that abnormalities in dendritic spines could contribute to epileptogenesis and seizure generation. The delicate balance between excitatory and inhibitory forces in the brain is critical, and any disturbance in this equilibrium can lead to hyperexcitability and seizures (Žiburkus et al., 2013). Changes in the structure or function of dendritic spines can impact the processing of synaptic inputs and, subsequently, the overall excitability of neurons. While a loss of dendritic spines and associated excitatory inputs may initially suggest reduced cellular excitability, the ultimate

impact on neuronal circuit outputs is likely more intricate. For instance, if the loss of excitatory inputs predominantly affects inhibitory networks, the net effect on circuit output may be increased excitability and a propensity for seizures.

To investigate the effect of the identified mutations on the expression and function of *PLPPR5*, the characterisation of the wild type (WT) and mutants in COS7 (monkey cells exhibiting fibroblast morphology, commonly used for cell biology studies) and HEK293T (human cells that are capable of efficient protein expression) lines was undertaken in this study. The aim was to investigate the effect of the *PLPPR5* patient mutations on protein stability, on the generation of filopodia-like extensions and to determine their effect on cellular metabolism through the studies of mitochondrial morphology and function.

4.2 Materials and Methods

4.2.1 Materials

4.2.1.1 Reagents and consumables

All chemicals used in this study are listed in Table 4.1, kits in Table 4.2, and consumables used for cell culture in Table 4.3. All reagents were stored according to the manufacturer's instructions in room temperature or at 4ºC.

Table 4.1: List of chemicals and their suppliers

Chemical	**Supplier**	**Cat. #**
4x Nupage LDS Loading Dye	Invitrogen	NP0007
Calcein Blue, AM-1mg	Life Technologies	C1429
Calcium Chloride (1M)	Sigma-Aldrich	21115
Cycloheximide Solution (10% In DMSO Sterile)	Cambridge Bioscience	C084
Dimethyl Sulfoxide	Sigma-Aldrich	D2650
Dithiothreitol (DTT)	Life Technologies Ltd	R0861
Ethidium Bromide	Sigma-Aldrich	1116080030
Glycerol	Sigma-Aldrich	49781
L-Carnitine, 99+%	Alfa Aesar	A17618.09
Magnesium Chloride (1M)	Sigma-Aldrich	M1028
MG132 (Carbobenzoxy-Leu-Leu-Leucinal)	Selleck Chemicals	S2619
Seahorse XF Palmitate-BSA FAO Substrate	Agilent	102720-100
SlowFade Gold antifade reagent	Invitrogen	S36937
Phenylmethanesulfonyl Fluoride Solution (PMSF; 0.1M)	Merck Life Science	93482
Trypan Blue Dye	Bio-Rad	1450021
Tween20	Sigma-Aldrich	P7949

Table 4.2: List of kits and consumable products

Kit or consumable	**Supplier**	**Cat. #**
100bp DNA ladder	New England Biolabs	N3231
Agarose	Sigma-Aldrich	A9539
BamHI-HF	New England Biolabs	R3136
Bsu36I	New England Biolabs	R0524
Chameleon Duo Pre-Stained Protein Ladder	Li-COR	928-60000
Chemically competent DH5α *E. Coli* cells	Invitrogen	C404003
EcoRI-HF	New England Biolabs	R3101
Genejet Plasmid Miniprep Kit	Thermo fisher Scientific	K0503
iQ SYBR green Supermix	Bio-Rad	170-8882
Kpn1	New England Biolabs	R3142
LB-agar	Thermo fisher Scientific	22700041
Lennox LB medium	Sigma-Aldrich	L7275
Mem-PER Plus Membrane Protein Extraction Kit	Life Technologies Ltd	89842
Nitrocellulose membranes (Immun-Blot 0.2µm pore size)	Bio-Rad	1704158
Nucleobond Xtra Midi kit	MACHEREY-NAGEL	740410.5
NuPage Novex precast 4-12% Bis-Tris polyacrylamide gels	Invitrogen	NP0321
PCR and DNA Cleanup kit	New England Biolabs	T1030S
Pierce 660nm protein assay kit	Thermo fisher Scientific	22662
Pngase F	New England Biolabs	P0704S
Q5 High-Fidelity DNA Polymerase	New England Biolabs	M0491
Qiaprep Spin Miniprep Kit	Qiagen	27104
Qiaquick Gel Extraction Kit	Qiagen	2870
Quantitect Reverse Transcription Kit	Qiagen	205311
Qubit^{tm} dsdna HS Assay Kit	Thermo fisher Scientific	Q32851
Quikchange Lightning Site-Directed Mutagenesis Kit	Agilent Technologies	210518
Rnasezap rnase Decontamination Solution	Thermo fisher Scientific	R2020
Rneasy Mini Kit	Qiagen	74104
S.O.C. medium	Invitrogen	15544034
Seahorse XF Cell Mito Stress Test Kit	Agilent Technologies	103015-100
Seahorse XFe96 fluxpak mini	Agilent Technologies	102601-100
Shrimp Alkaline Phosphatase (rsap)	New England Biolabs	M0371
Subcellular Protein Fractionation Kit for Cultured Cells	Thermo Scientific™	78840
T4 DNA Ligase	New England Biolabs	M0202

Table 4.3: Cell culture consumables

Product	**Supplier**	**Cat. #**
0.25% Trypsin-EDTA	Sigma-Aldrich	T3924
35-Mm Glass-Bottom Confocal Dishes	WPI	FD35-100
African Green Monkey Kidney Fibroblast-Like COS7 Cell Line – a kind gift from Prof. Ira Milosevic	ATCC	CRL-1651
Counting Slide	Bio-Rad	1450003
Dulbecco's Modified Eagle's Medium	Sigma-Aldrich	D6546
Fetal Bovine Serum Heat Inactivated	Merck Life Science	F9665
Fluorobrite™ DMEM	Gibco	A1896701
Glutamax Supplement	Thermo Scientific	35050087
Human Embryonic Kidney 293 Cells (HEK293-T) - a kind gift from Dr. Catarina Oliveira	ATCC	CRL-3216
L-Glutamine Solution (200mm, Sterile-Filtered)	Sigma-Aldrich	G7513
Lipofectamine 3000	Invitrogen	L300000
OPTI-MEM (Reduced Serum Medium)	Gibco	31985062
Penicillin-Streptomycin, With 10,000 Units Penicillin And 10 Mg	Sigma-Aldrich	P0781
Seahorse Xfe96 Fluxpack	Agilent	102601-100
XF 1.0M Glucose Solution	Agilent	103577-100
XF 100mm Pyruvate Solution	Agilent	103578-100
XF 200mm Glutamine Solution	Agilent	103579-100
XF DMEM Medium (pH 7.4)	Agilent	103575-100

4.2.1.2 Solutions

All solutions and buffers used in this study are listed in Table 4.4 with instructions of preparations where relevant.

Table 4.4: Solutions

Solution	**Components/recipe**	**supplier**	**cat. #**
Poly-D-Lysine	1 mg/ml Poly-D-lysine in dH_2O	Merck Life Science	P1024
Fixative solution	4% PFA was supplemented with $MgCl_2$ (final concentration: 1mM) and $CaCl_2$ (final concentration: 0.5mM)		
Phosphate Buffered Saline (PBS)	Phosphate Buffered Saline (PBS) pH 7.4 (1X)	Gibco	10010-015
Primary antibody solution for WB	Add Tween20 to Intercept Blocking Buffer for a final concentration of 0.2% Tween20	LI-COR	927-70001
Secondary antibody solution for WB	Add Tween20 to Intercept Blocking Buffer for a final concentration of 0.2% Tween20. Do not add SDS.	LI-COR	927-70001
Blocking solution for WB	Intercept Blocking Buffer (PBS)	LI-COR	927-70001
TAE buffer	Add 20ml UltraPure DNA Typing Grade TAE buffer (50X) to 980ml milliQ water	Gibco	24710
Running Buffer	add 50 ml MOPS SDS running buffer (20X) to 950ml milliQ water	Novex	B0001
LB Broth (LENNOX)	Add each 1.1g tablet in 48.3mL milli-Q water. Autoclave to sterilise. Allow to cool before adding the appropriate antibiotic.	Sigma-Aldrich	L7275
Cell mito stress test assay medium	Prepare 50ml of the assay medium by adding 50 ml XF DMEM Medium (pH 7.4), 500 µl 1 M Glucose (final concentration 10 mM), 500 µl XF 200 mM Glutamine Solution (final concentration 2 mM), and 500µl XF 100 mM Pyruvate (final concentration 1 mM)	Agilent	103680-100

4.2.1.3 Antibodies

All primary and secondary antibodies used for Western blot analyses are listed in Table 4.5

Table 4.5: Primary and secondary antibodies used for Western blot

1° Ab species	**Antigen**	**Dilution**	**Size (kDa)**	**Supplier**	**Cat. #**
Mouse	Anti-Flag	1:5000	Expected ~38kda corresponding to PLPPR5	Sigma-Aldrich	F1804
Rabbit	Anti-EGFR	1:1000	130	Rockland Immunochemicals	100-401-149
2° Ab species	**Antigen**	**Dilution**	**Info**	**Supplier**	**Cat. #**
Donkey	Anti-rabbit	1:5000	IRDye 680RD, 0.5mg	Li-Cor	926-68073
Goat	Anti-mouse	1:5000	IRDye 800CW, 0.5mg	Li-Cor	926-32210

4.2.1.4 Plasmids

All plasmids used in cell transfection for fixed or live cell imaging are listed in Table 4.6.

Table 4.6: Plasmids

Plasmids	**Supplier**	**Cat. #**
pcDNA3.1-PLPPR5-3xFlag	NovoPro	NPT127297
mito-mCherry	Gift from Prof. Ira Milosevic	-
ER-mScarlet	Addgene	137805
pmCherry-N1-GalT	Addgene	87327
LAMP1-RFP	Gift from Prof. Ira Milosevic	-

4.2.1.5 Equipment

Table 4.7: List of Equipment

Equipment	**Supplier**	**Cat. #**
Branson Sonifier® ultrasonic cell disruptor model S-450D	Marshall Scientific	B450
CFX96 Touch Real-Time PCR machine	Bio-Rad	1845097
Nanodrop 2000 Spectrophotometer	Thermofisher Scientific	ND-2000
Odyssey CLx Imager	Li-COR Bioscience	
Point scanning confocal microscope (Zeiss LSM900 Airyscan 2)	Zeiss	LSM900
spinning disc confocal microscope (Olympus spinSR SoRa with high content imaging)	Olympus Life Science	-
TC20 automated cell counter	Bio-Rad	1450102
Trans-Blot Turbo Transfer System	Bio-Rad	1704150
Xcell surelock Mini-Cell System	Invitrogen	EI0002
BD Fortessa X20	BD Biosciences	-

4.2.1.5 Software

Table 4.8: List of Software program

Software	**Supplier**
Empiria Studio software	Li-COR
Image Studio Software	Li-COR
SnapGene	Dotmatics
Zen Blue	ZEISS
CellSens	Olympus Life Science
GraphPad Prism	Dotmatics
FlowJo 10.9	TreeStar
Fiji ImageJ	

4.2.2 Method

4.2.2.1 Plasmid DNA preparation

Human *PLPPR5* (NM_001037317) cDNA/ORF in pcDNA3.1-3xFlag vector (Figure 4.1) was ordered from NovoPro. The lyophilised plasmid was resuspended in 100µl sterile water at 37°C for 10 minutes in a shaking plate (300x RPM) to dissolve the DNA. For plasmid amplification, 1µl resuspended plasmid was added to 50µl chemically competent DH5α *E. coli* cells for 20 minutes on ice, followed by heat shock at 42°C for 90 seconds, then incubated on ice for further 5 minutes. The transformed cells were then cultured in 250µl prewarmed S.O.C. medium at 37°C with shaking at 225 rpm for 1 hour. The recovered transformed cells were plated in Ampicillin-enriched LB-agar plates for overnight incubation at 37°C. Single colonies were used for inoculating overnight culture with Lennox LB medium containing Ampicillin with a final concentration of 50 µg/mL. Cells were harvested after 16-hour incubation for plasmid purification using the GeneJET Plasmid Miniprep Kit according to the manufacturer's manual. The plasmid sequence was verified by DNA sequencing, which was performed by Eurofins Genomics.

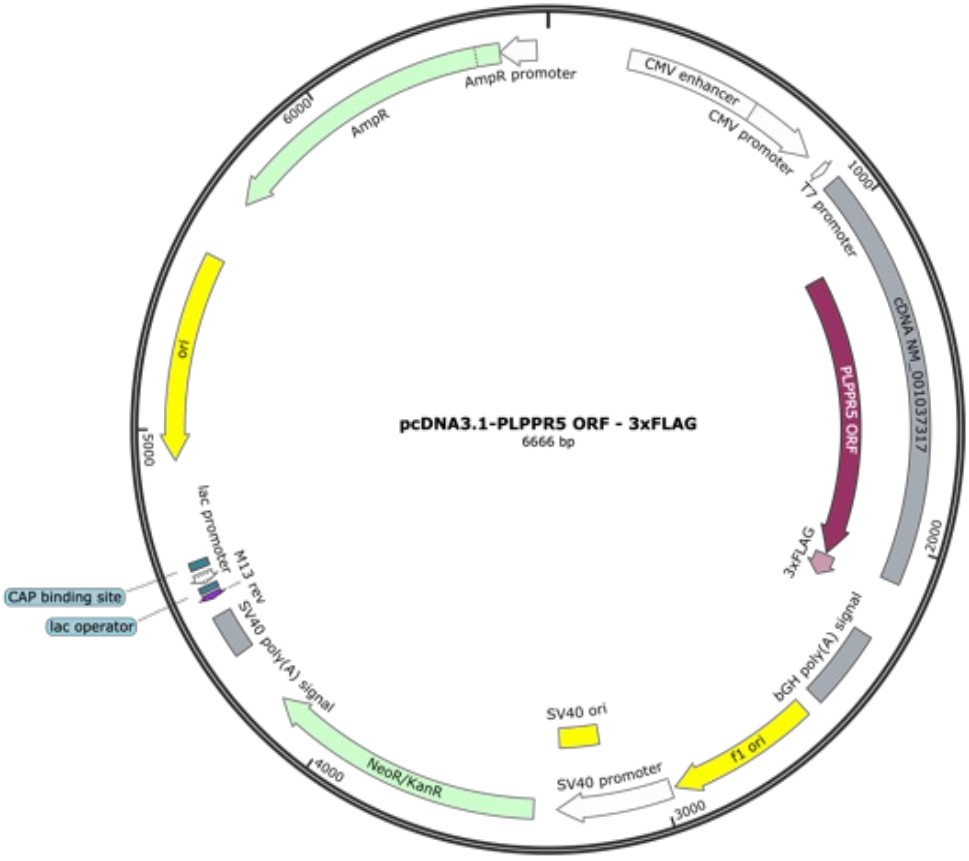


Figure 4.1: pcDNA3.1 vector with PLPPR5-3xFLAG.The plasmid was obtained from NovoPro (Cat. #NPT127297)

4.2.2.2 Site-directed mutagenesis

Site-directed mutagenesis based on PCR was used to generate the patients' identified mutations in *PLPPR5* (NM_001037317.2:c.349C>T, p.R117C and NM_001037317.2:c.37C>T, p.L13F). Oligos containing the specific mutations were designed using SnapGene (for c.37C>T: 5'-CGCTCACCAGCAGCATGTTCTATTTCCAGATGGTG-3', and for c. 349C>T: 5'-AAACCCGCTGGTGTGCCGAACTGTCCG-3'), then ordered from Eurofins Genomics. The primers were resuspended using double-distilled water (ddH_2O) to obtain 100ng/µl.

The mutant strand was synthesised by PCR using QuikChange Lightning Site-Directed Mutagenesis Kit. The PCR was setup in a 25µl reaction mixture containing 2.5µl 10×QuikChange Lightning Reaction Buffer, 0.75µl QuikSolution reagent, 30ng dsDNA template (plasmid), 1µl the mutagenesis primer (100ng/µl), 1µl dNTP mix, 5µl 200µM dNTPs, 1µl QuikChange Lightning Enzyme Mix, and PCR-grade water to a final volume of 25µl. The cycling parameters were set to 1 cycle at 95°C for 2 minutes, followed by 30 cycles of 95°C for 20 seconds, 55°C for 30 seconds, and 65°C for 3 minutes 30 seconds (30 seconds/kb of the ~7Kb plasmid), and x1 cycle at 65°C for 5 minutes. Once PCR finished, 1µl the *DpnI* enzyme was added to the PCR product and incubated at 37°C for 5 minutes. *DpnI* will digest the parental methylated and hemi-methylated DNA, leaving only the mutated DNA intact.

To increase the transformation efficiency, 2µl β-mercaptoethanol were added to XL10-Gold ultracompetent cells prior to the addition of 2µl the *DpnI*-treated DNA. The treated cells were incubated on ice for 30 minutes then Heat-pulsed in a 42°C water bath for 30 seconds then incubated on ice for 2 minutes. The cells were incubated in at 37°C for 1 hour with shaking at 225x RPM after the addition 0.5ml preheated LB broth to 42°C in a water bath. The cell suspension was briefly centrifuged at 2000xg for 1 minute to remove supernatant leaving 100µl for cell resuspension to be plated on Ampicillin-enriched LB-agar plates and incubated overnight at 37°C. Single colonies were used for overnight culture and plasmid was purified using the GeneJET Plasmid Miniprep Kit according to manufacturer's protocol. The presence of the mutations was verified by sequencing the purified plasmid DNA using Eurofins Genomics Sanger sequencing service and visualised using SnapGene (Figure 4.2).

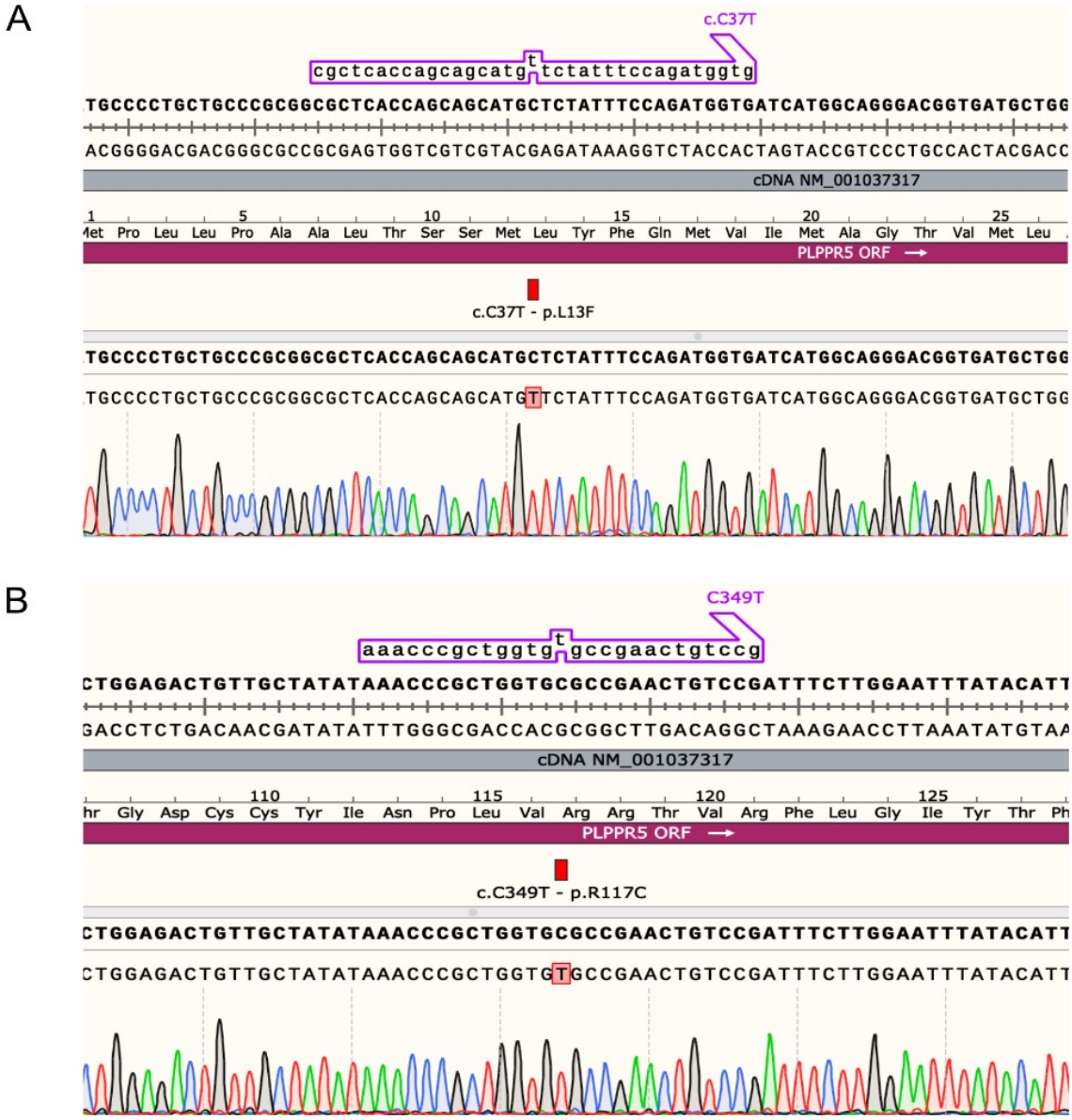


Figure 4.2: Verification of successful SDM to introduce PLPPR5 variants. Sanger traces showing the presence of the mutations generated from the SDM step and the sequence of the mutagenic primers used for generating each of the mutations.

4.2.2.3 Generating constructs with PLPPR5 and fluorescent tag fusion proteins for visualisation

4.2.2.3.1 Primer design

Plasmids containing mClover3 and mScarlet tags to be used as DNA templates were generously provided by Dr. Phalguni Rath from iPSC facility core in the Wellcome Centre for Human Genetics. The mClover3 and mScarlet PCR primers (Table 4.9) were designed to generate KpnI and BamHI restriction sites overhangs for mClover3 and KpnI and EcoRI overhangs for mScarlet. The primers were synthesised by Eurofins Genomics and dissolved in deionised distilled H_2O.

Table 4.9: Cloning primer sequences were modified to generate restriction sites for KpnI/EcoRI/BamHI (red colour) on the 5' and 3'ends of the mClover3 and mScarlet sequences.

Primer name	**Primer sequence**
mScarlet-KpnI	5'-GGGGTACCCCCATGGTGAGCAAGGGCGAG-3'
mScarlet-EcoRI	5'-CGAATTCTACTTGTACAGCT-3'
mClover3-KpnI	5'-ATGGTACCCTGGAGGTGCTGTTCCAGG-3'
mClover3-BamHI	5'-CCGGATCCTTATCATGCAGGTGCTACCTG-3'

4.2.2.3.2 Amplification of mClover3 and mScarlet

To perform the PCR, a 50µl reaction mixture was prepared containing 10µl (5X) Q5 Reaction Buffer, 1µl 10mM dNTPs, 10µM both forward and reverse primers, 0.5µl Q5 High-Fidelity DNA Polymerase, and water up to 50µl. The PCR reaction used 100pg plasmids as DNA templates. The thermal cycle program consisted of an initial denaturation step at 98°C for 30 seconds, followed by 35 cycles at 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, with a final extension step at 72°C for 2 minutes. PCR products were detected using agarose gel electrophoresis. The QIAquick Gel Extraction Kit was used to purify PCR products, following the manufacturer's instructions, and obtaining 25µl eluates. The extracted PCR product alongside 100bp DNA ladder was analysed using 1% agarose gel electrophoresis. The concentration of PCR DNA was determined using NanoDrop 2000 Spectrophotometer.

After the modification of mClover3 and mScarlet sequences to include the appropriate restriction sites, they were digested accordingly to create sticky ends (Figure 4.3). The vectors pcDNA3.1-PLPPR5-3xFlag containing either the WT or mutant version of *PLPPR5* were also digested either with KpnI and BamHI or with KpnI and EcoRI to create compatible sticky ends with which the modified tag sequence was to be ligated. The digestion reaction (20µl) was setup as follows: 1µg plasmid DNA, 1µl Kpn1 and 1µl BamHI or EcoRI, 2µl 10X buffer, and water to make up 20µl. The reaction mixture was incubated for 1 hour at 37°C then 1µl Shrimp Alkaline Phosphatase was added and incubated for further 30 minutes 37°C, and the digestion reactions terminated at 65°C for 20 minutes. The digested PCR product and vectors were electrophoresed on 1% agarose gel containing Ethidium Bromide (EtBr) for 60 minutes at 80 volts then extracted using the QIAquick Gel Extraction Kit. The complete purification and correct sizes for the inserts and vectors were confirmed using DNA gel electrophoresis on 1% agarose gel containing EtBr as previously described.

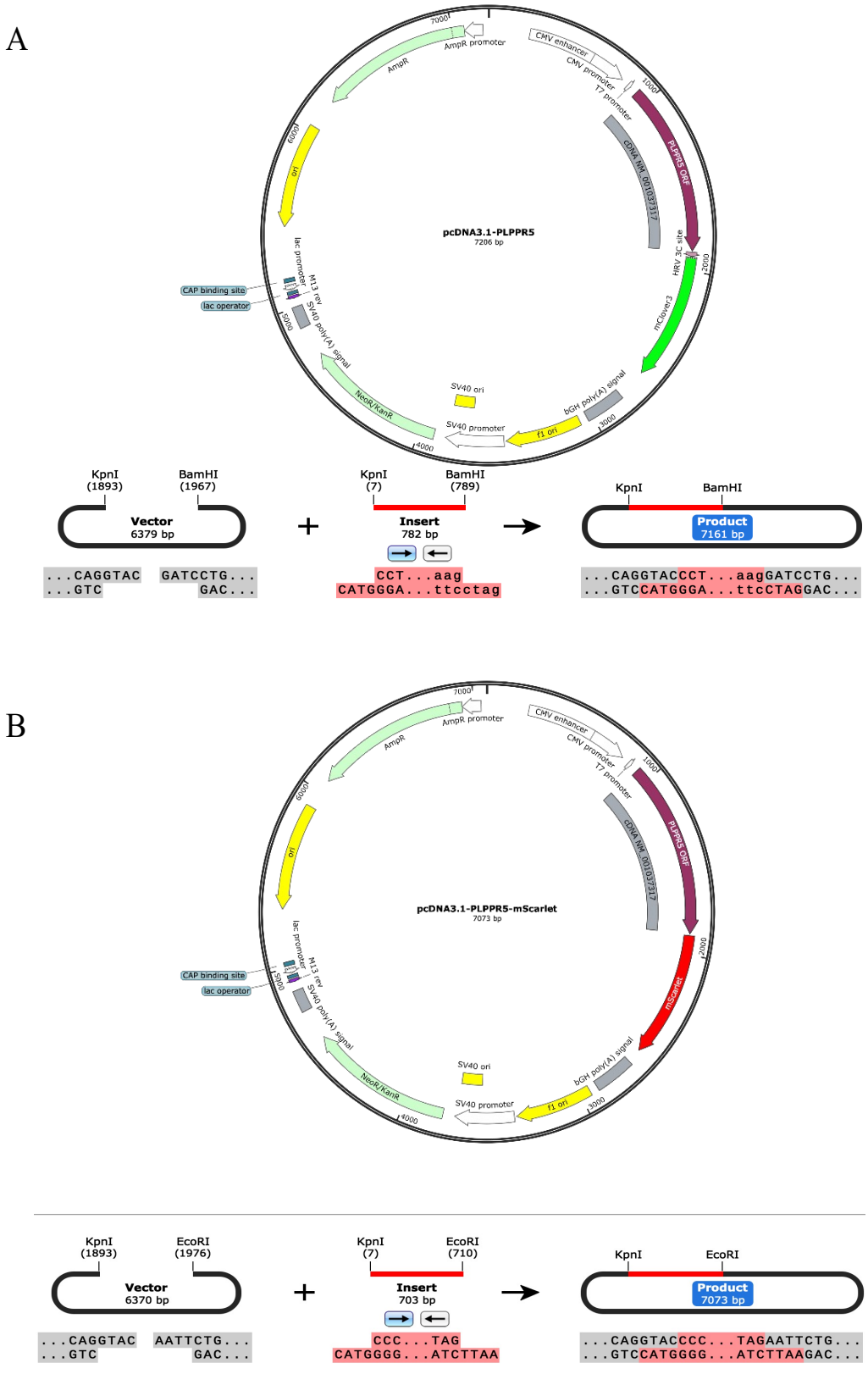


Figure 4.3: Strategy to introduce fluorescent tags into pcDNA3.1-PLPPR5.Schematic Diagram Showing Cloning of mClover3 (A) and mScarlet (B) into pcDNA3.1-PLPPR5 vectors (strategy used to generate PLPPR5-mClover3 or PLPPR5-mScarlet fusion proteins once expressed in the cells after transfection)

4.2.2.3.3 Ligation of pcDNA3.1-PLPPR5 Vectors and mClover3 or mScarlet Inserts

The concentrations of the purified PCR products and vectors were determined using Qubit 2.0 fluorometer with Qubit™ dsDNA HS Assay Kit according to the manufacturer's instructions. These values were used to calculate the amounts needed for 1:3 vector: insert ligation reaction using NEBioCalculator (https://nebiocalculator.neb.com/#!/ligation).

T4 DNA Ligase was used setup 10µl reaction by mixing 1µl T4 DNA Ligase Buffer (10X), 75 ng vector DNA, 25ng insert DNA, 1µl T4 DNA Ligase, and water to make up 10µl then incubated at room temperature for 1 hour.

After the ligation reaction, 5µl of each reaction mixture was transformed into 50µl DH5α *E. coli* competent cells by incubating on ice for 30 minutes. After 30 minutes, the reactions were heat shocked for 30 seconds at 42°C and immediately incubated on ice for 5 minutes. Each transformed mixture was spread evenly on prewarmed Ampicillin-enriched LB-agar plates and incubated overnight.

4.2.2.3.4 Analysing Transformants

Six single colonies were selected from each plate to inoculate 3ml Ampicillin containing LB medium. Each tube was incubated for 16 hours before making 50% glycerol stocks of the selected transformed colonies (stored at -80°C) and harvesting the cells for DNA isolation by Miniprep with the QIAprep Spin Miniprep Kit by following the manufacturer's instructions.

In order to confirm that these clones contained the correct inserts of expected size, the selected clones were digested using restriction enzymes that cut within the insert towards one of the ends, which confirms both the size and orientation of the insert. For pcDNA3.1-PLPPR5-mClover3 and pcDNA3.1-PLPPR5-mScarlet clones, 1µl SacII was used for digestion as previously described. The expected sizes from the digestion of pcDNA3.1-PLPPR5-mClover3 are 6.1Kb and 1.1Kb, and 5.7Kb and 1.3Kb for pcDNA3.1-PLPPR5-mScarlet (Figure 4.4). These clones were also sent for sanger sequencing by Eurofins Genomics to verify the presence of mClover3/ mScarlet inserts in frame with *PLPPR5* generating fusion protein. Results from sequencing were analysed with SnapGene to confirm ligation of mClover3/ mScarlet inserts in pcDNA3.1-PLPPR5 vectors; all restriction sites were identified. Intact start codons of inserts were confirmed to be in reading frame with *PLPPR5*.

To prepare for subsequent experiments, bacterial overnight cultures were generated for Midiprep using the NucleoBond Xtra Midi kit and for glycerol stocks of pcDNA3.1-PLPPR5-mClover3 constructs (WT, c.37C>T, and c.349C>T) and pcDNA3.1-PLPPR5-mScarlet constructs (WT, c.37C>T, and c.349C>T).

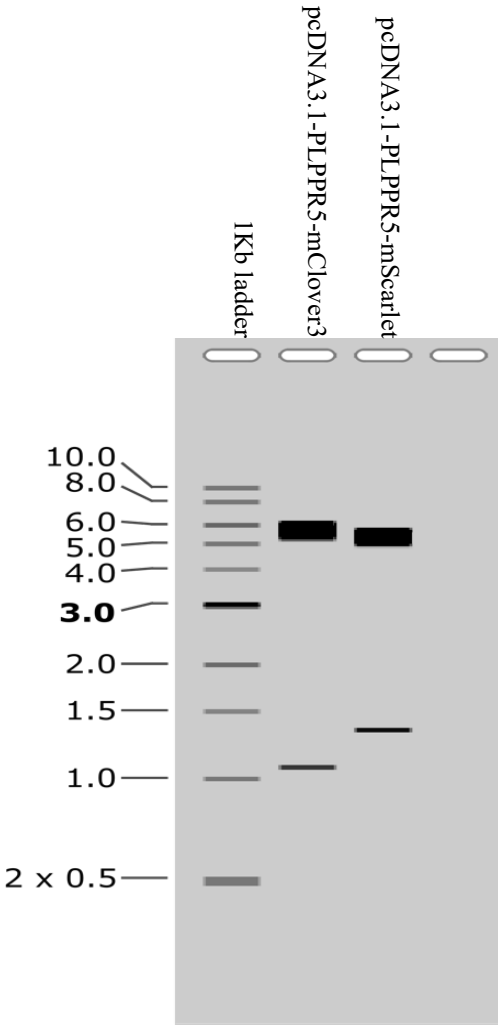


Figure 4.4: Cloning verification strategy using restriction digestion.Agarose gel (1%) simulation for restriction digestion strategy using SacII to confirm the correct size and orientation of the insert into pcDNA3.1-PLPPR5 vectors

4.2.2.4 Cell culture

Human embryonic kidney 293 cells (HEK293-T- Table 4.3), and African green monkey kidney fibroblast-like COS7 cell line (COS7- Table 4.3) were maintained in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 2mM L-Glutamine, and 1% penicillin/streptomycin. All cell lines used in this study are kept free of mycoplasma contamination and maintained at 37°C, 5% CO_2 unless otherwise stated.

4.2.2.4.1 passaging protocol

1. Aspirate old medium from T75 flask.
2. Wash once with sterile PBS to remove all traces of FBS and aspirate.
3. Add 3ml warm 0.25% trypsin-EDTA and return flask to incubator (37°C) for 3-5 minutes.
4. Inactivate trypsin by adding 9ml warm culture medium to the cell-suspension and transfer the total volume to a 15ml falcon tube - centrifuge 5 minutes at 200xg.
5. Aspirate medium and resuspend cell-pellet in fresh culture medium.
6. Split the cells with 1:5 for maintenance or count the cells when using for transfection. For cell counting, 10µl trypan blue dye is added to 10µl cell suspension. load 10µl of the sample onto a chamber of the counting slide and insert into TC20 automated cell counter for accurate count.
7. Return cell aliquots to new T75 flasks or plate desired amount as required.

4.2.2.5 Transient transfection

For transfection, cells were dissociated using 0.25% trypsin-EDTA, counted and seeded at 15-25×10^3 cells/cm^2 and reverse transfected using Lipofectamine 3000 following the manufacturer's protocol. Plasmid transfections were performed using 400ng DNA per 6x10^4 cells and 1:2 DNA to lipofectamine 3000 ratio diluted in OPTI-MEM. Cell number and lipofectamine 3000 reagents were adjusted accordingly to the well size used.

For all transfection experiments in this study, the following amounts of DNA were used per 6x10^4 cells (individually or combined for co-transfection): 400ng for pcDNA3.1-PLPPR5 constructs; and 100ng for mito-mCherry or ER-mScarlet or pmCherry-N1-GalT or pEGFP-Sec23A. DNA to lipofectamine 3000 ratio was determined as 1:2 and diluted in OPTI-MEM.

The cells were incubated with the medium containing the transfection mix for 24 hours before changing with fresh medium and incubating for further 6 hours. For transfection experiments to be used for fixed or live cell imaging, 35mm glass-bottom confocal dishes were used for plating cells at 6x10^4 cells/dish to be transfected and imaged. All DNA plasmids used in this work are listed in Table 4.6.

4.2.2.6 Quantitative RT-PCR

RNA techniques were performed using filter tips, molecular biology grade DNase, and RNase-free water on a bench cleansed with RNaseZap RNase Decontamination Solution and DNAZap PCR DNA Degradation Solutions to prevent nucleic acid or RNase contamination.

4.2.2.6.1 Sample preparation, RNA extraction, and cDNA synthesis

HEK-293T cells transfected to overexpress *PLPPR5* (WT or mutants) were washed with ice-cold PBS, scraped with cell scrapers, and collected on 1.5 Eppendorf microtubes and total RNA was extracted with commercially available RNeasy Mini Kit. RNA pellets were dissolved in RNAse-free water and the concentration was determined using NanoDrop 2000 spectrophotometer. Constant amount of 1000 ng total RNA was reverse–transcribed to cDNA using QuantiTect Reverse Transcription Kit according to the manufacturer's instructions.

4.2.2.6.2 Real-time polymerase chain reaction (qRT-PCR).

Gene expression levels of each target were analysed by qRT-PCR using SYBR-green.

The master-mix was prepared with iQ SYBR green Supermix according to the manufacturer's instructions. Most of the PCR reactions were carried out in total volume of 10µl or the reaction was alternatively enhanced to 20µl. the cycling protocol included 3 minutes of initial denaturation and enzyme activation at 95°C followed by 40 cycles of 15 seconds at 95°C for denaturation, 30 seconds at 55-60°C annealing temperature and extension and a final extension step of 30 seconds at 72°C. The program was setup at the CFX96 Touch Real-Time PCR machine. Ct-values were analysed using the ΔΔCt method and target genes were normalised to *GAPDH* or beta-Actin housekeeping genes.

4.2.2.7 Protein Analysis Methods

HEK293-T cells transiently overexpressing *PLPPR5* (WT or mutants) were analysed for protein expression by western blot.

4.2.2.7.1 Cell lysis and protein extraction

All steps were performed on ice to minimise degradation of protein and maintain activity. Plasma membrane fraction was isolated from HEK293T cells expressing pcDNA3.1-PLPPR5-3xFLAG using Subcellular Protein Fractionation Kit (Thermo Scientific™). Confluent cells in Ø10-cm dishes were washed 3x in ice-cold PBS then scraped using cold plastic cell scraper and collect in microcentrifuge tubes then centrifuge at 500 × g for 5 minutes. Ice-cold cytoplasmic extraction buffer was added to the cells pellet and incubated at 4°C for 10 minutes with gentle mixing. The cells were centrifuged at 500 × g for 5 minutes and the supernatant (cytoplasmic extract) was immediately transfer to a clean pre-chilled tube. Ice-cold membrane extraction buffer containing protease inhibitors was to the pellet and agitated on the highest vortex setting for 5 seconds before incubation at 4°C for 10 minutes with gentle mixing. The tubes were then centrifuged at 3000 × g for 5 minutes and the supernatant (membrane extract) was immediately transferred to a clean pre-chilled tube on ice.

4.2.2.7.2 Quantification of lysate

To estimate protein concentrations, Pierce 660nm protein assay kit was used, and the concentration was determined using NanoDrop Spectrophotometer 2000 according to the manufacturer's protocol. BSA-standard curve (0-20µg protein) was prepared from a stock solution of 2 mg/ml.

4.2.2.7.3 Western blot analysis

Protein expression was determined by standard Western blotting techniques. Briefly, mix 5-10µg plasma membrane fraction of each sample to the appropriate volume of 4x NuPage LDS loading dye and 1ul dithiothreitol (DTT) to stabilise protein. Samples were boiled for 5 minutes at 95°C, centrifuged and loaded onto 4-12% Bis-Tris polyacrylamide gels for electrophoretic separation. Molecular weights were estimated by comparison with Pre-stained Protein Ladder. Mini-gels were run at 100 V for 90-120 minutes using the XCell SureLock Mini-Cell System.

Following electrophoresis, proteins were transferred from onto nitrocellulose membranes (Immun-Blot 0.2µm pore size) using the Trans-Blot Turbo Transfer System blotting apparatus. Transfer of proteins was carried out by applying a current set at 25V and 1.0A for 60 minutes.

The nitrocellulose membranes were incubated in Li-Cor Odyssey Intercept (PBS) blocking buffer for 60 minutes at room temperature to block nonspecific binding sites. Thereafter, membranes were probed with primary antibody mouse anti-Flag diluted in Odyssey blocking buffer at 1:5000 (Sigma-Aldrich) and rabbit anti-alpha EGFR (1:1000; Rockland) overnight at 4°C. The membranes were washed three times for 10 minutes in Odyssey blocker +0.1% Tween20. Incubate the membrane with secondary antibody (1:5000; IRDye 800CW goat anti-mouse IgG and IRDye 680RD donkey anti-rabbit, LI-COR) for 1 hour at room temperature. Finally, wash membranes 3 times 10 minutes in Odyssey blocker+0.1% Tween20. The signals were detected on Odyssey CLx imager and analysed with Empiria Studio software.

4.2.2.8 Fixed cell imaging by Total internal reflection fluorescence (TIRF) microscopy

COS7 cells were grown on glass-bottom confocal dishes and transfected as described in section 4.2.2.5. After 30 hours incubation, cells were fixed with a pre-warmed fixative solution prepared as described in Table 4.4 for 1 hour at 37°C. The cells were then washed three times with PBS and kept in PBS until being imaged. The fixed cells were subsequently imaged using Zeiss Elyra TIRF PALM super-resolution microscope in the TIRF mode utilising the Zeiss ZEN Black image acquisition software with super-resolution module. A 60x objective lens was used with the laser beam setup at an angle of 65 to ensure true TIRF images of the plasma membrane is obtained.

4.2.2.8.1 Measuring fluorescence intensity from TIRF images

Cellular fluorescence intensity and signal were assessed using ImageJ. Six regions of interest (ROIs) were selected per cell, and one ROI next to the cell for determining the background signal. The same frame was used for all measurements to ensure consistency. The area, mean grey value, and integrated density measurements were extracted for each ROI.

To calculate the corrected total cell fluorescence (CTCF) for individual ROIs, the formula CTCF = Integrated Density – (Area of Selected Cell x Mean grey value for the background reading) was employed. Each CTCF value represented the corrected fluorescence intensity for a specific ROI. These values were subsequently used to determine the mean fluorescence intensity per cell.

To calculate the CTCF for the mean values, the following formula was employed: CTCF = Mean Integrated Density of the cell minus (the mean Area of the Cell multiplied by the Mean Grey Value for the background reading). This CTCF value indicates the corrected average fluorescence intensity per cell. Subsequently, the average mean and SD of fluorescence intensity were computed for the WT then used to generate the Z-scores transforming all corrected fluorescence intensities for the cells in each condition. These scores were calculated by subtracting the mean of the WT then divide by the SD.

Kruskal-Wallis test was used to check for significance difference between the conditions using the transformed values.

4.2.2.9 Live cell imaging by confocal microscopy

Cells were grown on glass-bottom confocal dishes. For all live cell imaging, cells were imaged in live-cell FluoroBrit DMEM medium supplemented with 1% GlutaMAX in live cell chamber supplied with 5% CO2 at 37 °C.

The confocal dishes were loaded to a laser point scanning confocal microscope (LSM900, Zeiss, Germany) equipped with multiple excitation lasers (405-nm, 488-nm, 561-nm, 647-nm) and Airyscan 2 detector. Cells were imaged with the 63x 1.2NA Water objective using the 488-nm laser for mClover3 or GFP, and 561-nm for mScarlet or mCherry.

Live cell imaging by Olympus SpinSR SoRa equipped with a 50um pinhole disc for confocal imaging, and a SoRa super-resolution disc, allowing you to switch imaging modalities. This enables rapid acquisition of large areas of interest, even when including z stacking. Cells on confocal dishes were loaded to the Olympus microscope with 60x 1.2NA sillicone objective, four lasers (405-nm, 488-nm, 561-nm, 647-nm) with corresponding filters.

4.2.2.9.1 Mitochondrial network using confocal images

Confocal images of cells labelled for mitochondria were analysed for morphological and network differences between WT and mutants. The images were analysed in ImageJ using Mito-morphology macro as described previously (Dagda et al., 2009). The macro traces the mitochondrial outline and measures the average mitochondrial perimeter, solidity, and circularity, which are well recognised indicators of mitochondrial fission and fusion. Perimeter is determined by counting the number of pixels that form the boundary of a mitochondrial region, solidity quantifies shape complexity, with higher values indicating a more compact shape of the mitochondria, and circularity indicates the roundness of mitochondria (Westrate et al., 2014). To quantify the differences in mitochondrial network topology between COS7 cells expressing WT *PLPPR5* and those expressing *PLPPR5* mutations, an average of ten randomly selected cells per repeat experiment were analysed. The average value per cell for each parameter was obtained by dividing the total value by the number of mitochondria. Mann Whitney's non-parametric test was used for statistical analysis.

4.2.2.10 Mitochondrial membrane potential

4.2.2.10.1 Flow cytometry

Flow cytometry was carried out using a BD Fortessa X20 flow cytometer (BD Biosciences) at the default setting (PE channel). Single, fixed cells were gated using FSC and SSC dot plot to select for GFP positive single cells. Mean fluorescence intensity of Tetramethyl-Rhodamine Methyl ester (TMRM) was determined on the gated GFP positive cells (Figure 4.5). Data analysis was performed using the FlowJo 10.9 software.

Prior to flow cytometry analysis, cells were transfected with PLPPR5-eGFP (WT/p.L13F/p.R117C) constructs and maintained for 48 hours. The cells were then treated with 100 nM TMRM (Sigma) for 30 min at 37ºC. Cells were washed twice with PBS then trypsinized and quenched with 1 mL DMEM. Cells were centrifuged at 300 rcf for 5 min, then fixed with 4% PFA for 20 minutes at room temperature, followed by washing and centrifugation at 300 rcf for 5 min. Finally, the cells were resuspended in 5 ml PBS, passed through cell strainer cap of falcon round-bottom polystyrene tubes to obtain single cells, centrifuged again and resuspended in 350 µl PBS. The fluorescence signals of TMRM (excitation 552/emission 579 nm) were measured and gates adjusted using a positive control (2 µM oligomycin for 15 minutes) and negative controls (20 µM FCCP for 15 minutes, and unstained un-transfected/transfected cells).

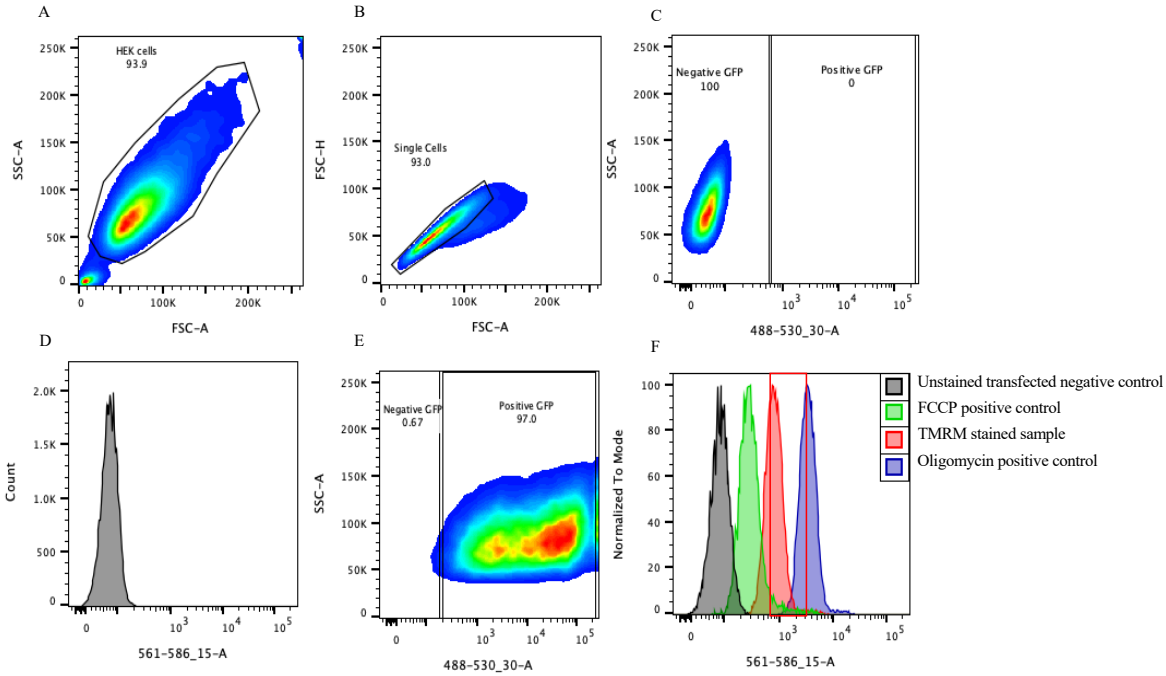


Figure 4.5: Gating strategy to determine true TMRM signal. A) Display all acquired HEK293T cells and distinguish from debris. B) gating to identify single cells population. C) Gating to identify GFP negative and positive single cells. D) TMRM intensity (PE channel) in unstained cells showing fluorescence considered as negative signal. E) Fraction of transfected cells showing positive signal for GFP. All samples showed >90% positive signal for GFP. F) Representative distribution of ΔΨm cells unstained/stained with TMRM in the presence or absence of inhibitors. Unstained transfected cells (grey peak) show that transfection does not interfere with TMRM signal. Green peak stained with TMRM in presence of FCCP determine the lower boundary for TMRM staining. Red peak represents GFP-positive cells stained with TMRM in the absence of inhibitors. Purple peak represents enhanced TMRM staining due to the presence of Oligomycin.

4.2.2.11 Oxygen consumption rate (OCR) measurements using Seahorse XF Cell Mito Stress Test

Agilent Technologies' Seahorse XFe96 Bioanalyzer was utilised to conduct extracellular flux assays. In this process, a live cell monolayer was seeded into Seahorse XFe96 microplates, and the medium around the cells was analysed for extracellular acidification rate (ECAR) and oxygen consumption rate (OCR) using Seahorse XFe96 FluxPack (Agilent). These consist of 96 sensors and were placed about 200µm above the cells, which results in a transient microchamber. Fluorophores capable of detecting changes in oxygen and pH levels were utilised to gauge OCR and ECAR, respectively. Metabolic parameters were computed through the administration of metabolism-modulating compounds such as Oligomycin, Carbonyl cyanide-p-trifluoromethoxyphenylhydrazone (FCCP), and Rot/Antimycin A through injecting directly into the medium of the cells.

4.2.2.11.1 Analysis of mitochondrial respiration in HEK-293T cells overexpressing PLPPR5 WT/mutant

HEK-293T cells were seeded in 6-well plates (300,000 cells/well) in growth medium for 24hr the transfected with pcDNA3.1-PLPPR5-mClover3 (WT or mutants) using lipofectamine 3000 according to manufacturer's protocol. The XFe96 Cell Culture Microplate were coated with PDL (Table 4.4) and incubated for 1 hour at 37°C then washed twice with PBS and a final wash with DMEM. The cells (30,000 cells/well as determined in Figure 4.7) were re-plated into the coated XFe96 Cell Culture Microplate 6 hours post-transfection in growth medium overnight before conducting the Seahorse extracellular flux assay. The Seahorse XFe96 FluxPacks were hydrated on the day prior to the assay with 200µL Seahorse XF Calibrant per well, the FluxPack was sealed with Parafilm to minimise evaporation of the calibrant and incubated at atmospheric CO_2 levels in a 37°C incubator overnight.

On the day of the assay, cells were checked to confirm cell health, morphology, seeding uniformity, purity (no contamination), and adherence. Prepare cell mito stress test assay medium (Table 4.4). One hour prior to the assay, cells were washed twice with the assay medium gently to remove any trace of FBS and the growth medium was exchanged for the assay medium. Cells were checked under the microscope to ensure that cells were not disturbed or washed away. and incubated for 1hr

in atmospheric CO_2 incubator at 37 °C. During this incubation period, the assay cartridge was loaded with XF Cell Mito Stress Test compounds (final concentrations: 1µM oligomycin, 1µM FCCP, 0.5µM rotenone/antimycin A) and calibration started. Cells in the XFe96 Cell Culture Microplate were placed in the Seahorse XFe Bioanalyzer and extracellular flux analysis was performed. Serial injections of the following compounds: oligomycin (1µM, injection port A), FCCP (1µM, injection port B), and combined antimycin A and rotenone (0.5µM, injection port C) from the XF Cell Mito Stress Test Kit (Agilent) was added to each well (Table 4.10). Oxygen consumption rate and extracellular acidification rate of cells were measured in at least 3 replicate wells generating an oxygen consumption rate profile as represented in Figure 4.6. All Seahorse data was normalised using the live/dead cell assay described in 4.2.2.11.3.

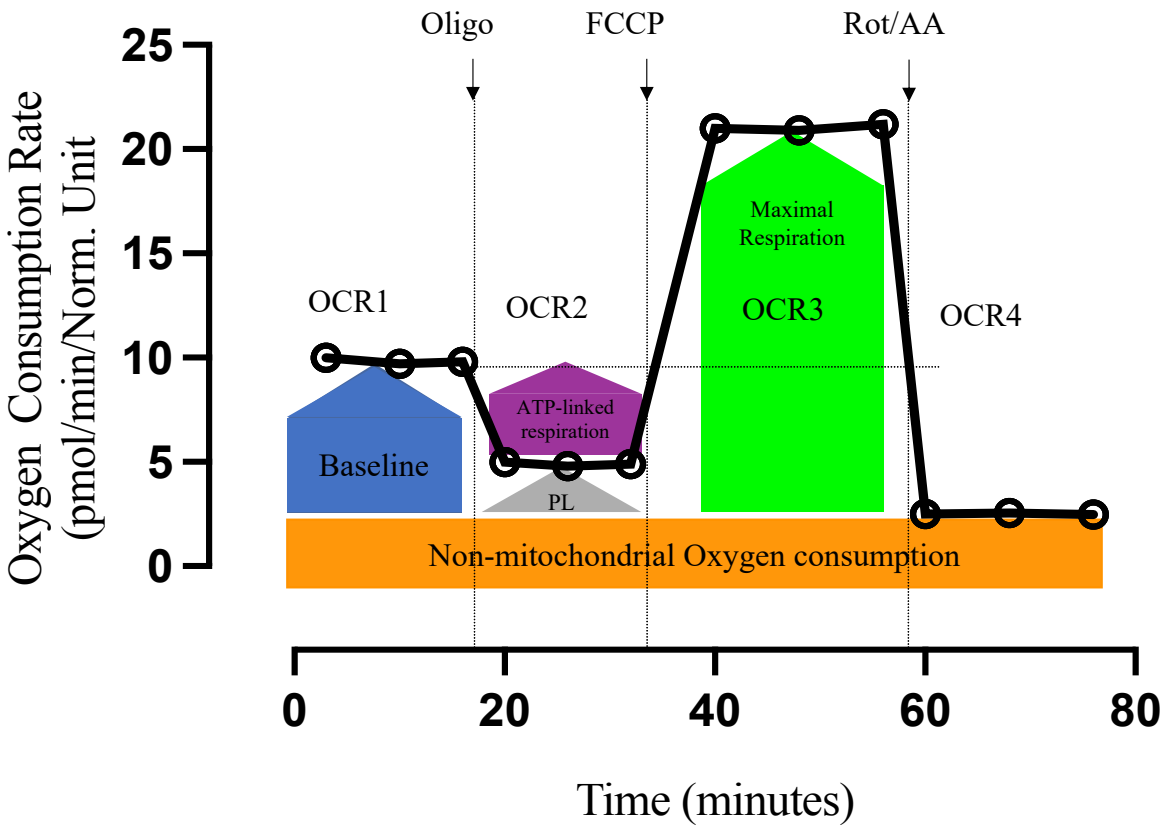


Figure 4.6: Illustration of the Mito-Stress test assay and OCR profile. The representative trace shows the oxygen consumption rate (OCR) measured by a Seahorse XFe96 Bioanalyzer. This assay involves baseline readings (OCR1) and measurements after subsequent injections of oligomycin (OCR2), FCCP (OCR3), and combined rotenone/antimycin A (Rot/AA; OCR4), enabling the determination of various parameters including basal OCR, ATP-linked OCR, maximal OCR, spare respiratory capacity, proton leak (PL), and non-mitochondrial OCR

Table 4.10: Functions of drugs on inner mitochondrial membrane

Compound	Function
Oligomycin	Inhibition of ATP synthase of the electron transport chain by blocking its proton channel (F_o subunit)
FCCP	Permeabilization of mitochondrial membrane leading to? transporting protons across the mitochondrial inner membrane
Rot/Antimycin A	Inhibition of Complexes I and III of the electron transport chain

4.2.2.11.2 Cell number optimisation

HEK293T cells were maintained as described as described in section 4.2.2.4. One day prior to the assay, cells were seeded in a XFe96 microplates at six different seeding densities with 4 replicates and incubated overnight. The mito stress test was performed as described in section 4.2.2.11.2. This allowed us to determine that 30,000 cells/well was the optimal seeding density for this assay (Figure 4.6)

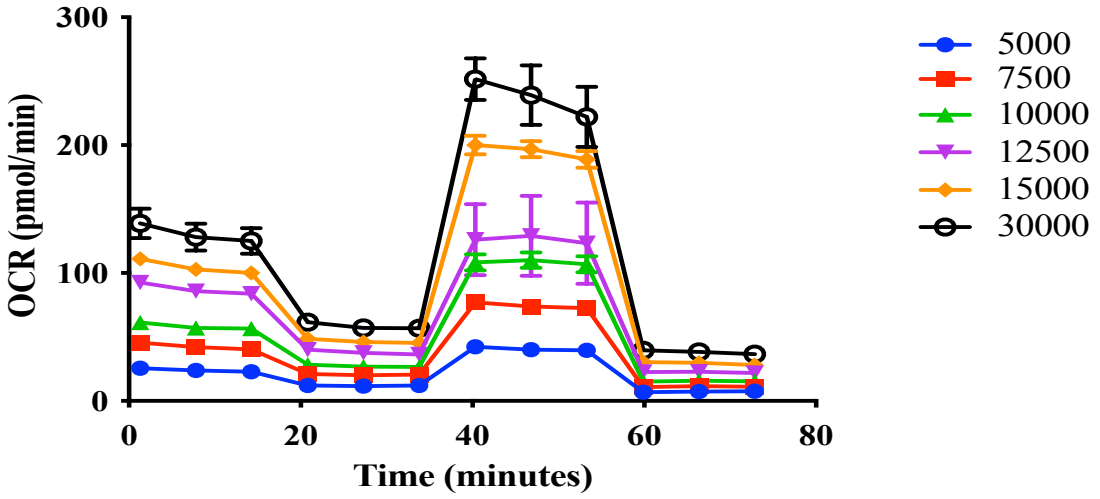


Figure 4.7: Cell number optimisation for Seahorse assay. Different cell number was used to determine the seeding density for informative OCR values between 100-200 pmol/min. For HEK293T cells, 30,000 cells displayed the largest difference after oligomycin and FCCP injections

4.2.2.11.3 Calcein Live/Dead cell assay

After completing the Seahorse run, the XFe96 sensor cartridge was ejected, and the culture medium was removed from the XFe cell culture plate. A master mix was prepared, which included 7.5ml PBS, 15µL ethidium (Invitrogen, Thermo Fisher Scientific), and 3.75µL Calcein (Invitrogen, Thermo Fisher Scientific). The cells were gently washed with 200µL PBS, followed by the addition of 70µL of the prepared master mix to each well. The plate was immediately wrapped in foil to protect it from light. Subsequently, the plate was incubated on a rocker at room temperature for 30 minutes. After incubation, the fluorescence readings were taken at 645 nm and 530 nm wavelengths.

4.2.2.11.4 Seahorse Mito Stress Test data analysis

Mitochondrial respiratory function was determined using readings of Oxygen consumption rate (OCR). The baseline OCR measurements comprised oxygen consumption by the electron transport chain (ETC) during oxidative phosphorylation, as well as other oxidase enzymes. To determine ATP-linked OCR, oligomycin was injected first to inhibit ATP synthase. The second injection, FCCP, uncoupled ATP synthesis from the ETC by dissipating the proton gradient across the mitochondrial inner membrane and allowed uninhibited electron transfer, thereby enabling the achievement of maximum mitochondrial OCR. The third and fourth injections were rotenone and antimycin A, respectively, which inhibited Complexes I and III of the ETC, leading to the cessation of mitochondrial oxygen consumption. These injections facilitated the calculation of the proportion of baseline OCR that was attributable to mitochondrial respiration. For a full list of the calculations, please refer to Table 4.11

Table 4.11: OCR calculations of the proportion of baseline OCR that was attributable to mitochondrial respiration

Respiratory function parameter	Calculation
Basal respiration	*OCR1 - OCR4*
ATP-linked respiration	*OCR1 - OCR2*
Maximal respiration	*OCR3 – OCR4*
Spare respiratory capacity	*OCR3 – OCR1*
Proton Leak	*OCR2 -OCR4*
Non-mitochondrial respiration	*OCR4*

4.2.2.12 Metabolomics data

Whole blood samples in EDTA-treated tubes were collected from both affected siblings, clinically unaffected mother (p.L13F carrier), and clinically unaffected sister (p.L13F carrier). However, no sample was available from the clinically unaffected father. (p.R117C carrier). The collection was done in the outpatient clinic at the John Radcliffe Hospital, Oxford, UK. Upon collection, the samples were immediately processed for plasma separation by centrifugation for 15 minutes at 1,300 x g. The resulting supernatant is designated plasma and was aliquoted into chilled polypropylene tubes and snap-frozen. The samples were stored at -80°C.

An untargeted metabolomics analysis using liquid chromatography–gas chromatography–mass spectrometry was performed by Metabolon. Small molecules (between 50-1,500 Daltons (Da) in molecular weight) were identified in patient samples. This identification was performed using four types of high-performance Ultra Performance Liquid Chromatography (UPLC) instruments paired with Mass Spectrometry (Evans et al., 2014, 2009).

The identification of each molecule was confirmed against a proprietary chemical library consisting of accurate molecular weight/mass plus information on any adductation, in source fragmentation, and/or polymerization (typically dimers and trimers), retention time/index on the chromatography columns, and mass spectral fragmentation patterns (Dehaven et al., 2010). The overall process variability was assessed using stable isotope standards and duplicates of two EDTA plasma samples to monitor the assay's performance and for quality control.

Metabolon performed statistical analysis, and a report of the main findings was shared with us. The Z-score for each detected biochemical was calculated and compared against 300 sample adult fasted reference cohort. Clinical samples without Z-scores for specific biochemicals indicates either the biochemical was not detected in the reference cohort and /or it was not detected in the specific clinical sample. Missing Z-scores for either the 97.5th percentile and / or the 2.5th percentile of the reference cohort indicates the biochemical was not detected in enough subjects from the population database to calculate the 97.5th percentile and /or 2.5th percentile reference range for that specific biochemical.

4.3 Results

4.3.1 PLPPR5 is trafficked to the plasma membrane via galT-positive vesicles in HEK293T and COS7 cells

In order to establish the effect of the mutations on the *PLPPR5* gene expression and function, it is necessary to first characterise the expression, localisation and impact of the WT protein to ensure I get comparable results to what is previously known in the literature.

Mild overexpression studies have shown that PLPPR5 localises to the plasma membrane to variable degrees using different cell models including HeLa, Human embryonic kidney cells, P19, N1E-115 and Neuro2a neuroblastoma cell lines (Broggini et al., 2010; Coiro et al., 2014; Gross et al., 2022; Coevorden-Hameete et al., 2015). In this experiment, I carried out transfection studies using a PLPPR5 construct that I modified to include a green fluorescent tag mClover3. The purpose was to visualise the localisation of PLPPR5 within COS7 and HEK293T cells. The PLPPR5-mClover3 protein was found to be present mainly in two distinct locations: the intracellular regions (as punctate signal) and the plasma membrane. It induced extensive filopodia-like extensions in both cell lines, similar to what has been reported before (Coiro et al., 2014). Two different sizes of punctate signal were observed, suggesting two distinct trafficking vesicular populations. To determine the nature of these organelles and potential trafficking mechanisms, co-transfection with either LAMP1-RFP and pmCherry-N1-GalT, lysosomal and trans-Golgi vesicle markers, respectively.

The results revealed co-localisation between PLPPR5-mClover3 (Figure 4.8: A) and galT marker for trans-Golgi vesicles (Figure 4.8: panel B) but not lysosome associated membrane protein-1 (LAMP1) as shown in Figure 4.8: panel C. This finding indicates that PLPPR5-mClover3 is trafficked to the plasma membrane by leaving trans-Golgi in a steady stream (Supplementary video S4.1). This is consistent with what was observed previously by Coiro et al. 2014 study in which the α-clathrin vesicle marker was used and found to co-localise with PLPPR5. This current study further supports that PLPPR5 regulates plasma membrane dynamics and other cellular processes. Since PLPPR5 has not been established as a human disease-causative gene so far, it was imperative to investigate whether the identified mutations change the localisation of the protein in the cell and whether it follows the same pathway as the WT during trafficking to the plasma membrane.

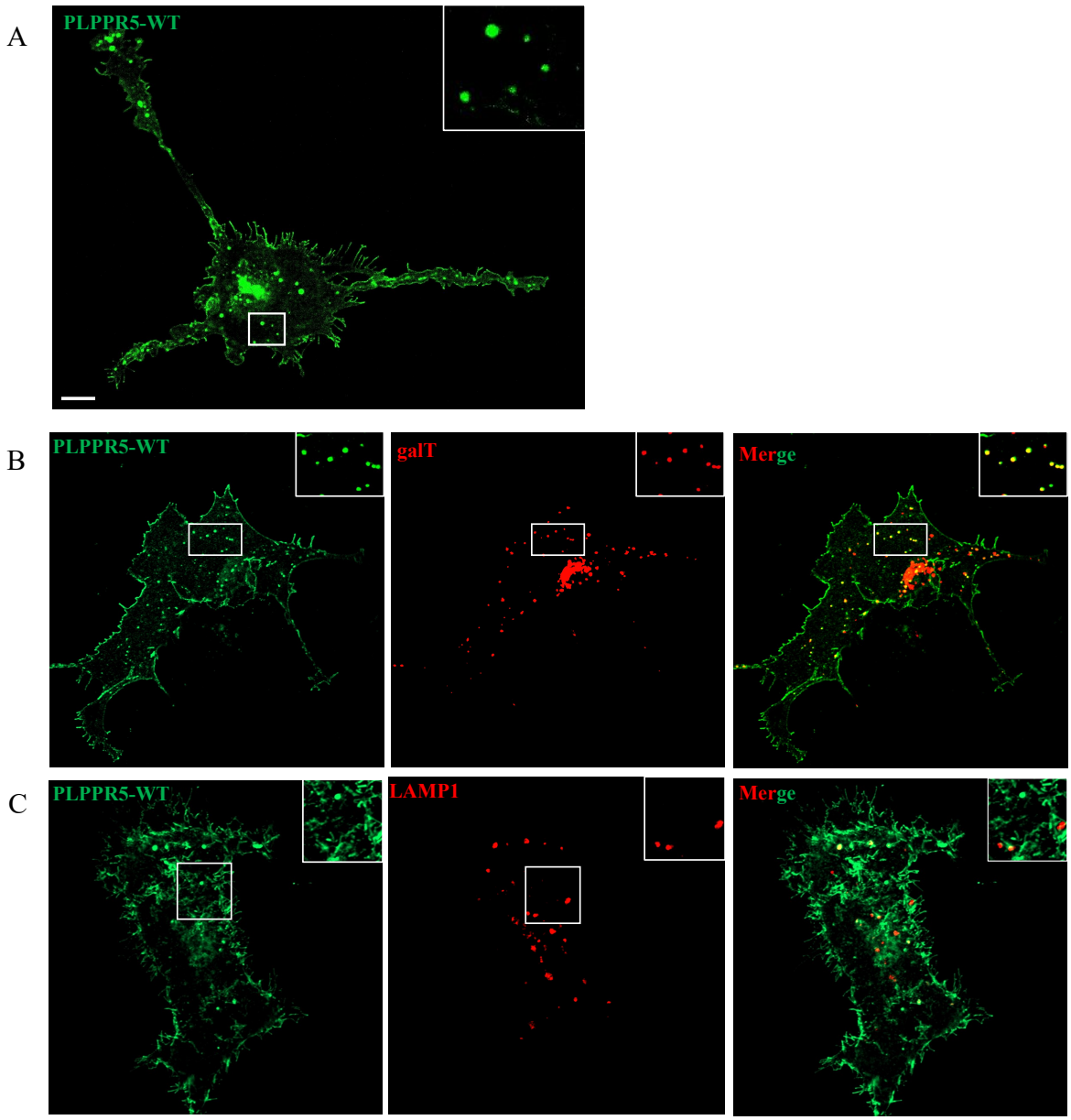


Figure 4.8: Characterisation of PLPPR5 WT in COS7 cells. A) PLPPR5 WT localises to the plasma membrane in COS7 cells and the finding is replicated in HEK293T cells (not shown here). The mClover3-tagged protein also shows different sized puncta (small and large) that are positive for the presence of PLPPR5 in both cell models pointing towards trafficking vesicles, which were further investigated for various markers. B) puncta that were positive for PLPPR5 were observed to be galT- positive, hence colocalising with a marker for the trans-Golgi membranes and trans-Golgi network, throughout the cytoplasm (often perinuclear) indicating that PLPPR5 is trafficked to the plasma membrane via trans-Golgi vesicles. C) LAMP1 marker was used to label lysosomes, which did not show co-localisation with PLPPR5. Scale bar: 10 µM

4.3.2 The patient-specific mutations impeded the trafficking of PLPPR5 to the plasma membrane, and affected the formation of filopodia-like extensions induced by PLPPR5

Compound heterozygous mutations in *PLPPR5* (p.L13F and p.R117C) identified in the affected members (discussed in 3.3.1) were investigated for their functional consequences on the protein localisation in the different cellular compartments and potential impact in its function. This was achieved through experiments comparing the behaviour of the WT and mutant PLPPR5-mClover3 constructs in both COS7 and HEK293T cells. Initially, I wanted to establish whether the variants affect the protein being trafficked to the plasma membrane; therefore, transfections of these constructs into cells were performed.

The signal was primarily detected in the plasma membrane in both mutants, as expected for PLPPR5, yet to somewhat lower level (Figure 4.9: A-C). Further, a visible difference in the induction of filopodia-like extension formation was observed in the p.R117C mutation compared to the WT or the p.L13F mutation. Moreover, the fluorescence intensity was visibly lower in the p.R117C mutation (Figure 4.9: C). This was also replicated in HEK293T cells, leading to similar observations, confirming that it is not cell model specific but a potential effect of the mutation.

To ascertain whether the observed difference in fluorescent intensity is genuine and possibly due to less protein reaching the plasma membrane in the case of p.R117C, transfected COS7 cells were imaged using Total internal reflection fluorescence (TIRF) microscope (Figure 4.10: A). This allowed measurement of fluorescence intensity of the plasma membrane in the WT and both mutants (Figure 4.10: B). Both mutations show that there is a significant reduction (p.L13F: $p=0.014$; p.R117C= $p<0.0001$) in the amount of protein reaching the plasma membrane with a greater degree of hinderance in the case of p.R117C mutation, which significantly different to both WT and p.L13F (Figure 4.10: C). This was further confirmed by performing western blot analysis on the plasma membrane fraction of the protein lysate from cells transfected with the WT-FLAG, p.L13F-FLAG, or p.R117C-FLAG constructs. The relative expression of both mutations was significantly reduced (p.L13F: $p=0.03$; p.R117C: $p=0.005$) in comparison to the WT (Figure 4.10: D) .

To further understand the dynamics of PLPPR5 mutants, it is vital to determine whether the translated protein is destined for degradation since reduced levels reach the plasma membrane. Even though we did not establish co-localisation between PLPPR5 (WT) and LAMP1-positive vesicles, it would be essential to look at the effect of the mutations on the number and size of LAMP1-positive vesicles as well as the expression of lysosomal-autophagy markers.

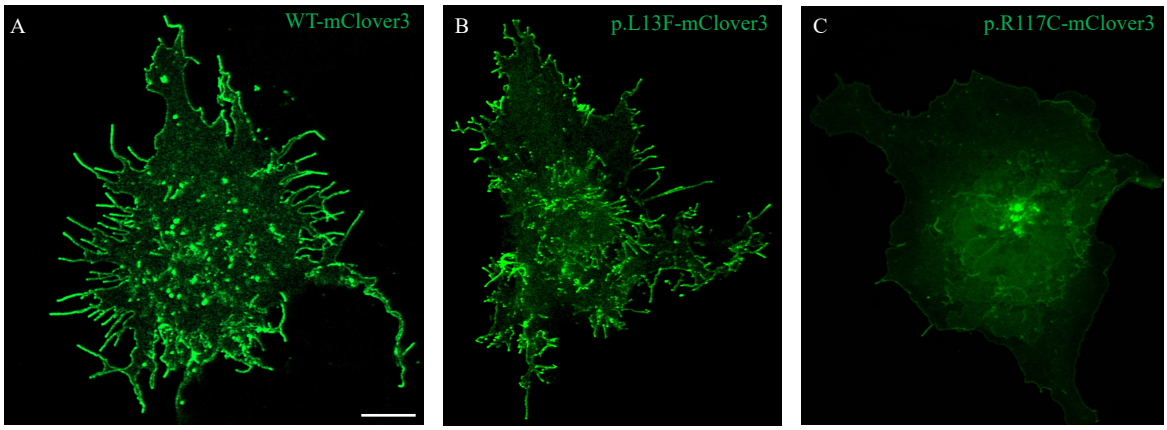


Figure 4.9: Effect of JS patient-specific mutations on PLPPR5 localisation in COS7 cells. A) PLPPR5-WT B) PLPPR5-p.L13F C) PLPPR5-p.R117. There are no striking differences observed in localisation to the plasma membrane and in inducing filopodia-like extensions between the WT and mutant p.L13F, however, mutant p.R117C (C) seems to fail to produce those extensions. Moreover, the signal PLPPR5-mClover3 signal is much lower in the mutant p.R117C than in the WT and mutant p.L13F. The same results were replicated in HEK293T cells (not shown here). Scale bar: 10 µM.

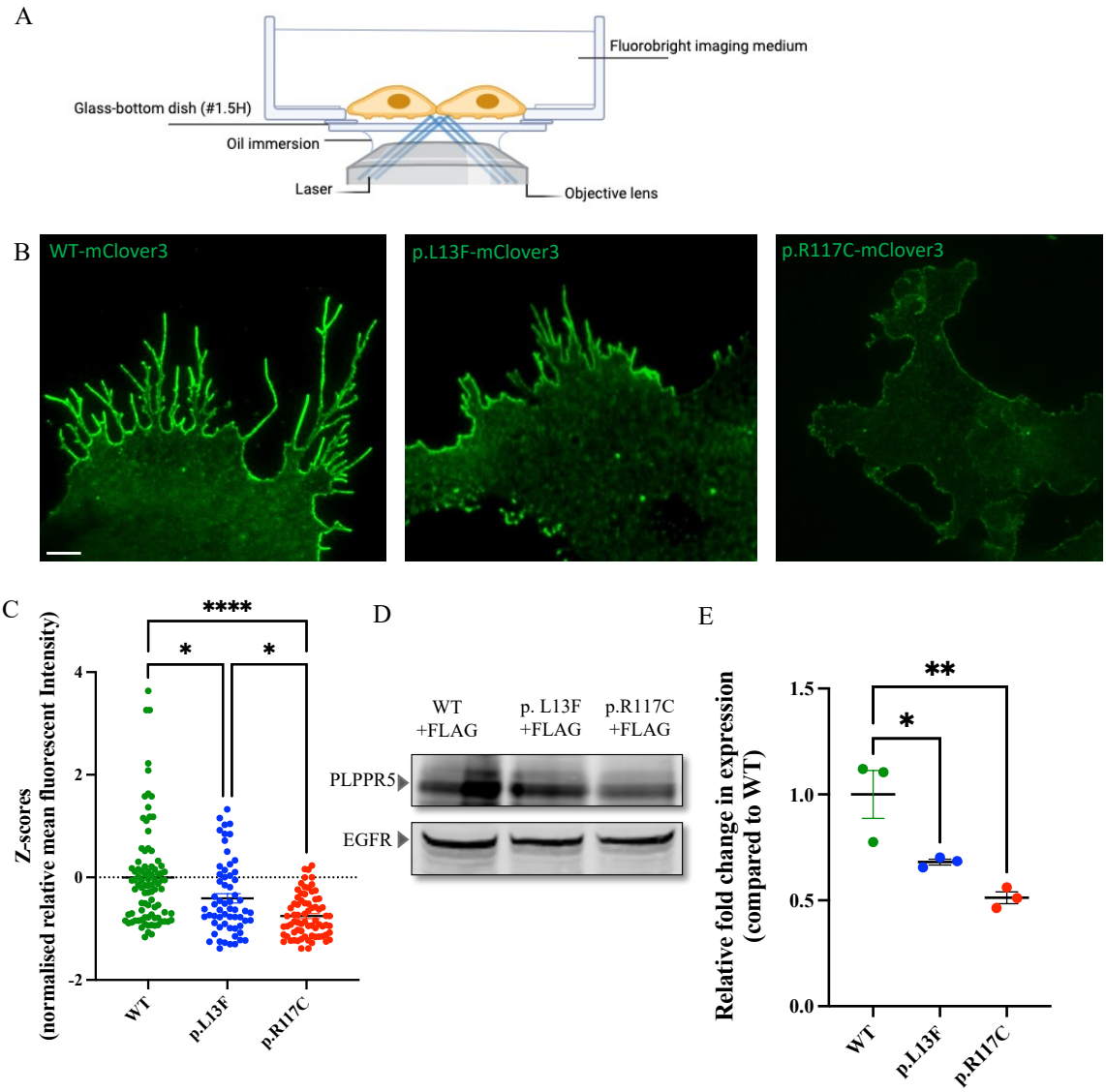


*Figure 4.10: Total Internal Reflection Fluorescence (TIRF) analysis. TIRF confirms the effect of the mutations in obstructing the protein reaching the plasma membrane when compared to the WT. A) glass-bottom dishes were used to image transfected COS7 cells using Elyra TIRF microscope in which the laser beams were set at 66º reflecting of the surface of the cells. B) Plasma membrane of COS7 cells transfected with PLPPR5 constructs (WT or mutant) tagged with mClover3. Scale bar: 10 µM C) Relative fluorescence intensity is significantly reduced in both mutations compared to WT (p.L13F = p 0.014; p.R117C p <0.0001, Kruskal-Wallis test). Three independent transfections were used to image an average of 65 cells. D-E) Expression level of PLPPR5 in the plasma membrane fraction. Western blot analysis of fractional plasma membrane lysates from WT (lane 1), p.L13F (lane 2), and p.R117C (lane 3) by using Anti-Flag (lower panel) and EGFR (upper panel). Relative fold change of PLPPR5 was calculated for n = 3 independent replicates, using normalized values to the mean (graph shows fold-change mean±S.E.M, * = p 0.03; **=p 0.005, one-way ANOVA test). Odyssey® CLx; resolution: 169 µm; Intensity: auto mode. Image Studio™ software was used for protein band detection.*

4.3.3 PLPPR5 mutations p.L13F and p.R117C lead to an increased number LAMP1-positive vesicles

Since we have shown in the previous experiments that the mutations result in less protein reaching the plasma membrane, we were curious to investigate the trafficking of mutant protein in the cells. COS7 cells were co-transfected with PLPPR5 (WT/mutant) and LAMP1 then imaged using confocal microscope for lysosomal analysis (Figure 4.11: panels A-C). Increased count of LAMP1-positive vesicles per cell in those over-expressing either of the mutations (p.L13F or p.R117C) when compared to WT (Figure 4.11: panel D) is noted. This would indicate an increased lysosomal content in cells over-expressing the mutations.

To elucidate this increase, expression levels of lysosome and autophagosome markers were characterised by quantitative reverse transcription-polymerase chain reaction (qRT-PCR). Two lysosome-related genes and two autophagy-related genes were investigated. For lysosomes, transcription factor EB (TFEB), which is a master regulator of many lysosomal genes leading to the induction lysosomal biogenesis (Sardiello et al., 2009); and LAMP1 a gene encoding a lysosomal membrane protein were selected. For autophagy, autophagy-related proteins (ATG) such as ATG-7 and ATG-12 were used as markers. ATG-7 is an essential autophagy effector enzyme, and ATG-12 is a ubiquitin-like protein involved in autophagy vesicles formation and participates in mitophagy (Alway, 2021). Cells expressing the PLPPR5-FLAG constructs with the mutations showed significantly increased levels of lysosomal and autophagy-related genes expression (Figure 4.12) at baseline under no stress when compared to the WT. This upregulation is consistent with the results showing the increase in LAMP1-positive vesicles from confocal images.

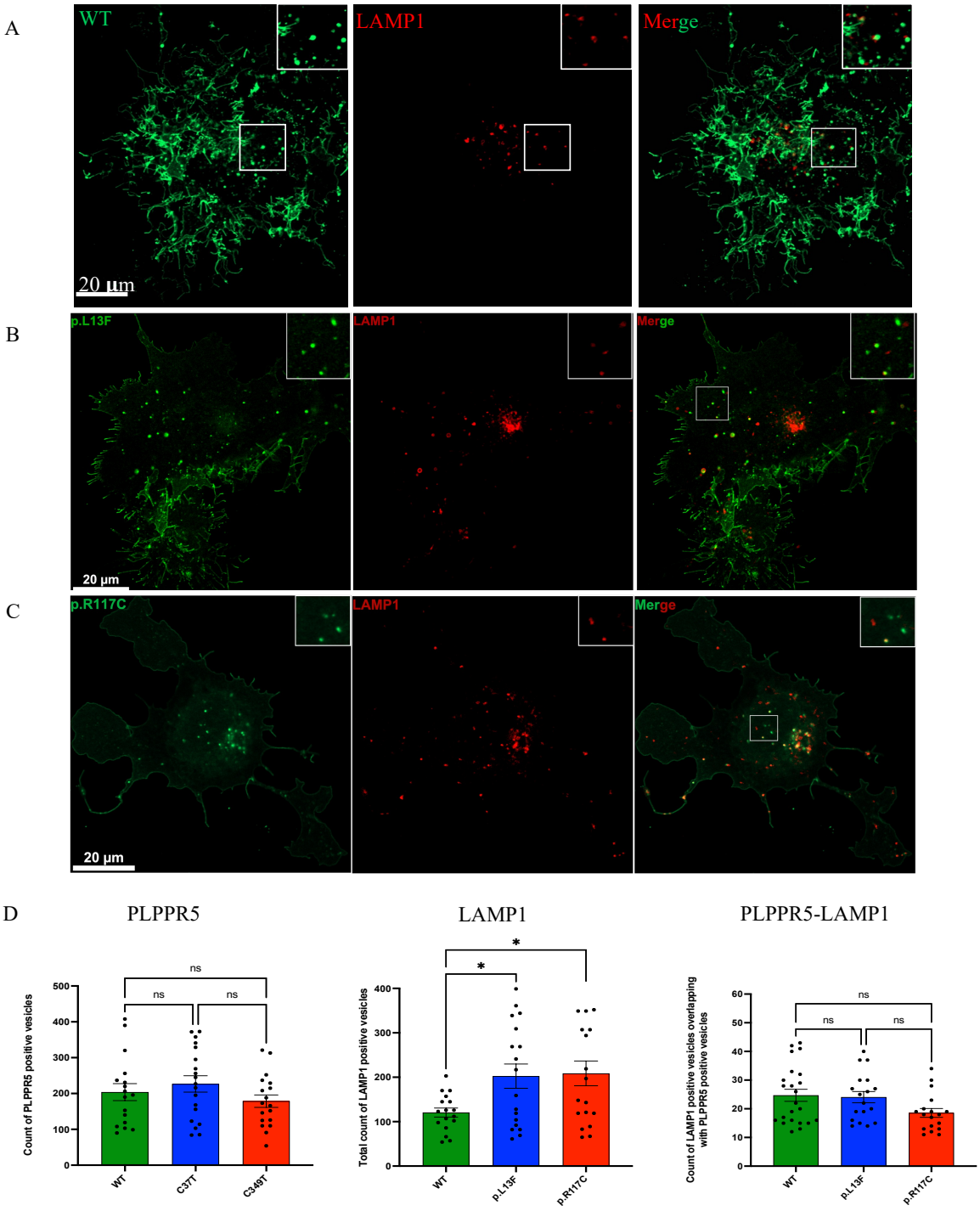


Figure 4.11: Lysosomal characterisation in COS7 cells expressing PLPPR5-mClover3. A) PLPPR5-WT does not co-localise with LAMP1-positive vesicles, therefore, puncta that are positive for PLPPR5-mClover3 are confirmed not to be present in the lysosomes. B) PLPPR5 mutant (p.L13F) C) PLPPR5 mutant (p.R117C). D) Quantification of the number of PLPPR5-positive vesicles, LAMP1-positive vesicles, and vesicles with overlapping signal. Quantitative results are mean ±s.e.m. One-way ANOVA and Bonferroni's multiple comparisons test. No significant difference can be detected in the count of vesicles positive for PLPPR5 or those with overlapping signal for both PLPPR5 and LAMP1. However, there is a significant increase in the total count of LAMP1- positive vesicles in both mutants when compared to the WT.

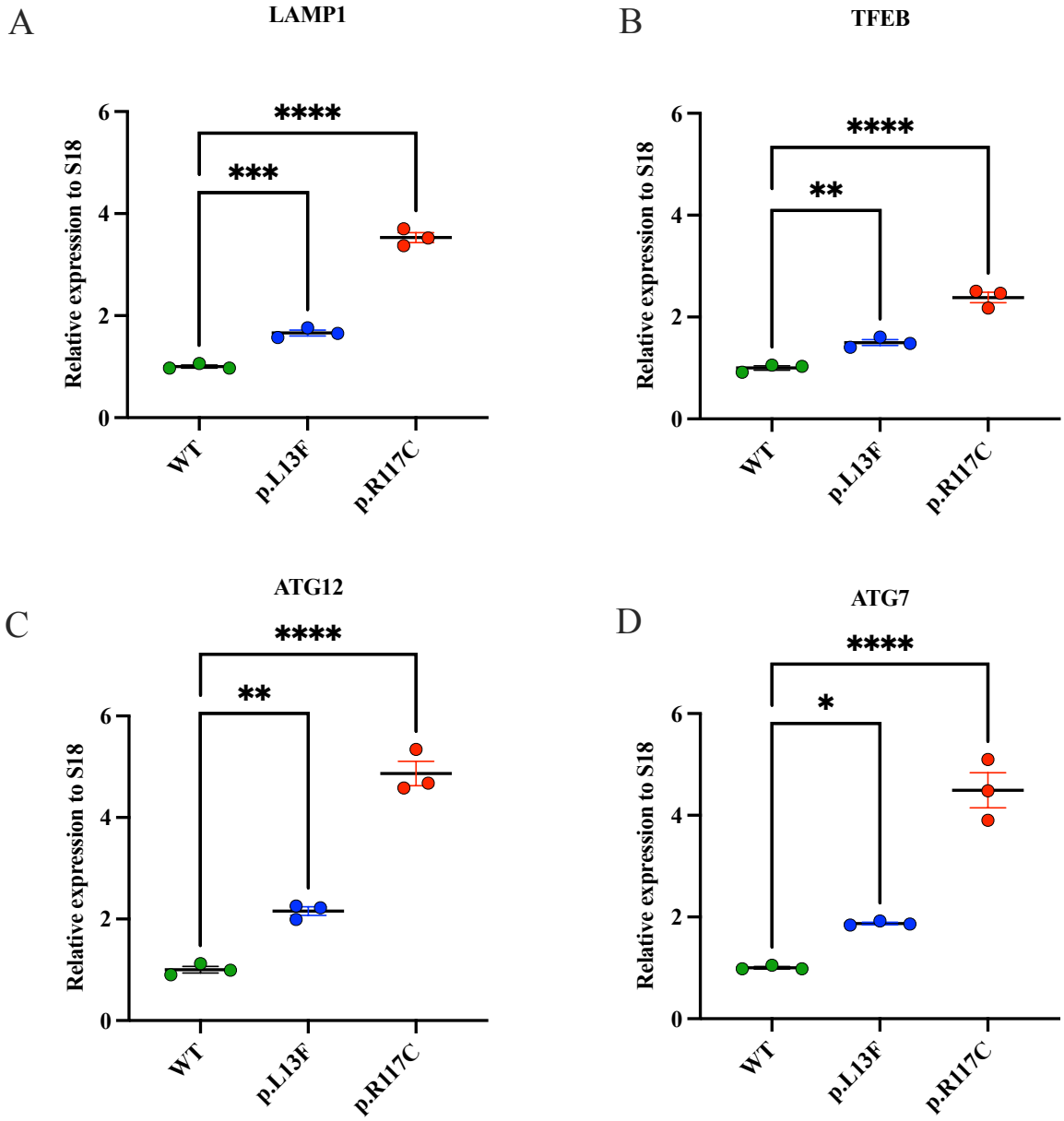


*Figure 4.12: Lysosomal and autophagosomal markers analysis show a significantly increased expression in PLPPR5 mutants. A-B) lysosomal markers LAMP1 and TFEB, C-D) autophagosomal markers ATG12 and ATG7 show significant increase in gene expression measured by qPCR in the mutants (p.L13F and p.R117C) when compared to the WT. (n=3). Statistical analysis was performed using One-way ANOVA and displayed as Mean±S.E.M. * denotes p=0.03, **denotes p=0.005, *** denotes p=0.0007, and **** denotes p <0.0001.*

4.3.4 Impaired mitochondrial dynamics in cells overexpressing either PLPPR5 mutations (p.L13F or p.R117C)

One specialised type of autophagy that has been recently shown to be upregulated in the *PLPPR5*$^{-/-}$ KO mouse model is mitophagy (Wang et al., 2022). This is a process of breaking down dysfunctional or surplus mitochondria within lysosomes (Sica et al., 2016). It is influenced by the dynamic fission-fusion processes of mitochondria altering their morphology, size, and number (Zungu et al., 2011). Therefore, we wondered if the mitochondrial dynamics might be impaired in the cells expressing *PLPPR5* mutations.

Morphological features of mitochondria can be used as predictors for the fission and fusion events. The establish characteristics are the mitochondrial perimeter, solidity, and circularity, which can be used to characterise the shape of mitochondria. There is a positive correlation between the perimeter of mitochondria and their likelihood of undergoing fission events. Likewise, mitochondrial solidity, which refers to a more compact shape, is positively correlated with mitochondria that are about to undergo fusion events (Westrate et al., 2014). Therefore, a more fused, convoluted, and irregularly shaped mitochondrial network would exhibit a larger perimeter but smaller solidity and circularity.

To assess whether the mutations could lead to abnormal mitochondrial network topology when compared to the WT, COS7 cells were co-transfected with PLPPR5-mClover3 and mito-mCherry as described in section 4.2.2.5 then subsequently imaged. The WT PLPPR5-expressing cells exhibited a typical highly interconnected mitochondrial network, indicating mitochondrial fusion (Figure 4.13: A). In contrast, cells expressing PLPPR5 mutations displayed a higher degree of fragmentation, with the p.R117C mutation showing more pronounced fragmentation (Figure 4.13: B-C). Statistical analysis using Mann Whitney's non-parametric test revealed a significant decrease in the average mitochondrial perimeter (adjusted p-value < 0.0001) for both mutations compared to the WT (Figure 4.13: D). In terms of average mitochondrial circularity, there was a significant increase (adjusted p-value =0.0271) in p.R117C-expressing cells compared to the WT (Figure 4.13: E). Similar results were observed in terms of mitochondrial solidity, with a significant increase in average solidity (adjusted p-value =0.027) in cells expressing the p.R117C mutation, while cells expressing the

p.L13F mutation showed a slight increase, but it did not reach statistical significance compared to the WT (Figure 4.13: F).

To rule out whether the dysregulation in the mitochondrial network topology could be due to aberrant transcription of mitochondrial genes, the expression of mitochondrial RNA polymerase (*POLRMT*) was examined using qRT-PCR. However, there was not significant difference in expression between the WT, p.L13F, or p.R117C (Figure 4.13: G).

The observed impairment in mitochondrial network prompted the question of whether the mutations have an impact on the mitochondrial membrane potential (ΔΨm). This was assessed by staining with the potentiometric fluorescent TMRM dye in WT and cells overexpressing mutations (p.L13F or p.R117C) using flow cytometry (Figure 4.13: H). The mean fluorescence intensity was measured in an average of 50,000 single cells (n=3) that were selected based on showing positive mClover3 signal. Both mutations showed a reduced mean fluorescence intensity compared to the WT, however, only p.R117C exhibited significantly lower ΔΨm under baseline conditions (Figure 4.13: I).

From these collective data showing impairment of the mitochondrial morphology and membrane potential, it was essential to assess whether the mitochondrial metabolic respiration was affected. Therefore, additional characterisation of the oxidative phosphorylation in both WT and cells expressing p.L13F or p.R117C mutations was performed using Seahorse metabolic analyser to measure oxygen consumption rate (OCR).

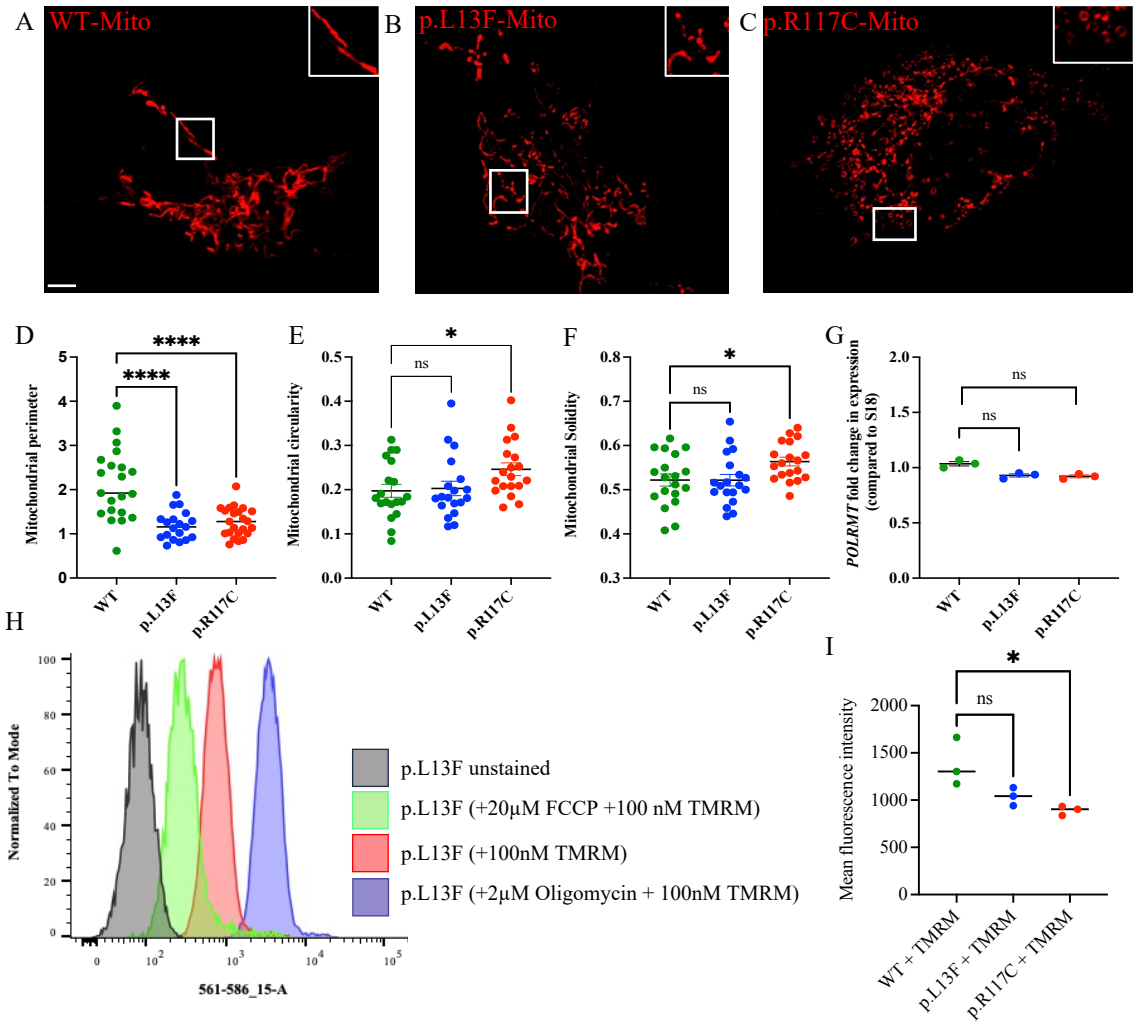


*Figure 4.13: Mitochondrial analysis in COS7 cells expressing PLPPR5 mutations show significant mitochondrial fragmentations compared to the WT. (A-C) Representative images of mitochondrial network topology were captured in COS7 cells expressing PLPPR5 WT or mutations (p.L13F or p.R117C), along with mito-cherry plasmid for mitochondrial morphology visualisation. Images were acquired using a Zeiss LSM900 AiryScan confocal microscope. Scale bar: 10µm. (D) The average mitochondrial perimeter was significantly different among the WT, p.L13F, and p.R117C mutants (p <0.0001). E) The average mitochondrial solidity showed a significant difference among the WT and p.R117C mutants (p=0.0271) but not p.L13F. Similarly, F) the average mitochondrial circularity exhibited a significant difference among the WT, and p.R117C mutant (p=0.0257) but does not reach significance in p.L13F. These parameters were quantified using a specialised mitochondrial morphology macro in ImageJ that evaluates mitochondrial network topology based on relevant parameters. G) Relative expression of POLRMT was determined using qPCR, which revealed no significant difference in expression between WT, p.L13F, or p.R117C. H) True TMRM staining signal was ascertained using three different controls (unstained cells for background noise, cells treated with 20uM FCCP+ 100nM TMRM as negative controls, cells treated with 2uM oligomycin + 100nM TMRM as positive controls). I)The mean fluorescence intensity reveals a reduced potential in both mutations that reaches the significance levels for p.R117C when compared to WT (p=0.024) but not for p.L13F (n= 3 biological replicates with 3 replicates per biological replicate). Statistical analysis was performed using Mann Whitney's non-parametric test. The datasets did not follow a normal distribution, as determined by at least two normality tests conducted in GraphPad Prism. The study consisted of two independent experiments, with 10 cells analysed per transfection experiment. The data are presented as mean ± SEM. ns indicates non-significant differences; * denotes p<0.05; ** denotes p<0.01.*

4.3.5 Mitochondrial bioenergetic profile in cells overexpressing either of the PLPPR5 mutations (p.L13F or p.R117C) indicates a lowered mitochondrial fitness under baseline conditions when compared to the WT

Mitochondria has a vital role in cellular metabolism, the generation of ROS, and the regulation of apoptosis. Consequently, evaluating mitochondrial functions offers invaluable understanding into the physiology and bioenergetics of cells under baseline or during stresses due to external stimuli or genetic defects in a relevant gene to the mitochondrial function.

The mitochondrial oxidative phosphorylation and glycolysis are key metabolic pathways in the brain, while it has a limited capacity for β-oxidation of fatty acid (Yang et al., 1987).

Mitochondrial dysfunction is often linked to diseases and is characterised by the lack of ATP production and maintenance of the mitochondria. It often presents as alterations in mitochondrial dynamics, morphology, size, ROS generation, and apoptosis. The primary ROS generation sites within the ETC are complexes I and III, where electrons may leak through the membrane, reacting with oxygen to form superoxide ion ($O_2\bullet^-$). Most superoxide ions are transformed into hydrogen peroxide (H_2O_2) through natural or enzymatic dismutation by superoxide dismutase (SOD), leading to their inactivation. However, the highly reactive hydroxyl radical (•OH) can be generated from H_2O_2, which can trigger oxidative stress, causing cellular damage (Tangvarasittichai, 2015). Oxidative stress affects mitochondrial function by causing lipid peroxidation of the internal mitochondrial membrane, leading to proton leakage, loss of electrochemical gradient, and impaired ETC or oxidative phosphorylation (OXPHOS) activity (Zhao et al., 2019). Oxidative stress prompts cytochrome c release into the cytoplasm, triggering apoptosis. The proximity of mitochondrial DNA (mtDNA) to ROS production sites and its lack of protective histones renders it susceptible to damage. Oxidative damage to mtDNA leads to single or double-stranded DNA breaks, and inability to repair the damaged mtDNA results in its segregation and subsequent mitophagy. Mitochondrial dynamics are linked to bioenergetics, with fusion enhancing ATP production and fission impairing OXPHOS and promoting ROS generation (Zorov et al., 2019). The balance between fusion and fission adapts mitochondria to different metabolic demands. Excessive ROS production compromises cellular physiology (Annesley & Fisher, 2019).

To characterise the impact of the *PLPPR5* mutations on mitochondrial function, I examined the OCR in HEK293T cells overexpressing the WT or mutations (p.L13F or p.R117C). This was achieved using the XFe96 Analyzer, which allows for measurements of OCRs under various conditions such as basal medium, during inhibition of OXPHOS (ATP-synthase, complex I and III of electron transport chain inhibition), and under uncoupled respiration chain (Figure 4.14).

By analysing the OCR trace, we can observe distinct changes in the cellular oxygen consumption at different stages of the Mito-Stress test (Figure 4.14: A). The difference between the basal OCR measurements and the OCR readings following Rotenone/Antimycin A injection reflects the overall mitochondrial basal respiratory activity due to electron transport chain under normal conditions. For these measurements, both mutations showed ~5-10% reduction in basal respiration, and that was not a significant difference to the cells expressing WT (Figure 4.14: B).

Following the injection of oligomycin, there was a reduction in OCR, indicating the inhibition of ATP production by blocking the ATP synthase. This difference between the basal OCR readings and those after the oligomycin injections represents the ATP-linked OCR, which corresponds to the OCR specifically associated with ATP synthesis. Here, we observe a significant difference between the cells expressing p.R117C compared to the WT but not in the case of p.L13F (Figure 4.14: C).

The subsequent injection of FCCP leads to an increase in OCR, surpassing the baseline level to the double. This increase signifies the maximal OCR, indicating the mitochondrial respiratory capacity in the absence of ATP synthesis limitation due to uncoupling of the ETC. There is no significant changes between the cells expressing WT or mutations in this parameter (Figure 4.14: D). The proportion of OCR in the ETC, but not driven from ATPase proton pumping, with respect to all non-ATPase driven OCR shows the proton leakage across the membrane. This calculation shows that both mutations have no significant difference to WT but trending to have slightly higher proton leak compared to the WT (Figure 4.14: E). Finally, the addition of combined rotenone/antimycin A (Rot/AA) further alters the OCR trace. Rotenone inhibits complex I, while antimycin A blocks complex III of the electron transport chain (ETC). By comparing the maximal OCR with the basal OCR, we can determine the spare respiratory capacity. This parameter represents the potential for additional ATP production during periods of increased energy demand. For this reading, both mutations show no significant differences to the WT but trending to have a lowered capacity (Figure

4.14: F). The resulting decrease in OCR following the injection represents the non-mitochondrial OCR, as it is unrelated to mitochondrial respiration. This also shows no significant differences compared to the WT (Figure 4.14: G). Therefore, the overall functional capacity of the mitochondria does not seem to be prominently impacted under baseline conditions in cells overexpressing the mutations compared to the WT. However, given the observed patterns, it would be important to test how well mutants cope with stressors such as long-chain fatty acids.

Long chain fatty acids (LCFAs) have to be converted into CoA-thioesters before being translocated as acyl-carnitines across the inner membrane of mitochondria, however, the brain mitochondria are unable to breakdown very long chain fatty acid (VLCFAs) that have ≥22 carbon atoms. Yet, there is an established susceptibility of the mitochondrial physiology to VLCFAs (Fourcade et al., 2014; Galea et al., 2012), which have to be partly broken down by peroxisomal β-oxidation before being released to the mitochondria (Kemp et al., 2012; Wanders & Waterham, 2006). When this cooperative degradation of the long-chain fatty acids is impaired, it could have detrimental effect on cellular energy metabolism and mitochondrial function. These LCFAs partly impair ATP regeneration, leading to a reduction in cellular energy levels. Furthermore, they can induce depolarisation of mitochondria, disrupting the electrochemical gradient necessary for ATP synthesis. This depolarisation can sensitise mitochondria to undergo permeability transition, a process associated with cell death. Additionally, elevated levels of LCFAs can stimulate both mitochondria-associated and non-mitochondrial oxidative stress, contributing to cellular damage and dysfunction inducing various pathological processes within the mitochondria, ultimately impairing cellular function (Schönfeld & Wojtczak, 2008).

An exploratory untargeted metabolomics analysis blind to the patient's phenotypes was conducted by Metabolon. While this did not show conclusive results shared between the affected siblings, it has revealed two significant differences in the biochemical profile of the affected female compared to the unaffected mother and unaffected sister. These differences were the levels of saturated and monounsaturated fatty acid (Table 4.12) and the phosphatidylethanolamines, in which most of these three lipid classes were above the 97.5th percentile of the expected range cohort. Additional potential biochemical perturbations noted were high glucocorticoid levels and increased creatine production.

The metabolomic findings point towards a possible impact on the function of the mitochondria under the presence of LCFAs. However, this was beyond the scope of the current study and needs to be further explored in future ones.

Table 4.12: Z-score table of significantly different biochemicals identified for each sample against a healthy fasted control of population 300 subjects. Samples highlighted in red indicate increase while the ones in green are marked decrease

		Reference Range 300 Adult Fasted						
Sub Pathway	**Biochemical Name**	**2.5th %**	**97.5th %**		**Affected male**	**Affected female**	**Mother**	**unaffected female**
Long Chain Saturated Fatty Acid	palmitate (16:0)	-1.7043	1.1237		0.0927	**1.1498**	0.7325	0.3516
	margarate (17:0)	-1.1934	0.9222		-0.6377	0.541	0.2821	-0.5711
	stearate (18:0)	-1.1588	0.7835		0.0607	**1.2239**	0.4056	-0.5166
	nonadecanoate (19:0)	-0.9856	0.7113		-0.2758	0.2733	0.3004	**-1.3968**
	arachidate (20:0)	-0.6505	0.9886		-0.3498	**0.9934**	0.1448	**-1.842**
Long Chain Monounsaturated Fatty Acid	myristoleate (14:1n5)	-2.7892	1.585		-0.2289	1.4817	0.5477	-0.374
	palmitoleate (16:1n7)	-2.5948	1.4255		0.2379	**1.632**	0.9289	0.5722
	10-heptadecenoate (17:1n7)	-2.034	1.2656		-0.0157	**1.2997**	0.8711	0.1407
	oleate/vaccenate (18:1)	-1.8995	1.0376		0.4271	**1.4119**	0.8187	0.8438
	10-nonadecenoate (19:1n9)	-1.1593	0.8285		0.012	**0.9516**	0.6409	-0.2652
	eicosenoate (20:1)	-1.1316	0.8815		0.173	**1.425**	**1.1037**	0.8627

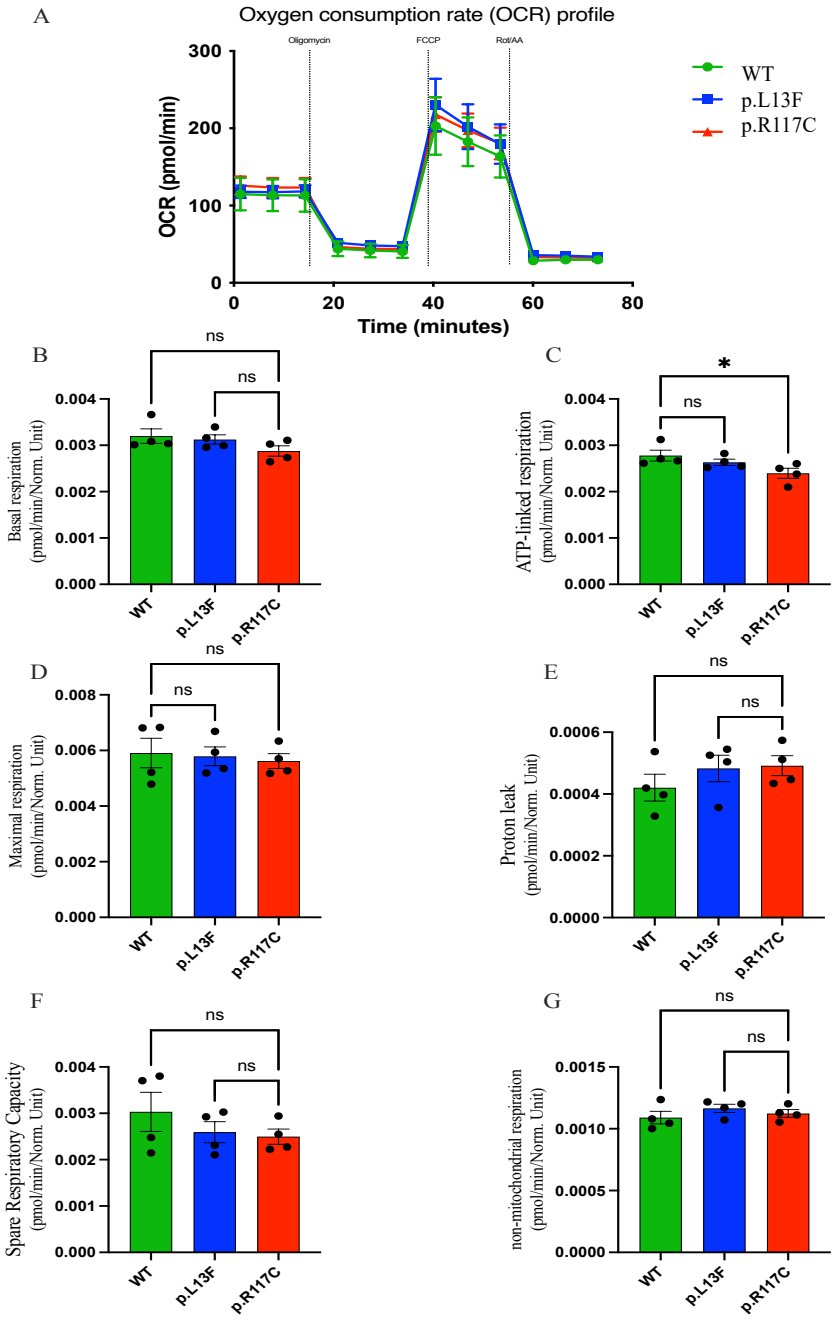


Figure 4.14: Oxygen Consumption Rate (OCR) profile in HEK293T cells expressing PLPPR5 (WT/mutant). A) the overall OCR profile used to evaluate the acute effect of mitochondrial targeted compounds, which provides valuable insights into mitochondrial respiratory function by measuring OCR at different stages and calculating parameters related to ATP production, and respiratory capacity. B-G) One-way ANOVA test reveals a significant difference between cells expressing the p.R117C and WT in ATP-linked respiration (p=0.04, n=3) but not p.L13F. Both mutants show a trend of lowered respiration linked to oxidative phosphorylation that does not reach the significance levels except for the ATP-linked respiration in p.R117C. This is reflected in a higher proton leak pattern, and non-mitochondrial respiration levels.

4.4 Discussion

This study was initiated upon finding rare compound heterozygous variants in *PLPPR5* from WGS data of two affected siblings with JS. The variants co-segregated with the phenotype when both unaffected parents and an unaffected female sibling were tested. The protein informatic analysis of the variants revealed that both are predicted to affect the protein structure and function. The p.L13F, an N-terminal mutation facing the cytosol, was predicted to be near cavity/pocket that potentially binds phospholipid heads from AlphaFold-2 model. Moreover, the p.R117C is predicted to disrupt an ER-exit signalling motif. Together with the function of *PLPPR5*, rarity of the variants, and co-segregation with phenotypes, *PLPPR5* makes a prime candidate for being a potentially novel disease-causing gene. Hence, further functional studies were conducted. These investigations have revealed significant differences in key cellular functions between mutants and WT PLPPR5. Both variants resulted in an upregulated expression of key autophagy genes, increase in lysosomal number, and significant mitochondrial morphological and functional impairment. Furthermore, both variants showed a reduced capacity to generate filopodia-like extensions that was markedly reduced in the case of p.R117C when compared to the WT.

The first aim of this study was to characterise delivery of WT and mutated PLPPR5 to the plasma membrane. The gene has been shown in several studies to be involved in dendritic spine formation in cortical neurons and in inducing filopodia-like extension formation in non-neuronal cell lines (Broggini et al., 2010; Coiro et al., 2014; Gross et al., 2022). Upon transfecting COS7 cells with WT or mutant constructs, the initial observation did not reveal a noticeable disparity in filopodia-like extensions between the WT and the p.L13F mutation. However, a closer examination of fluorescence intensity exposed a significant impairment in protein trafficking to the plasma membrane. This hinderance was further accentuated in the presence of p.R117C, resulting in a marked reduction of filopodia-like extensions in COS7 cells, possibly due to less mutated protein making it to the plasma membrane and sites where filopodia forms. These extensions are of interest to the observed phenotype in the patients as formed filopodia mature into dendritic spines in neurons.

Dendritic spines are tiny extensions found along the neuronal dendrites and form the main sites for excitatory synapses. They have dynamic structure and morphology as well as varying lengths and densities, which influence the neuronal synaptic activity and neuronal connectivity at different developmental stages (Parker et al., 2020; Penzes et al., 2011; Arellano et al., 2007). Dendritic spine changes is frequently observed as the earliest manifestation in hippocampal pyramidal neurons and dentate granule cells in patients with epilepsy where fewer or absent spines are seen (Rossini et al., 2021; Belichenko & Dahlström, 1995; Blümcke et al., 1999; Isokawa & Levesque, 1991; Scheibel et al., 1974; Von-Campe et al., 1997). Similar observations of dendritic spine loss have been reported in neocortical pyramidal neurons in patients with partial epilepsy (Multani et al., 1994). Furthermore, animal models of epilepsy that have acute seizures or chronic epilepsy have exhibited similar trends of dendritic spine loss (Ampuero et al., 2007; Isokawa, 1998; Nishizuka et al., 1991; Willmore et al., 1980). While the mechanism by which dendritic spine loss contribute to epileptogenesis is not clearly established, one would argue that the disruption of finely-tuned neuronal networks would promote seizures.

In addition to the failure in inducing filopodia-like extensions, this study also demonstrated increased expression of autophagy-related genes and elevated number of lysosomes in cells overexpressing the variants. Autophagy is a cellular process that involves the degradation and recycling of unnecessary or dysfunctional cellular components, such as damaged proteins or organelles. It is a tightly regulated process in which cells generate double-membrane structures called autophagosomes to encapsulate the targeted cellular material. These autophagosomes then fuse with lysosomes, forming autolysosomes, where the enclosed components are broken down by lysosomal enzymes. Autophagy serves as a crucial mechanism for maintaining cellular homeostasis, removing cellular waste, providing energy during periods of nutrient deprivation, and facilitating cellular remodelling and adaptation to various stress conditions (Dikic & Elazar, 2018). The significance of autophagy is increasingly being recognised in neural development (Lv et al., 2014), synaptic function (Hoffmann et al., 2019), and the establishment of healthy neuronal connectivity (Dragich et al., 2016). Consequently, it is not unexpected that aberrant autophagy has been observed in a range of neurodevelopmental disorders (Menzies et al., 2015), frequently accompanied by early-onset epilepsy such as in TBCK-related syndrome (Ortiz-González et al., 2018).

Dysfunctional autophagy has been established in association with various genetic variants linked to neurodegenerative diseases, which affect different stages of the autophagic process (Frake et al., 2015). These variants impact the regulation of autophagy initiation, the sequestration of target substances into autophagosomes, trafficking of autophagosomes to lysosomes, and the capacity of lysosomes for degradation.

In the case of the *PLPPR5* variants, the observed upregulation in autophagy-lysosomal related genes does not necessarily reflect an enhanced degradation of aberrant protein due to the variants. The reduced mutated protein trafficked to the plasma membrane suggest it is being subjected to degradation, but the significant increase in autophagy markers and lysosomal count in cells expressing the variants compared to WT indicate an impaired mechanism regulating the balance between activation/inactivation of autophagy.

Many studies established a link between epilepsy and autophagy through the activation of the mammalian Target of Rapamycin (mTOR) pathway. Initially it was observed that rapamycin, which inhibits the mTOR pathway and subsequently leads to the activation of autophagy, regulates seizures in patients with mTORpathy epilepsies (Giorgi et al., 2015). This provided an insight of a possible mechanism for epilepsy through defective autophagy. This observation was supported by genetic variants in genes linked to mTOR regulation such as TSC Complex Subunit 1 and 2 (*TSC1 and TSC2*) leading to hyperactivation of mTOR and inhibition of Autophagy (Park et al., 2018; Lipton & Sahin, 2014). *PLPPR5* is a binding partner of *PLPPR1*, which has been shown to activate *mTOR* and its regulator Phosphatase and Tensin homolog; *PTEN* (P. Yu et al., 2015). This interaction was found to be independent of the capacity of *PLPPR1* and its binding partners such as *PLPPR5* to form filopodia-like extension (P. Yu et al., 2015). Given that mTOR is a negative regulator for the initiation of autophagy (Laplante & Sabatini, 2012), it would be expected that variants in PLPPRs inhibit *mTOR* and in turn upregulate the expression of autophagy-related genes without compromising the degradation capacity of the lysosomes. While in this study I have not tested mTOR expression, it would be essential to test the level of the active phosphorylated-mTOR in cells expressing variants and/or in patients' samples to support or negate this hypothesis.

Another possibility is that *PLPPR5* does not directly affect mTOR phosphorylation but acts downstream during one of the autophagic vesicle maturation steps and the variants lead to a

disruption of the feedback loop that eventually activates mTOR and inhibits autophagy. In this possibility, the aberrant protein accumulates in the cytosol without reaching its final destination in the plasma membrane impacting the degradation capacity of the lysosomes. This is observed in patients with variants in TBC1 Domain Containing Kinase (*TBCK*), which causes TBCK-related syndrome. In patients with this syndrome, the upregulated autophagy is not responsive to pro-autophagic stimuli (Ortiz-González et al., 2018). Moreover, despite the observed upregulation, there is no concurrent increase in the degradation of glycosylated proteins or their accumulation in neurons (Beck-Wödl et al., 2018). This discrepancy suggests an impairment in the regulation of autophagy, leading to a storage disease in which epilepsy is a frequent phenotype of the patients (Teinert et al., 2020). This is further supported with reports on CLN3 disease, which is a lysosomal storage condition with severe neurodegeneration and early-onset seizures. In CLN3 disease, there is an increased lysosomal protein expression but defective degradation capacity (Chear et al., 2022). Interestingly, a similar pattern of upregulated autophagy was found in patients with Juvenile Myoclonic Epilepsy (JME), which has a great overlap with JS (Tanaka et al., 2018).

The finding in this study is in line with these recent reports suggesting that the variants in *PLPPR5* might be leading to inactivation of mTOR pathway as it acts upstream of mTOR and PTEN (P. Yu et al., 2015), which in turns induces autophagy without the enhanced degradation capacity.

Mitophagy is specialised form of autophagy that has been suggested by Wang et al., 2022 to be upregulated in *PLPPR5* KO mice. While I did not look at mitophagy in this study, I have explored the effect of the variants on the mitochondrial dynamics and function.

Mitochondrial dysfunction has been linked to the development of seizures and epilepsies through various mechanisms including: Ca_2^+ buffering (Sanganahalli et al., 2013), impaired ATP production, increased generation of ROS, which in turns induces oxidation of ion channels and neurotransmitter transporters (Trotti et al., 1998), as well as increased excitability due to a decrease in the membrane potential (Hardie et al., 2012).

The results from the mitochondrial study on the *PLPPR5* variants (p.L13F and p.R117C) revealed that while PLPPR5 does not localise in the mitochondria, it has an effect on their morphology and network. The mitochondria show a greater degree of fragmentation in the mutants compared to the

WT. This was accompanied by a marked decrease in the mitochondrial membrane potential especially in the case of p.R117C variant.

The observed decrease in membrane potential under normal conditions is likely associated with the presence of compromised mitochondria. This is expected to negatively impact the viability and functions of the cell, as the mitochondrial membrane potential plays a critical role in driving ATP synthesis (Klingenberg & Rottenberg, 1977). Sustaining the mitochondrial membrane potential is crucial for the internal movement of cations like Ca_2^+ (Gunter & Pfeiffer, 1990), and it is essential for the uptake of various mitochondrial proteins (Shariff et al., 2004; Krayl et al., 2007). Mechanisms responsible for upholding the quality of mitochondria, including processes like mitochondrial fragmentation (Ishihara et al., 2003) and the elimination of depolarized mitochondria through mitophagy (Youle & Narendra, 2011; Pickles et al., 2018), are thus indispensable. These mechanisms are likely impaired in cells expressing the PLPPR5 variant (p.R117C), which might elucidate the observed presence of damaged and fragmented mitochondria in these cells. However, mitophagy studies would need to be conducted to confirm this speculation.

Overall, this study provides insights into how variants in *PLPPR5* might contribute to epilepsy development, suggesting that these mechanisms involve failure to induce filopodia formation, as well as the disruption of both autophagy and mitochondrial functions.

4.5 Limitations of the current work

The study investigated the variants separately, even though the patient has a compound heterozygous state. This limitation may not accurately represent the combined effects of these variants, as the interactions between them could be different from their individual effects. Understanding the synergistic effects of compound variants could provide more insight.

The results were based on overexpression studies. While it is informative, it is difficult to eliminate the potential effects of endogenous WT when overexpressing the mutant. A more accurate understanding of the variant's impact would require studies in edited cell lines or after knockdown of the endogenously expressed *PLPPR5*.

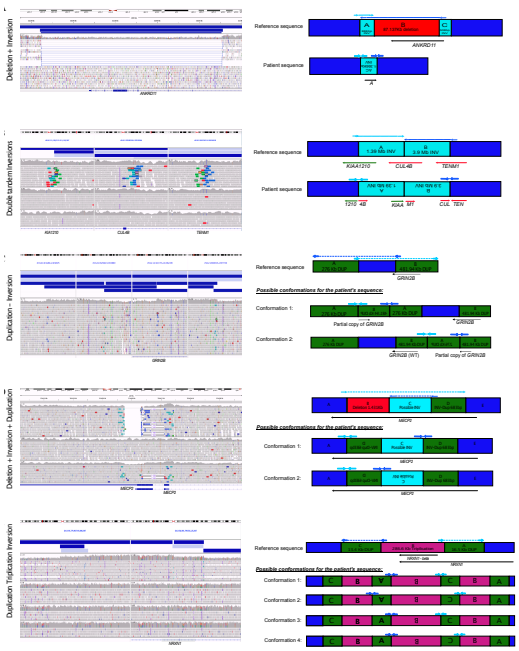
Moreover, the use of transient expression for studying the effect of the variants provides only a short window for investigations. Transient expression may not accurately capture long-term effects, stability, or changes that might occur over time. Stable cell lines can provide more insight into how expression of the variants affect cells over extended periods. However, despite careful efforts we were unable to establish stable cell lines for p.R117C, which presented an obstacle to using stable cell lines potentially missing out on insights into the variants' effects.

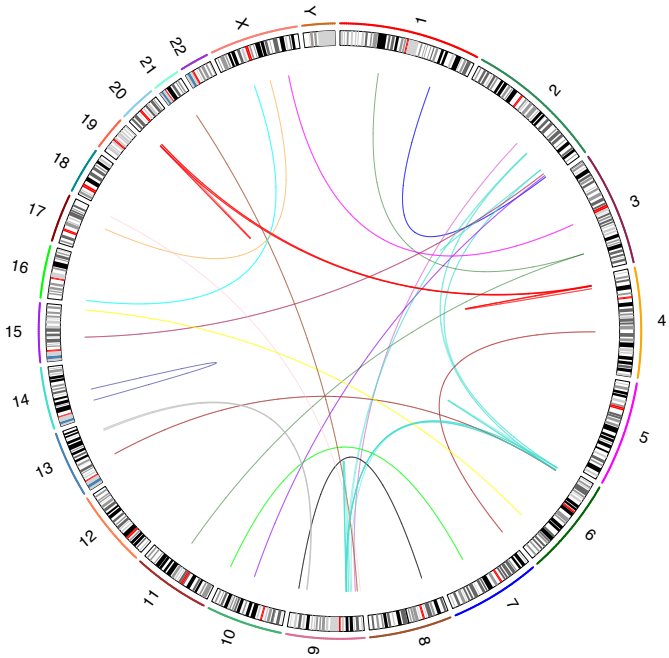
Furthermore, the study conducted experiments in non-neuronal cell lines. To definitively link the variants to epilepsy, it would be beneficial to investigate them in a neuronal cell line. Neuronal cells have unique properties and functions, and their response to variants may differ from that of non-neuronal cells.

In summary, while the study provides valuable insights into the effects of specific variants on *PLPPR5*, these limitations highlight the need for further research to better understand their complex nature and effect in inducing seizures leading to epilepsy. Future studies that address these limitations, such as investigating compound variants and using neuronal cell lines, may provide a more comprehensive understanding of the molecular mechanisms involved.

Graphical abstract chapter 5

Systematic analysis of structural variants including cxSVs, and BNDs in WGS data of individuals with neurological disorders specifically neurodevelopmental ones, or previously known karyotype abnormalities significantly increases the diagnostic yield, leading to rapid identification of the causal genes/ variants, thereby ending the diagnostic odyssey for many.





Chapter 5. Structural Variants: Rapidly increasing the diagnostic yield of Whole-Genome Sequencing

5.1: Introduction

SVs are large genomic rearrangements involving 50 base-pairs or more (Sudmant et al., 2015). These alterations range in size leading to different variations that affect the gene's structure and function. SVs include a class of variants that lead to copy number changes, such as deletions, duplications, and unbalanced translocations. In contrast, inversions and balanced translocations are copy-neutral (Sudmant et al., 2015; Figure 5.1).

In recent years, more SVs have been detected that do not fit in canonical SV categories as they have two or more SV types in combination. These complex non-canonical SVs cover heterogeneous rearrangements ranging from inversions flanked by CNVs to chromosomal shattering and reorganisation (Chiang et al., 2012; Kloosterman et al., 2011). The most severe, ultra-rare form of these complex SV is chromoanagenesis (Holland & Cleveland, 2012), which encompasses three different mechanisms: chromothripsis[1] (Stephens et al., 2011), chromoanasynthesis[2] (Liu et al., 2011), and chromoplexy[3] (Baca et al., 2013). The most described type is chromothripsis, which was initially reported in cancer cases; however, later studies discovered both balanced and unbalanced germline chromothripsis. These discoveries have shed light on a spectrum of SVs in human genomes that were not previously appreciated.

[1] A catastrophic event in which 10s to 100s of chromosomal rearrangements have occurred in a one-off cellular crisis

[2] A chromosomal reconstitution that involves different copy number profiles reflecting the involvement of cellular DNA replicative/repair process for maintaining genome stability such as (Microhomology-mediated breakage induced repair (MMBIR) / Fork Stalling and Template Switching (FoSTeS))

[3] DNA translocations and deletions that arise in a highly interdependent manner and occurs in cancers only.

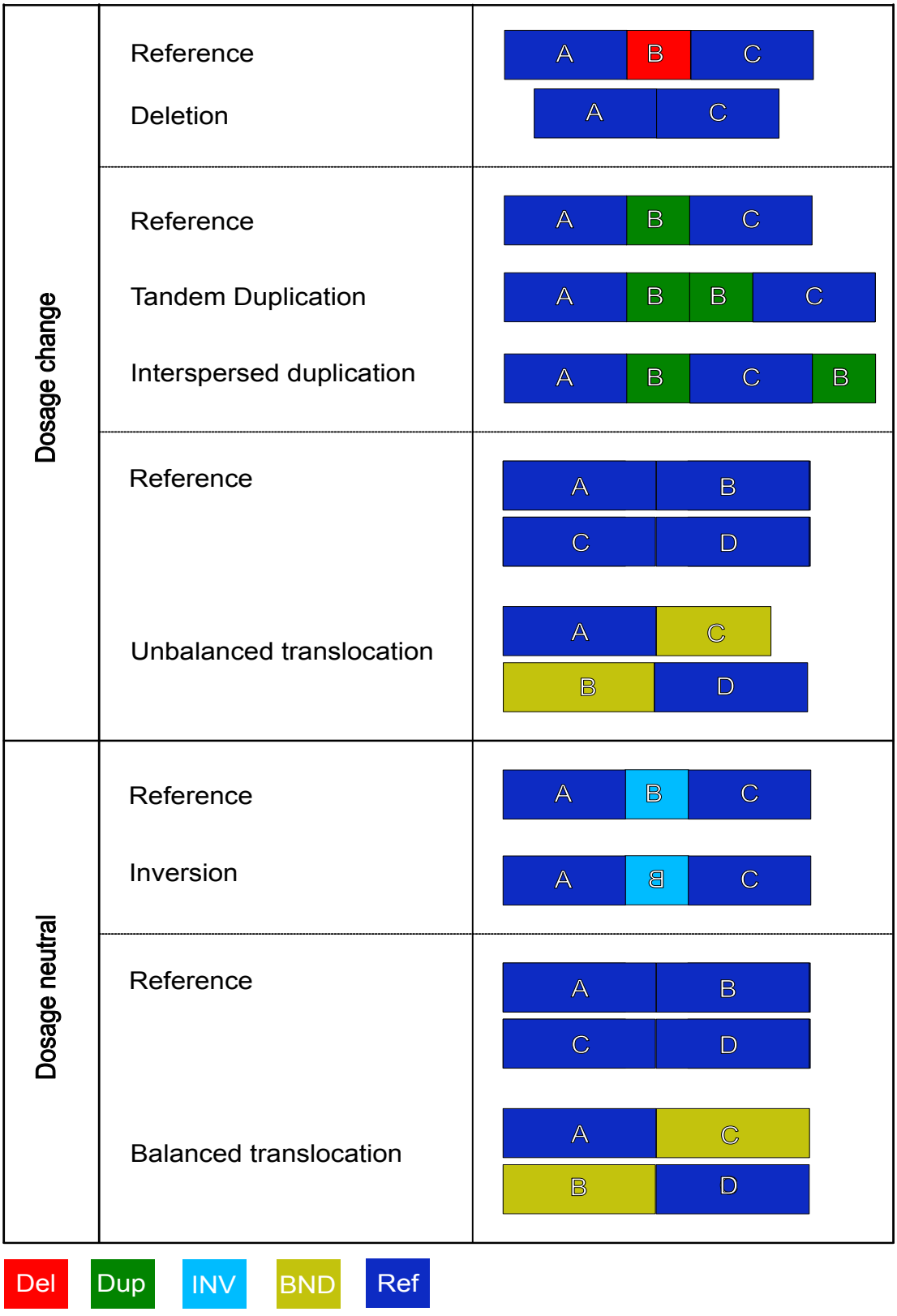


Figure 5.1: Conventional structural variant types classified by their effect on dosage alterations. "Unbalanced" SVs lead to loss/gain of genetic material such as CNVs; "Balanced" changes such as inversions and balanced translocations do not involve dosage alterations. Del= Deletion, Dup = Duplication, INV= Inversion, BND= Breakend, Ref= reference sequence.

5.1.1 Challenges to SV detection

Despite changing more nucleotides per genome compared to SNVs and small indels (Sudmant et al., 2015), little is known about the influence of SVs on genomic structure and function. This challenge is due to several contributing factors ranging from limited sample size and low resolution of routinely used detection technologies to using suboptimal models utilising DNA from cell lines in population studies (Perry et al., 2008; Sudmant et al., 2015). Also, there is no gold standard for algorithms used to identify structural variants, as no single algorithm can capture all types of SVs with high precision and sensitivity. Moreover, SVs are extremely common across the genome as making it difficult to develop filtering processes that can eliminate the vast majority for consideration whilst retaining genuine rare, high impact ones.

The limited sample size is being continuously addressed by significant population studies such as the 1000 Genomes Project and gnomAD collating exome and genome sequencing (WES and WGS) data from various large sequencing projects. These databases facilitate the calculation of allele frequency information, thus aiding in filtering and prioritising identified SVs in patients and inform the possibility of them being involved in disease development. The 1000 Genomes Project (with a sample size of 2,504 genomes at 7x coverage) indicated that there are about 3,441 SVs per genome (Sudmant et al., 2015), which was recently revised to 7,439 SVs per genome from studies performed using gnomAD WGS data (Collins et al., 2020). Furthermore, these projects have addressed the limitations associated with using cell lines and their impact on studying population genetics in relation to development of disease. Although cell line studies have been valuable for biomedical research, various analyses have revealed that the use of cell lines can lead to inaccuracies in estimating allele frequency at the population level, resulting in poor-quality data (Rodriguez et al., 2021).

A significant obstacle that has been a hindrance to SV identification has been the low resolution of detection methods such as chromosome G-banding, FISH, and array-CGH (Kannan & Zilfalil, 2009). While chromosomal G-banding allows for detecting chromosomal structural abnormalities

with a resolution of ~2-5 Mb, it very much depends on the chromosomal region and the degree of chromosome condensation (Drets & Shaw, 1971).

Soon after, FISH became a gold standard providing a massive clinical diagnostic utility as it gives a better resolution than G-banding. It allows for identifying hard-to-detect structural variants by G-banding and disease-relevant microdeletion/duplication, such as chromosome 22q11 deletion in DiGeorge syndrome or duplication of chromosome 17p11.2 in Potocki-Lupski syndrome. However, the technique depends on fluorescent probes, which restricts the number of probes used in any experiment due to the limited fluorescent filters in the microscope systems, increasing the technical and financial burden. Moreover, the potential crosstalk between adjacent probes leads to non-specific signals and a higher chance of false positive results (Arppe et al., 2017).

The third widely used technique is array-based technologies, which are broadly split into two categories: array-CGH and SNP arrays. The latter has the advantage of robustly identifying regions of homozygosity (ROH), enhanced capacity to identify various forms of mosaicism, including monosomies and trisomies, and uniparental disomy (Biesecker & Spinner, 2013). The obtained genotype data can also allow GWAS studies to be performed in parallel with CNV detection if run in suitably large cohorts. The increasing resolution for CNV detection has led to improved diagnostic rates across rare diseases and prompted the recommendation to switch from karyotyping to array-based methods as a first-line genetic test (Ahn et al., 2010). However, array-based methods are effectively blind to SVs that are copy-neutral. This subgroup includes balanced translocations, inversions, and more complex rearrangements; consequently, these rarer types of SV may go undetected.

With the continuous advancement of NGS, such as WES and WGS, the resolution obstacle has been overcome. However, new challenges have been introduced in the face of SV identification due to the lack of gold standard algorithms for calling all SV types with high precision and sensitivity. The commonly used tools for SVs and CNVs (and which were used in the 100,000 Genomes Project) are Manta and Canvas, which target medium- to large-sized CNVs, thus a higher chance of missing variants in the lower detection boundary closer to 50bp. While Canvas uses read coverage, restricting

its potential to detect only duplications and deletions, Manta uses paired-end and split read mapping information, allowing the detection of deletions, duplications, translocations, insertions, and inversions (Chen et al., 2015; Roller et al., 2016).

5.1.2 Structural variants in disease

The significance of large structural variants in neurological and neurodevelopmental disorders such as autism spectrum disorder (C. X. Yap et al., 2021), intellectual disability (Cooper et al., 2011), and epilepsy (Olson et al., 2014), has been appreciated for over four decades (Ledbetter et al., 1981) and extensive efforts have been focused on narrowing down the critical regions and genes involved in the phenotypes. The example of *UBE3A* in Angelman syndrome (MIM # 105830) has been a remarkable success in such efforts (Matsuura et al., 1997); however, the majority of these disorders were difficult to pin down to a single gene. Moreover, LCRs represent a major predisposing factor to nearly all recurrent SVs (both deletions and duplications). These LCRs often contain multiple genes and it can be challenging to disentangle the contributions of the individual genes within the LCR, unless heterozygous LoF of one of the genes yields the same/similar phenotype to the recurrent rearrangement. To this date, structural variants remain a challenging group of variants and are usually missed by WGS analytical pipelines. In a recent study, *de novo* intragenic SVs contributed 35% of pathogenic variants identified in the neurological and neuromuscular disorders investigated (Truty et al., 2019). Moreover, specific types of SVs were found to be associated with a few neurodegenerative diseases. An example of such SVs is the deletion of T cell receptor alpha and gamma loci regions linked to multiple sclerosis and duplications/triplications of *SNCA* in Parkinson's disease (Sato et al., 2015).

Translocations have been predominantly associated with cancer, as they are frequently observed structural variations in somatic cell cancer (Li et al., 2020). However, they have also been linked to genetic diseases (Sudmant et al., 2015). The identification and characterisation of translocations through techniques like karyotyping and fine-mapping have proven to be effective in uncovering new associations between disease and genes. Several notable examples demonstrate the impact of translocation studies in discovering disease-causing genes. In a study conducted in 1991,

translocations with breakpoints in 7p13 were found in three families with Greig cephalopolysyndactyly syndrome (MIM #175700), leading to the identification of *GLI3* as the critical gene (Vortkamp et al., 1991). Similarly, a study conducted over two decades ago focused on two individuals with severe orofacial dyspraxia (MIM #602081) and identified *de novo* translocations involving 7q31, which led to the discovery of *FOXP2* as the underlying gene (Lai et al., 2000). These examples underscore the significance of translocations as a type of structural variation in disease development and highlight their role in advancing our understanding of genetic disease mechanisms.

Many of the frequently reported translocation breakpoints in the literature are usually found in regions that contain repetitive genomic sequences such as palindromic AT-rich repeats posing a challenge to accurately determining the exact breakpoints. These repetitive sequences facilitate the formation of secondary structures such as hairpins and cruciforms leading to double strand breaks resulting in structural rearrangements and translocations (Kurahashi et al., 2000; Kurahashi et al., 2003; Kurahashi et al., 2006; Sheridan et al., 2010; Kato et al., 2012). On the other hand, translocation breakpoints that lie within non-repetitive genomic loci can be detected by massively parallel sequencing methodologies. This has been proven true by exome sequencing capturing a significant subset of cryptic SVs (Gardner et al., 2021), however, the gradual transition to WGS should mean that SV breakpoints can be captured more robustly in future.

In many instances, individuals who carry a balanced translocation, which has an occurrence of one in 600 individuals (Van-Dyke et al., 1983), exhibit no apparent health issues. As a result, this rearrangement often goes unnoticed unless the carriers encounter fertility problems or experience recurrent miscarriages (Warburton, 1991). Moreover, if a balanced translocation is identified in cytogenetic bands that do not directly correspond to the patient's observed phenotype or symptoms, it is often noted as an abnormality but not extensively investigated further. In such cases, the translocation may be considered an incidental finding without a direct connection to the patient's specific condition. However, it is important to note that each clinical case is unique, and the decision on whether to explore the translocation further depends on the specific circumstances, the patient's

medical history, and the judgment of the healthcare provider involved in the case. Moreover, it should be noted that different investigators may employ different methods to characterise translocation breakpoints, and there is currently no agreement on the optimal approach. This lack of consensus highlights the need for a systematic analysis of the utility of short-read sequencing data in characterising translocation breakpoints. Such an analysis would greatly benefit projects like the 100KGP, a national programme aiming to provide diagnoses to patients and integrate WGS into the NHS to reach its full potential (Turnbull et al., 2018).

In this study, I have reviewed WGS data from 71,408 participants recruited to the rare disease arm of the 100KGP, including a cohort of 22 families recruited to the 100KGP under "Balanced translocations with an unusual phenotype". The aim is to:

1. Increase the diagnostic yield in epilepsy, using a gene-oriented approach to identify complex structural variants.
2. Assess the effectiveness of 150 bp paired-read genome sequencing in characterising the precise breakpoint junctions.
3. Address the question of how common it is that translocations which appear balanced under the microscope are part of more complex rearrangements.
4. Identify genes disrupted by this set of rearrangements and possible novel disease-gene candidates.

5.2: Method

5.2.1 CNV/SV pipeline of the 100KGP

SVs were called by the GEL CNV pipeline using Canvas and Manta algorithms and merged into single 'SV.vcf' files; however, only Canvas calls were considered in the reporting process. This process was all performed by the 100KGP pipeline.

5.2.2 Structural variant analysis using SVRare

For structural variant analysis, I reviewed SVs using annotated per-gene/ per-family reports generated by the SVRare pipeline, edition 17th November 2021 (Yu et al., 2021). SVRare is a MySQL database with an aggregate of 554,060,126 SVs identified in 71,408 participants from 33,924 families in the rare disease domain of 100KGP. The SVs in the reports include deletions, duplications, and inversions called by Canvas and Manta software. The allele frequency for rarity was determined by considering SVs with a similarity score >80% as being the same event, which was calculated using the fraction of the overlap (intersection) over the total length (union). This clustering allows for establishing allele frequency for SVs, enabling prioritisation of rare SVs by filtering out variants with a frequency >1% of participants in GEL. Another feature that aids the analysis of gene reports is the annotation of SVs with familial segregation and prediction of protein-coding disruption (Yu et al., 2021). The annotated SVs per gene reports are available to GeCIP members in the "re_gecip/shared_allGeCIPs/JingYu-SVquery" directory within the GEL RE.

5.2.2.1 Filtering structural variants for gene-oriented analysis

The gene-oriented analysis focused on 71 genes (Table 5.1), which are listed as green genes in PanelApp for the genetic epilepsy panel as well as have high supporting evidence for haploinsufficiency (intolerance to loss of DNA dosage) or triplosensitivity (intolerance to gain of DNA dosage) (Collins et al., 2022). The haploinsufficiency disease gene list with scores indicating sufficient to high evidence supporting the haploinsufficiency (HI score ≥ 3) was downloaded from the Clinical Genome Resource (17th October 2022). Haploinsufficiency indicators include pLI and

loss-of-function observed/expected upper bound fraction (LOEUF) metrics (Lek et al., 2016; Karczewski et al., 2020). They measure a gene's tolerance to loss-of-function based on the frequency of protein truncating variants, such as frameshift, splice donor, splice acceptor, and stop-gain variants, recorded in control databases. However, pLI score is influenced by the size of the gene and the depth of sequencing (Ziegler et al., 2019).

Table 5.1: Expert curated and approved autosomal dominant genes linked to genetic epilepsy syndromes. A list of 71 green genes in Genetic epilepsy syndrome panel v3.0 that are also have sufficient to high supporting evidence for Haploinsufficiency or Triplosensitivity. The gene panel and ClinGen list were downloaded on 17th of October 2022

Gene symbol	Approved name	HGNC ID	Location	pLi	HI Score	pTriplo
ANKRD11	ankyrin repeat domain containing 11	HGNC:21316	16q24.3	1.00	3.00	0.23
ARHGEF9	Cdc42 guanine nucleotide exchange factor 9	HGNC:14561	Xq11.1	1.00	3.00	-
ARID1B	AT-rich interaction domain 1B	HGNC:18040	6q25.3	1.00	3.00	0.98
ARX	aristaless related homeobox	HGNC:18060	Xp21.3	0.91	3.00	-
ATP7A	ATPase copper transporting alpha	HGNC:869	Xq21.1	1.00	3.00	-
ATRX	ATRX chromatin remodeler	HGNC:886	Xq21.1	1.00	3.00	-
CACNA1A	calcium voltage-gated channel subunit alpha1 A	HGNC:1388	19p13.13	1.00	3.00	1
CASK	calcium/calmodulin dependent serine protein kinase	HGNC:1497	Xp11.4	1.00	3.00	-
CDKL5	cyclin dependent kinase like 5	HGNC:11411	Xp22.13	1.00	3.00	-
CHD2	chromodomain helicase DNA binding protein 2	HGNC:1917	15q26.1	1.00	3.00	0.99
CLTC	clathrin heavy chain	HGNC:2092	17q23.1	1.00	3.00	1
CNKSR2	connector enhancer of kinase suppressor of Ras 2	HGNC:19701	Xp22.12	1.00	3.00	-
CREBBP	CREB binding protein	HGNC:2348	16p13.3	1.00	3.00	1
CUL4B	cullin 4B	HGNC:2555	Xq24	1.00	3.00	-
DCX	doublecortin	HGNC:2714	Xq23	0.30	3.00	-
DDX3X	DEAD-box helicase 3 X-linked	HGNC:2745	Xp11.4	1.00	3.00	-
DLL1	delta like canonical Notch ligand 1	HGNC:2908	6q27	1.00	3.00	0.52
DYRK1A	dual specificity tyrosine phosphorylation regulated kinase 1A	HGNC:3091	21q22.13	1.00	3.00	0.98
EFTUD2	elongation factor Tu GTP binding domain containing 2	HGNC:30858	17q21.31	1.00	3.00	1
EHMT1	euchromatic histone lysine methyltransferase 1	HGNC:24650	9q34.3	1.00	3.00	1
FLNA	filamin A	HGNC:3754	Xq28	1.00	3.00	-
FOXG1	forkhead box G1	HGNC:3811	14q12	0.94	3.00	0.97
GRIN2A	glutamate ionotropic receptor NMDA type subunit 2A	HGNC:4585	16p13.2	1.00	3.00	1
GRIN2B	glutamate ionotropic receptor NMDA type subunit 2B	HGNC:4586	12p13.1	1.00	3.00	1
IKBKG	inhibitor of nuclear factor kappa B kinase regulatory subunit gamma	HGNC:5961	Xq28	0.54	3.00	-
IQSEC2	IQ motif and Sec7 domain ArfGEF 2	HGNC:29059	Xp11.22	0.99	3.00	-
KCNQ2	potassium voltage-gated channel subfamily Q member 2	HGNC:6296	20q13.33	1.00	3.00	0.83
KMT2E	lysine methyltransferase 2E (inactive)	HGNC:18541	7q22.3	1.00	3.00	0.99
MBD5	methyl-CpG binding domain protein 5	HGNC:20444	2q23.2	1.00	3.00	0.26
MECP2	methyl-CpG binding protein 2	HGNC:6990	Xq28	0.89	3.00	-
MEF2C	myocyte enhancer factor 2C	HGNC:6996	5q14.3	0.97	3.00	1
NBEA	neurobeachin	HGNC:7648	13q13.3	1.00	3.00	1
NEXMIF	neurite extension and migration factor	HGNC:29433	Xq13.3	1.00	3.00	-
NR4A2	nuclear receptor subfamily 4 group A member 2	HGNC:7981	2q22-q23	1.00	3.00	0.88
NRXN1	neurexin 1	HGNC:8008	2p16.3	1.00	3.00	0.54
NSD1	nuclear receptor binding SET domain protein 1	HGNC:14234	5q35.3	1.00	3.00	1
NSDHL	NAD(P) dependent steroid dehydrogenase-like	HGNC:13398	Xq28	0.96	3.00	-
NUS1	NUS1 dehydrodolichyl diphosphate synthase subunit	HGNC:21042	6q22.1	0.98	3.00	0.74
OPHN1	oligophrenin 1	HGNC:8148	Xq12	1.00	3.00	-
PAFAH1B1	platelet activating factor acetylhydrolase 1b regulatory subunit 1	HGNC:8574	17p13.3	1.00	3.00	1
PCDH19	protocadherin 19	HGNC:14270	Xq22.1	1.00	3.00	-
PDHA1	pyruvate dehydrogenase E1 subunit alpha 1	HGNC:8806	Xp22.12	1.00	3.00	-
PRRT2	proline rich transmembrane protein 2	HGNC:30500	16p11.2	0.59	3.00	0.39
PTEN	phosphatase and tensin homolog	HGNC:9588	10q23.31	0.26	3.00	0.99
PURA	purine rich element binding protein A	HGNC:9701	5q31.3	0.94	3.00	1
SCN1A	sodium voltage-gated channel alpha subunit 1	HGNC:10585	2q24.3	1.00	3.00	0.95
SCN2A	sodium voltage-gated channel alpha subunit 2	HGNC:10588	2q24.3	1.00	3.00	1
SETBP1	SET binding protein 1	HGNC:15573	18q12.3	0.15	3.00	0.99
SETD1A	SET domain containing 1A, histone lysine methyltransferase	HGNC:29010	16p11.2	1.00	3.00	1
SETD5	SET domain containing 5	HGNC:25566	3p25.3	1.00	3.00	0.96
SLC16A2	solute carrier family 16 member 2	HGNC:10923	Xq13.2	0.99	3.00	-
SLC2A1	solute carrier family 2 member 1	HGNC:11005	1p34.2	0.99	3.00	-
SLC35A2	solute carrier family 35 member A2	HGNC:11022	Xp11.23	0.72	3.00	-
SLC6A1	solute carrier family 6 member 1	HGNC:11042	3p25.3	1.00	3.00	0.85
SLC6A8	solute carrier family 6 member 8	HGNC:11055	Xq28	1.00	3.00	-
SLC9A6	solute carrier family 9 member A6	HGNC:11079	Xq26.3	1.00	3.00	-
SMC1A	structural maintenance of chromosomes 1A	HGNC:11111	Xp11.22	1.00	3.00	-
SMS	spermine synthase	HGNC:11123	Xp22.11	0.97	3.00	-
STXBP1	syntaxin binding protein 1	HGNC:11444	9q34.11	1.00	3.00	0.98
SYN1	synapsin I	HGNC:11494	Xp11.3-p11.23	0.99	3.00	-
SYNGAP1	synaptic Ras GTPase activating protein 1	HGNC:11497	6p21.32	1.00	3.00	0.98
TBL1XR1	TBL1X/Y related 1	HGNC:29529	3q26.32	1.00	3.00	0.7
TCF4	transcription factor 4	HGNC:11634	18q21.2	1.00	3.00	0.99
TSC1	TSC complex subunit 1	HGNC:12362	9q34	1.00	3.00	0.93
TSC2	TSC complex subunit 2	HGNC:12363	16p13.3	1.00	3.00	0.43
UBE2A	ubiquitin conjugating enzyme E2 A	HGNC:12472	Xq24	0.81	3.00	-
UBE3A	ubiquitin protein ligase E3A	HGNC:12496	15q11.2	1.00	3.00	0.93
WDR45	WD repeat domain 45	HGNC:28912	Xp11.23	0.99	3.00	-
ZBTB18	zinc finger and BTB domain containing 18	HGNC:13030	1q44	1.00	3.00	0.92
ZDHHC9	zinc finger DHHC-type palmitoyltransferase 9	HGNC:18475	Xq26.1	0.74	3.00	-
ZEB2	zinc finger E-box binding homeobox 2	HGNC:14881	2q22.3	1.00	3.00	0.96

The individual SVRare gene reports were inspected and filtered for rare variants with Allele frequency (AF) ≤ 8 participants, which was determined based on family structure in GEL (Table 5.2) to ensure no qualifying SVs shared by multiple affected individuals were missed. The gene reports have extensive annotations including the genes disrupted by the SV breakpoints or included within the SV, which was used for filtering for SVs that disrupt the coding region. This filtering criteria was removed for analysis of gene reports where disruption of intergenic regulatory elements and 3D chromatin conformation is a known mechanism. These include *ARX, FOXG1, MEF2C*, and *SLC2A1* (D'haene & Vergult, 2021). This filtering yielded an average of 22 rare variants per green-rated epilepsy gene, ranging in size from 59 bp to 155 Mb. Each of these structural variants were manually reviewed by inspecting the read alignments on IGV (v2.11.9) to determine whether it is genuine, and how it is likely to disrupt the gene The visibility range threshold setting was increased to 100 kb to allow a larger window to view the alignments. Moreover, the "SV.vcf" file was loaded into IGV with the feature "visibility window size" set to 0 to view the called SVs by Canvas and Manta across the loaded chromosome. The inheritance of the SV was assessed where parents' data were available, and the phenotypic information of the participant were reviewed to check if it is consistent with the known phenotypes for the disrupted gene in OMIM/literature. Finally, the "ExitQ" forms and submitted diagnostic discovery table to check if the case was solved.

The verified variants, which were found to truly disrupt the coding region of the gene or a known regulatory element in the four specified genes, were then submitted for review by the clinical team and recruiting clinician. Once the clinical team has approved it, the variant is Sanger validated in the clinical laboratory for reporting to the family, or in the research setting and reported as research finding (Figure 5.2: A - SV analysis approaches).

5.2.2.2 Filtering structural variants in the karyotypic abnormality cohort

The analysis described here is of 22 participants in GEL, all with known karyotypic abnormalities: 19/22 families were recruited to GEL under "Balanced translocation with unusual phenotype", and the remaining three families were recruited under the normalised disease term "Intellectual disability" and included in the cohort due to direct contact of the recruiting clinician informing the presence of a karyotypic abnormality in their patient.

While the karyotype abnormality was noted in all families, only 2/22 had an entry in the 100KGP database with the karyotype results. Therefore, forms for contacting the clinicians were submitted to obtain the missing information. Karyotype results were received for 15/20 cases; therefore, one of two processes were used to identify Breakend (BND) variants for the cases based on the knowledge of the previous karyotype results. The coordinates for BNDs were extracted from the "SV.vcf" file of each patient. The total number ranged from 620 to 3060 BND calls per genome, depending on the build used for mapping. The samples mapped to GRCh37 had a lower number of BNDs, between 620-782 per genome, than those mapped to GRCh38, with 2494-3060 calls per genome. Two inclusion criteria were required for BNDs to be considered true positives and help reduce the number of false positives. These include identifying two adjacent BND calls less than 1 Mb apart in the same chromosome; and reciprocal calls between any two chromosomes involved in the translocation. However, if no adjacent BNDs were identified, then SVRare report for the family was used to determine if any large SVs were present. That could explain the two BNDs being further apart than 1 Mb.

Table 5.2: Family structure and participant counts within families in GEL

Family size	Count
1	12966
2	6576
3	12633
4	1445
5	234
6	52
7	11
8	5
9	0
10	0
11	1

For participants where the karyotype results were known, the BND extraction was done for the chromosomes involved in the rearrangement, which yielded between 2-8 positions to check visually in IGV. However, where the clinician did not disclose the karyotype results, the SVRare report of the family was reviewed first for rare structural variants. If a rare event was identified, it was inspected visually for nearby BNDs, and usually appeared to be part of a more complex translocation event. If no rare SVs were identified from the family SVRare report, an iteration of chromosomal combination analysis using custom scripts was performed to identify chromosome pairs involved in the translocation (Figure 5.2: B – SV analysis approaches). DNA samples were obtained from four participants for validation of identified SVs breakpoints using PCR (Table 2.2- validation primers #5-8). Subsequently, the PCR product was purified (Method section 2.6.1) and sent to GENEWIZ for Sanger Sequencing. For complex or challenging to validate SVs, fresh samples were collected from the participants in the outpatient clinic and sent for long-read sequencing using either PacBio or Oxford Nanopore Technology as indicated in the specific sections. Fresh whole blood samples were required for high molecular weight DNA extraction, which was performed by the laboratory conducting the sequencing.

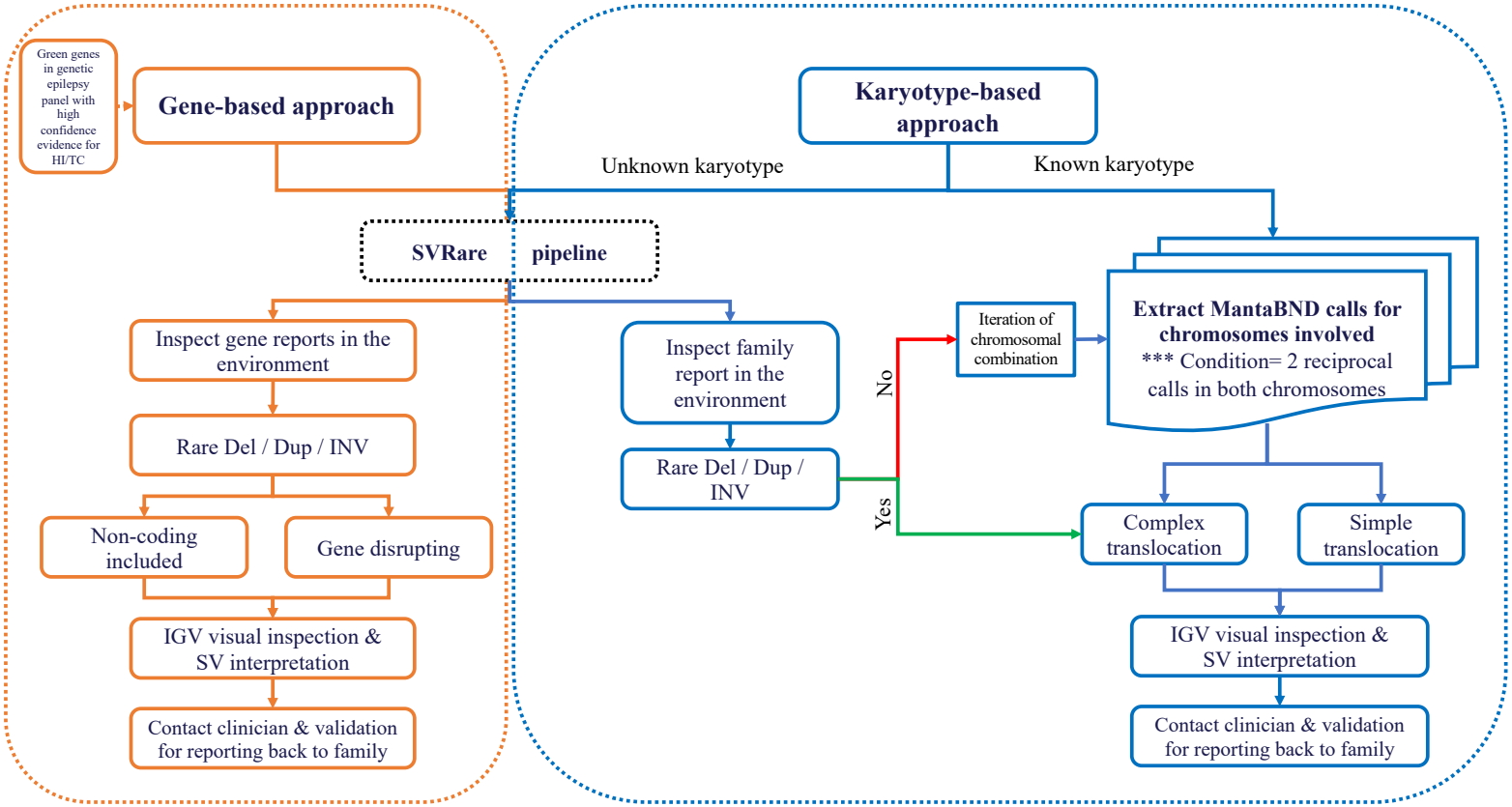


Figure 5.2: Identification of rare structural variants with SVRare pipeline. Dosage sensitive (Haploinsufficiency and Triplosensitivity) genes that are green in genetic epilepsy gene panel were selected for A) the gene-based approach (orange panel) providing diagnostic findings for unsolved cases in 100KGP that were missed using the GEL pipeline. B) The karyotype-based approach (blue panel) was used for a cohort with known karyotypic abnormality. The custom scripts utilised bcftools query and linux commands utilising the various INFO tags to extract reciprocal BNDs between the two reported chromosomes or for iteration of chromosomal combinations.

5.4 Results

5.4.1 Discovery of complex structural variants using a gene-oriented approach rapidly increases the diagnostic yield of WGS

The analysis considered SVs from gene reports generated by SVRare from 71,408 participants from 33,923 families recruited to the rare disease arm of the 100KGP (main programme release_v15_2022-05-26). The participants in the rare-disease domain were recruited under various normalised disease terms (221 terms), including neurological and neurodevelopmental disorders. Since only high-confidence haploinsufficient autosomal dominant genes in the genetic epilepsy panel were considered, rare SVs disrupting the protein-coding or known regulatory regions were considered pathogenic. Moreover, one would expect the participants with pathogenic SVs in these genes to have either neurodevelopmental disorders or have a blended phenotype with symptoms that can be partly explained by the presence of these SVs.

Variant filtering using AF ≤ 8 identified an average of 22 rare SVs per gene report, which were then screened for disruption of gene function or a known regulatory region. The alignments were then inspected in IGV, which confirmed a total of 56 SVs in 34/71 genes as truly disruptive and were subsequently reported as diagnostic variants for the phenotype of the probands via the diagnostic discovery pathway in the 100KGP RE (Figure 5.3).

Of the 56 variants, 11 were complex (19.6%); six were inversions (10.7%), 38 were deletions (67.9%), and one duplication (1.8%), all in unsolved participants with VUS or no proposed variants in "ExitQ" forms in GEL. These SVs were not detected by the currently used GEL CNV pipeline, which relies on Canvas calls to report diagnostic variants to families.

The 100KGP pipeline reported only deletions and duplications (read-depth based CNVs) above 10 kb (with PASS filter status from Canvas) to the genomic laboratory hubs. However, this was introduced part way through the analysis for the project, therefore, not all families had CNVs reported (personal communication- Dr. Susan Walker). This restriction to Canvas calls could explain the missed diagnoses involving simple inversions (6/56), which are copy-neutral. Moreover, the segment size for 6/56 SVs identified ranged between 59 – 675 bp, which is below the robust

detection limits for Canvas (>1 kb), hence was missed, an example is shown in section 5.4.2. For 12/56 SVs, the sizes ranged between 3.54- 9.2 kb, which is below the reporting threshold for the 100KGP pipeline. For the complex SVs involving deletions/duplications, Canvas calls were slightly over /under-estimated. This was mostly associated with segment complexity in which there are repetitive elements in the surrounding sequence, which could explain the occurrence of such SV but can also complicate CNV detection. In these instances, Manta calls reflected the true size of each separate event involved in the complex SV, while Canvas struggled to accurately determine the SV boundaries especially in regions of low complexity such as that observed in individual 5_5 discussed in section 5.4.1. Another potential reason was coverage variability, which is one of the primary factors that can affect CNV detection by Canvas.

The different SV types present various challenges for detection, interpretation, and validation leading to a reduced number of SVs being reported in large cohorts. In this chapter, I will delve into the complexities of detecting, characterising, and validating SVs in WGS data and explore some obstacles hindering the diagnostic yield of WGS analyses.

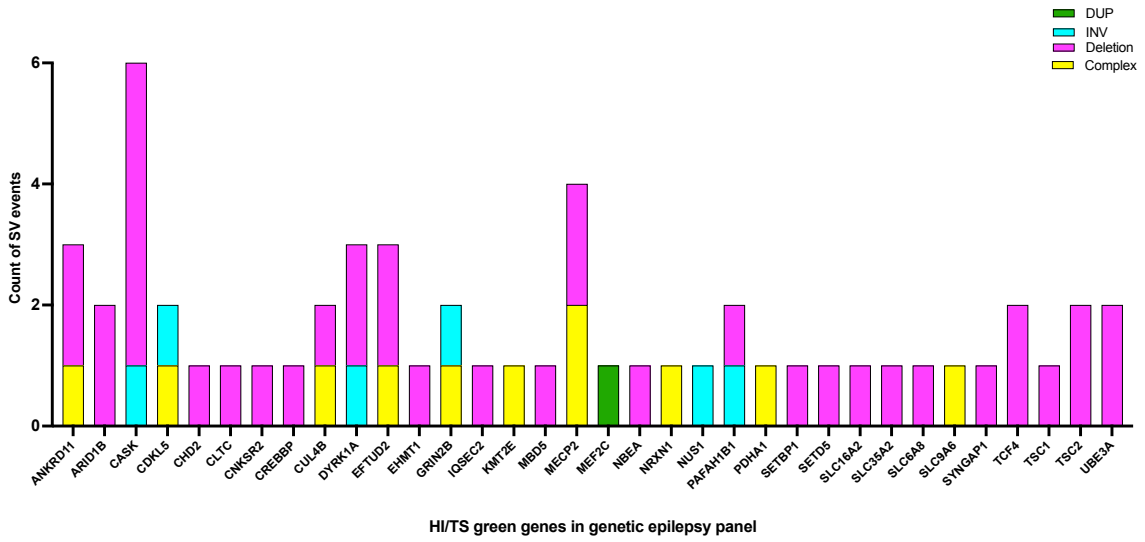


Figure 5.3: Overall summary of SV findings in Seventy-one dosage-sensitive Haploinsufficient and Triplosensitive genes that are green in genetic epilepsy gene panel. They were examined for rare structural variants including deletions, duplications, and inversions. Only genes where a rare event has been identified is shown in this graph. Eleven complex variants (yellow bars) were identified in ten neurodevelopmental genes from independent families.

5.4.2 Clinically relevant structural variants hiding in plain sight

While the majority of deletions identified in this study could have been reported by the GEL pipeline using Canvas, a small deletion of 59 bp in the first coding exon of AT-Rich Interaction Domain 1B (*ARID1B*; linked to Coffin-Siris syndrome1 #135900) was only called by Manta. The deletion is present in an individual (**I1_5** - Table S5.1) recruited under Intellectual disability with various neurodevelopmental HPO terms were recorded in the RE. These included delayed gross motor development, global developmental delay, coarse facial features, delayed speech and language development, generalised hypotonia, abnormality of the outer ear, delayed fine motor development, autistic behaviour, intellectual disability, macrocephaly.

The 59 bp deletion is predicted to cause a frameshift and an early termination (p.A547Gfs95*), potentially leading to haploinsufficiency. This SV is below the detection threshold of Canvas (accurate for SV size > 1 kb). Moreover, it was labelled as "BadReads" by Platypus as it falls within a low complexity region of GCCGCC repeats; therefore, the deletion was not included in the prioritised variant list for the tiering table. The genomic sequence context within this region would present a challenge to designing primers for the validation of the deletion using PCR. This is due to the presence of various repeats in the region, giving only a small window for placing primers that could yield unique PCR products, further confounded by the high GC content reaching >80%. Furthermore, MLPA would most likely miss the deletion, given that both left and right probe oligonucleotides are positioned upstream of the deletion (Figure 5.4), however, this was not tested. Nonetheless, the clinical team in GEL as well as the recruiting clinician reviewed and considered this variant as causative for the patient's phenotype, and it has been scheduled for validation by the clinical laboratory.

This example highlights the potential limitations of relying on the Canvas algorithm alone to report clinically significant variants especially those with sizes falling below the lower limits of detection. It also underscores the importance of further developing and enhancing existing algorithms specifically designed to detect and accurately report medium-sized variants ranging from > 50 bp to <1 kb.

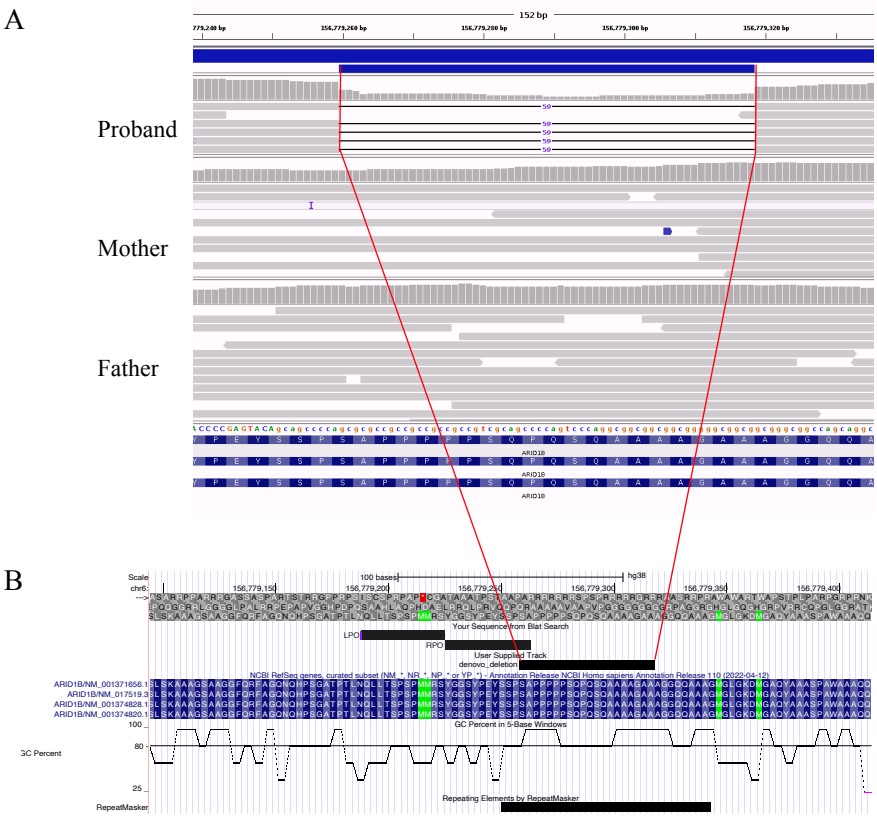


Figure 5.4: A missed de novo small deletion in ARID1B A. Read alignment showing a de novo deletion of 59bp on exon 1 of ARID1B absent from both parents. B. The 59bp deletion falls within repeat region that has GC content of >80% both making validation using PCR challenging to design unique primers. Also shows the position of left and right probe oligonucleotides (LPO, RPO) of the MRC Holland MPLA kits usually used by clinical genetics laboratories for detecction and validation of clinical findings.

The second category of missed variants that the current GEL pipeline would not pick up due to its reliance on Canvas for reporting is inversions, which are copy-neutral structural variants. In the 100KGP pilot study (Genomes-Project-Pilot-Investigators, 2022) involving 4660 participants, no inversions were reported; however, in the small number of genes investigated in this study, several simple inversions have been identified in *CASK, CDKL5, DYRK1A*, *GRIN2B*, *NUS1*, *PAFAH1B1* and *FOXG1*. Further ten inversions have been identified as part of more complex structural variants, which were reported as separated adjacent events called by Canvas and Manta.

5.4.3 Potential clinically relevant complex structural variants (cxSVs) in patients with neurodevelopmental disorder/ phenotype:

During the inspection of gene reports, I have identified 11 individuals from independent families with cxSVs disrupting one of ten genes (*ANKRD11, CDKL5, CUL4B, EFTUD2, GRIN2B, KMT2E, MECP2, NRXN1, PDHA1, SLC9A6*) that are approved for the Genetic Epilepsy gene panel and have high scores for being intolerant to loss of function (Table 5.1). The complexity of the variants varied from inversions-deletions/duplications to duplication-triplication-inversion-duplication. Five exemplar cases are highlighted in the following section to discuss the effect of the various complexities on detection and interpretation.

5.4.3.1 Exemplar cases of SVs of varying complexities:

Individual 2_5 (I2_5): A 16-year-old affected male was recruited under "intellectual disability" as a singleton case and parents were reported as healthy non-consanguineous. He presented with intellectual disability, delayed gross and fine motor development, autistic behaviour, delayed speech and language development, as well as aortic root dilatation. A rare deleterious missense (NM_138615.3:c.3287C>T; p.P1096L; CADD= 26.6) in DExH-Box Helicase 30 (*DHX30*) was reported as VUS in the ExitQ. *DHX30* is linked to Neurodevelopmental disorder with severe motor impairment and absent language (OMIM #617804). Nonetheless, to date, only *de novo* variants in *DHX30* within conserved functional motifs have been reported in the literature. In this case, the origin of this missense cannot be ascertained from WGS data within the RE due to lack of parental

WGS data and this determination was also not made by the clinical laboratory. A protein structure-based analysis on this SNV was performed using the MichelaNGLo-VENUS tool (https://michelanglo.sgc.ox.ac.uk/venus), which utilised the AlphaFold2 model based on Q7L2E3. It predicts p.P1096L to be structurally neutral with ΔΔG of 0.1 kcal/mol, suggesting that it does not affect the stability of the protein or disrupt a functional motif.

On the other hand, SV analysis identified a large deletion of ~86.14 kb (NC_000016.10(NM_013275.6):c.-59-13745_*23795del), which removes the entire coding sequence of Ankyrin Repeat Domain Containing 11 (*ANKRD11*) as well as ~23 kb downstream of the gene. It only spares part of the 5-prime Untranslated Region (5'UTR; Figure 5.5: A) and is further complexed by an inverted intragenic region of ~1.5 kb upstream of Zinc Finger Protein 778 (*ZNF778*), which is not known to be linked to any disease and does not have an entry in OMIM. Haploinsufficiency of *ANKRD11* has been previously associated with autosomal dominant KBG syndrome with a loss-of-function mechanism (OMIM #148050). A recent clinical study showed that some patients with KBG syndrome and heterozygous LoF variants in *ANKRD11* also exhibited cardiac features such as aortic root dilatation alongside the neurodevelopmental phenotype (Parenti et al., 2021). This finding has led to the revision of the molecular diagnosis of this patient to KBG syndrome based on the *ANKRD11* finding; however, we cannot discard the possible contribution of the "VUS" variant in *DHX30*. It is vital that studies report new cases where the syndrome involves multisystemic features such as this, to enable the delineation of genotype-phenotype correlation. This would lead to a better understanding of the contribution of each variant towards the observed heterogeneous phenotypes across different individuals with a similar diagnosis, yet slightly different clinical manifestations.

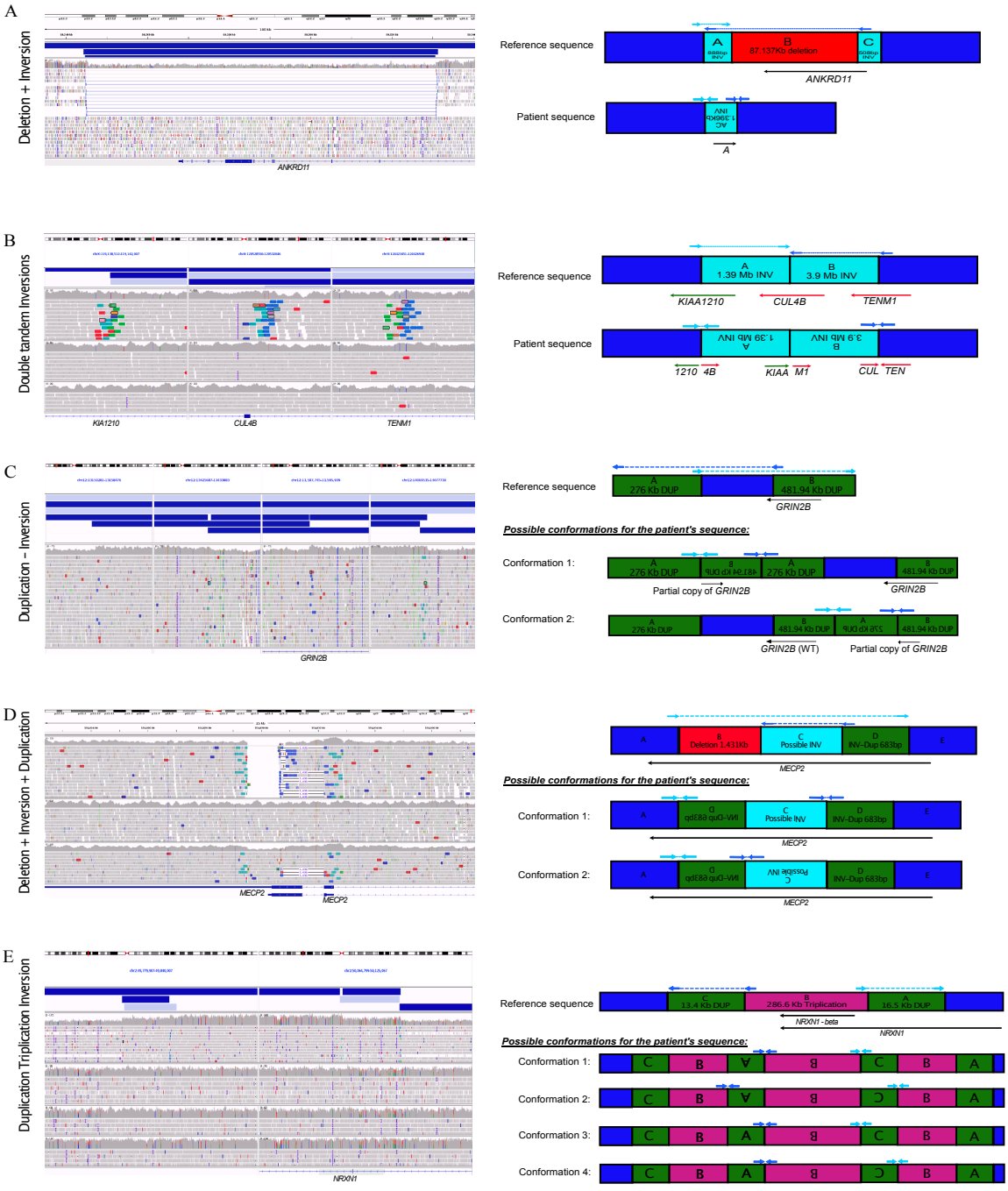


Figure 5.5: Representation of complex structural variant classes identified in individuals with neurodevelopmental disorders in the 100KGP. These variants are the results of rearrangements involving two or more types of SVs, which is shown as they appear in short read Illumina data viewed on IGV (left side) and in a linear schematic of each class of variant (right side). All identified variants have been assessed through the clinical team of 100KGP as well as by the recruiting clinician overseeing the patient. Validation of diagnostic SVs are carried out in clinical laboratories by dideoxy sequencing, MLPA, or a-CGH. A) SVs involving paired inversions flanking a deletion have been observed in two patients disrupting ANKRD11, or EFTUD. The example illustrated here is for the SV in ANKRD11 described in patient 2. B) An unusual double tandem inversion in one patient interrupting CUL4B. C) Non-tandem inverted duplication in one patient affecting GRIN2B. Another class is inverted tandem duplications (not shown here) in SLC9A6 in a male patient. D) Duplication with a deletion (including inverted duplications with deletion) has been detected in MECP2 in three patients, one of which is a translocated 13 kb inverted duplication from 19q13.43 inserted into MECP2 3'-UTR, another partial duplication of IRAK1into the C-terminal region of MECP2, and the third is the affected male exemplar case shown in D inherited from an unaffected mother. E) Duplication-triplication/inv-duplication has been observed in an individual affecting NRXN1 and PDHA1. All SVs are detailed in Appendix Table S5.1.

Individual 3_5 (I3_5): is a 16-year-old male recruited under "intellectual disability" as part of a child-parent trio with an affected father diagnosed with intellectual disability and a healthy mother. This suggests a dominant inheritance pattern of a possible variant being inherited from the father. However, the proband has two *de novo* adjacent inversions (NC_000023.11(NM_001394962.1):c.-1402895_-12407inv and NC_000023.11(NM_001079872.2):c.-3864488_2440-356inv) that disrupt Cullin 4B (*CUL4B*), Teneurin Transmembrane Protein 1 (*TENM1*), and *KIAA1210* at the breakpoints (Figure 5.5: B). While both *CUL4B* and *TENM1* have a high pLI score of 1, only *CUL4B* has been previously linked to disease, X-linked syndromic intellectual developmental disorder, the Cabezas type (OMIM #300354). This unusual inversion type could explain the phenotype present in the proband, including global developmental delay, motor delay, seizures, delayed speech and language development, and intellectual disability. The father has a slightly different presentation to the proband including macrocephaly, hyperpigmented nevi, and autistic behaviour in addition to the HPOs present in the proband.

Closer inspection of the genomic sequence reveals the presence of multiple repeat regions surrounding the breakpoints. One of the breakpoints disrupts *TENM1* and lies within a Long-Interspersed Element 1 (LINE-1/L1) retrotransposon sequence with a minus strand orientation. The second breakpoint, which affects *CUL4B*, is not directly within a repeat region but is flanked by Short-Interspersed Element (SINE) sequences, including ALU elements, and a LINE-1/L1 element (plus strand orientation) located 804 bp away from the breakpoint. This is similar to the genomic sequence around the third breakpoint, impacting *KIAA1210*, where a LINE-1/L1 element is situated 394 bp away with a plus strand orientation. These elements can act as regions of homology between non-allelic but homologous sequences, resulting in structural variation events through non-allelic homologous recombination (NAHR). The opposing orientation of these elements at the breakpoints increases the likelihood of inversions over deletions and duplications.

The double inversions add another complexity layer towards interpretation and resolving the correct orientation of the variant, which would present a difficulty when trying to validate. This is an example of the effect of copy-neutral structural variants, which detection algorithms such as Canvas will inevitably miss.

Individual 4_5 (I4_5): is a 13-year-old affected male proband recruited to the 100KGP as a singleton case at the age of 9 under a diagnosis of intellectual disability. The proband had the following reported HPO terms: delayed fine motor development, autistic behaviour, delayed gross motor development, global developmental delay, attention deficit hyperactivity disorder, delayed speech and language development, inability to walk, small for gestational age, and intellectual disability.

The SNV analysis by the GEL pipeline returned negative results, with no variants identified as causative in the intellectual disability panel (version 2.787). Similarly, the CNV analysis using the GEL pipeline failed to identify any potential causative variants.

The SV analysis using SVRare identified two inverted non-tandem duplications: A) a smaller duplication (NC_000012.12(NM_001423.3):c.-42779_*215391invdup) of 276 kb that encompasses Epithelial Membrane Protein 1 (*EMP1*) and Long Intergenic Non-Protein Coding RNA 1559 (*LINC01559*) and B) a larger segment (NC_000012.12(NM_000834.5):c.-93001_2010+17237invdup) of 481.94 kb, which duplicates exons 1-9/13 of Glutamate Ionotropic Receptor NMDA Type Subunit 2B (*GRIN2B*). Of the three genes, only *GRIN2B* is linked to disease, autosomal dominant Intellectual developmental disorder 6 with or without seizures (OMIM #613970) and is highly sensitive to both haploinsufficiency and an increase in dosage (Karczewski, et al., 2020).

This example is of particular interest as short-read alignments were suitable for identifying the breakpoints; however, two conformations can explain the split-reads, affecting the clinical relevance and interpretation of this variant (Figure 5.5: C).

Conformation 1 could potentially suggest that the larger segment (B), containing a partial copy of *GRIN2B*, has undergone inversion and subsequent insertion between the duplicated segment (A). In this scenario, clinical relevance might not be anticipated, as the partial copy could likely undergo nonsense-mediated decay. As a result, the configuration would suggest the presence of two intact *GRIN2B* copies, potentially negating any occurrences of haploinsufficiency or triplosensitivity. Consequently, this SV might be interpreted as benign. Conversely, conformation 2 suggests that the small, duplicated segment (A) has undergone inversion and subsequent integration within the larger duplicated segment (B) ~93 kb upstream of *GRIN2B*.

All breakpoints lie within or near to a LINE/L2 element either on the plus or minus orientation. Both breakpoints for the smaller duplication (GRCh38(chr12):g.13154136_13430082) lie within LINE/L2 elements in opposing orientation, which could lead to an inverted duplication through NAHR. On the other hand, one of the breakpoints for the larger duplication (GRCh38(chr12):g.13591366) is 392 bp from LINE/L2 element (minus orientation) and the second breakpoint (GRCh38(chr12):g.14073895) within the LINE/L2 (minus orientation), which would favour the occurrence of duplication through NAHR over an inversion event.

While configuration 2 does not disrupt the gene directly, it could be similar to a previously reported deletion that was considered as likely pathogenic variant in dbVar (GRCh38(chr12):g.14116150_14703048del; nssv3442642). The reported deletion is ~135 kb upstream of *GRIN2B* in a patient with autistic behaviour. It encompasses three genes (*ATF7IP*: activating transcription factor 7 interacting protein; *GUCY2C*: guanylate cyclase 2C; and *PLBD1*: phospholipase B domain containing 1), of which only *GUCY2C* is linked to disease; Diarrhoea 6 (OMIM # 601330) and none would explain the patient's phenotype. The closest disease gene with matching phenotype is *GRIN2B* and the deletion could be removing essential long-range regulatory elements.

Considering the second alternative configuration in our patient, the disruption of long-range upstream regulatory elements could be a notable implication.

GRIN2B expression is tightly controlled by various long-range looping extending over 0.5 Mb, facilitating the interactions between Activator Protein-1 (AP-1) and Nuclear Respiratory Factor 1 (NRF-1) transcription factors and TALE-based transcriptional activators distal to *GRIN2B* with proximal sequences (Bharadwaj et al., 2014).

For this individual, resolving the correct conformation would affect the clinical interpretation, however, the clinical team in GEL and the recruiting clinician, reviewed the SV as a potential candidate. Disruption to *GRIN2B* expression could explain the phenotype of the proband, which align with previously reported cases of *GRIN2B* disruption (Freunscht et al., 2013; Wolfe et al., 2017)

The sample from this individual has been prioritised for long-read sequencing, but no data were available for review at the time of submitting this thesis. Long-read sequencing would provide a

crucial orthogonal technique for resolving the conformation as it can provide longer sequence reads compared to the 150 bp short-read sequencing. This would enable the mapping of reads across repetitive and complex genomic regions, such as the breakpoints of this SV implicating *GRIN2B* in this patient. By generating longer reads that span the breakpoints, it can help definitively determine the accurate conformation of the SV and characterise its exact structure with a clearer view of the genomic architecture surrounding the SV breakpoints.

The prevalence of *GRIN2B*-related neurodevelopmental disorder is unknown and fewer than 100 cases have been reported in the literature, with individuals heterozygous for pathogenic missense variant tending to exhibit more severe phenotypes than null variants (Platzer & Lemke, 2021). Therefore, it is important to validate and report variants affecting *GRIN2B* to allow for better characterisation of the phenotype-genotype correlation.

Individual 5_5 (I5_5): is a 10-year-old male recruited to the 100KGP under the "intellectual disability" diagnosis category as part of a duo with an unaffected mother. The proband presents with microcephaly, delayed speech and language development, clinodactyly of the fourth and fifth fingers, bruxism, delayed fine motor development, global developmental delay, intellectual disability, developmental regression, feeding difficulties, failure to thrive and inability to walk. The GEL SNV/CNV analyses for the proband were negative, but the SVRare analysis identified a maternally inherited inversion (NC_000023.11(NM_004992.4):c.969_1139inv) in methyl-CpG-binding protein-2 (*MECP2*), which is linked to Rett syndrome (OMIM #312750). This is an x-linked dominant disorder that occurs almost exclusively in females and leads to embryonic lethality in hemizygous males. However, a rescue of the premature lethality occurs in those with somatic mosaicism or who have an extra copy of the X-chromosome (Klinefelter's), but they present with atypical Rett syndrome expressing a spectrum of different phenotypes to a classic Rett syndrome (Moog et al., 2003).

Inspection of the read alignment of I5_5 revealed a more complex structure to the SV (inversion-duplication-deletion; Figure 5.5: D). There is ~1.34 kb deletion with a 140 bp retained segment within the deletion and the SV is further complexed by a 681 bp duplication (NM_004992.4:c.27-332_376dup). The deletion was called as two deletions by Manta (1341 bp, and 180 bp), which possibly was due to the presence of the retained 140 bp segment challenging the algorithm. This

complicated precise identification of the 5' boundary of the deletion, which was ambiguous with two possible end positions either at NM_004992.4:p.L323 or NM_004992.4:p.L383 (Figure 5.6: A). Considering that the heterozygous carrier mother is unaffected and knowledge about heterozygous variants in MeCP2 affecting females and lethal to karyotypically normal males, the cxSV might seem irrelevant to the phenotype in this patient at a first glance. However, it clearly disrupts *MECP2* and the survival of the male proband could potentially be attributed to one of the two possible conformations that affect the gene's expression to varying degrees. Conformation 1 predicts disruption in the last coding exon (4/4) at NM_004992.4:p.L323 or NM_004992.4:p.L383, which would affect the last ~25% of the coding region as well as the 3' UTR. This conformation would possibly preserve some expression and function, which would explain the mother being unaffected and an affected male. Conformation 2 on the other hand, predicts that the disruption is at the end of exon 3 after NM_004992.4:p.I125, resulting in early termination and a possible complete loss of gene expression, which is expected to be lethal in males (Figure 5.6: B).

MECP2 encodes for a multifunctional protein that has the ability to repress transcription of genes through the binding to the methylated regions of the genome, it was also found to mediate some of its effect through the activation of some genes and binding to unmethylated DNA (Yasui et al., 2007; Chahrour et al., 2008). Forty-six *MECP2* mutations (missense either germline or somatic mosaic, in-frame deletions, frameshift, stop gain, and large duplications) have been reported in a study screening 2697 males with neurological phenotypes. The mutations were either familial in multiple affected males inherited from an affected mother or unaffected with skewed X-chromosome inactivation or were in sporadic cases arising *de novo* (Villard, 2007). The male patients exhibited varying severity of symptoms that are linked to classical or atypical Rett syndrome, neonatal encephalopathy, or progressive encephalopathy (Neul et al., 2019).

A closer inspection of the protein-truncating mutational spectrum in males, the earliest truncating mutation in *MECP2* in males is NM_004992.4:p.G163fs* in a male with severe neonatal encephalopathy (Geerdink et al., 2002). This amino acid position is immediately outside of the Methyl-CpG Binding Domain (MBD), which requires the minimal stretch of amino acids between 78 and 162 (NM_004992.4) to enable its DNA binding to the methylated CpGs (Nan et al., 1993). Moreover, it has been shown that an intact MBD is required for facilitating the inter-domain

interactions in a functional *MECP2* such as the N-terminal domain, the intervening domain, and the C-terminal domain alpha and beta (Ghosh et al., 2008). Therefore, truncating mutations that lie outside of the minimal sequence required for *MECP2* binding to the methylated CpGs may preserve some functional aspects. In such instances, male embryos may be rescued, but exhibit severe phenotypes.

Therefore, in this individual (I5_5), conformation 1 would explain the survival of a hemizygous male postnatally and potentially be consistent with the inheritance pattern from an unaffected or mildly affected mother. On the other hand, conformation 2 would not explain the survival of a hemizygous male or the inheritance of cxSV from an unaffected mother. Based on this finding, the clinician performed further clinical examination of the mother, which showed no learning difficulties. Therefore, resolving the conformation of this SV would be essential for clinical relevance and interpretation, hence DNA sample was obtained.

A few challenging aspects of designing primers for validating this cxSV were faced. These were the high GC content and ambiguity of the exact start position of segment C from short-read alignments (Figure 5.6: A). PCR and Sanger sequencing validation of BNDs (1-4) was performed using various combinations of primers, as shown in (Figure 5.6: B). Only primers listed in Table 2.2 yielded a PCR product, and the Sanger sequence trace confirmed the coordinates being consistent with WGS and the inversion of the duplicated segment D (NC_000023.11(NM_004992.4):c.*859 on the +ve strand, and NC_000023.11(NM_004992.4):c.27-333 on the -ve strand). However, the PCRs using primers spanning the other BNDs (2-4) failed to produce a product.

Short-read data hold great potential for detecting breakpoints and uncovering the complexities associated with SVs. However, such data fail to accurately decipher and provide conclusive results for regions of low complexity in the genome. In such cases, long-read sequencing maybe a better choice for determining the structure as it is not affected by these regions (Logsdon et al., 2020).

In this case of complex rearrangement affecting *MECP2*, PacBio long-read sequencing was performed as part of the Oxford-Birmingham-Exeter long-read sequencing collaboration and the data were deposited into the GEL RE. The PacBio data indeed confirmed that P4 has conformation 1 (Figure 5.7), highlighting the superior performance of long-read sequencing compared to short-read sequencing in this case. This demonstrates the long-read sequencing ability to accurately

characterise large-sized structural variants, which significantly impacts clinical interpretation and has the potential to enhance the diagnostic yield. Long-read sequencing proves to be superior to short-read sequencing in cases where the phenotype suggests specific disease genes or regions, but no immediate clear variants can be detected using other methods.

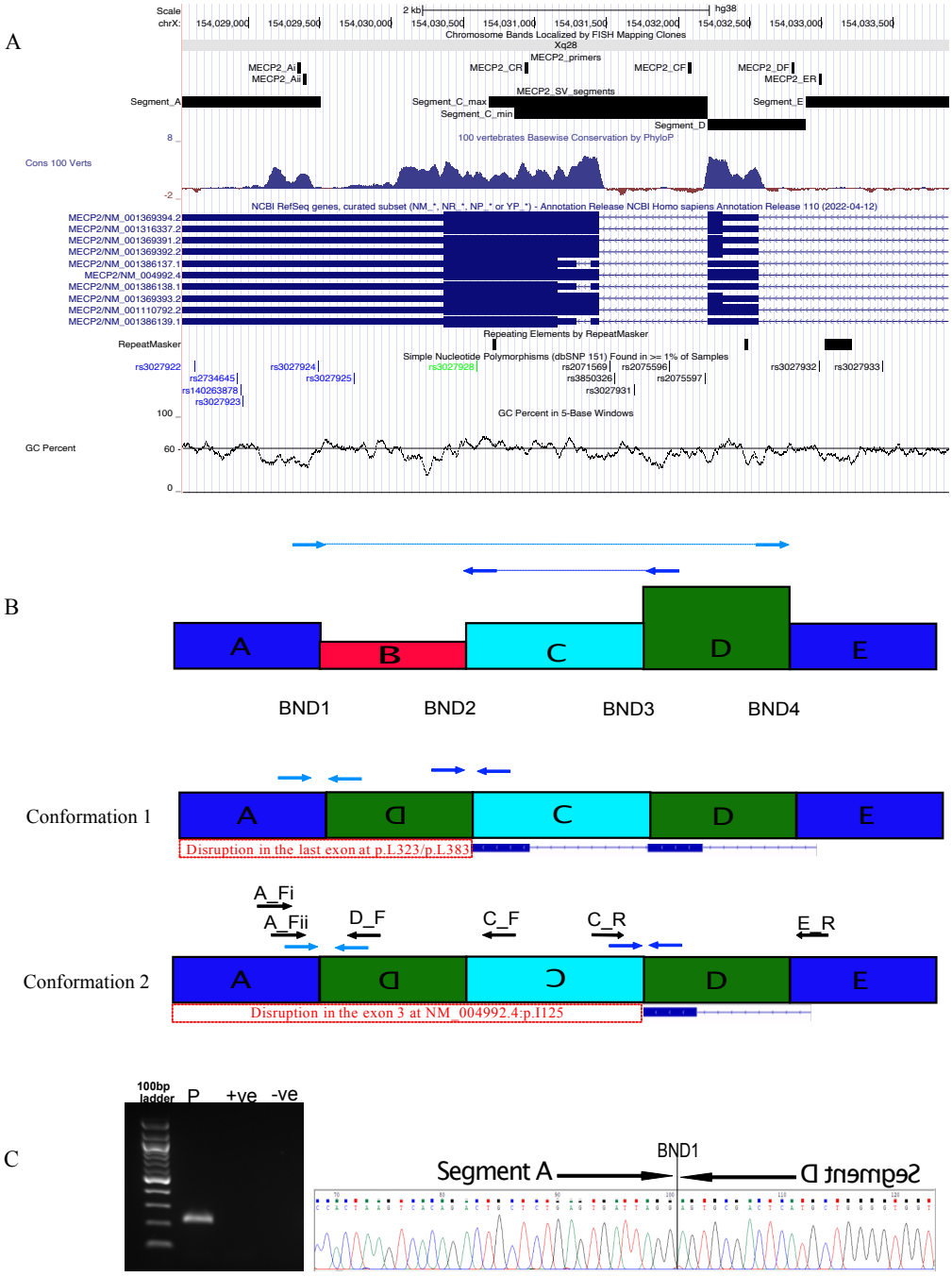
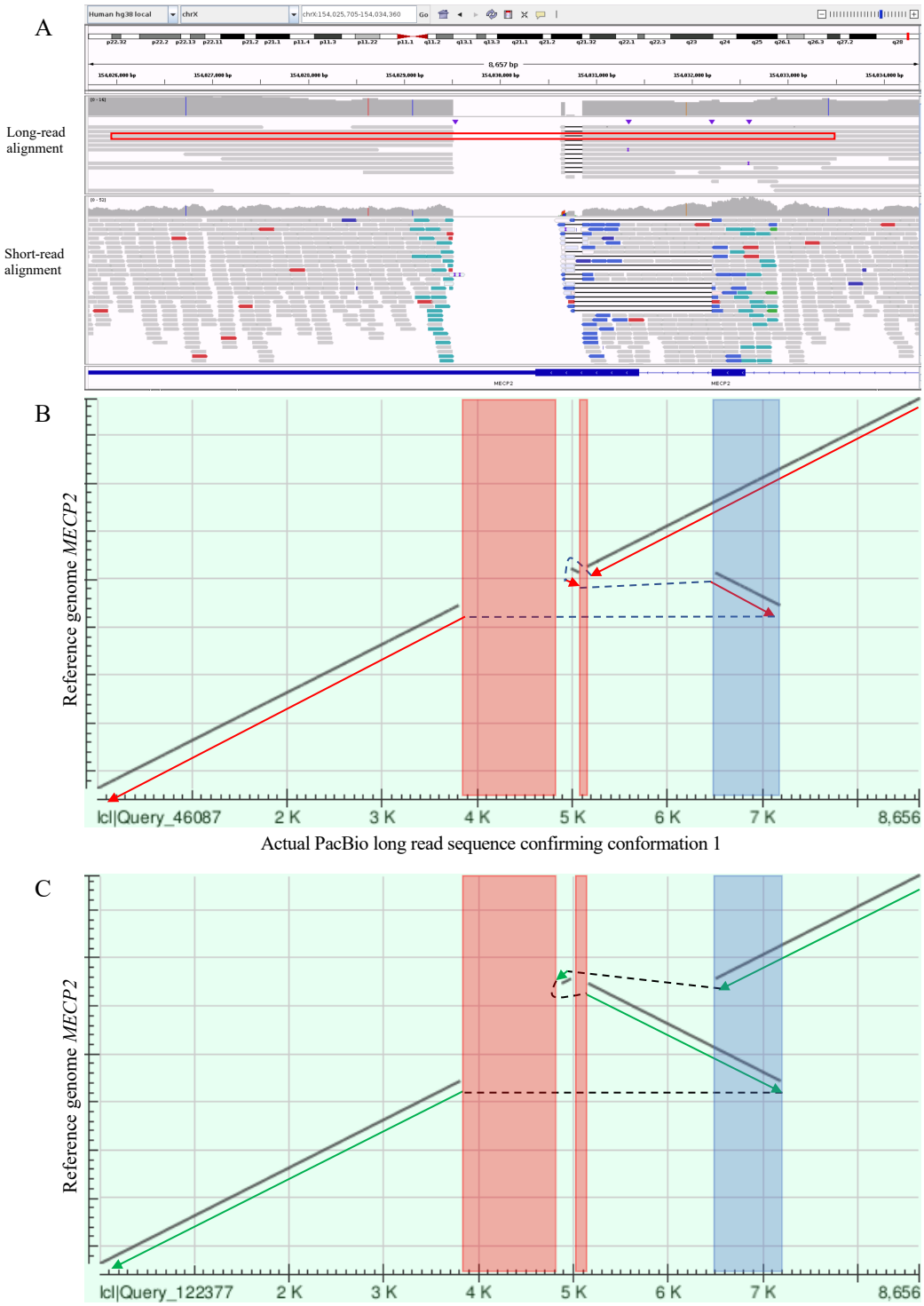


Figure 5.6: Challenges to validation of cxSV in MECP2 (Deletion-inversion-duplication). A) UCSC genome browser track showing the uncertainty of the length of segment C, which makes it difficult to determine where to position BND2. Moreover, the GC content of the gene is relatively high with some repeat elements at BND4 (available at https://genome.ucsc.edu/s/mhashim/MECP2_cxSV). B) a simple linear schematic showing two possible conformations extrapolated from the short-read data as well as the position of the primers relative to the different BNDs. Segment B is deleted in the proband and heterozygous for the deletion in the mother. C) While various PCR reactions have been performed, only MECP2_Afii + MECP2_DF (Table 2.2) has given a product, which was subsequently sanger sequenced confirming the sequence at BND1 and the orientation of the inverted duplicated segment D. P = proband, +ve= genomic DNA from a healthy participant (from GEL) who is not a carrier of the inversion, -ve= negative control (no DNA template) for the PCR reaction.



Actual PacBio long read sequence confirming conformation 1

Simulated long read sequence illustrating conformation 2

Figure 5.7: PacBio validation of cxSV in MECP2 in a male proband with atypical Rett syndrome. A) PacBio long-read alignments (top panel) and Illumina short-read alignment (bottom alignment). The >20 kb read from PacBio highlighted in red spans the complete length of cxSV and was used to generate the dot plot in panel B depicting the genomic sequence in the proband truncating beyond critical MBD sequence (p.L323) and the artificial modified sequence to simulate possible conformation 1 with the earlier truncation (p.I125) in panel C. B) Dot plot generated using >20 kb long read from PacBio, which confirms that the proband has a complex rearrangement that truncates the protein at p.L323 (NM_004992.4).

Individual 6_5 (I6_5): An 18-year-old female with moderate intellectual disability, global developmental delay, and macrocephaly, along with delayed speech and language development, was recruited to the 100KGP in 2016 under "Intellectual disability". Whole-genome sequencing was performed on the proband and both unaffected non-consanguineous parents. Intellectual disability *in silico* panel (version 1.158) was applied. This example has been described in (Hashim et al., 2023) Compound heterozygous variants in *SLC17A5* (NM_012434.5), a gene associated with Salla disease were reported in the "ExitQ" form, but the case was marked as unsolved. The reported variants were a maternally inherited NM_012434.5(SLC17A5):c.406A>G (p.Lys136Glu) well-known pathogenic alteration in ClinVar (VCV000021493.31), and a paternally inherited variant (VCV000781754.30; NM_012434.5(SLC17A5):c.85G>A; p.Ala29Thr) that is likely benign.

The SVRare analysis identified a paternally inherited direct tandem duplication (NC_000009.12(NM_001371194.2):c.-1906387_414+1034dup), however, the significance is unknown as the father is unaffected. Moreover, the analysis identified several adjacent *de novo* SVs that disrupt *KMT2E* in the proband. Manta identified two inversions and a deletion, while Canvas called a deletion of 7,705 bp. A closer inspection of the read alignments surrounding the SV using IGV (v2.11.9) showed a more complex nature to the variant, including three deletions removing exons 18-24 (NM_182931.3) with two retained internal segments including part of exon 19, one of which is inverted (Figure 5.8: A). The exact breakpoints for the SV in P5 are listed in Table 5.3. Although the breakpoints of this rearrangement lie downstream of the SET functional domain (NM_182931.3:c.967_1353; p.323-451), six other variants that lie further towards the 3'-end of the gene that truncate the protein further downstream of this SV mutation have been reported (C. Velmans et al., 2022). Therefore, the SV was highly likely to be pathogenic, hence a form for contacting the clinician was submitted to obtain DNA sample for validation of the breakpoints and check whether the patient phenotypes fit with ODLURO (OMIM #618512). This allowed for further details on the phenotype of the proband and results of previous tests performed to be obtained.

The proband was first assessed at 11 years of age, following a referral to clinical genetics due to moderate learning difficulties, she was ~3 years behind her peers in education and required educational support. Additionally, she displayed autistic spectrum traits. Her postnatal course was marked by macrocephaly, with her occipitofrontal head circumference (OFC) measuring in the 50th

centile at birth but rising to the 98th centile by nine months. Notably, she presented with a large tongue. During a physical examination at 11 years, her height was recorded as 155cm (between the 91st and 98th centiles), while her OFC measured 58cm (considerably above the 95th centile; the 95th centile at this age is 55cm).

At the age of 18, her height was plotted between the 98th and 99.6th centiles, and she exhibited distinctive facial features such as a prominent forehead and a square nasal tip. Beyond these physical characteristics, the patient was undergoing treatment for acne and reported instances of headaches and low mood. Her medical history also documented gastrointestinal symptoms, including constipation, which led to a referral for investigations concerning possible gluten and lactose intolerance.

Prior to WGS in the 100KGP, various investigations were performed, including Agilent ISCA 60K oligo array analysis. The array data identified a rare paternally inherited 2.15 Mb duplication of 9q22.2q22.31. The significance of this duplication encompassing: Semaphorin 4D (*SEMA4D*), Growth Arrest and DNA Damage-Inducible Gamma (*GADD45G*), DIRAS Family GTPase 2 (*DIRAS2*), Spleen Associated Tyrosine Kinase (*SYK*), and AU RNA Binding Methylglutaconyl-CoA Hydratase (*AUH*) was unclear. Of the five genes, *SYK* and *AUH* are OMIM morbid genes associated with Immunodeficiency 82 with systemic inflammation (OMIM #619381) and 3-methyl glutaconic acidurua type 1 (OMIM #250950), respectively. The large OFC at 11 years prompted Sanger sequencing and MLPA analysis of *PTEN* but no pathogenic variants were identified.

Moreover, the initial finding of a compound heterozygous variants in *SLC17A5* prompted testing of levels of urine sialic acid in the clinical laboratory and was found to be normal, therefore, excluding the variants from being a likely cause of the phenotype.

Primers were designed to capture all three junctions in a single PCR and thus help confirm the orientation and structure of this cxSV (Figure 5.8: B). All three breakpoints were amplified successfully. Sanger sequencing verified the sequence, confirming the DEL-DEL-INV-DEL as a single structural *de novo* mutational event (Figure 5.8: C).

Table 5.3: KMT2E SVs identified in I6_5.The coordinates as identified by the CNV algorithms and coordinates from IGV visual inspection of the read alignments.

		Deletions	Inversions
SVs according to Algorithm	**Manta**	NM_182931.3:c.2197-1181_2523del	NM_182931.3:c.2596+136_2597-129inv NM_182931.3:c.2596+224_3756-77inv
	Canvas	NM_182931.3:c.2196+354_3756-27del	-
SVs from IGV		NM_182931.3:c.2197-1181_2524del NM_182931.3:c.2596+137_2596+223del NM_182931.3:c.2597-127_3756-74del	NM_182931.3:c.2596+223_2597-128inv

In 2019, we published with collaborators 38 individuals from 36 families (O'Donnell-Luria et al., 2019b), which was followed by a second cohort (Velmans et al., 2021) and several case reports bringing the total to over 60 patients within two years of the first publication. Most patients had mild intellectual disability, and 25% met the criteria for autism. Common features included macrocephaly, hypotonia, functional gastrointestinal problems such as reflux, vomiting and bowel motility issues. Moreover, they had subtle facial gestalt. About 20% of seizures were easy to control except in those with a missense mutation.

Upon further consideration, the overseeing clinician confirmed that the clinical features in P5 were in keeping ODLURO syndrome. To date, all LoF variants described in the literature are frameshift, splicing variants, nonsense, microdeletions of a full copy of *KMT2E* or as part of a larger cxSV (Imaizumi et al., 2020). However, no cryptic cxSV within *KMT2E* has been described yet.

This variant is the first cryptic cxSV within *KMT2E* falling below the resolution of arrays. Furthermore, MLPA verification of this variant would have been unsuccessful due to the limited number and positions of the probes across the gene. The genomic context surrounding the variant presents a challenge for validation due to high GC content and multiple repeat regions, especially at the proximal end of the variant (Figure 5.8: B). The overestimation of deletion length by Canvas and the miscalling of deletion as an inversion by Manta further confounded the primer design. Visual inspection of the paired-read alignments was thus critical to successful primer design.

This example highlights the importance of revisiting unsolved affected individuals using better SV prioritisation tools and updated gene panels. This is evident here as the original *KMT2E* study was published three years after the initial analysis of the WGS data for the proband. Finally, there is an ongoing need for a more comprehensive assessment of structural variation in clinical WGS programmes to ensure cryptic and complex SVs are not overlooked.

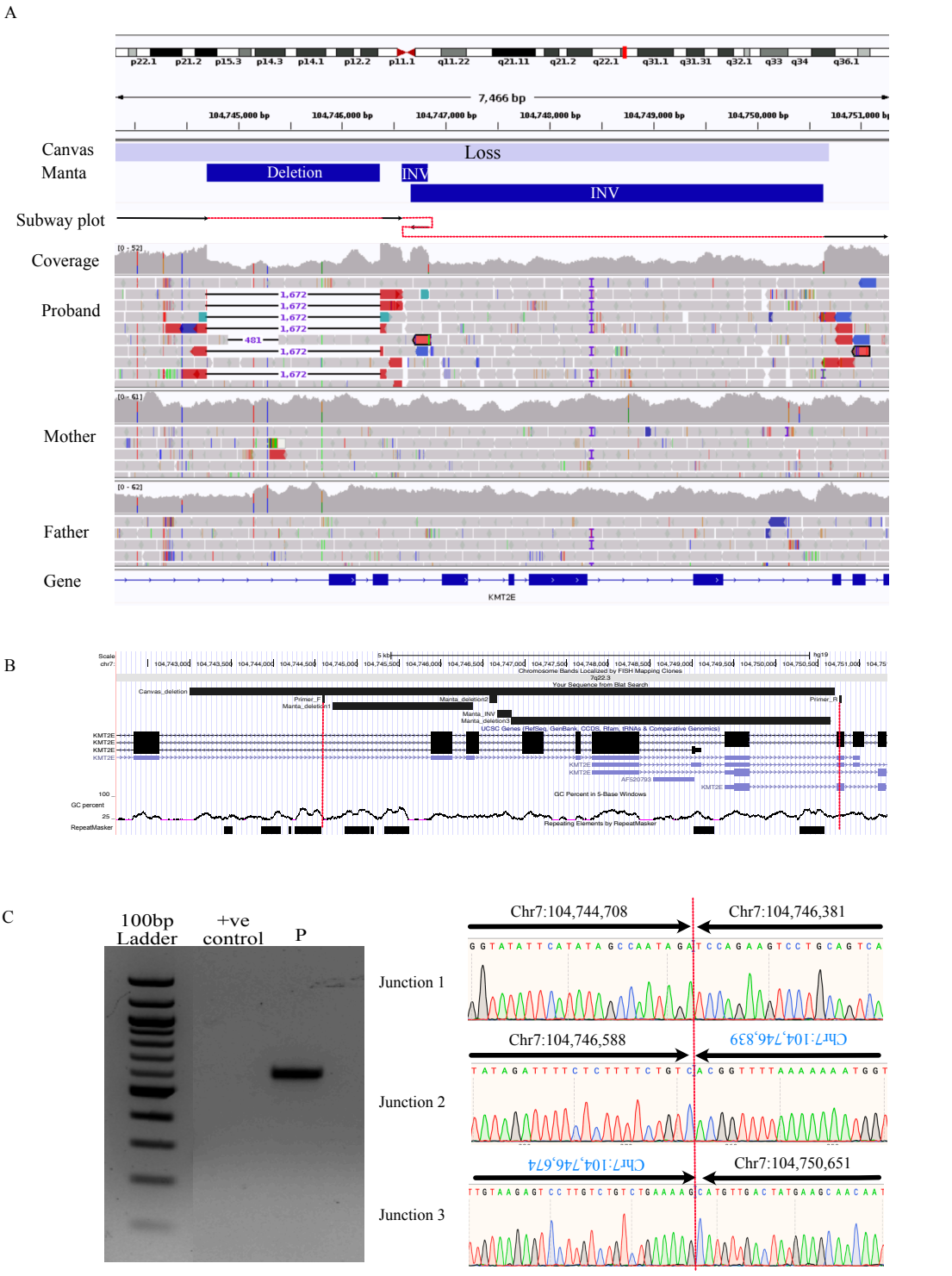


Figure 5.8: Confirmation of the first KMT2E cryptic structural variant leading to a revised diagnosis to ODLURO syndrome. A. Short-read alignment data confirms the presence of a de novo DEL-DEL-INV-DEL complex SV called as separate events by Canvas and Manta. Both tools fail to capture the full complexity of the variant, however, Manta calls show accurate breakpoints, while Canvas over calls the boundaries of the deletion. B. Primers used for validation of the sequence and orientation of the SV- UCSC session http://hgw1.soe.ucsc.edu/s/mhashim/KMT2E. C. Agarose gel shows a unique band in the proband (P) but no PCR product in a control DNA. Electropherograms of Sanger sequence shows the expected three junctions and confirm that the identified SVs are all in the same allele. Arrows indicate the direction of primers; and genomic positions marking the junctions are indicated on top. Position numbers appear in upside down manner (blue colour) indicate antisense sequences of the inverted segment.

5.4.4 Systematic analysis of cases with karyotypic abnormalities including apparently balanced translocations from the 100KGP

Apparently balanced translocations (ABTs) are a type of structural variant that leads to chromosomal rearrangement while preserving the nucleotide count as no net gain or loss of genetic material should have taken place (Chowdhury et al., 2020). Most individuals who carry ABTs do not show any visible clinical features. However, it is estimated that 6-10% of *de novo* ABT carriers will exhibit specific clinical characteristics (Warburton, 1991). Additionally, recent studies suggest that *de novo* carriers of the ABT have an increased comorbidities, with a risk as high as 27% (Halgren et al., 2018).

Several molecular mechanisms have been proposed for how ABTs can lead to clinical phenotypes observed in the carriers. One possible explanation is the direct disruption of haploinsufficient genes that do not tolerate changes in dosage (Bonaglia et al., 2001). ABT breakpoints can also occur in intergenic regions, disrupting the regulatory regions that control nearby genes' expression through long-range position effects; LRPE (Bonaglia et al., 2001). This disruption could occur by completely removing or displacing regulatory elements such as promoters, enhancers, or silencers, leading to the dysregulation of the target genes. Another mechanism in which positional effect manifest is through the disruption of highly-conserved TADs interrupting the formation of chromatin loops that facilitate the physical contact between genes and their long distant regulatory elements (Nora et al., 2012; Rao et al., 2014). Another possible explanation for the observed phenotype is through the presence of cryptic genomic imbalances or complex chromosomal rearrangements (CCRs) associated with the translocation (Patsalis et al., 2004; Sismani et al., 2008). Fusion genes, albeit rarer, may also be generated by translocations, potentially impeding the production of a functional protein or leading to the formation of pathogenic proteins (Córdova-Fletes et al., 2015). It is noteworthy that ABTs might not always be linked to the phenotypes presented by the carrier (Aristidou et al., 2017). However, once a karyotypic abnormality such as ABT is detected, a detailed characterisation of the breakpoints is essential to identify whether they directly disrupt a disease-causing gene or the regulatory region of a nearby one (Tommerup, 1993). WGS is a powerful tool

for establishing the potential molecular mechanism underlying the pathogenesis of the observed phenotype in affected ABT carriers (Redin et al., 2017). Therefore, further research is necessary to fully understand ABT's clinical implications and develop appropriate management strategies for individuals with these rearrangements.

One of the main advantages of large WGS national programmes such as the 100KGP is the ability to assemble cohorts with karyotypic abnormalities linked to clinical features that karyotyping results cannot explain directly. In this study, I have identified 22 participants (IDs: I7_5 – I28_5; males=10, females=12) with karyotypic abnormalities in the rare disease arm of the project and were all marked as unsolved. The age of the participants ranged between 10-52 years old, and each endured a minimum of 6 years of a diagnostic odyssey.

The main goals of this analysis were: i) to evaluate the ability of 150-base-pair paired-read genome sequencing to pinpoint the precise breakpoints and disease-causing genes disrupted by a translocation, ii) to determine the frequency of translocations that appear balanced under the microscope but are part of a more complex rearrangement, and iii) identify any novel disease-causing genes or mechanisms.

While the presence of karyotypic abnormality was noted in the 22 individuals from previous genetic testing performed, only two karyotype results were recorded in the GEL RE. However, through contacting the clinicians, I was able to obtain information on karyotype results for 15/20 remaining individuals. Since five individuals were missing this information, analyses blind to chromosomes involved in the translocations were performed using the approach described in section 5.2.2.

Out of 22 individuals, 13 were confirmed *de novo* based on karyotype results or from WGS data, two segregated with phenotypic abnormality within a family, and two were maternally inherited but the maternal phenotype is unknown; however, the inheritance information was unavailable for both parents in six. The participants in this cohort presented with clinical phenotypes ranging from system-specific to multisystem disorders as well as neurodevelopmental disorders such as intellectual disability. While this cohort was selected for the presence of karyotypic abnormality

regardless of the phenotype, the most common features were neurological defects present in 90% of participants and included developmental delays, intellectual disability, and autistic behaviours (Figure 5.9). The high incidence of neurological phenotypes is likely due to the large proportion of participants (>20%) with neurological and neurodevelopmental disorders in the rare disease arm of the 100KGP.

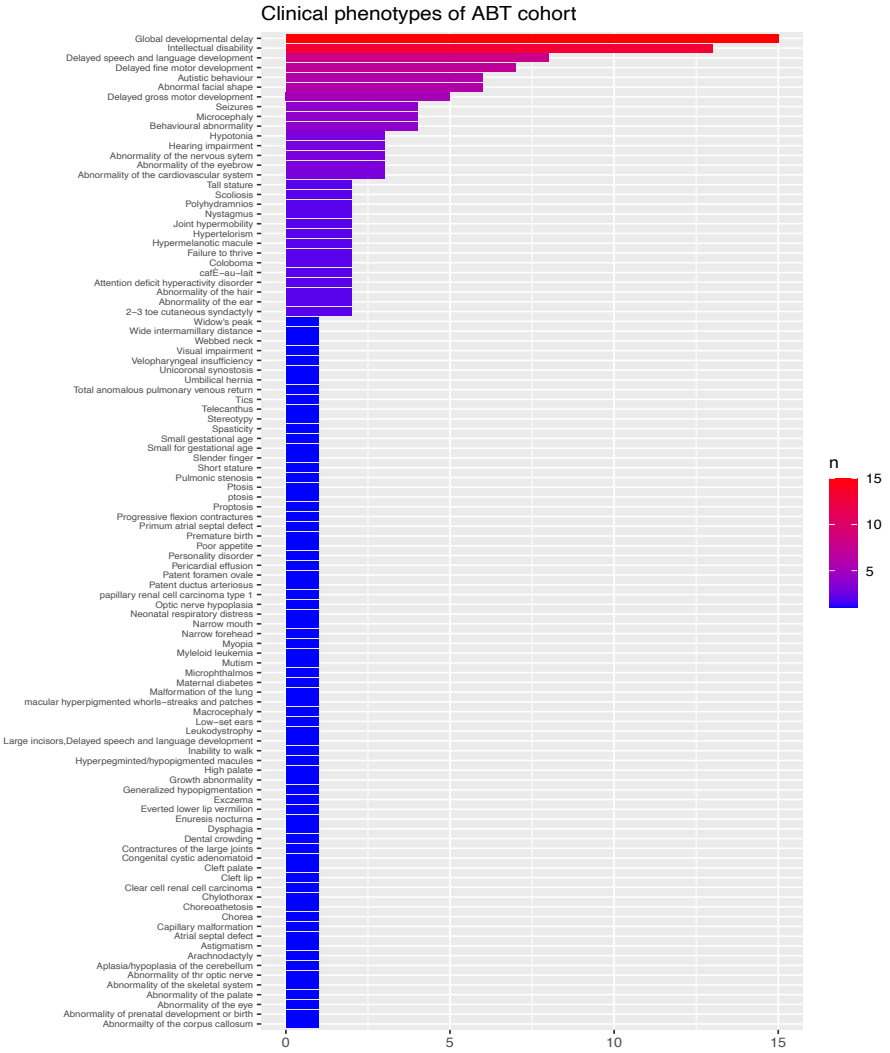


Figure 5.9: Clinical phenotypes as reported in the 100KGP database for the participants in the cohort.

The breakpoints were successfully identified in 20 individuals (known karyotype=16/17; unknown karyotype=4/5), Figure 5.10: A which led to molecular diagnoses in 9/20 individuals through determining the disease-causing gene or mechanism (Figure 5.10: B).

The precise detection of BNDs (n = 67 junctions) through WGS has led to revision of the breakpoint's position to a neighbouring cytogenetic sub-band (at least one sub-band) in 14/16 individuals in which the karyotype was reported, and BNDs identified from WGS data (Figure 5.10: C). Moreover, the karyotype reports indicated a simple reciprocal balanced translocation with two breakpoints and no associated genomic imbalances in 16/17. However, seven out of those with ABTs (n=16) were found to harbour cryptic complexity below the detection threshold for either karyotype, array or FISH. These complexities varied, ranging from instances of low complexity in which one additional genetic alteration, including gains, losses, and inversions, were associated with the primary translocation event, to instances of high complexity (Figure 5.10: D). Those of high complexities involved the presence of >3 breakpoints of which two CCR were consistent with the phenomena of chromothripsis or chromoanasynthesis, which involve extensive chromosomal shattering and random joining of the fragments from one or more chromosomes. These phenomena were determined based on a minimum of 8 junctions, the inferred DNA replication, and gains or losses associated with the event (Chatron et al., 2020).

There were four breakpoints in which the position of the mate-read was ambiguous from the short-read data, possibly lying within regions of low complexity such as the telomeric region of the chromosome as in case 22 (Appendix Table S5.2). Due to the process of releasing participants' samples to research laboratories, validation by dideoxy sequencing had to be conducted in the corresponding NHS diagnostic laboratory. However, these labs usually have a long turnover period for validation of results. Therefore, I was able to obtain the DNA sample for one individual for validation and the dideoxy-sequences of the breakpoints for a second in this cohort. Fortunately, these sequences were in concordance with the expected sequences based on visual alignment checks. Therefore, using the junction sequences, I found that approximately half of the breakpoints (32/67) showed no homology and presumably driven by non-homologous end joining (NHEJ). Moreover, a substantial number of junctions displayed microhomology of 2-7 bp at the breakpoints suggesting that microhomology-mediated break-induced replication (MMBIR) play a role in the occurrence of translocations (Zhang et al., 2009).

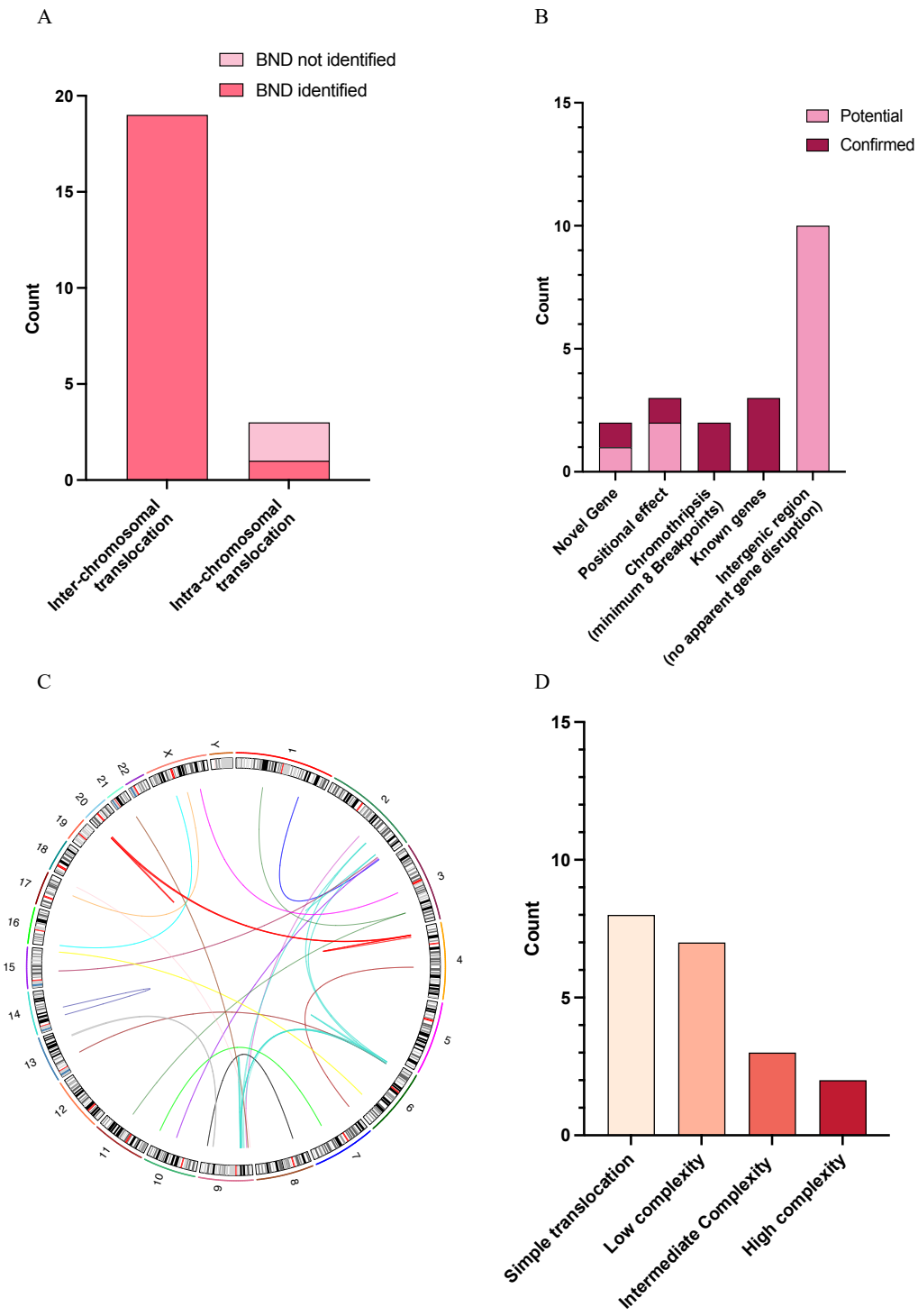


Figure 5.10: Overview summary results karyotypic abnormality cohort in the 100KGP. A) Types of translocations included are inter-chromosomal ones involving at least 2 chromosomes and intra-chromosomal ones involving large paracentric inversions detected by karyotype. B) Disease-causing genes or mechanisms identified in the individuals with karyotypic abnormalities. Known/validated genes disrupted (dark red) and pink coloured bars show either potential disease genes that are yet to be functionally validated or BNDs falling within intergenic regions without clear disruption to genes or TADs. C) Circos plot representing all breakpoints identified across the cohort. Each unique colour represents all BNDs in an individual case. D) the range of complexities in the cohort that were cryptic to the karyotype. Simple translocations involve 2 BNDs, low complexity involves <8 BNDs and one other type of SVs linked to the translocation, intermediate complexity involved <8 BNDs and 2 other types of SVs as part of the event, and high complexity involves >8 BNDs as well as other types of SVs.

The mechanism by which the translocations in this cohort have contributed towards the phenotype of the patients involved direct disruption of a gene linked to the specific patient phenotype. These were missed previously due to the low resolution of methods used. The ability to precisely characterise breakpoints allowed the identification of two novel disease genes, three known disease-causing genes that were missed, and two individuals with highly complex chromosomal rearrangement. Other breakpoints within intergenic regions were examined for nearby genes and possible disruption of long-range regulatory elements such as TADs, which might lead to the disruption of gene regulation hence disease development. In three individuals, the breakpoints disrupted either a long-range regulatory element (downstream of *FOXG1*) or a highly conserved regulatory long non-coding RNA loci, lnc*NR2F1* and lnc*NR2F2*.

5.4.5 PPP2R2A and ECPAS as potential novel disease genes directly disrupted by translocations in individuals with neurodevelopmental disorders.

One of the main objectives of using WGS in clinical or research settings is to facilitate novel disease-causing gene discovery, which would improve the diagnostic yield of WGS and provide an answer for persons with an unknown diagnosis.

While individual "novel" genes can be identified rapidly using the current advances in WGS variant detection and prioritisation tools, establishing causality is usually impeded by the requirement of not only identifying multiple, unrelated individuals with variants disrupting the same gene in a similar pattern, but also sharing similar phenotypes. Another route to validation is functional studies to provide evidence for gene causality using animal or cellular models. The disruption of genes that are intolerant to loss of function is a prediction of a haploinsufficiency mechanism for disease.

In this cohort, two individuals with ABT had breakpoints which disrupted potentially new disease-causing genes: Protein phosphatase 2 regulatory subunit B alpha (*PPP2R2A*) and Ecm29 Proteasome Adaptor and Scaffold Protein (*ECPAS*), which have pLI scores of 0.99 and 1, respectively.

5.4.5.1 PPP2R2A: A New Addition to the PPP2A Family of Neurodevelopmental Disorder-Causing Genes, Solving a 33-Year diagnostic odyssey for a Multigenerational Family

Individual 7_5: is a 13-year-old female with microcephaly and intellectual disability. She was recruited to the 100KGP in 2016 as a singleton under "Balanced translocation with an unusual phenotype" (Appendix Table S5.2). The karyotype result indicated a maternally balanced translocation between chromosomes 8 and 9, t(8;9)(p12;q34.11). Although this is a singleton case, the familial clinical history showed that the translocation is present in six affected members, all showing microcephaly and intellectual difficulty. On the other hand, six tested family members did not carry the translocation and did not show any clinical phenotypes confirming the segregation of the ABT with the clinical phenotypes (Figure 5.11). The translocation was first detected in the family by karyotype in 1989, but despite being followed up in the same clinic, no gene has been linked to the observed phenotype in the carriers since then.

WGS data analysis (SNVs and CNVs) did not identify any potential candidate genes. Inspection of the breakpoints of the segregating translocation showed that one of the breakpoints disrupts *PPP2R2A* in intron 2 (NM_002717.4:c.82+15396), hence displacing the first coding exon as well as the upstream regulatory region from chromosome 8 to chromosome 9.

Protein phosphatase type 2A (PP2A) is a major serine/threonine complex enzyme that is composed of three subunits: the C subunit, responsible for catalysing reactions; the A subunit, which acts as a scaffold; and the B subunit, which regulates the activity of the enzyme and controls substrate specificity (Cundell et al., 2016). Nineteen different genes encode these subunits, with two encoding C, two encoding A, and 15 encoding B subunits (Reynhout & Janssens, 2019). The B subunit is crucial in determining PPP2A's functions within cells and tissues (Cundell et al., 2016).

Over 60 cases in different subunits of the enzyme have been reported linked to neurodevelopmental disorders. These subunits include *PPP2R5D* (Houge et al., 2015), *PPP2R5C, PPP2R5B* (Loveday et al., 2015), *PPP2R2C* (Backx et al., 2010), *PPP2R1A* (Lenaerts et al., 2021), and *PPP2CA* with *PPP2R5D* having the highest reported number (n= 25) of *de novo* pathogenic variants. However, *PPP2R2A* has not been linked to disease yet.

PPP2R2A encodes for PP2A-B55α regulatory subunit. It is involved in the dephosphorylation of β-catenin during Wnt signalling, potentially impacting embryonic development (Zhang, et al., 2009) . In mice, homozygous deletion of *Ppp2r2a* caused embryonic lethality in most embryos; neural tube defects, limb defects and impaired epidermal barrier formation were observed in $Ppp2r2a^{-/-}$ embryos collected at different gestational stages. However, heterozygous $Ppp2r2a^{+/-}$ grew and developed normally. They were monitored for 24 months but no anatomical or behavioural abnormalities were noted (Panicker et al., 2020).

In a recent study describing the phenotypic spectrum of PPP2R1A-related neurodevelopmental disorder, some of the discussed variants (p.Pro179Leu, p.Arg182Trp, and p.Arg258His) impaired binding to the B55α subunit, leading to microcephaly and intellectual disability in these patients (Lenaerts et al., 2021). Through collaborations with international groups, another seven individuals with missense (n=5) or LoF variants (n=2) in *PPP2R2A* were identified. This collaboration was established through GeneMatcher (Sobreira et al., 2015), which offers the opportunity to connect with scientists, clinicians, and patients in search of persons with similar genotypes and phenotypes. All individuals present with microcephaly and intellectual disability, however, the ones with missense variants exhibited a more severe phenotype.

The familial co-segregation of the translocation with the phenotype, the additional individuals from independent families, intolerance of *PPP2R2A* to loss-of-function, the phenotype of mouse KO, and the presence of neurodevelopmental disorder in other related genes strongly implicates *PPP2R2A* is a disease-causative gene in this family, thus ending a 33-year diagnostic odyssey. Furthermore, it adds a novel disease gene for neurodevelopmental disorders, that can be screened in undiagnosed individuals and potentially increases the diagnostic yield of WGS in individuals with NDD.

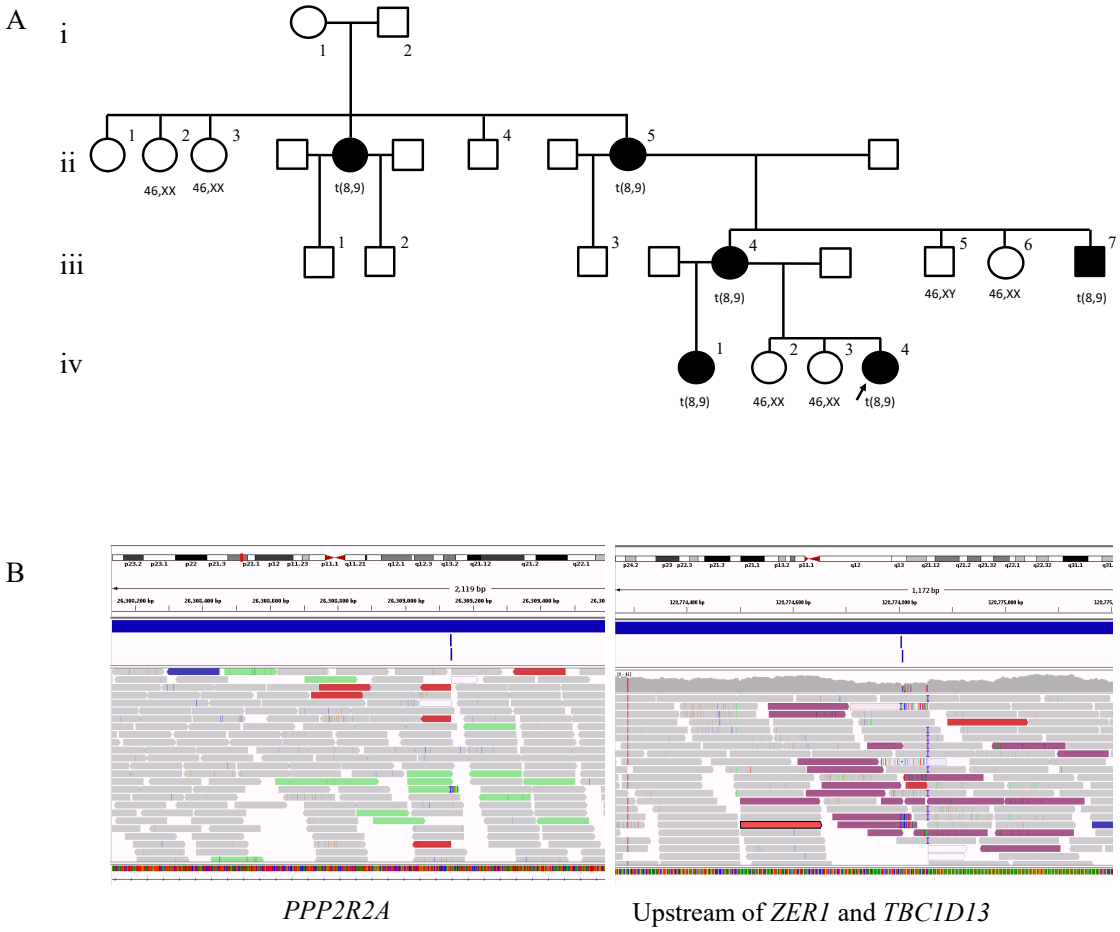


Figure 5.11: A segregating balanced translocation disrupting PPP2R2A in a multigeneration family. A) the pedigree showing the karyotype result of 12 tested members of which 6 are affected carriers and 6 healthy non-carriers. B) short-read Illumina data alignment showing the breakpoints in WGS data of individual iv.4 disrupting PPP2R2A in chromosome 8, the second breakpoint in chromosome 9 is in an intergenic region upstream of ZER1 and TBC1D13

5.4.5.2 Identification of ECPAS as a Potential Novel Disease Gene: Implications for Further Research

Individual 8_5: is a 50-year-old female presenting with tics, behavioural abnormality, personality disorder, and attention deficit hyperactivity disorder (Appendix Table S5.2). Initial testing included karyotyping, which showed an apparently balanced translocation t(9;13)(q32;q31). WGS confirmed this presence of the translocation, but the breakpoints have been reassigned to t(9,13)(q31.3,q31.2), and additional complexity was detected. The BNDs disrupt three protein-coding genes: Sorting Nexin Family Member 30 (*SNX30*), Ecm29 Proteasome Adaptor and Scaffold Protein (*ECPAS*), and Melanoma Antigen, family D, 4 (*MAGED4;* alias *KIAA1859*) (Figure 5.12: A). Two inverted segments of ~1.2 Mb in chr9 and ~300 kb translocated from chr9 and inserted into chr13 were identified with no deletions or duplications as part of this complex rearrangement (Figure 5.12: B). None of the three disrupted genes has been linked to disease, but *ECPAS* has a pLI score of 1, while *SNX30* and *KIAA1859* have pLI scores of 0.59 and 0.32, respectively.

Phenotypes for KO mouse models for *Ecpas* were available in the Mouse Genome Database. The homozygous mouse phenotypes are reminiscent to those observed in the proband, however, the heterozygous KO mouse did not exhibit apparent differences to the WT. The mouse model phenotypes included behavioural abnormalities such as abnormal vocalisation, decreased thigmotaxis – a response to touch stimuli by movement or change in orientation, as well as hyperactivity (MGI:6461232). The resemblance between the mouse and the patient suggests that *ECPAS* could be a promising candidate for the phenotype in this case. However, the possibility that *SNX30* and *MAGED4* also play a role cannot be ruled out. The pursuit of collecting other individuals is ongoing and will be complemented by future functional studies on human iPSCs.

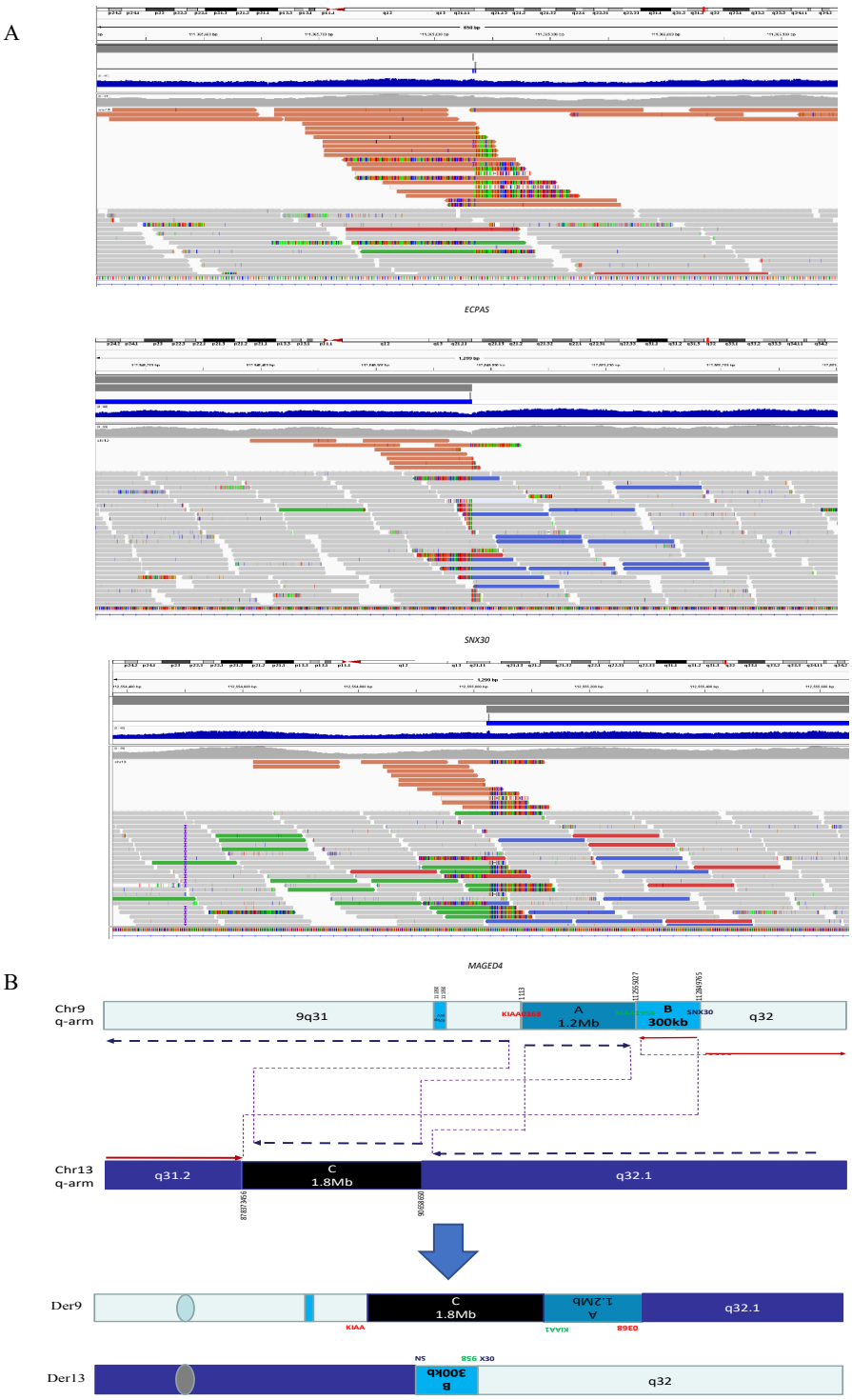


Figure 5.12: Complex rearrangement in apparently balanced translocation disrupting three protein-coding genes. A) Short-read alignments showing the BNDs involved in the translocation disrupting KIAA0368 (alias ECPAS), SNX30, and KIAA1859 (alias MAGED4). B) Linear schematic of segments in the q-arm of chr9 and chr13 and how they are arranged to satisfy the split-reads shown in the alignment, in which part of each of the disrupted genes have been either translocated to a different chromosome or inverted, split, and moved away from the rest of the gene.

5.4.6 Known neurodevelopmental disorder-associated genes disrupted by de novo translocations

Previous studies showed that a large proportion of patients with developmental disorders harbouring translocations (26.6%), the breakpoints affected previously defined disease genes (Redin et al., 2017). The disruption of these genes is more commonly reported due to missense or LoF variants and due to translocations to a much lesser extent. Therefore, the exact characterisation of translocations can help refine genotype-phenotype correlations as they often result in the fusion of genetic elements from different genes, potentially generating novel chimeric proteins or disrupting regulatory elements. Therefore, understanding the functional consequences of these events can provide insights into disease mechanisms and phenotypic heterogeneity observed in patients.

The analysis of the 22 individuals of balanced translocation revealed disruptions of three known genes that were missed by the conventional methods in clinical laboratories. Specifically, *SHANK2, NBEA*, and *CUX1* in three patients recruited to the 100KGP in 2016 under "Disease group: balanced translocation with an unusual phenotype" as part of a child-parent-parent trio.

In a 15-year-old female patient (Appendix Table S5.2, individual 9_5), a *de novo* simple reciprocal translocation involving chromosomes 9 and 11 was identified where the SH3 And Multiple Ankyrin Repeat Domains 2 (*SHANK2*) has been disrupted by one of the BNDs. *SHANK2* is linked to Autism susceptibility 17 (OMIM #613436). It is a scaffolding protein that forms part of the postsynaptic density and functions as an anchor for protein complexes of N-methyl-D-aspartate receptor (NMDR), α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor (AMPAR), and metabotropic glutamate receptors; mGluRs (Sheng & Kim, 2000). These complexes are involved in the regulation of signalling pathways of transcription and translational control (Santini & Klann, 2014).

For the past decade, various studies reported *de novo* LoF variants in *SHANK2*, including deletions (Berkel et al., 2010), nonsense (Zhou et al., 2019), and frameshift variants (Guo et al., 2018) as being highly penetrant and commonly associated with a diagnosis of autism or autism spectrum disorder, intellectual disability (mild to moderate), developmental delay, and speech and language developmental delay. Furthermore, studies have also reported additional features such as sleep

disorders, attention-deficit/hyperactivity disorder, hyperactivity, and anxiety (Caumes et al., 2020). Revisiting the proband's phenotype in light of this finding confirms that they are related to *SHANK2* haploinsufficiency.

The second is a 35-year-old female proband (Appendix Table S5.2, I10_5), in whom WGS data identified the precise breakpoints of the previously reported balanced translocation from karyotyping. It has been since then recognised as a *de novo* unbalanced translocation as small deletions are present at each breakpoint in chromosomes 6 and 13. In chromosome 6, the BND interrupts Ataxin 1 (*ATXN1*) in intron 5 (NM_000332). The gene has a pLI score of 0.97 and is depleted for LoF variants in gnomAD (only 3 LoF variants observed, expected 24.6). It has been linked to autosomal dominant spinocerebellar ataxia 1 (OMIM #164400), a neurodegenerative disease with a late onset in the third or fourth decade of life. Although the proband is in the middle of the third decade, the phenotypes are of childhood onset and do not match the typical phenotypes of spinocerebellar ataxia 1.

Nonetheless, the implication of this gene cannot be disregarded given that the translocation will cause a clear *ATXN1* loss-of-function, yet it does not explain the observed phenotypes of the patient. On the other hand, the breakpoints in chromosome 13 lie in intron 38/58 of Neurobeachin (*NBEA*), which is involved in regulating synaptic structure and function (Mulhern et al., 2018). Moreover, haploinsufficiency of *NBEA* has been linked to autosomal dominant NDD with or without early-onset generalised epilepsy (OMIM #619157). As a result, the proband's phenotypes are now believed most likely to be explained by heterozygous disruption of NBEA function.

Finally, a newly described disease-causing gene *CUX1* has been identified as the likely cause of neurodevelopmental disorder phenotypes in a 20-year-old female (Appendix Table S5.2, I11_5). During prenatal checks, an increased risk of trisomy was indicated on screening, but parents declined invasive testing followed by uncomplicated birth at term (+9 days). The proband presented with mild intellectual disability and delayed speech development, with first speech between the ages of 2-3 years. Delayed motor development was also observed, and she first walked just under the age of 2 years; however, she caught up on her speech and motor developmental milestones but complained of low mood. Muscular hypotonia and uncontrolled generalised non-motor absence seizures are the main neurological symptoms, with no abnormalities in the brain MRI imaging. Additional

phenotypes include mild scoliosis, joint hypermobility, pes cavus, and some dysmorphic facial features, including deep-set eyes, full straight eyebrows, and slightly cupped simple ears in structure. Prior to WGS, Affymetrix CytoSNP 750K array and karyotype were performed. While the array showed a normal copy number, the karyotype results indicated a *de novo* ABT t(7;10)(q22.1;q26.11), but no potentially disrupted genes within the affected cytogenetic bands explain the phenotypes. WGS was performed on the proband and both unaffected non-consanguineous parents. *In silico* panels including Intellectual disability (v1.648), familial genetic generalised epilepsies (v1.22), Epileptic encephalopathy (v1.114), mitochondrial disorders (v1.64), Undiagnosed metabolic disorders (v.172), and Unexplained skeletal dysplasia (v.178) were applied, but no tier 1 or 2 variants were identified. Using the chromosomal information provided by the clinical team, the structural variant call file "SV.vcf" of the proband was interrogated for "PASS" Manta reciprocal BND calls between chr7 and chr10. Moreover, the family SVRare report was examined for any rare SVs in the proband, which identified two rare deletions that share the exact coordinates as the BNDs, indicating a degree of complexity to this translocation redefining it as "unbalanced translocation". The BNDs extracted from the vcf file were inspected on IGV (Figure 5.13: A) and confirmed true. It is important to note that BNDs are difficult for prioritization tools to identify robustly and often result in a high rate of false positive detections. In chr10, the two breakpoints are intergenic, with one breakpoint (chr10:129249183, build 38) located ~217 kb upstream of Methylguanine methyltransferase (*MGMT*) and ~600 kb downstream Early B-Cell Factor 3 (*EBF3*). The second breakpoint (chr10:129162520) is located ~3.3 Mb upstream of Uroporphyrinogen III synthase (*UROS*), a disease-associated gene linked to autosomal recessive Porphyria, congenital erythropoietic (OMIM #263700). Both BNDs mark the boundaries of ~86.6 kb deletion, but no apparent gene is affected directly or indirectly by this deletion.

Although *MGMT* is not associated with any disease, *EBF3* has been linked to autosomal dominant Hypotonia, Ataxia, and Delayed Development Syndrome (OMIM #617330). However, the affected region has no reported long-range regulatory elements, transcription factor binding sites, or highly conserved regions that could be disrupted by this deletion or translocation. Despite this, the possibility that *EBF3* is involved cannot be completely ruled out. On the other hand, the two breakpoints identified in chr7 disrupt Cut-Like Homeobox 1 (*CUX1*). One of the breakpoints

(NC_000007.14(NM_181552.4):c.883) falls within exon 11, leading to a frameshift and early protein truncation (p.Glu307_Glu1055delinsThrLeuGluAlaSerPheSerProAlaLysTrpAla*). The second BND falls within intron 14 (NC_000007.14:g.10222531). It also marks one end of a ~46 kb deletion that removes part of exon 11 and exons 12-14 (Figure 5.13: B).

CUX1 encodes a vertebrate homolog of Drosophila *Cut* that is conserved in mammals. The gene encodes two different proteins generated through alternative splicing: cut-like homeobox 1 (CUX1) and cut alternately spliced protein (CASP), previously called CUTL1 (Sansregret & Nepveu, 2008). *CUX1* is a transcription factor containing four DNA binding motifs, three CUT repeats, and a homeodomain, whereas CASP is a Golgi protein that shares the N-terminal region with CUX1 but none of the CUT repeats nor the homeodomain (Gillingham et al., 2002). CUX1 proteins comprise, among others, a full-length p200 isoform. p200 CUX1 protein acts as a transcriptional repressor, whereas shorter CUX1 isoforms can activate or suppress transcription (Hulea & Nepveu, 2012; Sansregret & Nepveu, 2008).

CUX1 has a pLI score of 1, and its haploinsufficiency has been recently linked to autosomal dominant global developmental delay with or without impaired intellectual development (OMIM #618330). The initial report describes nine individuals with heterozygous truncating variants in *CUX1* (Platzer et al., 2018). In line with the involvement of *CUX1* in human brain functions, all individuals had delayed speech development; most exhibited delayed motor development and intellectual disability, with some showing catch-up on their speech and motor development (Platzer et al., 2018). However, this was not specific enough to make a conclusive decision on the diagnosis of this individual.

At the time of this study, a second cohort of 18 individuals, in which I11_5 was included, allowed for further refinement of the phenotypes linked to *CUX1* haploinsufficiency (Oppermann et al., 2023). The proband's phenotype is in keeping with global developmental delay with or without impaired intellectual development linked to *CUX1* haploinsufficiency.

In light of this finding, a retrospective examination of the array data confirmed the presence of ~46 kb deletion in *CUX1*, which was previously missed. The possible reasons for missing this diagnosis from the initial array data were scrutinised and identified due to having only 15 probes across the deletion region. The signal is lost when normalised across 50 markers and becomes below the

significance threshold. Furthermore, when the array was performed, *CUX1* was not linked to disease; therefore, even if it were prioritised, the deletion would have been considered a VUS according to ACMG guidelines. This case has contributed to the diagnostic yield of undiagnosed cases and led to the reassessment of unsolved cases with available Array data in clinical laboratories. Moreover, the concerned clinical laboratory has since revised the analysis strategy and threshold used in its analysis pipeline.

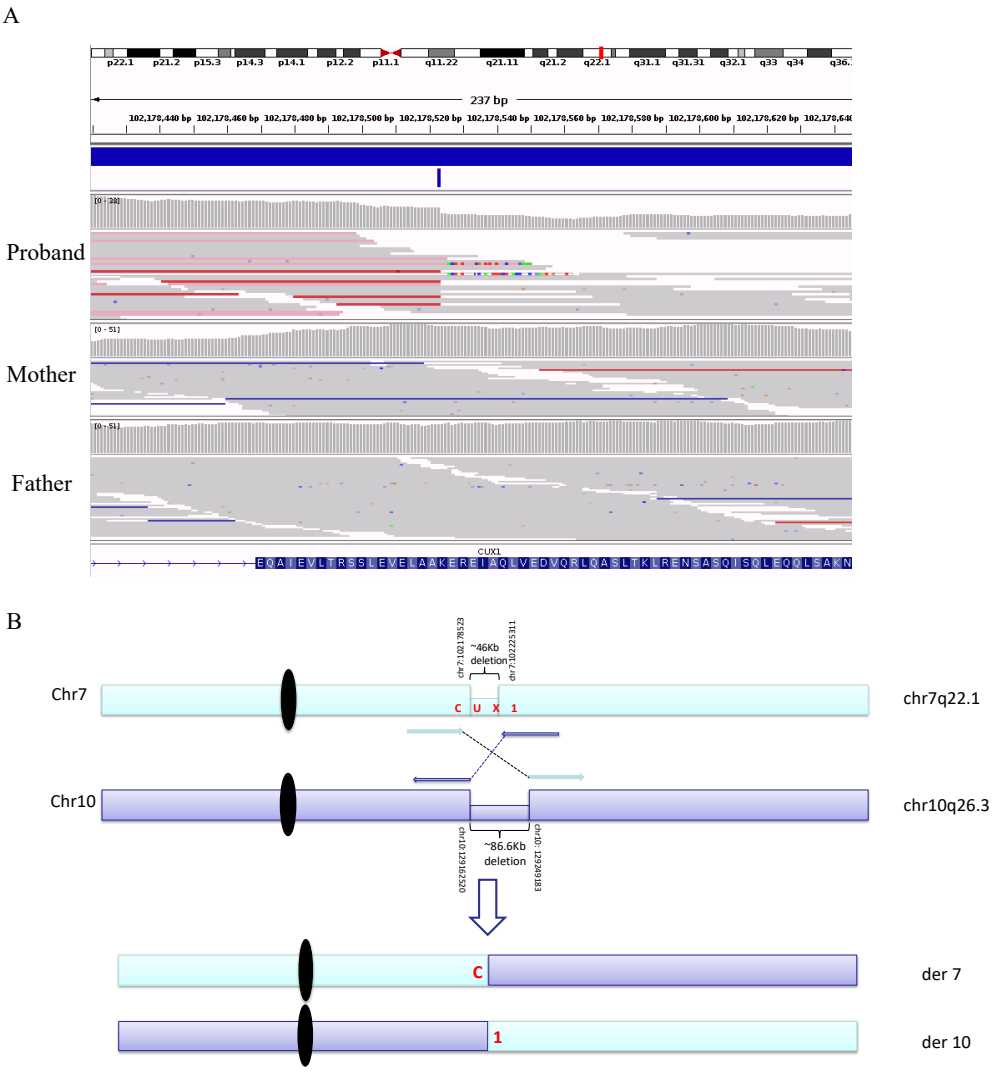


Figure 5.13: WGS data reveals cryptic complexities in CUX1 including deletions occurring at both breakends in this unbalanced translocation reported as balanced by karyotype. A) Short-read alignments as viewed on IGV (squished mode) shows a de novo translocation that is absent in both parents. Reads (pink colour) indicate that the mate-read maps to a different chromosome (in this case chromosome 10) and shows that one of the breakpoints in chromosome 7 (chr7:102178523, build 38) falls within exon 11/22 of CUX1 (NM_001202543.2). The coverage falls at the site of the breakpoints indicating the presence of a deletion that removes exons part of exon 11 and exons 12-14, which is supported by multiple reads (red colour). B) A schematic diagram showing the rearrangements involved in this unbalanced translocation. Both breakpoints in chromosome 7 disrupt CUX1 with a 46 kb deletion between the two BNDs. In chromosome 10, both breakpoints are intergenic with a ~ 86.6 kb deletion that does not include any known gene.

5.4.7 Positional effect through the disruption of chromatin architecture and lnc-RNAs

Breakpoints of translocations usually fall within the non-coding regions of the genome, making it challenging to understand their impact on human disease. The complexity of gene regulation within these non-coding regions, the incomplete annotation and the dependence on 3D genome conformation further complicate this task.

In NDDs, a growing number of SVs in non-coding regions are identified as possible causes of the disorder. Therefore, it is crucial to understand their effect on gene regulation for improved genetic diagnosis, given the highly regulated dynamic spatiotemporal gene expression during all stages of neural development (Nord et al., 2015). One possible effect of the SV is the disruption of long non-coding RNAs (lncRNAs). These are a group of non-coding RNAs that are longer than 200 nucleotides and are involved in gene regulation through various mechanisms acting in *cis* and *trans* (Briggs et al., 2015). They can act as scaffolds, which help organise and recruit specific regulatory protein complexes to specific genomic loci. They can also act as decoys by sequestering, leading to the inactivation of TFs and microRNAs (Briggs et al., 2015).

While changes in the expression of lncRNAs have been linked to neurodevelopmental disorders (Wang et al., 2015), there are relatively few cases in which disruptions to these lncRNAs have been proven to cause disease development, as opposed to simply being associated with the disease. Research involving differential expression analysis during mouse neuronal development has identified a potential role for the highly conserved lnc-NR2F1 in neurodevelopment (Brown et al., 2009). This lncRNA is transcribed in the opposite direction from the *NR2F1* gene and is found to overlap with a focal deletion commonly seen in patients with ASD/ID. This association was further confirmed in a case harbouring a paternally inherited translocation highlighting the potential clinical relevance in neurodevelopmental disorders. Studies in mouse models have shown that *lnc-Nr2f1* promotes neuronal maturation pathways in a distinct manner from its neighbouring gene *NR2F1*. Furthermore, RNA Immunoprecipitation sequencing assays looking at histone H3 enrichment of associated lncRNA alongside chromatin localisation assays have revealed that *lnc-NR2F1* regulates target genes through in *trans* binding (Ang et al., 2019).

Individual 12_5 (I12_5) is a 16-year-old affected female recruited to 100KGP as part of a duo with an unaffected mother in 2016 under the disease term "Intellectual disability" (Appendix Table S5.2). The proband had several HPO terms reported: velopharyngeal insufficiency, moderate intellectual disability, delayed speech and language development, global developmental delay, delayed gross motor and fine motor skills, brachycephaly, congenital horizontal nystagmus, congenital strabismus, generalised hypopigmentation, congenital cystic adenomatoid malformation of the lung, and 2-3 toe cutaneous syndactyly. Moreover, the records in the RE indicated that the proband has undergone a-CGH, karyotype, and FISH prior to WGS. The tests results reported four imbalances in the proband but not the specific findings of these tests. However, inspecting the WGS data and the SVRare family report revealed a maternally inherited unbalanced translocation t(9,17)(p24.3,q25.3), as well as a gain of ~363 kb on chromosome 2, and a loss of ~664 kb on chromosome 5 that occur in the paternal allele and not present in the mother. The inherited unbalanced translocation involved a loss of the p-ter of chromosome 9 and duplication of the q-ter of chromosome 17. Inspection of the regions of losses or gains for genes linked to disease or intolerant to dosage change identified eight genes across all regions with *NR2F1* and *lncNR2F1* as the primary candidates (Table 5.4).

Four studies have reported deletions in 5q15 with the minimal region shared between them encompassing *NR2F1* and *lncNR2F1* (Ang et al., 2019; Brown et al., 2009; Cardoso et al., 2008; Malan et al., 2006). The shared phenotypes include developmental delay, facial dysmorphism, hearing loss, hypotonia, and ophthalmological abnormalities. However, extensive functional studies on *NR2F1* and *lnc-NR2F1* in differentiated neural progenitor cell from human stem cells as well as in mouse embryonic stem (mES) cells suggest the contribution of the *lnc-NR2F1* rather than the coding gene, *NR2F1*, towards the observed phenotypes. The functional studies assessed both gain and loss of function as well as chromatin localisation of *lnc-NR2F1* and *NR2F1*. The results showed that *lnc-NR2F1* acts in trans to affect gene expression. Moreover, the phenotype of the *lnc-Nr2f1* KO can be rescued by overexpression of *lnc-Nr2f1* mRNA, which does not affect *Nr2f1* expression in *lnc-Nr2f1* KO cells. Furthermore, ChIRP-sequencing [4] of *lnc-Nr2f1* in differentiated neurons from

[4] Chromatin Isolation by RNA Purification Sequencing

mES cells show that *lnc-Nr2f1* does not bind to the promoter region of *Nr2f1* (Ang et al., 2019). In light of this, loss of *lncNR2F1* can be considered the main candidate to explain the proband's phenotypes but the role of haploinsufficiency of *NR2F1* cannot be ignored.

Table 5.4: List of regions with loss or gain as detected by WGS in I12_5. The losses and gains in chromosomes 9 and 17 were considered as benign as they are inherited from an unaffected mother

CNV	CHR	Coordinates	Gene symbol	Gene name	OMIM disease	pLI	TS
Gain	**2**	13,382-375,335	-	-	-	-	-
Loss	**5**	93,535,098-94,199,119	***NR2F1-AS1***	NR2F1 antisense RNA 1	-	-	-
			NR2F1	nuclear receptor subfamily 2 group F member 1	Bosch-Boonstra-Schaaf optic atrophy syndrome, AD, OMIM # 615722	0.99	0.93
			KIAA0825	KIAA0825	Polydactyly, postaxial, type A10, AR, OMIM #618498	0	0.14
Loss	**9**	112,741-1,804,549	*DOCK8*	dedicator of cytokinesis 8	Hyper-IgE recurrent infection syndrome, AR, OMIM # 243700	0	0.07
			KANK1	KN motif and ankyrin repeat domains 1	Cerebral palsy, spastic quadriplegic, 2, OMIM #612900	0	0.10
Gain	**17**	82,334,106-83,204,179	*CYBC1*	Cytochrome b-245 chaperone 1	Chronic granulomatous disease 5, autosomal recessive	0	0
			WDR45B	WD repeat domain 45B	Neurodevelopmental disorder with spastic quadriplegia and brain abnormalities with or without seizures (AR), OMIM #617977	0.16	0.29
			ZNF750	zinc finger protein 750	?Seborrhea-like dermatitis with psoriasiform elements, AD OMIM #610227	0.87	0.45
			TBCD	Tubulin folding cofactor D	Encephalopathy, progressive, early-onset, with brain atrophy and thin corpus callosum, AR, OMIM #617193	0	0.45

CNV= Copy number variant; AR=Autosomal recessive; AD=Autosomal dominant; pLI=Probability of intolerance to loss-of-function; TS= Triplosensitivity score.

Another example of a mechanism that affects gene expression is altering chromatin's three-dimensional organisation. The increase in using Hi-C techniques to analyse the 3D chromatin structure has allowed for understanding the mechanism by which various long-range regulatory elements interact in a non-linear manner to regulate gene expression. One method is forming

enhancer-promotor loops (Schoenfelder & Fraser, 2019) restricted to topologically associated domains separated by sites for CCCTC-binding factor insulator elements and short interspersed nuclear elements (Dixon et al., 2012). A classic example is the disruption of Forkhead box G1 (*FOXG1*). It encodes a transcriptional repressor protein known as a winged-helix protein, which is essential for the developing telencephalon in early embryonic development. Variants disrupting *FOXG1* directly or its regulatory region within a 2 Mb TAD result in a severe atypical form of Rett syndrome known as the congenital variant. Studies have continuously tried to narrow down the smallest region required for manifesting phenotype. This smallest region of overlap was finally refined to ~124 kb downstream of *FOXG1* using Hi-C and UMI-4C[5] techniques in 2021 (D'haene et al., 2021).

Individual 13_5 (I13_5): is a 16-year-old male presented with severe phenotypes, including microcephaly, abnormality of the corpus callosum, global developmental delay, delayed gross and fine motor development, delayed speech and language development, intellectual disability, seizures, inability to walk, central hypotonia, chorea, choreoathetosis, and spasticity (Appendix Table S5.2). A diagnosis of Rett syndrome, the congenital variant, was suspected, and targeted sequencing of *FOXG1* did not identify any coding variants. Other tests conducted prior to WGS included a-CGH and karyotype, which showed a normal copy number of the genome from a-CGH. However, the karyotype showed a large paracentric inversion involving cytogenetic bands q12 and q22 of chromosome 14 (46, XY, inv(14)(q12q22)). From WGS data, the cytogenetic bands encompassing the *de novo* 22.98 Mb inversion were refined to inv(14)(q12q22.1) with one of the BNDs ~75 kb downstream of *FOXG1* (Figure 5.14: A-B).

Based on this finding, DNA sample was obtained and the inversion was validated using dideoxy sequencing of the purified PCR product (primers in Table 2.2), which revealed a total of 21 bp deletion on the ends of the inversion (Figure 5.14: C). Accurately identifying and validating structural variants, specifically non-coding ones, are crucial for correct clinical interpretation.

[5] Hi-C =High-throughput Chromosome Conformation Capture; UMI-4C = Unique Molecular Identifier - Circular Chromosome Conformation Capture , which. is a variation of the Chromosome Conformation Capture (3C) technique that aims to provide a higher-resolution view of chromatin interactions by incorporating unique molecular identifiers (UMIs)

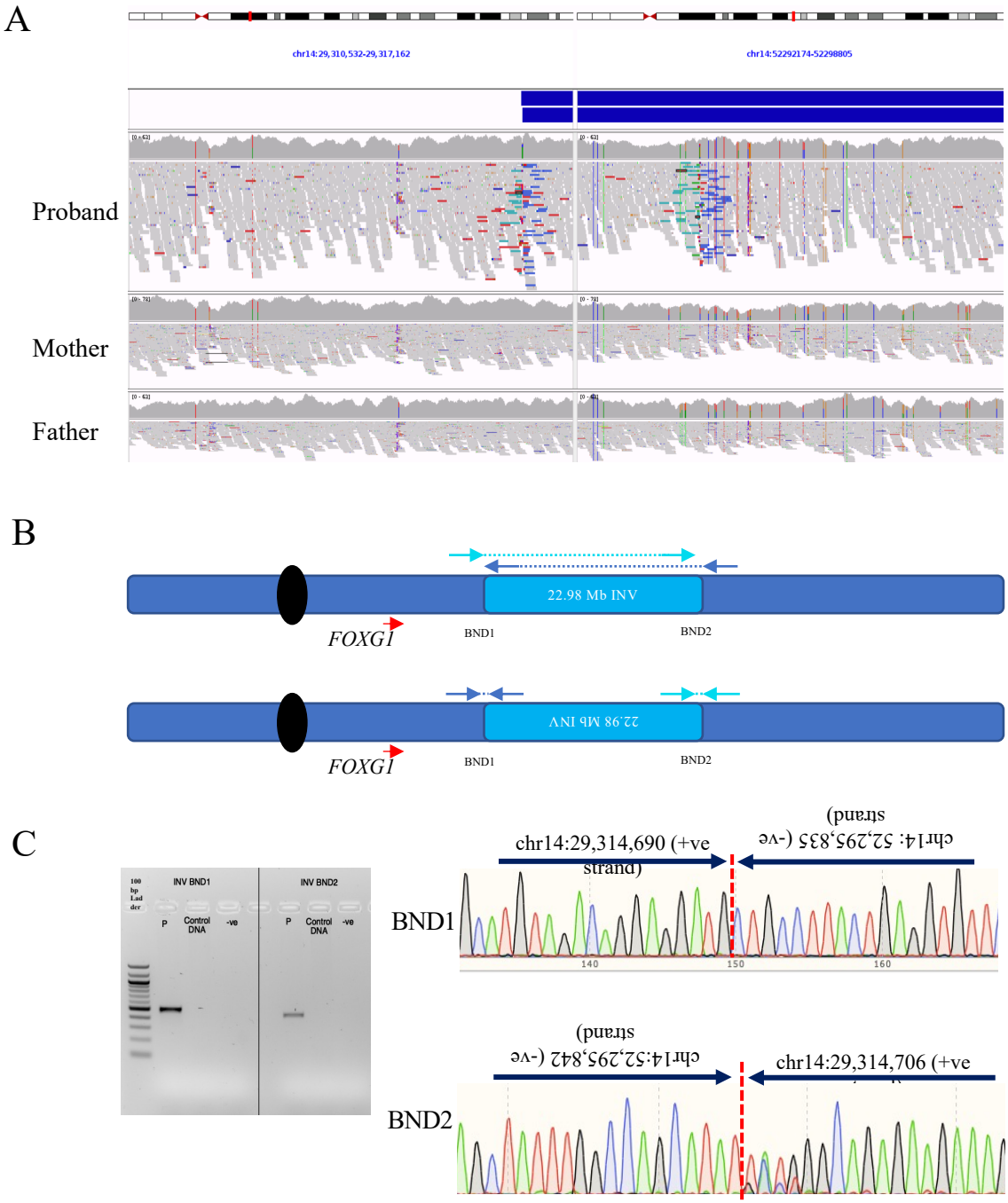


Figure 5.14: characterisation of a large inversion downstream of FOXG1. A) short-read alignment shows a 22.98 Mb de novo inversion ~75 kb downstream of FOXG1 in the proband that is absent in both parents. B) a linear schematic of the inversion and its position in relation to FOXG1. C) Validation of the SV. The agarose gel shows a band specific to the proband that has been subsequently dideoxy-sequenced. The Sanger traces indicates the presence of small deletions on both side of the inversion 15 bp (GRCh38(chr14):g.28845485_28845499del) and 6 bp (GRCh38(chr14):g.:51829120_51829125del).

5.4.8 Germline chromothripsis and chromoanasynthesis: Complex and Challenging Diagnosis

CCRs occur due to structural variants involving two or more chromosomes exchanging segments resulting in three or more breakpoints in each chromosome (Pellestor et al., 2011). However, this does not exclude that CCRs can occur in one chromosome just as complex (Poot & Haaf, 2015). The occurrence of *de novo* chromosomal rearrangements larger than 100 kb is estimated to be around 1.2×10^{-2} per meiosis, which is likely to decrease as the rearrangement size increases and becomes less compatible with life (Poot & Haaf, 2015). Earlier studies could not capture the extent of chromosomal rearrangements with low-resolution technologies. However, with the advancement of arrays and WGS, a study in 2011 was able to capture the massive extent of genomic rearrangements occurring in cancer tissue showing shattering and random reassembly of chromosomes introducing the term "chromothripsis" (Stephens et al., 2011). This CCR type was later also identified as part of constitutional rearrangements in patients with various developmental abnormalities (Kloosterman et al., 2011). The mechanism by which this type of rearrangement occurs has been speculated to involve aberrant DNA repair mechanisms, including microhomology-mediated break-induced replication and fork stalling and template switching (Carvalho & Lupski, 2016).

The mechanism by which chromothripsis occurs could lead to copy-neutral CCRs involving translocations and inversions but could also result in losses of genomic material. However, other CCRs with similar signatures to chromothripsis were found to have copy-number gain that the same mechanism cannot explain. Therefore, the latter cases with copy-number gain have been termed "chromoanasynthesis", which could also involve copy-number loss alongside the gain (Zhang et al., 2013). Nonetheless, the distinction between chromothripsis and chromoanasynthesis is yet to be fully elucidated. To further increase the complexity associated with these rearrangements, other cases with constitutional CCRs have been identified which do not fit into any of the groups mentioned above (Masset et al., 2016).

Interestingly, most *de novo* germline chromothripsis or chromoanasynthesis occur on the paternal chromosome, keeping with the hypothesis that there is high susceptibility to DNA damage and low repair capability during spermatogenesis (Pellestor et al., 2014).

In this study, two participants were recruited as part of a child-parent-parent trio to the 100KGP under “Balanced translocation with unusual phenotype” and were found to harbour highly complex chromosomal rearrangements. The karyotype result was available for one proband (case 8- Appendix Table S5.2), while I had to identify the breakpoints for case 9 (Appendix Table S5.2) blind to the karyotype report.

Individual 14_5 (I14_5): is a 24-year-old female presenting with moderate intellectual disability, autism, global developmental delay, scoliosis, cafe-au-lait patches, and a hypermelanotic macule. Before WGS, various tests were performed, including a-CGH, karyotype, and FISH for 22(q11 and q13). This identified an ABT t(4,20)(p15.2,p11.23) and der(22)(D22Z3+'D14/22Z1+, D22S9-,609C6-). While this case was one of two in which the karyotype results reflected the same cytogenetic bands confirmed via WGS, a higher complexity than anticipated was revealed, with a total of 11 junctions connecting fragments from chromosomes 4 and 20. These BNDs involved the boundaries of six deleted segments and one duplicated region, suggesting the occurrence of the chromoanasynthesis phenomenon (Figure 5.15). Four genes are disrupted by 5/11 junctions, while the remainder BNDs are in intergenic regions. The disrupted genes are Phospholipase C Beta 4 (*PLCB4*), p21 (RAC1) Activated Kinase 5 (*PAK5*), Synaptosome Associated Protein 25 (*SNAP25*), and ESF1 nucleolar pre-rRNA processing protein homolog (*ESF1*). Only *PLCB4* and *SNAP25* are associated with diseases linked to haploinsufficiency, Auriculocondylar Syndrome 2 (OMIM #614669) and Myasthenic syndrome congenital 18 (OMIM #616330), respectively. All breakpoints have been validated by dideoxy sequencing by the Wessex clinical genetics laboratory.

While it is tempting to assign a specific gene as the primary causative gene, CCRs exert a deleterious effect on several genes via different mechanisms. These include a direct effect via changing the dosage of gene expression due to gains, losses, or interrupting the in *cis* or in *trans* regulatory machinery. The indirect effect of structural variations can occur through the formation of fusion genes and changes in the position of chromosomes within the nucleus. Chromosomes with large rearrangements, such as translocations, occupy different territories within the nucleus than their wild type counterpart (Wang et al., 2022), which can affect the level of chromatin condensation and, in turn, gene expression (Matarazzo et al., 2007).

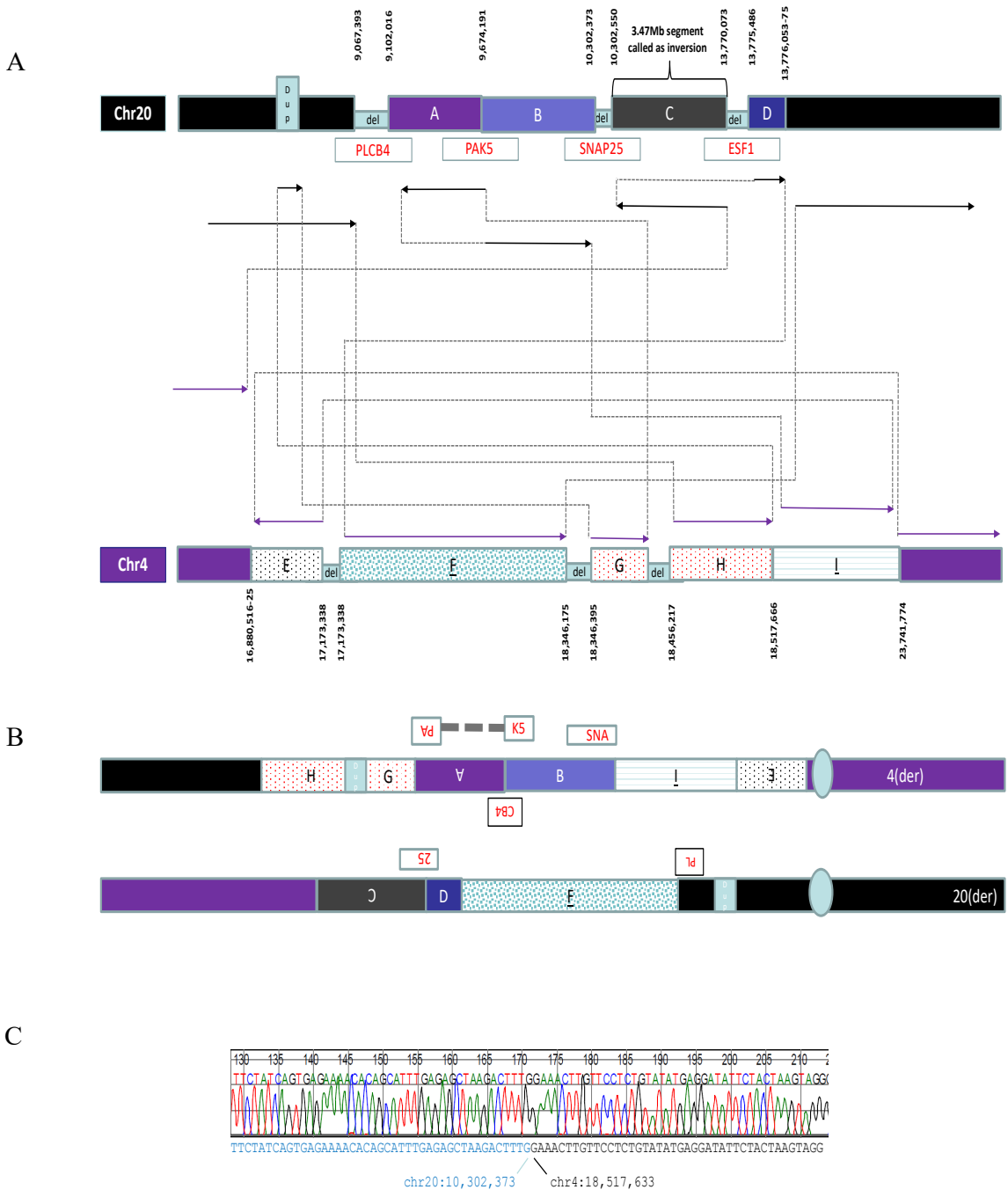


Figure 5.15: Linear schematic of possible chromoanasynthesis event in a female proband with developmental defects. A) Subway plot showing the breakpoints on chromosomes involved in the rearrangement and genes disrupted. The subway plot was generated manually assembling each breakpoints B) schematic of the derivative chromosomes after rearrangement. C) Sanger sequence trace for one of the breakpoints disrupting SNAP25. All breakpoints were dideoxy sequencing validated.

Individual 15_5 (I15_5): is a 22-year-old female presenting with failure to thrive in infancy, delayed fine and gross motor development, intellectual disability, severe global developmental delay, delayed speech and language development, macrocephaly, mutism, autistic behaviour, myopia, astigmatism, and high palate. Although karyotyping results indicated a balanced translocation, the report was unavailable to direct the analysis. A rare *de novo* ~1.3 Mb inversion in chromosome 2 was identified via the SVRare pipeline, and upon visualisation of alignments on IGV, BNDs mapping to chromosome 6 flanking the inversions were identified. Interrogating the SV.vcf file of the proband following the process described in methods, quickly unravelled highly complex rearrangements involving chromosomes 2, 6 and 9. No copy-number changes were identified, and only copy-neutral variants were detected as part of the rearrangement indicating signatures of chromothripsis with 19 breakpoints, and an inferred mechanism of NHEJ for the majority (Figure 5.16).

This CCR presented a few challenges to resolving the configuration. One of the main obstacles was the inability to identify the mate-read for two BNDs precisely. Only one BND was identified in GRCh37(chr9):g.19420038 and GRCh37(chr6):g.15785388 with an ambiguous mate-read that does not map to a unique position in the reference genome build 38. A BLAST search of the mate-read sequence against the T2T reference genome revealed that both sequences align to the same location on chromosome 9 within a highly repetitive region, but the results are not definitive. One way to overcome this obstacle facing short-read sequencing is by performing long-read sequencing, therefore, fresh DNA samples for high molecular weight extraction were obtained for Nanopore sequencing. However, the sequencing data were unavailable at the time of submission of thesis.

Another challenge is the scarcity of algorithms that can accurately identify and reconstruct the structure of shattered chromosomes that have undergone extensive rearrangement. As a result, manual reassembly of these events may not fully capture the complexity of the structural change.

Moreover, this type of CCR poses a significant challenge when interpreting the clinical implications of the findings. Along with the genes amplified, fused, or disrupted by the breakpoints, it is also essential to consider the positional effects that can contribute to human disease. These position effects are caused by changes in chromatin organization and can affect transcriptional regulatory networks. However, recent technological advances have made it possible to predict these position effects by modelling chromatin topology changes and their impact on gene expression (Zepeda-

Mendoza et al., 2017; Bianco et al., 2018; Schwessinger et al., 2020). This triage through modelling can reduce the amount of analytical work required in the lab and provide additional candidates to explain the observed phenotype. It is important to remember that multiple genes can contribute to the phenotype and that the contribution of each chromoanagenesis junction should be analysed and integrated into the final diagnosis.

Gaining insight into the prevalence, effects on local and distant regions, and functional implications of highly CCRs is crucial for comprehending the intricacies of human genetics. Such knowledge will be instrumental in identifying genes and genomic regions that are clinically significant. The increasing use of clinical genome sequencing is expected to uncover more information about the untapped potential of complex structural variations and their impact on human health and disease.

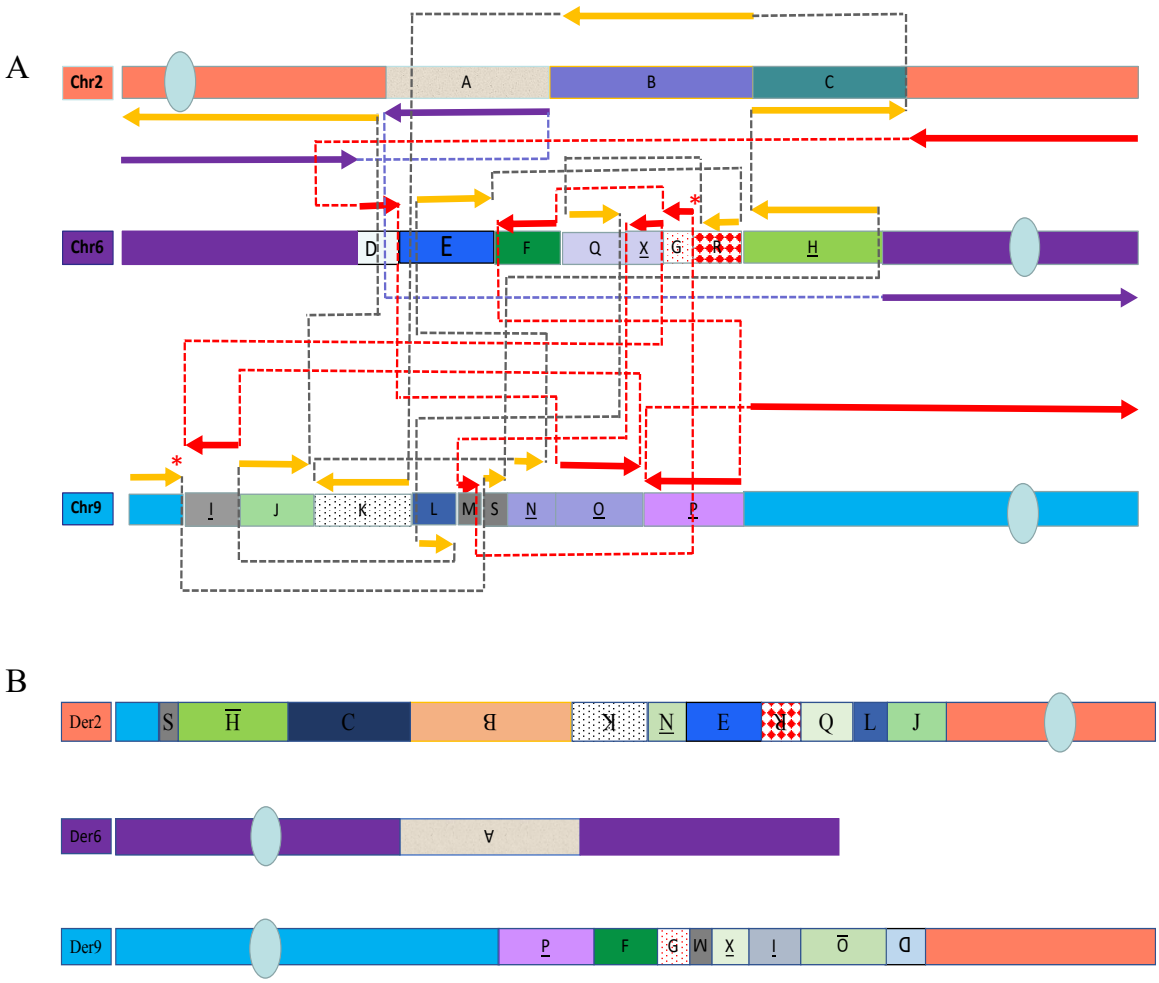


*Figure 5.16: Linear schematic of the rearrangements in chromothripsis in I15_5. A) Subway plot showing the process by which the shattered chromosomal fragments were rearranged. There are more than 20 breakpoints involved in the process. B) Schematic reconstruction of derivative chromosomes after reorganisation. The * denotes BNDs for which the position of mate-read could not be discerned directly from short-read data mapped to reference genome build 38 as they fall within highly repetitive genomic regions.*

5.5 Discussion

In conclusion, the work described in this chapter has delved into the clinical utility of WGS in identifying structural variations, specifically those that may be missed by traditional methods. The study has emphasized that WGS should be a standard part of the diagnostic workup for patients with developmental defects, especially those with unexplained genetic disorders. However, this study has also highlighted some limitations that require attention in future research. The aim is to continue to optimize the diagnostic yield of WGS and to provide more accurate and detailed information about the impact of structural variations on human health and disease.

One limitation identified in this chapter is the lack of fully utilising detection and prioritisation algorithms targeting medium-sized SVs larger than 50 bp but smaller than 10 kb, such as Manta, in the reporting pipeline adopted by the 100KGP. Currently Canvas is used which misses these medium-sized SVs as it is affected by several factors such as the SV size and coverage depth. Therefore, it is recommended that future reviews of the reporting pipeline should incorporate pipelines such as SVRare that mitigate many of the challenges facing SV reporting due to the large number of Manta calls per genome for reporting and potentially avoiding missed diagnoses. The SVRare pipeline used for SV prioritisation in this chapter utilises datasets from large sequencing projects of normal variation such as gnomAD and take advantage of large number of participants in the 100KGP to generate AF. This helps in providing more comprehensive, well-annotated SV calls to prioritise rare variants.

Another limitation facing the diagnosis of patients, especially those with multisystemic blended phenotypes, is the lack of comprehensive, informative phenotypic terms. The incomplete information would affect which gene panels are applied to the WGS data of the patient for analysis. It is crucial to report new cases of structural variations in known genes to better delineate genotype-phenotype correlations. This characterisation will aid in identifying the specific contribution of each variant to the phenotype and provide more detailed information about the impact of SVs on human health and disease. Moreover, WGS data of unsolved cases should be automated for reanalysis with updated gene panels with newly described disease-causing genes, to prevent subjecting patients to avoidable longer diagnostic odyssey.

Additionally, many inversions either simple or as part of a cxSV have been identified in this study, highlighting that they are under-explored in analytical pipelines as they are copy-neutral. Moreover, special consideration is required for primer design to ensure reliable and accurate validation of this particular group of SVs. Unlike traditional primer design, where one forward and one reverse primer are used, both primer pairs in the validation of inversions must have the same orientation. This design aspect is crucial in achieving proper amplification and detecting the presence of the inversion in the sample.

Short-read sequencing can detect variants robustly, especially deletions that do not fall. Within highly repetitive regions. However, its performance deteriorates with duplications larger than 200 bp and for complex variants. Their conformation cannot be unequivocally deduced as reads cannot span either ends of the duplication or the multiple breakpoints of the cxSV, which is be vital for clinical interpretation. Therefore, determining the correct conformation of such SVs in clinical settings would require a multifaceted approach depending on the type of SV. These methods include using long-read sequencing technologies such as Oxford Nanopore and PacBio, which can help address some of the limitations of short-read sequencing. Another effective method for identifying and confirming SVs is genome optical mapping, like BioNano. This approach is not constrained by the same limitations as sequencing technologies, particularly regarding repetitive genome regions, which can pose challenges for both short- and long-read sequencing methods. Currently, the main obstacles facing their implementation in clinical settings are the cost and long-read sequence algorithms being under development and having a high error rate. However, the ongoing advancements in cost-effectiveness and error reduction could potentially facilitate these technologies to be implemented on a larger scale.

Furthermore, it is vital to consider using bioinformatics tools to accurately detect catastrophic events and reconstruct the structure of shattered chromosomes when analysing sequencing data from patients with developmental defects and not be restricted to cancer genomes. In recent years, it has become evident that such complex rearrangements play a significant role in the development of rare diseases, and these tools should be applied to identify this underrepresented group of pathogenic SVs.

Overall, this chapter has highlighted the importance of WGS in identifying a range of structural variations and the need for continued research to fully understand these events' implications in human genetics. This is demonstrated by the identification of several hard-to-detect SVs with direct clinical implications. These included informing of reproductive options available, closure for parents of deceased patients, and better understanding of the patient's phenotype. An example is the cxSV in *KMT2E*, which led to the revision of the diagnosis to ODLURO syndrome in the patient. Furthermore, facilitated for a better delineation of phenotype-genotype correlation of recently reported disease-causing genes as in the case of *CUX1*. The discovery of various novel disease-gene associations such as *PPP2R2A* and proposing other potential candidates for future functional studies. The disruption of *PPP2R2A* did not only bring an end to a very long diagnostic odyssey of 33 years in a multigenerational kinship but also facilitated for the identification of several other patients from independent families. By addressing the limitations identified in this chapter and implementing the recommendations for improvement, the diagnostic yield of WGS can be further improved for patients with rare neurodevelopmental disorders or epilepsies.

Chapter 6. Discussion and conclusion

This chapter concludes the thesis, highlighting the overall summary of results, directions for future work, and concluding remarks.

6.1 Summary of results

The work of this thesis tried to address some of the fundamental issues facing diagnosis in rare neurodevelopmental disorders. These include the lack of knowledge of causative genes, hence identifying novel disease genes and under-explored structural variants becomes imperative. The main aims are to: (i) identify novel disease-gene associations, (ii) functionally validate the novel disease-causing variants using various molecular techniques and cellular models, and finally (iii) detect and characterise structural variants.

In Chapter 3, I utilised the Variant Explorer pipeline, which combines multiple complementary algorithms and prioritisation tools to interrogate the different variant types. While the pipeline is not unique, the integration of multiple tools enabled a comprehensive examination of both the coding and non-coding segments of the genome. This integration strengthened the confidence in identified variants by consolidating evidence from separate prioritisation tools that rely on the patient's specific phenotype profile, including tools such as Exomiser and GADO. The use of this approach for patients from the HICF2 and 100KGP programmes who had been pre-screened for suspected genes in clinical laboratories led to the discovery of several potential novel disease genes that have now been either confirmed or proposed for further functional investigation. These genes include *KMT2E, PLPPR5,* and *MGRN1.* Further two novel disease-causing gene associations including *PPP2R2A* and *ECPAS* were identified in chapter 5. In these instances, data-sharing have played an important role in facilitating for the rapid clinical translation of these finding. Data-sharing allows researchers from different labs to cross-validate evidence for a gene's association with the disease, as seen with *KMT2E*, and refine disease phenotypes, as in the case of *CUX1*. This accelerated translation to clinical applications enabled molecular diagnosis for individuals with previously undiagnosed conditions.

Moreover, several previously missed diagnoses involving newly described genes, such as *DOCK7* and *CUX1,* have been identified through this study. A possible reason for such missed diagnoses is the lag time between reporting newly described disease-causing genes and updating targeted gene panels used for genetic screening in clinical laboratories. The delays between reporting and updating the panels that could range from few months to several years. For instance, *CUX1* was initially described as a causative gene for non-syndromic global developmental delay in July 2018 (Platzer et al., 2018b), however, it was not approved as a "green" gene in PanelApp for intellectual disability panel until July 2019 and for the "Severe Paediatric Disorders" panel in February 2020. Therefore, patients screened during this period would not have a report for variants in this gene until their data had been reanalysed. However, in the case of *DOCK7*, the process was lengthier as the gene was first described for Global developmental delay with or without impaired intellectual development in 2014 (Perrault et al., 2014) and approved for the "Intellectual Disability" panel in February 2016, for the "Early onset or syndromic epilepsy" in November 2017, and for the "Severe Paediatric Disorders" panel in February 2020. Therefore, depending on which panel was applied to the analysis of these patients, the delay in finding the molecular diagnosis will be impacted. This demonstrates a strong need for a more efficient pipeline for flagging genes with high evidence for expedited inclusion in *in silico* panels and emphasises the importance of periodic re-analysis for unresolved cases.

In Chapter 3, I also described the successful resolution of two challenging cases by pinpointing elusive SVs. Detecting such variants from WGS data can be particularly difficult. High GC-rich regions often result in uneven read depth, while the short sequencing reads can pose challenges in accurately mapping reads to highly repetitive genomic regions. To address these complexities, the Variant Explorer pipeline utilises a combination of algorithms based on distinct theoretical models. This approach significantly mitigated the risk of false-positive (FP) CNVs without compromising sensitivity, ultimately leading to the identification of a previously overlooked diagnostic deletion in the *ARX* gene and a subsequent revision of the clinical diagnosis for the patient. Nonetheless, there are several FPs variants in *ARX* in 100KGP involving the first exon which highlights this difficult GC rich and repetitive region. Missed diagnoses often entail prolonged diagnostic journeys, which is a recurring theme among rare disease patients recruited to the HICF2 and 100KGP programmes.

Even with access to genome sequencing data, many patients experienced diagnostic odysseys lasting for more than six years. This underscores the need for improved prioritisation of structural variants and sheds light on the challenges associated with validating complex structural variations.

Moreover, while prediction tools give a very good estimate of the effect of splicing variants, it is important to carry out functional validation to confirm their effect. The choice of validation method can significantly influence the results, as shown in the case of the canonical splice donor site affecting *MGRN1*. In this instance, validation was conducted through minigene assay, Sanger sequencing of RT-PCR products, and RNA-seq. Each of these methods contributed towards the understanding of the effects of the variant. The simplest and most efficient method is the sequencing of RT-PCR product; however, it relies on the availability of the patient's sample. Moreover, it is a PCR based method, which has the natural tendency to preferentially amplify shorter fragments more efficiently (Shagin et al., 1999). On the other hand, minigene assays alleviated the need for the patient's sample (N. Sharma et al., 2014) but limited the genomic context being tested for the splicing effect. Both techniques supported the prediction of exon-skipping as result of the variant, which was also seen in RNA-seq data. In addition, RNA-seq revealed an intron retention effect of the splicing variant that was not seen by the other two methods proving a superior method for splicing assessment over both RT-PCR and minigene (Lord & Baralle, 2021).

In Chapter 4, the association of *PLPPR5* with JS was further explored. Here, I performed extensive functional validations of the effect of compound heterozygous variants in *PLPPR5* identified in a JS patient. The experiments confirmed localisation of the protein to the plasma membrane, which was previously known (Broggini et al., 2010; Coiro et al., 2014; Gross et al., 2022) and suggested that PLPPR5 is trafficked to the plasma membrane through galT-positive Golgi vesicles. The patient variants were found to reduce the amount of protein reaching the plasma membrane and the number of filopodia-like extensions. Another novel association established in this study is the link between PLPPR5 and autophagy. I have shown that the presence of variants was associated with an increased expression of lysosomal and autophagosomal markers. Furthermore, the variants have significantly impaired the morphology, network and function of mitochondria. These results together show a complex cellular interaction contributing to the development of JS in the patient.

Interestingly, it is plausible that a similar mechanism to that observed with *PLPPR5* can be invoked for recessive variants in *MGRN1* leading to the syndrome. However, further exploration of this association will be necessary in future investigations. Nonetheless, the function of *MGRN1* points towards impaired autophagy in the cells as well as mitochondrial dysfunction as shown in the literature from studies on KO mouse models (Gunn et al., 2019). Taken together, the results from cell biological investigations of *PLPPR5* and *MGRN1* suggest a mechanism for JS involving an interplay between impaired autophagy and mitochondrial homeostasis leading to the phenotype exhibited by the patients. Interestingly, a similar pattern of upregulated autophagy was found in patients with JME, which has a strong phenotypic overlap with JS (Tanaka et al., 2018).

In Chapter 5, I further investigated the contribution of structural variants to the development of neurodevelopmental disorders and epilepsy. Here, I looked at both copy-changing and copy-neutral variants using a novel pipeline SVRare developed by Dr Jing Yu specifically for large databases such as the 100KGP. In the initial work, I undertook a gene-based approach focusing on genes with compelling evidence for haploinsufficiency from ClinGen database that are also approved in PanelApp for genetic epilepsy. This restricted the analysis to 71 genes in which any gene disruptive variant would be considered pathogenic in patients with phenotypes consistent with the reported disease. This resulted in the identification of 66 diagnostic SVs, including complex ones, in unsolved cases by the 100KGP pipeline emphasising the improved prioritisation rate of SVRare pipeline relying on Manta calls for such events. However, it also highlighted that short-read sequencing data, while a powerful tool, has its limitations in resolving the structure of complex variants and that long-read sequencing is needed for resolving ambiguities to help enhance clinical interpretation. The example of a maternally inherited complex SV in *MECP2* affecting the last exon emphasizes the importance of the last coding exon, previously reported in the literature (Gardner et al., 2021; Redin et al., 2017; Lebo et al., 2001).

The molecular findings of two pathogenic variants in *DHX30* and *ANKRD11* in a patient with various neurodevelopmental phenotypes as well as cardiac abnormalities is one of many cases encountered. While *ANKRD11* could explain the presence of cardiac phenotypes (Parenti et al., 2021), both genes could contribute to the neurodevelopmental ones underscoring the need for comprehensive genetic

assessments and the close interaction between researchers and clinicians to guide analyses and diagnoses.

Another aspect examined in this chapter is a focused analysis on translocations. This was undertaken to evaluate i) the efficacy of using whole-genome sequencing to identify precise breakpoint junctions and ii) catalogue the range of complexity associated with what appears to be balanced translocations under the microscope.

Breakpoints were identified in 20/22 cases, leading to the discovery of disease-causing genes or mechanisms in 10 cases. Additional cryptic complexities, detected in seven cases, included two with suspected chromothripsis/chromoanasynthesis. For a couple of breakpoints in the complex cases, alignment to the reference genomes (build 38) presented uncertainty in mapping of the mate read. Consequently, Manta was unable to accurately identify the second breakpoint. Upon examination, it became evident that the mate read's sequence corresponded to a low-complexity repetitive region that corresponded to a gap in the used reference genome. Therefore, utilising the T2T (Telomere-to-Telomere) genome provided valuable insights. It offered a clearer indication of the region associated with the second breakpoint, thereby facilitating the resolution of the potential structure of the complex rearrangement.

Chromosomal abnormalities were reassigned to at least one sub-cytogenetic bands in 93% of cases, which is similar to results observed in a larger study on 273 subjects (Redin et al., 2017). This observation illustrates the inherent limitations of karyotyping, which can be attributed to several factors. These include the technical challenges in obtaining metaphase spreads and variability among different laboratory technicians or cytogeneticists, leading to subtle variations in band interpretation. These differences are further exacerbated by suboptimal chromosomal morphology, often associated with complex karyotypes or the presence of cryptic rearrangements. Additionally, the limited resolution of karyotyping can potentially introduce misinterpretations or yield false-negative results, particularly those involving near the chromosome ends (Bridge, 2008). The band mis-assignment from karyotyping may reflect missed opportunities for FISH testing of strong candidate genes.

Beyond sequencing data, complementary approaches such as RNAseq, protein informatics analysis, and TAD analysis offer deeper insights into the functional consequences of genetic variants. This was demonstrated by the analysis of suspicious balanced translocations where the breakpoints lie

within intergenic regions without direct disruption of genes. An example is *FOXG1* which was suspected by the clinician to be the candidate gene for a suspected Rett syndrome diagnosis in the patient but in which no variant was detected in the coding region. Whilst karyotyping did reveal a large inversion, the cytogenetic band was not accurately identified and led to the diagnosis being missed. In this case, TAD analysis in combination with similar cases reported in literature (Mencarelli et al., 2009; Ellaway et al., 2013; Perche et al., 2013; Takagi et al., 2013) showed that *FOXG1* is indeed the underlying cause of the phenotype through a positional effect.

The work in Chapter 5 emphasises the utility of WGS in resolving cases with a prior karyotype abnormality. Moreover, it highlights a wide range of cryptic complexity associated with translocations that were missed with other methods.

6.2 Research impact

In addition to increasing the diagnostic yield for patients, this thesis explores and adds to the broader research questions relating to NDDs and epilepsies.

An important contribution is the identification of novel disease-causing genes for neurodevelopmental disorders and epilepsy. According to the recent study on approximately 31K trios from research exome sequencing data, an estimate of 1,000 developmental disorder haploinsufficient genes are yet to be discovered (Kaplanis et al., 2020). This study contributed to the initial reporting of *KMT2E* as a disease-causing gene for a neurodevelopmental disorder (ODLURO), which is now an approved gene in PanelApp for screening the WGS data of patients with neurodevelopmental phenotypes. Furthermore, over 60 patients have now been reported in the literature and further seven patients received a diagnosis based on finding a variant in *KMT2E* in the 100KGP.

Another novel finding was that variants that disrupt *PPP2R2A* are a potential cause of ID phenotype. The disruption of this haploinsufficient gene in several members of an extending family co-segregated with the clinical phenotype observed and provided an immediate diagnosis for the family, ending a long diagnostic odyssey of 33 years. Moreover, it instigated a further screening of SNVs and SVs in this gene across the 100KGP as well as other several databases such as the DDD, and the 31K trios from research exome sequencing data identifying further three cases with two sharing

a recurrent stop gain mutation (NM_002717.4:c.457; p.R153*). Through international collaboration established via GeneMatcher, a further eight patients were identified to harbour *de novo* variants in *PPP2R2A* (six missense and two LoF variants). This provides further evidence for causality, however, patients with missense variants tend to exhibit a more severe phenotype than those with LoF variants. Nonetheless, a better understanding of the phenotype-genotype is required and currently being examined by the clinical collaborators. Such observation of a more severe form of the disorder due to missense variants compared protein-truncating variants can be explained by a dominant negative effect in which the presence of a mutated form of the protein maintaining partial function or interrupting the function of the WT protein might be more deleterious to the cells than the complete absence of a functional protein. Missense variants in *CHD1* lead to neurodevelopmental phenotypes while a deletion does not is an example of such dominant-negative mechanism (Pilarowski et al., 2018). Predicting the impact and interpreting missense or non-LoF variants poses greater challenges compared to LoF variants (Gerasimavicius et al., 2022). Consequently, there is a pressing demand for improved prediction tools to assess the consequences of non-LoF variants. The release of AlphFold2 (Jumper et al., 2021), a machine-learning three-dimensional protein structure prediction program based sequence information, has largely impacted the structural biology field and research. It facilitated for other tools such as Venus (Ferla et al., 2022) to predict the effect of non-LoF variants on the protein structure and function beyond destabilisation of the structure helping scientist to formulate hypotheses on the effect of variants identified.

In the interim, it becomes particularly crucial to consider the possibility of a dominant negative effect, especially when dealing with dominant disorders or when encountering heterozygous variants in a gene that is typically associated with a recessive disorder that would otherwise be easily dismissed. An established example is *ITPR1* heterozygous mutation in which cause spinocerebellar ataxia (McEntagart et al., 2016) and LoF (Gerber et al., 2016) variants cause Gillespie Syndrome. Another example is heterozygous variants in *GJB2* (connexin26) gene, which encodes one of six subunits that oligomerise to form a hemichannel. In this scenario, mutant protein oligomerises with the functional ones rendering the heterodimer non-functional, which in turn affects the formation

and function of the heteromeric Gap-Junction channels leading to deafness and hearing loss (Martínez et al., 2009).

Another contribution of this work is the refinement of phenotypes linked to newly described genes such as *CUX1.* Studies delineating the phenotype-genotype correlation are important in highlighting core features caused by disruption of the gene but also indicate the possibility of variants with variable penetrance in these genes. Transmitted variants from mildly affected or seemingly healthy parents to a severely affected child are usually dismissed as VUS, benign or not causative making interpretation of such variants difficult. A re-evaluation of carrier parents will be required as well as investigating modifier effects of other genes in the proband's genome that could contribute to the severity of their phenotypes.

Another result from my thesis pertains to the cell biology of *PLPPR5*, which enhances our understanding of the gene function and suggests a potential pathomechanism of JS. While KO *PLPPR5* has been studied in mouse models (Wang et al., 2021; Gross et al., 2022; Wang et al., 2022), there have been no investigations into human variants linked to disease to date. Therefore, this work represents the first functional validation of human variants in a gene that has not been implicated in any human disease so far.

The concluding significant contribution of this thesis revolves around the clinical applicability of WGS in detecting structural variations. I highlighted two categories of relatively under-investigated variants within the current diagnostic frameworks, both of which bear substantial clinical significance in the context of neurodevelopmental disorders: inversions and balanced translocations, given their copy-neutral nature.

The work in this thesis also demonstrates the importance of close interaction between clinicians, scientists and bioinformaticians in clinical and research settings. The provision of accurate HPO terms, and the results from any prior genetic testing facilitated for a better guidance for the analysis of genomic data. Moreover, the rapid advancement within the genomic medicine field required a close interaction between bioinformatics tool developers and the genomics scientists analysing the WGS data to improve analytical pipelines. This is shown here as different tools were used for the analysis and re-analysis of cases. Furthermore, the close communications with clinicians after the first pass analysis to propose candidates, which may be accepted or refuted, and the clinicians might

suggest specific candidate genes or pathways they suspect may be involved. Proposal of candidate genes might lead to clinical re-evaluation and further assessments for phenotypes that might have gone unnoticed until suggested by the molecular finding. This iterative process can be difficult to achieve in large-scale national programmes and with clinical laboratories being overwhelmed with existing testing or being limited in resources to explore variants outside known disease genes, the research community would offer a more suitable platform for exhausting the analysis of genomic data.

6.3 Limitations and barriers

Despite extensive efforts, there were several barriers to the validation of many identified structural variants. These included delays in validation of variants in research settings or in clinical laboratories. The latter was mainly due to the substantial load on clinical laboratories, which in several cases was further confounded by inaccessibility to bam files in order to design primers or lack of expertise in validating the less commonly encountered variants such as inversions that does not follow the standard primer designs. One of the main obstacles to validation in research settings was access to patient samples. For a large fraction of this study, the effect COVID19 pandemic meant that in-person clinical appointments were suspended resulting in delays of over 18 months to obtaining samples. Moreover, difficulty in obtaining samples from healthy individuals due to lack of ethical procedures being in place hampered the establishment of appropriate controls for RNAseq studies. Availability of 7500 RNAseq datasets, soon to be made available through the GEL RE should offer a solution to this, at least when examining cases from 100kGP.

Working with genomic data in the GEL RE added delays in contacting the recruiting clinician leading or indeed failure to establish the contact, despite repeated attempts. The reasons for this are due to the time gap between the recruitment, of patients which occurred between 2015-2018, and the potential research finding between 2020-2023 during which many clinical geneticists had retired, moved clinics, or patients had transferred from paediatric to adult clinics, a transition which required further consent to be obtained which were not manageable during the period of this thesis. Similar experience of other researchers using the 100KGP database have reported a low response rate of 20% (Best et al., 2022).

An example for this is the case of *MGRN1*, which was considered to be a strong potential disease-causing gene. Despite extensive efforts to find other patients with recessive variants in *MGRN1*, only one additional patient with homozygous missense variant was identified but the lack of comprehensive phenotypes and failed attempts to contact the clinician hindered progressing this association further without conducting functional studies to determine its involvement in epilepsy. The failed attempts to contact the clinician prevented the possibility of establishing patient-derived iPSC line.

6.4 Future work

The work and findings in this thesis lay the grounds for several potential follow up studies. One of the main next steps is to further characterise the effect of the balanced translocation where no clear candidates have been identified. This will be pursued through interrogation of RNA-seq data that are being deposited gradually by the 100KGP within the RE. This will give a deeper understanding of the effect of such large genomic changes on the whole transcriptome and looking at differentially expressed genes impacted by the translocations. Moreover, a plan for using long-read sequencing platforms such as PacBio, Nanopore, or BioNano has been initiated for unsolved cases with complexities linked to SVs that could not be resolved from the short-read sequencing data. The use of long-read sequencing should not only be limited to unresolved complex SVs but to be adopted as the primary approach in future once variant calling algorithms/ pipelines have reached a sufficient level of sensitivity and specificity.

The next phase of this study entails the functional validation of *MGRN1*'s role in epilepsy development. This will encompass the generation of a knock-in iPSC line using the CRISPR/Cas9 genome editing technique. Following this, a series of rigorous assessments for genomic stability and potential off-target effects will be conducted. These measures are vital to ensure that any subsequent observations are attributed solely to the introduced mutation and not influenced by the CRISPR/Cas9 editing process. Positively edited clones will be selected for neuronal differentiation into cortical neurons (2-D culture) and cortical assembloids that would better depict the 3-D interactions of cells in their microenvironment. The differentiated cortical neurons will be assayed at different maturation stages to establish the optimal *MGRN1* expression timepoint for informing the effect of the mutation

on the lysosomal and mitochondrial function. The expression in the WT and mutant cells will be compared using qPCR as well as through immunofluorescence staining of cells with antibodies against various neuronal markers alongside MGRN1. Once the optimal expression timepoint has been determined, lysosomal and mitochondrial morphology and function will be examined. Furthermore, the ubiquitinome will be studied to determine possible targets of MGRN1 for degradative and non-degradative ubiquitination. Furthermore, electrophysiological assessment using patch-clamp technique on mature neurons will be conducted to firmly establish the link between *MGRN1* and epileptic activity in mutant neurons.

In conclusion, genome sequencing is being adopted for neurodevelopmental disorders in clinical settings as a valuable technology for delivering genetic diagnoses. Nevertheless, limitations arise when the analysis is restricted solely to *in silico* gene panels or the coding regions of the genome, potentially overlooking pathogenic structural, splice site, and deep intronic variants. To fully exploit the diagnostic potential of genome sequencing, it is imperative to conduct a comprehensive analysis of the entire genome sequence. This approach ensures that the diagnostic yield of genome sequencing is maximised.

References

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*(6), 974–984. https://doi.org/10.1101/gr.114876.110

Acedo, A., Sanz, D. J., Durán, M., Infante, M., Pérez-Cabornero, L., Miner, C., & Velasco, E. A. (2012). Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes. *Breast Cancer Research : BCR*, *14*(3), R87. https://doi.org/10.1186/bcr3202

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013a). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *Chapter 7*, Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013b). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *Chapter 7*, Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76

Ahn, J. W., Mann, K., Walsh, S., Shehab, M., Hoang, S., Docherty, Z., Mohammed, S., & Mackie Ogilvie, C. (2010). Validation and implementation of array comparative genomic hybridisation as a first line test in place of postnatal karyotyping for genome imbalance. *Molecular Cytogenetics*, *3*, 9–9. PubMed. https://doi.org/10.1186/1755-8166-3-9

Aicardi. (1999). *Aicardi syndrome: Old and new findings*. *14*, 5–8.

Aicardi, Lefebvre, J, & LERIQUEK, A. (1965). *A new syndrome: Spasms in flexion, callosal agenesis, ocular abnormalities*. *19*((6), 609.

Alway, S. E. (2021). Linking mitochondrial dysfunction to sarcopenia. In *Sarcopenia* (pp. 1–58). Elsevier. https://doi.org/10.1016/B978-0-12-822146-4.00004-1

Ampuero, E., Dagnino-Subiabre, A., Sandoval, R., Zepeda-Carreño, R., Sandoval, S., Viedma, A., Aboitiz, F., Orrego, F., & Wyneken, U. (2007). Status epilepticus induces region-specific changes in dendritic spines, dendritic length and TrkB protein content of rat brain cortex. *Brain Research*, *1150*, 225–238. https://doi.org/10.1016/j.brainres.2007.02.089

Anderson, S., Menten, B., Kogelenberg, M. von, Robertson, S., Waginger, M., Mentzel, H.-J., Brandl, U., Skirl, G., & Willems, P. (2009). Aicardi Syndrome in a Male Patient. *Neuropediatrics*, *40*(01), 39–42. https://doi.org/10.1055/s-0029-1220760

Ang, C. E., Ma, Q., Wapinski, O. L., Fan, S., Flynn, R. A., Lee, Q. Y., Coe, B., Onoguchi, M., Olmos, V. H., Do, B. T., Dukes-Rimsky, L., Xu, J., Tanabe, K., Wang, L., Elling, U., Penninger, J. M., Zhao, Y., Qu, K., Eichler, E. E., … Chang, H. Y. (2019). The novel lncRNA lnc-NR2F1 is pro-neurogenic and mutated in human neurodevelopmental disorders. *eLife*, *8*, e41770. PubMed. https://doi.org/10.7554/eLife.41770

Annesley, S. J., & Fisher, P. R. (2019). Mitochondria in Health and Disease. *Cells*, *8*(7), 680. https://doi.org/10.3390/cells8070680

Aoyagi, Y., Kobayashi, H., Tanaka, K., Ozawa, T., Nitta, H., & Tsuji, S. (1999). A de novo splice donor site mutation causes in-frame deletion of 14 amino acids in the proteolipid protein in Pelizaeus-Merzbacher disease. *Annals of Neurology*, *46*(1), 112–115. https://doi.org/10.1002/1531-8249(199907)46:1<112::aid-ana16>3.0.co;2-u

APPG. (2016). *Undiagnosed genetic conditions and the impact of genome sequencing*. https://www.raredisease.org.uk/wp-content/ uploads/sites/7/2018/04/final-swan-graduate-report .pdf

Arellano, J. I., Benavides-Piccione, R., Defelipe, J., & Yuste, R. (2007). Ultrastructure of dendritic spines: Correlation between synaptic and spine morphologies. *Frontiers in Neuroscience*, *1*(1), 131–143. https://doi.org/10.3389/neuro.01.1.1.010.2007

Aristidou, C., Koufaris, C., Theodosiou, A., Bak, M., Mehrjouy, M. M., Behjati, F., Tanteles, G., Christophidou-Anastasiadou, V., Tommerup, N., & Sismani, C. (2017). Accurate Breakpoint Mapping in Apparently Balanced Translocation Families with Discordant Phenotypes Using Whole Genome Mate-Pair Sequencing. *PLOS ONE*, *12*(1), e0169935. https://doi.org/10.1371/journal.pone.0169935

Arppe, R., Carro-Temboury, M. R., Hempel, C., Vosch, T., & Just Sørensen, T. (2017). Investigating dye performance and crosstalk in fluorescence enabled bioimaging using a model system. *PloS One*, *12*(11), e0188359–e0188359. PubMed. https://doi.org/10.1371/journal.pone.0188359

Aspromonte, M. C., Bellini, M., Gasparini, A., Carraro, M., Bettella, E., Polli, R., Cesca, F., Bigoni, S., Boni, S., Carlet, O., Negrin, S., Mammi, I., Milani, D., Peron, A., Sartori, S., Toldo, I., Soli, F., Turolla, L., Stanzial, F., … Leonardi, E. (2019). Characterization of intellectual disability and autism comorbidity through gene panel sequencing. *Human Mutation*, *40*(9), 1346–1363. https://doi.org/10.1002/humu.23822

Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T. Y., Ghandi, M., Van Allen, E., Kryukov, G. V., Sboner, A., Theurillat, J.-P., Soong, T. D., Nickerson, E., Auclair, D., Tewari, A., Beltran, H., … Garraway, L. A. (2013). Punctuated evolution of prostate cancer genomes. *Cell*, *153*(3), 666–677. PubMed. https://doi.org/10.1016/j.cell.2013.03.021

Backx, L., Vermeesch, J., Pijkels, E., de Ravel, T., Seuntjens, E., & Van Esch, H. (2010). PPP2R2C, a gene disrupted in autosomal dominant intellectual disability. *European Journal of Medical Genetics*, *53*(5), 239–243. https://doi.org/10.1016/j.ejmg.2010.06.006

Bagher, P., Jiao, J., Owen Smith, C., Cota, C. D., & Gunn, T. M. (2006). Characterization of Mahogunin Ring Finger-1 expression in mice. *Pigment Cell Research*, *19*(6), 635–643. https://doi.org/10.1111/j.1600-0749.2006.00340.x

Bai, B., Guo, Y.-R., Zhang, Y.-H., Jin, C.-C., Zhang, J.-M., Chen, H., & Zhu, B.-S. (2019). Novel DOCK7 mutations in a Chinese patient with early infantile epileptic encephalopathy 23. *Chinese Medical Journal*, *132*(5), 600–603. https://doi.org/10.1097/CM9.0000000000000100

Baieli, S., Pavone, L., Meli, C., Fiumara, A., & Coleman, M. (2003). Autism and phenylketonuria. *Journal of Autism and Developmental Disorders*, *33*(2), 201–204. https://doi.org/10.1023/a:1022999712639

Ballabio, A., & Andria, G. (1992). Deletions and translocations involving the distal short arm of the human X chromosome: Review and hypotheses. *Human Molecular Genetics*, *1*(4), 221–227. https://doi.org/10.1093/hmg/1.4.221

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews. Genetics*, *12*(11), 745–755. https://doi.org/10.1038/nrg3031

Barcia, G., Fleming, M. R., Deligniere, A., Gazula, V.-R., Brown, M. R., Langouet, M., Chen, H., Kronengold, J., Abhyankar, A., Cilio, R., Nitschke, P., Kaminska, A., Boddaert, N., Casanova, J.-L., Desguerre, I., Munnich, A., Dulac, O., Kaczmarek, L. K., Colleaux, L., & Nabbout, R. (2012). De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. *Nature Genetics*, *44*(11), Article 11. https://doi.org/10.1038/ng.2441

Barros-Silva, D., Marques, C. J., Henrique, R., & Jerónimo, C. (2018). Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. *Genes*, *9*(9). https://doi.org/10.3390/genes9090429

Bartolini, E. (2021). Inherited Developmental and Epileptic Encephalopathies. *Neurology International*, *13*(4), 555–568. https://doi.org/10.3390/neurolint13040055

Bearden, D., Strong, A., Ehnot, J., DiGiovine, M., Dlugos, D., & Goldberg, E. M. (2014). Targeted treatment of migrating partial seizures of infancy with quinidine. *Annals of Neurology*, *76*(3), 457–461. https://doi.org/10.1002/ana.24229

Beaudet, A. L. (2013). The utility of chromosomal microarray analysis in developmental and behavioral pediatrics. *Child Development*, *84*(1), 121–132. https://doi.org/10.1111/cdev.12050

Beck-Wödl, S., Harzer, K., Sturm, M., Buchert, R., Rieß, O., Mennel, H.-D., Latta, E., Pagenstecher, A., & Keber, U. (2018). Homozygous TBC1 domain-containing kinase (TBCK) mutation causes a novel lysosomal storage disease – a new type of neuronal ceroid lipofuscinosis (CLN15)? *Acta Neuropathologica Communications*, *6*, 145. https://doi.org/10.1186/s40478-018-0646-6

Beenhakker, M., & Ritger, M. (2019). One Seizure Please, Hold the Sprouts: The Role of Hippocampal Mossy Fiber Sprouting in Epilepsy. *Epilepsy Currents*, *19*(6), 414–416. https://doi.org/10.1177/1535759719876556

Bélanger, S. A., & Caron, J. (2018). Evaluation of the child with global developmental delay and intellectual disability. *Paediatrics & Child Health*, *23*(6), 403–410. https://doi.org/10.1093/pch/pxy093

Belichenko, P. V., & Dahlström, A. (1995). Studies on the 3-dimensional architecture of dendritic spines and varicosities in human cortex by confocal laser scanning microscopy and Lucifer Yellow microinjections. *Journal of Neuroscience Methods*, *57*(1), 55–61. https://doi.org/10.1016/0165-0270(94)00125-Z

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., Langley, R. J., Zhang, L., Lee, C. C., Schilkey, F. D., Sheth, V., Woodward, J. E., Peckham, H. E., Schroth, G. P., Kim, R. W., & Kingsmore, S. F. (2011). Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing. *Science Translational Medicine*, *3*(65), 65ra4. https://doi.org/10.1126/scitranslmed.3001756

Berkel, S., Marshall, C. R., Weiss, B., Howe, J., Roeth, R., Moog, U., Endris, V., Roberts, W., Szatmari, P., Pinto, D., Bonin, M., Riess, A., Engels, H., Sprengel, R., Scherer, S. W., & Rappold, G. A. (2010). Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nature Genetics*, *42*(6), 489–491. https://doi.org/10.1038/ng.589

Best, S., Inglehearn, C. F., Watson, C. M., Toomes, C., Wheway, G., & Johnson, C. A. (2022). Unlocking the potential of the UK 100,000 Genomes Project—Lessons learned from analysis of the "Congenital Malformations caused by Ciliopathies" cohort. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics*, *190*(1), 5–8. https://doi.org/10.1002/ajmg.c.31965

Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Research*, *1380*, 42–77. https://doi.org/10.1016/j.brainres.2010.11.078

Bettella, E., Di Rosa, G., Polli, R., Leonardi, E., Tortorella, G., Sartori, S., & Murgia, A. (2013). Early-onset epileptic encephalopathy in a girl carrying a truncating mutation of the ARX gene: Rethinking the ARX phenotype in females. *Clinical Genetics*, *84*(1), 82–85. https://doi.org/10.1111/cge.12034

Bharadwaj, R., Peter, C. J., Jiang, Y., Roussos, P., Vogel-Ciernia, A., Shen, E. Y., Mitchell, A. C., Mao, W., Whittle, C., Dincer, A., Jakovcevski, M., Pothula, V., Rasmussen, T. P., Giakoumaki, S. G., Bitsios, P., Sherif, A., Gardner, P. D., Ernst, P., Ghose, S., … Akbarian,

S. (2014). Conserved higher-order chromatin regulates NMDA receptor gene expression and cognition. *Neuron*, *84*(5), 997–1008. PubMed. https://doi.org/10.1016/j.neuron.2014.10.032

Bianco, S., Lupiáñez, D. G., Chiariello, A. M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., Mundlos, S., & Nicodemi, M. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. *Nature Genetics*, *50*(5), 662–667. https://doi.org/10.1038/s41588-018-0098-8

Bienvenu, T., Poirier, K., Friocourt, G., Bahi, N., Beaumont, D., Fauchereau, F., Ben Jeema, L., Zemni, R., Vinet, M.-C., Francis, F., Couvert, P., Gomot, M., Moraine, C., van Bokhoven, H., Kalscheuer, V., Frints, S., Gecz, J., Ohzaki, K., Chaabouni, H., … Chelly, J. (2002). ARX, a novel Prd-class-homeobox gene highly expressed in the telencephalon, is mutated in X-linked mental retardation. *Human Molecular Genetics*, *11*(8), 981–991. https://doi.org/10.1093/hmg/11.8.981

Biesecker, L. G. (2010). Exome sequencing makes medical genomics a reality. *Nature Genetics*, *42*(1), 13–14. https://doi.org/10.1038/ng0110-13

Biesecker, L. G., & Spinner, N. B. (2013). A genomic view of mosaicism and human disease. *Nature Reviews Genetics*, *14*(5), Article 5. https://doi.org/10.1038/nrg3424

Bionano Genomics. (2018). *Bionano Solve Theory of Operation: Structural Variant Calling.Document Number: 30110 (Revision K)*. https://bionanogenomics.com/wp-content/uploads/2018/04/30110-Bionano-Solve-Theory-of-Operation-Structural-Variant-Calling.pdf

Bittencourt, S., Covolan, L., Hamani, C., Longo, B. M., Faria, F. P., Freymuller, E., Ottersen, O. P., & Mello, L. E. (2015). Replacement of Asymmetric Synaptic Profiles in the Molecular Layer of Dentate Gyrus Following Cycloheximide in the Pilocarpine Model in Rats. *Frontiers in Psychiatry*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyt.2015.00157

Blackwood, E. M., & Kadonaga, J. T. (1998). Going the distance: A current view of enhancer action. *Science (New York, N.Y.)*, *281*(5373), 60–63. https://doi.org/10.1126/science.281.5373.60

Blümcke, I., Zuschratter, W., Schewe, J. C., Suter, B., Lie, A. A., Riederer, B. M., Meyer, B., Schramm, J., Elger, C. E., & Wiestler, O. D. (1999). Cellular pathology of hilar neurons in

Ammon's horn sclerosis. *The Journal of Comparative Neurology*, *414*(4), 437–453. https://doi.org/10.1002/(sici)1096-9861(19991129)414:4<437::aid-cne2>3.0.co;2-3

Boguski, M. S., Lowe, T. M., & Tolstoshev, C. M. (1993). dbEST--database for 'expressed sequence tags'. *Nature Genetics*, *4*(4), 332–333. https://doi.org/10.1038/ng0893-332

Bonaglia, M. C., Giorda, R., Borgatti, R., Felisari, G., Gagliardi, C., Selicorni, A., & Zuffardi, O. (2001). Disruption of the ProSAP2 gene in a t(12;22)(q24.1;q13.3) is associated with the 22q13.3 deletion syndrome. *American Journal of Human Genetics*, *69*(2), 261–268. https://doi.org/10.1086/321293

Bräuer, A. U., & Nitsch, R. (2008a). Plasticity-related genes (PRGs/LRPs): A brain-specific class of lysophospholipid-modifying proteins. *Biochimica Et Biophysica Acta*, *1781*(9), 595–600. https://doi.org/10.1016/j.bbalip.2008.04.004

Bräuer, A. U., & Nitsch, R. (2008b). Plasticity-related genes (PRGs/LRPs): A brain-specific class of lysophospholipid-modifying proteins. *Biochimica Et Biophysica Acta*, *1781*(9), 595–600. https://doi.org/10.1016/j.bbalip.2008.04.004

Bridge, J. A. (2008). Advantages and limitations of cytogenetic, molecular cytogenetic, and molecular diagnostic testing in mesenchymal neoplasms. *Journal of Orthopaedic Science*, *13*(3), 273–282. https://doi.org/10.1007/s00776-007-1215-1

Briggs, J. A., Wolvetang, E. J., Mattick, J. S., Rinn, J. L., & Barry, G. (2015). Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. *Neuron*, *88*(5), 861–877. https://doi.org/10.1016/j.neuron.2015.09.045

Brindley, D. N. (2004). Lipid phosphate phosphatases and related proteins: Signaling functions in development, cell division, and cancer. *Journal of Cellular Biochemistry*, *92*(5), 900–912. https://doi.org/10.1002/jcb.20126

Broggini, T., Nitsch, R., & Savaskan, N. E. (2010a). Plasticity-related Gene 5 (PRG5) Induces Filopodia and Neurite Growth and Impedes Lysophosphatidic Acid– and Nogo-A–mediated Axonal Retraction. *Molecular Biology of the Cell*, *21*(4), 521–537. https://doi.org/10.1091/mbc.E09-06-0506

Broggini, T., Nitsch, R., & Savaskan, N. E. (2010b). Plasticity-related Gene 5 (PRG5) Induces Filopodia and Neurite Growth and Impedes Lysophosphatidic Acid– and Nogo-A–mediated

Axonal Retraction. *Molecular Biology of the Cell*, *21*(4), 521–537. https://doi.org/10.1091/mbc.E09-06-0506

Brosig, A., Fuchs, J., Ipek, F., Kroon, C., Schrötter, S., Vadhvani, M., Polyzou, A., Ledderose, J., van Diepen, M., Holzhütter, H.-G., Trimbuch, T., Gimber, N., Schmoranzer, J., Lieberam, I., Rosenmund, C., Spahn, C., Scheerer, P., Szczepek, M., Leondaritis, G., & Eickholt, B. J. (2019). The Axonal Membrane Protein PRG2 Inhibits PTEN and Directs Growth to Branches. *Cell Reports*, *29*(7), 2028-2040.e8. https://doi.org/10.1016/j.celrep.2019.10.039

Brown, K. K., Alkuraya, F. S., Matos, M., Robertson, R. L., Kimonis, V. E., & Morton, C. C. (2009). NR2F1 deletion in a patient with a de novo paracentric inversion, inv(5)(q15q33.2), and syndromic deafness. *American Journal of Medical Genetics. Part A*, *149A*(5), 931–938. PubMed. https://doi.org/10.1002/ajmg.a.32764

Buckmaster, P. S. (2014). Does Mossy Fiber Sprouting Give Rise to the Epileptic State? In H. E. Scharfman & P. S. Buckmaster (Eds.), *Issues in Clinical Epileptology: A View from the Bench* (pp. 161–168). Springer Netherlands. https://doi.org/10.1007/978-94-017-8914-1_13

Buratti, E., Baralle, M., & Baralle, F. E. (2006). Defective splicing, disease and therapy: Searching for master checkpoints in exon definition. *Nucleic Acids Research*, *34*(12), 3494–3510. https://doi.org/10.1093/nar/gkl498

Bush, L. W., Bartoshesky, L. E., David, K. L., Wilfond, B., Williams, J. L., Holm, I. A., & ACMG SELI committee. (2018). Pediatric clinical exome/genome sequencing and the engagement process: Encouraging active conversation with the older child and adolescent: points to consider-a statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *20*(7), 692–694. https://doi.org/10.1038/gim.2018.36

Cardoso, C., Boys, A., Parrini, E., Mignon-Ravix, C., McMahon, J. M., Khantane, S., Bertini, E., Pallesi, E., Missirian, C., Zuffardi, O., Novara, F., Villard, L., Giglio, S., Chabrol, B., Slater, H. R., Moncla, A., Scheffer, I. E., & Guerrini, R. (2008). Periventricular heterotopia, mental retardation, and epilepsy associated with 5q14.3-q15 deletion. *Neurology*, *72*(9), 784–792. https://doi.org/10.1212/01.wnl.0000336339.08878.2d

Carson, A. R., Smith, E. N., Matsui, H., Brækkan, S. K., Jepsen, K., Hansen, J.-B., & Frazer, K. A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics*, *15*(1), 125. https://doi.org/10.1186/1471-2105-15-125

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Reviews Genetics*, *3*(4), Article 4. https://doi.org/10.1038/nrg775

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews. Genetics*, *17*(4), 224–238. PubMed. https://doi.org/10.1038/nrg.2015.25

Caulfield, Mark, Davies, Jim, Dennys, Martin, Elbahy, Leila, Fowler, Tom, Hill, Sue, & et, al. (2017). *The National Genomics Research and Healthcare Knowledgebase* [dataset]. figshare. https://doi.org/10.6084/m9.figshare.4530893.v6

Caumes, R., Smol, T., Thuillier, C., Balerdi, M., Lestienne-Roche, C., Manouvrier-Hanu, S., & Ghoumid, J. (2020). Phenotypic spectrum of SHANK2-related neurodevelopmental disorder. *European Journal of Medical Genetics*, *63*(12), 104072. https://doi.org/10.1016/j.ejmg.2020.104072

Chahrour, M., Jung, S. Y., Shaw, C., Zhou, X., Wong, S. T. C., Qin, J., & Zoghbi, H. Y. (2008). MeCP2, a Key Contributor to Neurological Disease, Activates and Represses Transcription. *Science (New York, N.Y.)*, *320*(5880), 1224–1229. https://doi.org/10.1126/science.1153252

Chatron, N., Giannuzzi, G., Rollat-Farnier, P.-A., Diguet, F., Porcu, E., Yammine, T., Uguen, K., Bellil, Z.-L., Zillhardt, J. L., Sorlin, A., Ader, F., Afenjar, A., Andrieux, J., Bardel, C., Calpena, E., Chantot-Bastaraud, S., Callier, P., Chelloug, N., Chopin, E., … Schluth-Bolard, C. (2020). *The enrichment of breakpoints in late-replicating chromatin provides novel insights into chromoanagenesis mechanisms*. https://doi.org/10.1101/2020.07.17.206771

Chatterjee, S., & Pal, J. K. (2009). Role of 5′-and 3′-untranslated regions of mRNAs in human diseases. *Biology of the Cell*, *101*(5), 251-262.

Chear, S., Perry, S., Wilson, R., Bindoff, A., Talbot, J., Ware, T. L., Grubman, A., Vickers, J. C., Pébay, A., Ruddle, J. B., King, A. E., Hewitt, A. W., & Cook, A. L. (2022). *Lysosomal*

alterations and decreased electrophysiological activity in CLN3 disease (966 bp deletion, E295K) patient-derived cortical neurons (p. 2022.04.28.489465). bioRxiv. https://doi.org/10.1101/2022.04.28.489465

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2015). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Cheng, J., Sahani, S., Hausrat, T. J., Yang, J.-W., Ji, H., Schmarowski, N., Endle, H., Liu, X., Li, Y., Böttche, R., Radyushkin, K., Maric, H. M., Hoerder-Suabedissen, A., Molnár, Z., Prouvot, P.-H., Trimbuch, T., Ninnemann, O., Huai, J., Fan, W., … Vogt, J. (2016). Precise Somatotopic Thalamocortical Axon Guidance Depends on LPA-Mediated PRG-2/Radixin Signaling. *Neuron*, *92*(1), 126–142. https://doi.org/10.1016/j.neuron.2016.08.035

Chess, S. (1977). Follow-up report on autism in congenital rubella. *Journal of Autism and Childhood Schizophrenia*, *7*(1), 69–81. https://doi.org/10.1007/BF01531116

Chiang, C., Jacobsen, J. C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., Mills, R. E., Kirby, A., Lindgren, A. M., Rudiger, S. R., McLaughlan, C. J., Bawden, C. S., Reid, S. J., Faull, R. L. M., Snell, R. G., Hall, I. M., Shen, Y., Ohsumi, T. K., Borowsky, M. L., … Talkowski, M. E. (2012). Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nature Genetics*, *44*(4), 390–397, S1. https://doi.org/10.1038/ng.2202

Chowdhury, M. R., Singh, A., & Dubey, S. (2020). Chapter 24—Role of cytogenetics and molecular genetics in human health and medicine. In A. S. Verma & A. Singh (Eds.), *Animal Biotechnology (Second Edition)* (pp. 481–501). Academic Press. https://doi.org/10.1016/B978-0-12-811710-1.00022-7

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Coban-Akdemir, Z., White, J. J., Song, X., Jhangiani, S. N., Fatih, J. M., Gambin, T., Bayram, Y., Chinn, I. K., Karaca, E., Punetha, J., Poli, C., Boerwinkle, E., Shaw, C. A., Orange, J. S., Gibbs, R. A., Lappalainen, T., Lupski, J. R., & Carvalho, C. M. B. (2018). Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *The American Journal of Human Genetics*, *103*(2), 171–187. https://doi.org/10.1016/j.ajhg.2018.06.009

Coiro, P., Stoenica, L., Strauss, U., & Bräuer, A. U. (2014a). Plasticity-related gene 5 promotes spine formation in murine hippocampal neurons. *The Journal of Biological Chemistry*, *289*(36), 24956–24970. https://doi.org/10.1074/jbc.M114.597880

Coiro, P., Stoenica, L., Strauss, U., & Bräuer, A. U. (2014b). Plasticity-related gene 5 promotes spine formation in murine hippocampal neurons. *The Journal of Biological Chemistry*, *289*(36), 24956–24970. https://doi.org/10.1074/jbc.M114.597880

Collaer, M. L., & Hines, M. (1995). Human behavioral sex differences: A role for gonadal hormones during early development? *Psychological Bulletin*, *118*(1), 55–107. https://doi.org/10.1037/0033-2909.118.1.55

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. PubMed. https://doi.org/10.1038/s41586-020-2287-8

Collins, R. L., Glessner, J. T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.-M., Ulirsch, J., Everett, S., Howrigan, D. P., Boone, P. M., Fu, J., Karczewski, K. J., Kellaris, G., Lowther, C., Lucente, D., Mohajeri, K., … Esko, T. (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell*, *185*(16), 3041-3055.e25. https://doi.org/10.1016/j.cell.2022.06.036

Collins, R. L., Glessner, J. T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.-M., Ulirsch, J., Everett, S., Howrigan, D. P., Boone, P. M., Fu, J., Karczewski, K. J., Kellaris, G., Lowther, C., Lucente, D., Mohajeri, K., … Talkowski, M.

E. (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell*, *185*(16), 3041-3055.e25. https://doi.org/10.1016/j.cell.2022.06.036

Committee to Evaluate the Supplemental Security Income Disability Program for Children with Mental Disorders, Board on the Health of Select Populations, Board on Children, Youth, and Families, Institute of Medicine, Division of Behavioral and Social Sciences and Education, & The National Academies of Sciences, Engineering, and Medicine. (2015). *Mental Disorders and Disabilities Among Low-Income Children* (T. F. Boat & J. T. Wu, Eds.). National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK332882/

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., … Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nature Genetics*, *43*(9), 838–846. PubMed. https://doi.org/10.1038/ng.909

Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J., & Nickerson, D. A. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature Methods*, *7*(4), 250–251. https://doi.org/10.1038/nmeth0410-250

Córdova-Fletes, C., Domínguez, M. G., Delint-Ramirez, I., Martínez-Rodríguez, H. G., Rivas-Estilla, A. M., Barros-Núñez, P., Ortiz-López, R., & Neira, V. A. (2015). A de novo t(10;19)(q22.3;q13.33) leads to ZMIZ1/PRR12 reciprocal fusion transcripts in a girl with intellectual disability and neuropsychiatric alterations. *Neurogenetics*, *16*(4), 287–298. https://doi.org/10.1007/s10048-015-0452-2

Costa, T., Greer, W., Rysiecki, G., Buncic, J. R., & Ray, P. N. (1997). Monozygotic twins discordant for Aicardi syndrome. *Journal of Medical Genetics*, *34*(8), 688–691. https://doi.org/10.1136/jmg.34.8.688

Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., Bolduc, V., Waddell, L. B., Sandaradura, S. A., O'Grady, G. L., Estrella, E., Reddy, H. M., Zhao, F., Weisburd, B., Karczewski, K. J., O'Donnell-Luria, A. H., Birnbaum, D., Sarkozy, A., Hu, Y., … MacArthur, D. G. (2017). Improving genetic diagnosis in Mendelian disease with

transcriptome sequencing. *Science Translational Medicine*, *9*(386), eaal5209. https://doi.org/10.1126/scitranslmed.aal5209

Cundell, M. J., Hutter, L. H., Nunes Bastos, R., Poser, E., Holder, J., Mohammed, S., Novak, B., & Barr, F. A. (2016). A PP2A-B55 recognition signal controls substrate dephosphorylation kinetics during mitotic exit. *The Journal of Cell Biology*, *214*(5), 539–554. PubMed. https://doi.org/10.1083/jcb.201606033

Dagda, R. K., Cherra, S. J., Kulich, S. M., Tandon, A., Park, D., & Chu, C. T. (2009). Loss of PINK1 Function Promotes Mitophagy through Effects on Oxidative Stress and Mitochondrial Fission*. *Journal of Biological Chemistry*, *284*(20), 13843–13855. https://doi.org/10.1074/jbc.M808515200

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Shih-Chen, F., Aleksic, B., Biscaldi, M., Bolton, P. F., Brownfeld, J. M., Cai, J., … Buxbaum, J. D. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, *515*(7526), 209–215. https://doi.org/10.1038/nature13772

de Sainte Agathe, J.-M., Filser, M., Isidor, B., Besnard, T., Gueguen, P., Perrin, A., Van Goethem, C., Verebi, C., Masingue, M., Rendu, J., Cossée, M., Bergougnoux, A., Frobert, L., Buratti, J., Lejeune, É., Le Guern, É., Pasquier, F., Clot, F., Kalatzis, V., … Baux, D. (2023). SpliceAI-visual: A free online tool to improve SpliceAI splicing variant interpretation. *Human Genomics*, *17*(1), 7. https://doi.org/10.1186/s40246-023-00451-1

Deelen, P., van Dam, S., Herkert, J. C., Karjalainen, J. M., Brugge, H., Abbott, K. M., van Diemen, C. C., van der Zwaag, P. A., Gerkes, E. H., Zonneveld-Huijssoon, E., Boer-Bergsma, J. J., Folkertsma, P., Gillett, T., van der Velde, K. J., Kanninga, R., van den Akker, P. C., Jan, S. Z., Hoorntje, E. T., te Rijdt, W. P., … Franke, L. (2019). Improving the diagnostic yield of

exome- sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-10649-4

Dehaven, C. D., Evans, A. M., Dai, H., & Lawton, K. A. (2010). Organization of GC/MS and LC/MS metabolomics data into chemical libraries. *Journal of Cheminformatics*, *2*(1), 9. https://doi.org/10.1186/1758-2946-2-9

Deng, Y., & Pan, W. (2018). Significance Testing for Allelic Heterogeneity. *Genetics*, *210*(1), 25–32. https://doi.org/10.1534/genetics.118.301111

Devinsky, O., Vezzani, A., O'Brien, T. J., Jette, N., Scheffer, I. E., de Curtis, M., & Perucca, P. (2018a). Epilepsy. *Nature Reviews Disease Primers*, *4*(1), Article 1. https://doi.org/10.1038/nrdp.2018.24

Devinsky, O., Vezzani, A., O'Brien, T. J., Jette, N., Scheffer, I. E., de Curtis, M., & Perucca, P. (2018b). Epilepsy. *Nature Reviews. Disease Primers*, *4*, 18024. https://doi.org/10.1038/nrdp.2018.24

D'haene, E., & Vergult, S. (2021). Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. *Genetics in Medicine*, *23*(1), Article 1. https://doi.org/10.1038/s41436-020-00974-1

D'haene, Vantomme, L., & Menten, B., Callewaert, B., De Baere, E., Vergult, Sarah. (2021). Structural variants disrupt a critical regulatory region downstream of FOXG1. *EMBO Workshop 'Enhanceropathies : Understanding Enhancer Function to Understand Human Disease', Abstracts*.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Dikic, I., & Elazar, Z. (2018). Mechanism and medical implications of mammalian autophagy. *Nature Reviews. Molecular Cell Biology*, *19*(6), 349–364. https://doi.org/10.1038/s41580-018-0003-4

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R.,

Thomson, J. A., & Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331–336. https://doi.org/10.1038/nature14222

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. PubMed. https://doi.org/10.1038/nature11082

Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., Scheffler, K., van Vugt, J. J. F. A., French, C., Sanchis-Juan, A., Ibáñez, K., Tucci, A., Lajoie, B. R., Veldink, J. H., Raymond, F. L., … Eberle, M. A. (2019). ExpansionHunter: A sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics (Oxford, England)*, *35*(22), 4754–4756. https://doi.org/10.1093/bioinformatics/btz431

Dong, S., Walker, M. F., Carriero, N. J., DiCola, M., Willsey, A. J., Ye, A. Y., Waqar, Z., Gonzalez, L. E., Overton, J. D., Frahm, S., Keaney, J. F., Teran, N. A., Dea, J., Mandell, J. D., Hus Bal, V., Sullivan, C. A., DiLullo, N. M., Khalil, R. O., Gockley, J., … Sanders, S. J. (2014). De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Reports*, *9*(1), 16–23. https://doi.org/10.1016/j.celrep.2014.08.068

Donnenfeld, A. E., Packer, R. J., Zackai, E. H., Chee, C. M., Sellinger, B., & Emanuel, B. S. (1989). Clinical, cytogenetic, and pedigree findings in 18 cases of Aicardi syndrome. *American Journal of Medical Genetics*, *32*(4), 461–467. https://doi.org/10.1002/ajmg.1320320405

Dragich, J. M., Kuwajima, T., Hirose-Ikeda, M., Yoon, M. S., Eenjes, E., Bosco, J. R., Fox, L. M., Lystad, A. H., Oo, T. F., Yarygina, O., Mita, T., Waguri, S., Ichimura, Y., Komatsu, M., Simonsen, A., Burke, R. E., Mason, C. A., & Yamamoto, A. (2016). Autophagy linked FYVE (Alfy/WDFY3) is required for establishing neuronal connectivity in the mammalian brain. *eLife*, *5*, e14810. https://doi.org/10.7554/eLife.14810

Drets, M. E., & Shaw, M. W. (1971). Specific banding patterns of human chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *68*(9), 2073–2077. PubMed. https://doi.org/10.1073/pnas.68.9.2073

Ebbert, M. T. W., Jensen, T. D., Jansen-West, K., Sens, J. P., Reddy, J. S., Ridge, P. G., Kauwe, J. S. K., Belzil, V., Pregent, L., Carrasquillo, M. M., Keene, D., Larson, E., Crane, P., Asmann,

Y. W., Ertekin-Taner, N., Younkin, S. G., Ross, O. A., Rademakers, R., Petrucelli, L., & Fryer, J. D. (2019). Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology*, *20*(1), 97. https://doi.org/10.1186/s13059-019-1707-2

Eble, T. N., Sutton, V. R., Sangi-Haghpeykar, H., Wang, X., Jin, W., Lewis, R. A., Fang, P., & Van den Veyver, I. B. (2009). Non-random X chromosome inactivation in Aicardi syndrome. *Human Genetics*, *125*(2), 211–216. https://doi.org/10.1007/s00439-008-0615-4

Ekşioğlu, Y. Z., Pong, A. W., & Takeoka, M. (2011). A novel mutation in the aristaless domain of the ARX gene leads to Ohtahara syndrome, global developmental delay, and ambiguous genitalia in males and neuropsychiatric disorders in females. *Epilepsia*, *52*(5), 984–992. https://doi.org/10.1111/j.1528-1167.2011.02980.x

Ellaway, C. J., Ho, G., Bettella, E., Knapman, A., Collins, F., Hackett, A., McKenzie, F., Darmanian, A., Peters, G. B., Fagan, K., & Christodoulou, J. (2013). 14q12 microdeletions excluding FOXG1 give rise to a congenital variant Rett syndrome-like phenotype. *European Journal of Human Genetics*, *21*(5), 522–527. https://doi.org/10.1038/ejhg.2012.208

Ellingford, J. M., Ahn, J. W., & Bagnall, R. D. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med 14, 73 (2022)

Emberger, W., Petek, E., Kroisel, P. M., Zierler, H., & Wagner, K. (2001). Clinical and molecular cytogenetic characterization of two patients with partial trisomy 1q41-qter: Further delineation of partial trisomy 1q syndrome. *American Journal of Medical Genetics*, *104*(4), 312–318. https://doi.org/10.1002/ajmg.10096

Epi4K & EPGP. (2013). De novo mutations in the classic epileptic encephalopathies. *Nature*, *501*(7466), 217–221. https://doi.org/10.1038/nature12439

Epi25 Collaborative. Electronic address: s.berkovic@unimelb.edu.au & Epi25 Collaborative. (2019). Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. *American Journal of Human Genetics*, *105*(2), 267–282. https://doi.org/10.1016/j.ajhg.2019.05.020

EURORDIS. (2007). *Survey of the Delay in Diagnosis for 8 Rare Diseases in Europe ('EURORDISCARE 2')*. http://www.eurordis.org/sites/default/files/publications/Fact_Sheet_Eurordiscare2.pdf

Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M., & Milgram, E. (2009). Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical Chemistry*, *81*(16), 6656–6667. https://doi.org/10.1021/ac901536h

Evans, Bridgewater Br, , Liu Q2, , Mitchell MW1, , Robinson RJ1, , Dai H, Stewart SJ1, , DeHaven CD1, & Miller LAD. (2014). High Resolution Mass Spectrometry Improves Data Quantity and Quality as Compared to Unit Mass Resolution Mass Spectrometry in High-Throughput Profiling Metabolomics. *Journal of Postgenomics Drug & Biomarker Development*, *04*(02). https://doi.org/10.4172/2153-0769.1000132

Faundes, V., Newman, W. G., Bernardini, L., Canham, N., Clayton-Smith, J., Dallapiccola, B., Davies, S. J., Demos, M. K., Goldman, A., Gill, H., Horton, R., Kerr, B., Kumar, D., Lehman, A., McKee, S., Morton, J., Parker, M. J., Rankin, J., Robertson, L., … Banka, S. (2018). Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. *American Journal of Human Genetics*, *102*(1), 175–187. https://doi.org/10.1016/j.ajhg.2017.11.013

Faust, G. G., & Hall, I. M. (2014). SAMBLASTER: Fast duplicate marking and structural variant read extraction. *Bioinformatics (Oxford, England)*, *30*(17), 2503–2505. https://doi.org/10.1093/bioinformatics/btu314

Federici, G., & Soddu, S. (2020). Variants of uncertain significance in the era of high-throughput genome sequencing: A lesson from breast and ovary cancers. *Journal of Experimental & Clinical Cancer Research*, *39*(1), 46. https://doi.org/10.1186/s13046-020-01554-6

Ferla, M. P., Pagnamenta, A. T., Damerell, D., Taylor, J. C., & Marsden, B. D. (2020). MichelaNglo: Sculpting protein views on web pages without coding. *Bioinformatics (Oxford, England)*, *36*(10), 3268–3270. https://doi.org/10.1093/bioinformatics/btaa104

Ferla, M. P., Pagnamenta, A. T., Koukouflis, L., Taylor, J. C., & Marsden, B. D. (2022). Venus: Elucidating the Impact of Amino Acid Variants on Protein Function Beyond Structure Destabilisation. *Journal of Molecular Biology*, *434*(11), 167567. https://doi.org/10.1016/j.jmb.2022.167567

Flore, L. A., & Milunsky, J. M. (2012). Updates in the genetic evaluation of the child with global developmental delay or intellectual disability. *Seminars in Pediatric Neurology*, *19*(4), 173–180. https://doi.org/10.1016/j.spen.2012.09.004

Fourcade, S., López-Erauskin, J., Ruiz, M., Ferrer, I., & Pujol, A. (2014). Mitochondrial dysfunction and oxidative damage cooperatively fuel axonal degeneration in X-linked adrenoleukodystrophy. *Biochimie*, *98*, 143–149. https://doi.org/10.1016/j.biochi.2013.09.012

Frake, R. A., Ricketts, T., Menzies, F. M., & Rubinsztein, D. C. (2015). Autophagy and neurodegeneration. *The Journal of Clinical Investigation*, *125*(1), 65–74. https://doi.org/10.1172/JCI73944

Freunscht, I., Popp, B., Blank, R., Endele, S., Moog, U., Petri, H., Prott, E.-C., Reis, A., Rübo, J., Zabel, B., Zenker, M., Hebebrand, J., & Wieczorek, D. (2013). Behavioral phenotype in five individuals with de novo mutations within the GRIN2B gene. *Behavioral and Brain Functions : BBF*, *9*, 20. https://doi.org/10.1186/1744-9081-9-20

Galea, E., Launay, N., Portero-Otin, M., Ruiz, M., Pamplona, R., Aubourg, P., Ferrer, I., & Pujol, A. (2012). Oxidative stress underlying axonal degeneration in adrenoleukodystrophy: A paradigm for multifactorial neurodegenerative diseases? *Biochimica Et Biophysica Acta*, *1822*(9), 1475–1488. https://doi.org/10.1016/j.bbadis.2012.02.005

Gardner, E. J., Sifrim, A., Lindsay, S. J., Prigmore, E., Rajan, D., Danecek, P., Gallone, G., Eberhardt, R. Y., Martin, H. C., Wright, C. F., FitzPatrick, D. R., Firth, H. V., & Hurles, M. E. (2021). Detecting cryptic clinically relevant structural variation in exome-sequencing data increases diagnostic yield for developmental disorders. *American Journal of Human Genetics*, *108*(11), 2186–2194. PubMed. https://doi.org/10.1016/j.ajhg.2021.09.010

Gaszner, M., & Felsenfeld, G. (2006). Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nature Reviews. Genetics*, *7*(9), 703–713. https://doi.org/10.1038/nrg1925

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., & Price, A. L. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nature Genetics*, *49*(10), 1421–1427. https://doi.org/10.1038/ng.3954

Geerdink, N., Rotteveel, J. J., Lammens, M., Sistermans, E. A., Heikens, G. T., Gabreëls, F. J. M., Mullaart, R. A., & Hamel, B. C. J. (2002). MECP2 Mutation in a Boy with Severe Neonatal Encephalopathy: Clinical, Neuropathological and Molecular Findings. *Neuropediatrics*, *33*(1), 33–36. https://doi.org/10.1055/s-2002-23598

Genomes-Project-Pilot-Investigators, T. (2022). 100,000 genomes pilot on rare-disease diagnosis in health care preliminary report. *Yearbook of Paediatric Endocrinology*. https://doi.org/10.1530/ey.19.15.16

Genovese, G., Handsaker, R. E., Li, H., Altemose, N., Lindgren, A. M., Chambert, K., Pasaniuc, B., Price, A. L., Reich, D., Morton, C. C., Pollak, M. R., Wilson, J. G., & McCarroll, S. A. (2013). Using population admixture to help complete maps of the human genome. *Nature Genetics*, *45*(4), 406–414, 414e1-2. https://doi.org/10.1038/ng.2565

Gerasimavicius, L., Livesey, B. J., & Marsh, J. A. (2022). Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nature Communications*, *13*, 3895. https://doi.org/10.1038/s41467-022-31686-6

Gerber, S., Alzayady, K. J., Burglen, L., Brémond-Gignac, D., Marchesin, V., Roche, O., Rio, M., Funalot, B., Calmon, R., Durr, A., Gil-da-Silva-Lopes, V. L., Ribeiro Bittar, M. F., Orssaud, C., Héron, B., Ayoub, E., Berquin, P., Bahi-Buisson, N., Bole, C., Masson, C., … Fares Taie, L. (2016). Recessive and Dominant De Novo ITPR1 Mutations Cause Gillespie Syndrome. *American Journal of Human Genetics*, *98*(5), 971–980. https://doi.org/10.1016/j.ajhg.2016.03.004

Ghosh, R. P., Horowitz-Scherer, R. A., Nikitina, T., Gierasch, L. M., & Woodcock, C. L. (2008). Rett Syndrome-causing Mutations in Human MeCP2 Result in Diverse Structural Changes That Impact Folding and DNA Interactions. *The Journal of Biological Chemistry*, *283*(29), 20523–20534. https://doi.org/10.1074/jbc.M803021200

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W. M., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., de Vries, B. B. A., Kleefstra, T., Brunner, H. G., … Veltman, J. A. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, *511*(7509), 344–347. https://doi.org/10.1038/nature13394

Gillingham, A. K., Pfeifer, A. C., & Munro, S. (2002). CASP, the alternatively spliced product of the gene encoding the CCAAT-displacement protein transcription factor, is a Golgi membrane protein related to giantin. *Molecular Biology of the Cell*, *13*(11), 3761–3774. PubMed. https://doi.org/10.1091/mbc.e02-06-0349

Giorgi, F. S., Biagioni, F., Lenzi, P., Frati, A., & Fornai, F. (2015). The role of autophagy in epileptogenesis and in epilepsy-induced neuronal alterations. *Journal of Neural Transmission*, *122*(6), 849–862. https://doi.org/10.1007/s00702-014-1312-1

Gòdia, M., Casellas, J., Ruiz-Herrera, A., Rodríguez-Gil, J. E., Castelló, A., Sánchez, A., & Clop, A. (2020). Whole genome sequencing identifies allelic ratio distortion in sperm involving genes related to spermatogenesis in a swine model. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *27*(5), dsaa019. https://doi.org/10.1093/dnares/dsaa019

Gonorazky, H. D., Naumenko, S., Ramani, A. K., Nelakuditi, V., Mashouri, P., Wang, P., Kao, D., Ohri, K., Viththiyapaskaran, S., Tarnopolsky, M. A., Mathews, K. D., Moore, S. A., Osorio, A. N., Villanova, D., Kemaladewi, D. U., Cohn, R. D., Brudno, M., & Dowling, J. J. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *American Journal of Human Genetics*, *104*(3), 466–483. https://doi.org/10.1016/j.ajhg.2019.01.012

Gouw, M., Michael, S., Sámano-Sánchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L. B., Davey, N. E., Deng, Z., Diella, F., Gürth, C.-M., Huber, A.-K., Kleinsorg, S., Schlegel, L. S., Palopoli, N., Roey, K. V., Altenberg, B., Reményi, A., … Gibson, T. J. (2018). The eukaryotic linear motif resource—2018 update. *Nucleic Acids Research*, *46*(D1), D428–D434. https://doi.org/10.1093/nar/gkx1077

Graungaard, A. H., & Skov, L. (2007). Why do we need a diagnosis? A qualitative study of parents' experiences, coping and needs, when the newborn child is severely disabled. *Child: Care, Health and Development*, *33*(3), 296–307. https://doi.org/10.1111/j.1365-2214.2006.00666.x

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., McGuire, A. L., Nussbaum, R. L., O'Daniel, J. M., Ormond, K. E., Rehm, H. L., Watson, M. S., Williams, M. S., & Biesecker, L. G. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine*, *15*(7), 565–574. https://doi.org/10.1038/gim.2013.73

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., McGuire, A. L., Nussbaum, R. L., O'Daniel, J. M., Ormond, K. E., Rehm, H. L., Watson, M. S., Williams, M. S., Biesecker, L. G., & American College of Medical Genetics and Genomics. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *15*(7), 565–574. https://doi.org/10.1038/gim.2013.73

Gross, A. M., Ajay, S. S., Rajan, V., Brown, C., Bluske, K., Burns, N. J., Chawla, A., Coffey, A. J., Malhotra, A., Scocchia, A., Thorpe, E., Dzidic, N., Hovanes, K., Sahoo, T., Dolzhenko, E., Lajoie, B., Khouzam, A., Chowdhury, S., Belmont, J., … Taft, R. J. (2019). Copy-number variants in clinical genome sequencing: Deployment and interpretation for rare and undiagnosed disease. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *21*(5), 1121–1130. https://doi.org/10.1038/s41436-018-0295-y

Gross, I., Brandt, N., Vonk, D., Köper, F., Wöhlbrand, L., Rabus, R., Witt, M., Heep, A., Plösch, T., Hipp, M. S., & Bräuer, A. U. (2022). Plasticity-Related Gene 5 Is Expressed in a Late Phase of Neurodifferentiation After Neuronal Cell-Fate Determination. *Frontiers in Cellular Neuroscience*, *16*. https://www.frontiersin.org/articles/10.3389/fncel.2022.797588

Gross, I., Tschigor, T., Salman, A. L., Yang, F., Luo, J., Vonk, D., Hipp, M. S., Neidhardt, J., & Bräuer, A. U. (2022). Systematic expression analysis of plasticity-related genes in mouse brain development brings PRG4 into play. *Developmental Dynamics*, *251*(4), 714–728. https://doi.org/10.1002/dvdy.428

Gunn, T. M., Silvius, D., Lester, A., & Gibbs, B. (2019). Chronic and age-dependent effects of the spongiform neurodegeneration-associated MGRN1 E3 ubiquitin ligase on mitochondrial homeostasis. *Mammalian Genome : Official Journal of the International Mammalian Genome Society*, *30*(5–6), 151–165. https://doi.org/10.1007/s00335-019-09802-7

Gunter, T. E., & Pfeiffer, D. R. (1990). Mechanisms by which mitochondria transport calcium. *American Journal of Physiology-Cell Physiology*, *258*(5), C755–C786. https://doi.org/10.1152/ajpcell.1990.258.5.C755

Guo, H., Wang, T., Wu, H., Long, M., Coe, B. P., Li, H., Xun, G., Ou, J., Chen, B., Duan, G., Bai, T., Zhao, N., Shen, Y., Li, Y., Wang, Y., Zhang, Y., Baker, C., Liu, Y., Pang, N., … Xia, K. (2018). Inherited and multiple de novo mutations in autism/developmental delay risk genes suggest a multifactorial model. *Molecular Autism*, *9*, 64–64. PubMed. https://doi.org/10.1186/s13229-018-0247-z

Haberlandt, E., Valovka, T., Janjic, T., Müller, T., Blatsios, G., Karall, D., & Janecke, A. R. (2021). Characteristic facial features and cortical blindness distinguish the DOCK7-related epileptic encephalopathy. *Molecular Genetics & Genomic Medicine*, *9*(3), e1607. https://doi.org/10.1002/mgg3.1607

Halgren, C., Nielsen, N. M., Nazaryan-Petersen, L., Silahtaroglu, A., Collins, R. L., Lowther, C., Kjaergaard, S., Frisch, M., Kirchhoff, M., Brøndum-Nielsen, K., Lind-Thomsen, A., Mang, Y., El-Schich, Z., Boring, C. A., Mehrjouy, M. M., Jensen, P. K. A., Fagerberg, C., Krogh, L. N., Hansen, J., … Bache, I. (2018). Risks and Recommendations in Prenatally Detected De Novo Balanced Chromosomal Rearrangements from Assessment of Long-Term Outcomes. *American Journal of Human Genetics*, *102*(6), 1090–1103. PubMed. https://doi.org/10.1016/j.ajhg.2018.04.005

Hardie, D. G., Ross, F. A., & Hawley, S. A. (2012). AMPK: A nutrient and energy sensor that maintains energy homeostasis. *Nature Reviews. Molecular Cell Biology*, *13*(4), 251–262. https://doi.org/10.1038/nrm3311

Harrison, S. M., & Rehm, H. L. (2019). Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. *Genome Medicine*, *11*(1), 72. https://doi.org/10.1186/s13073-019-0688-9

Hashim, M., Stewart, H., Yu, J., Banos-Pinero, B., Consortium, G. E. R., Pagnamenta, A. T., & Taylor, J. C. (2023). Genome sequencing identifies KMT2E-disrupting cryptic structural variant in a female with O'Donnell-Luria-Rodan syndrome. *Clinical Genetics*, *104*(3), 390–392. https://doi.org/10.1111/cge.14355

Heuser, M., Yap, D. B., Leung, M., de Algara, T. R., Tafech, A., McKinney, S., Dixon, J., Thresher, R., Colledge, B., Carlton, M., Humphries, R. K., & Aparicio, S. A. (2009). Loss of MLL5 results in pleiotropic hematopoietic defects, reduced neutrophil immune function, and extreme sensitivity to DNA demethylation. *Blood*, *113*(7), 1432–1443. https://doi.org/10.1182/blood-2008-06-162263

Hines. (2005). *Brain gender*. Oxford University Press.

Hoag, H. M., Taylor, S. A., Duncan, A. M., & Khalifa, M. M. (1997). Evidence that skewed X inactivation is not needed for the phenotypic expression of Aicardi syndrome. *Human Genetics*, *100*(3–4), 459–464. https://doi.org/10.1007/s004390050534

Hoang, N., Buchanan, J. A., & Scherer, S. W. (2018). Heterogeneity in clinical sequencing tests marketed for autism spectrum disorders. *Npj Genomic Medicine*, *3*(1), Article 1. https://doi.org/10.1038/s41525-018-0066-3

Hochstenbach, R., Slunga-Tallberg, A., Devlin, C., Floridia, G., de Alba, M. R., Bhola, S., Rack, K., & Hastings, R. (2017). Fading competency of cytogenetic diagnostic laboratories: The alarm bell has started to ring. *European Journal of Human Genetics: EJHG*, *25*(3), 273–274. https://doi.org/10.1038/ejhg.2016.177

Hoffmann, S., Orlando, M., Andrzejak, E., Bruns, C., Trimbuch, T., Rosenmund, C., Garner, C. C., & Ackermann, F. (2019). Light-Activated ROS Production Induces Synaptic Autophagy. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *39*(12), 2163–2183. https://doi.org/10.1523/JNEUROSCI.1317-18.2019

Holland, A. J., & Cleveland, D. W. (2012). Chromoanagenesis and cancer: Mechanisms and consequences of localized, complex chromosomal rearrangements. *Nature Medicine*, *18*(11), 1630–1638. PubMed. https://doi.org/10.1038/nm.2988

Hopkins, I. J., Humphrey, I., Keith, C. G., Susman, M., Webb, G. C., & Turner, E. K. (1979). The Aicardi Syndrome in a 47, XXY Male. *Journal of Paediatrics and Child Health*, *15*(4), 278–280. https://doi.org/10.1111/j.1440-1754.1979.tb01246.x

Houge, G., Haesen, D., Vissers, L. E. L. M., Mehta, S., Parker, M. J., Wright, M., Vogt, J., McKee, S., Tolmie, J. L., Cordeiro, N., Kleefstra, T., Willemsen, M. H., Reijnders, M. R. F., Berland, S., Hayman, E., Lahat, E., Brilstra, E. H., van Gassen, K. L. I., Zonneveld-Huijssoon, E., … Janssens, V. (2015). B56δ-related protein phosphatase 2A dysfunction identified in patients with intellectual disability. *The Journal of Clinical Investigation*, *125*(8), 3051–3062. PubMed. https://doi.org/10.1172/JCI79860

Huang, P.-S., Ban, Y.-E. A., Richter, F., Andre, I., Vernon, R., Schief, W. R., & Baker, D. (2011). RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLOS ONE*, *6*(8), e24109. https://doi.org/10.1371/journal.pone.0024109

Hulea, L., & Nepveu, A. (2012). CUX1 transcription factors: From biochemical activities and cell-based assays to mouse models and human diseases. *Gene*, *497*(1), 18–26. https://doi.org/10.1016/j.gene.2012.01.039

Illumina. (n.d.). *An Introduction to Next-Generation Sequencing Technology*.

Imaizumi, T., Yamamoto-Shimojima, K., Yanagishita, T., Ondo, Y., & Yamamoto, T. (2020). Analyses of breakpoint junctions of complex genomic rearrangements comprising multiple consecutive microdeletions by nanopore sequencing. *Journal of Human Genetics*, *65*(9), 735–741. https://doi.org/10.1038/s10038-020-0762-6

International League Against Epilepsy Consortium on Complex Epilepsies. Electronic address: epilepsy-austin@unimelb.edu.au. (2014). Genetic determinants of common epilepsies: A meta-analysis of genome-wide association studies. *The Lancet. Neurology*, *13*(9), 893–903. https://doi.org/10.1016/S1474-4422(14)70171-1

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., … Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare

Missense Variants. *American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Ishihara, N., Jofuku, A., Eura, Y., & Mihara, K. (2003). Regulation of mitochondrial morphology by membrane potential, and DRP1-dependent division and FZO1-dependent fusion reaction in mammalian cells. *Biochemical and Biophysical Research Communications*, *301*(4), 891–898. https://doi.org/10.1016/S0006-291X(03)00050-0

Isokawa, M. (1998). Remodeling dendritic spines in the rat pilocarpine model of temporal lobe epilepsy. *Neuroscience Letters*, *258*(2), 73–76. https://doi.org/10.1016/s0304-3940(98)00848-9

Isokawa, M., & Levesque, M. F. (1991). Increased NMDA responses and dendritic degeneration in human epileptic hippocampal neurons in slices. *Neuroscience Letters*, *132*(2), 212–216. https://doi.org/10.1016/0304-3940(91)90304-c

Itan, Y., Shang, L., Boisson, B., Patin, E., Bolze, A., Moncada-Vélez, M., Scott, E., Ciancanelli, M. J., Lafaille, F. G., Markle, J. G., Martinez-Barricarte, R., de Jong, S. J., Kong, X.-F., Nitschke, P., Belkadi, A., Bustamante, J., Puel, A., Boisson-Dupuis, S., Stenson, P. D., … Casanova, J.-L. (2015). The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences*, *112*(44), 13615–13620. https://doi.org/10.1073/pnas.1518646112

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015

JAX. (2022). *Comparing Genetic Tests*. The Jackson Laboratory. https://www.jax.org/education-and-learning/clinical-and-continuing-education/clinical-topics/cancer-resources/genetic-testing-technology-comparison

Jensen, E. (2014). Technical review: In situ hybridization. *Anatomical Record (Hoboken, N.J.: 2007)*, *297*(8), 1349–1353. https://doi.org/10.1002/ar.22944

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 7873. https://doi.org/10.1038/s41586-021-03819-2

Kahles, A., Ong, C. S., Zhong, Y., & Rätsch, G. (2016). SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics (Oxford, England)*, *32*(12), 1840–1847. https://doi.org/10.1093/bioinformatics/btw076

Kannan, T. P., & Zilfalil, B. A. (2009). Cytogenetics: Past, Present And Future. *The Malaysian Journal of Medical Sciences : MJMS*, *16*(2), 4–9.

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., McRae, J. F., Short, P. J., Torene, R. I., de Boer, E., Danecek, P., Gardner, E. J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., … Retterer, K. (2020a). Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*, *586*(7831), Article 7831. https://doi.org/10.1038/s41586-020-2832-5

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., McRae, J. F., Short, P. J., Torene, R. I., de Boer, E., Danecek, P., Gardner, E. J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., … Retterer, K. (2020b). Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*, *586*(7831), Article 7831. https://doi.org/10.1038/s41586-020-2832-5

Kapoor, D., Anand, A., Siddiqui, S., & Sharma, S. (2022). A novel pathogenic variant in DOCK 7 gene in an infant with dysmorphism, epileptic encephalopathy and cortical blindness. *Clinical Dysmorphology*, *31*(1), 39–41. https://doi.org/10.1097/MCD.0000000000000393

Karam, S. M., Riegel, M., Segal, S. L., Félix, T. M., Barros, A. J. D., Santos, I. S., Matijasevich, A., Giugliani, R., & Black, M. (2015). Genetic causes of intellectual disability in a birth cohort: A population-based study. *American Journal of Medical Genetics. Part a*, *167*(6), 1204–1214. https://doi.org/10.1002/ajmg.a.37011

Karczewski, K. J., Francioli, L. C., & MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Yearbook of Paediatric Endocrinology*. https://doi.org/10.1530/ey.17.14.3

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), Article 7809. https://doi.org/10.1038/s41586-020-2308-7

Karimzadeh, M., Ernst, C., Kundaje, A., & Hoffman, M. M. (2018). Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Research*, *46*(20), e120. https://doi.org/10.1093/nar/gky677

Kato, M., Das, S., Petras, K., Kitamura, K., Morohashi, K., Abuelo, D. N., Barr, M., Bonneau, D., Brady, A. F., Carpenter, N. J., Cipero, K. L., Frisone, F., Fukuda, T., Guerrini, R., Iida, E., Itoh, M., Lewanda, A. F., Nanba, Y., Oka, A., … Dobyns, W. B. (2004). Mutations of ARX are associated with striking pleiotropy and consistent genotype–phenotype correlation. *Human Mutation*, *23*(2), 147–159. https://doi.org/10.1002/humu.10310

Kato, T., Kurahashi, H., & Emanuel, B. S. (2012). Chromosomal translocations and palindromic AT-rich repeats. *Current Opinion in Genetics & Development*, *22*(3), 221–228. https://doi.org/10.1016/j.gde.2012.02.004

Katsanis, S. H., & Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nature Reviews. Genetics*, *14*(6), 415–426. https://doi.org/10.1038/nrg3493

Kemp, S., Berger, J., & Aubourg, P. (2012). X-linked adrenoleukodystrophy: Clinical, metabolic, genetic and pathophysiological aspects. *Biochimica Et Biophysica Acta*, *1822*(9), 1465–1474. https://doi.org/10.1016/j.bbadis.2012.03.012

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kleinjan, D. (1998). Position effect in human genetic disease. *Human Molecular Genetics*, *7*(10), 1611–1618. https://doi.org/10.1093/hmg/7.10.1611

Kleinjan, D. A., & van Heyningen, V. (2005). Long-range control of gene expression: Emerging mechanisms and disruption in disease. *American Journal of Human Genetics*, *76*(1), 8–32. PubMed. https://doi.org/10.1086/426833

Klima, H., Ullrich, K., Aslanidis, C., Fehringer, P., Lackner, K. J., & Schmitz, G. (1993). A splice junction mutation causes deletion of a 72-base exon from the mRNA for lysosomal acid lipase in a patient with cholesteryl ester storage disease. *Journal of Clinical Investigation*, *92*(6), 2713–2718.

Klingenberg, M., & Rottenberg, H. (1977). Relation between the Gradient of the ATP/ADP Ratio and the Membrane Potential across the Mitochondrial Membrane. *European Journal of Biochemistry*, *73*(1), 125–130. https://doi.org/10.1111/j.1432-1033.1977.tb11298.x

Kloosterman, W. P., Guryev, V., van Roosmalen, M., Duran, K. J., de Bruijn, E., Bakker, S. C. M., Letteboer, T., van Nesselrooij, B., Hochstenbach, R., Poot, M., & Cuppen, E. (2011). Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline†. *Human Molecular Genetics*, *20*(10), 1916–1924. https://doi.org/10.1093/hmg/ddr073

Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M., & Seelow, D. (2011). Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PloS One*, *6*(11), e28240. https://doi.org/10.1371/journal.pone.0028240

Kornienko & Pronin. (2009). *Diagnostic Neuroradiology*. Springer Science.

Kovac, S., Dinkova Kostova, A. T., Herrmann, A. M., Melzer, N., Meuth, S. G., & Gorji, A. (2017). Metabolic and Homeostatic Changes in Seizures and Acquired Epilepsy—Mitochondria, Calcium Dynamics and Reactive Oxygen Species. *International Journal of Molecular Sciences*, *18*(9), Article 9. https://doi.org/10.3390/ijms18091935

Krayl, M., Lim, J. H., Martin, F., Guiard, B., & Voos, W. (2007). A Cooperative Action of the ATP-Dependent Import Motor Complex and the Inner Membrane Potential Drives Mitochondrial Preprotein Import. *Molecular and Cellular Biology*, *27*(2), 411–425. https://doi.org/10.1128/MCB.01391-06

Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T. B., Graf, E., Schwarzmayr, T., Terrile, C., Koňaříková, E., Repp, B., Kastenmüller, G., Adamski, J.,

Lichtner, P., Leonhardt, C., Funalot, B., Donati, A., Tiranti, V., … Prokisch, H. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications*, *8*, 15824. https://doi.org/10.1038/ncomms15824

Kroner, B. L., Preiss, L. R., Ardini, M.-A., & Gaillard, W. D. (2008). New incidence, prevalence, and survival of Aicardi syndrome from 408 cases. *Journal of Child Neurology*, *23*(5), 531–535. https://doi.org/10.1177/0883073807309782

Kuang, Q., Purhonen, P., & Hebert, H. (2015). Structure of potassium channels. *Cellular and Molecular Life Sciences: CMLS*, *72*(19), 3677–3693. https://doi.org/10.1007/s00018-015-1948-5

Kuhlenbäumer, G., Hullmann, J., & Appenzeller, S. (2011). Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Human Mutation*, *32*(2), 144–151. https://doi.org/10.1002/humu.21400

Kuivaniemi, H., Sabol, C., Tromp, G., Sippola-Thiele, M., & Prockop, D. J. (1988). A 19-base pair deletion in the pro-alpha 2(I) gene of type I procollagen that causes in-frame RNA splicing from exon 10 to exon 12 in a proband with atypical osteogenesis imperfecta and in his asymptomatic mother. *The Journal of Biological Chemistry*, *263*(23), 11407–11413.

Kukimoto-Niino, M., Tsuda, K., Ihara, K., Mishima-Tsumagari, C., Honda, K., Ohsawa, N., & Shirouzu, M. (2019a). Structural Basis for the Dual Substrate Specificity of DOCK7 Guanine Nucleotide Exchange Factor. *Structure (London, England: 1993)*, *27*(5), 741-748.e3. https://doi.org/10.1016/j.str.2019.02.001

Kukimoto-Niino, M., Tsuda, K., Ihara, K., Mishima-Tsumagari, C., Honda, K., Ohsawa, N., & Shirouzu, M. (2019b). Structural Basis for the Dual Substrate Specificity of DOCK7 Guanine Nucleotide Exchange Factor. *Structure*, *27*(5), 741-748.e3. https://doi.org/10.1016/j.str.2019.02.001

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

Kurahashi, H., Inagaki, H., Ohye, T., Kogo, H., Kato, T., & Emanuel, B. S. (2006). Palindrome-mediated chromosomal translocations in humans. *DNA Repair*, *5*(9–10), 1136–1145. https://doi.org/10.1016/j.dnarep.2006.05.035

Kurahashi, H., Shaikh, T. H., Zackai, E. H., Celle, L., Driscoll, D. A., Budarf, M. L., & Emanuel, B. S. (2000). Tightly Clustered 11q23 and 22q11 Breakpoints Permit PCR-Based Detection of the Recurrent Constitutional t(11;22). *The American Journal of Human Genetics*, *67*(3), 763–768. https://doi.org/10.1086/303054

Kurahashi, H., Shaikh, T., Takata, M., Toda, T., & Emanuel, B. S. (2003). The Constitutional t(17;22): Another Translocation Mediated by Palindromic AT-Rich Repeats. *American Journal of Human Genetics*, *72*(3), 733–738.

Kwong, A. K.-Y., Ho, A. C.-C., Fung, C.-W., & Wong, V. C.-N. (2015). Analysis of mutations in 7 genes associated with neuronal excitability and synaptic transmission in a cohort of children with non-syndromic infantile epileptic encephalopathy. *PloS One*, *10*(5), e0126446. https://doi.org/10.1371/journal.pone.0126446

Lai, C. S., Fisher, S. E., Hurst, J. A., Levy, E. R., Hodgson, S., Fox, M., Jeremiah, S., Povey, S., Jamison, D. C., Green, E. D., Vargha-Khadem, F., & Monaco, A. P. (2000). The SPCH1 region on human 7q31: Genomic characterization of the critical interval and localization of translocations associated with speech and language disorder. *American Journal of Human Genetics*, *67*(2), 357–368. PubMed. https://doi.org/10.1086/303011

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862-868. https://doi.org/10.1093/nar/gkv1222

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), Article 4. https://doi.org/10.1038/nmeth.1923

Laplante, M., & Sabatini, D. M. (2012). mTOR Signaling in Growth Control and Disease. *Cell*, *149*(2), 274–293. https://doi.org/10.1016/j.cell.2012.03.017

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., & Church, D. M. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Research*, *41*(Database issue), D936-941. https://doi.org/10.1093/nar/gks1213

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), R84. https://doi.org/10.1186/gb-2014-15-6-r84

Lebo, R. V., Ikuta, T., Milunsky, J. M., & Milunsky, A. (2001). Rett syndrome from quintuple and triple deletions within the MECP2 deletion hotspot region. *Clinical Genetics*, *59*(6), 406–417. https://doi.org/10.1034/j.1399-0004.2001.590605.x

Ledbetter, D. H., Riccardi, V. M., Airhart, S. D., Strobel, R. J., Keenan, B. S., & Crawford, J. D. (1981). Deletions of chromosome 15 as a cause of the Prader-Willi syndrome. *The New England Journal of Medicine*, *304*(6), 325–329. https://doi.org/10.1056/NEJM198102053040604

Lee, B. H., Smith, T., & Paciorkowski, A. R. (2015). Autism spectrum disorder and epilepsy: Disorders with a shared biology. *Epilepsy & Behavior: E&B*, *47*, 191–201. https://doi.org/10.1016/j.yebeh.2015.03.017

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), Article 7616. https://doi.org/10.1038/nature19057

Lenaerts, L., Reynhout, S., Verbinnen, I., Laumonnier, F., Toutain, A., Bonnet-Brilhault, F., Hoorne, Y., Joss, S., Chassevent, A. K., Smith-Hicks, C., Loeys, B., Joset, P., Steindl, K., Rauch, A., Mehta, S. G., Chung, W. K., Devriendt, K., Holder, S. E., Jewett, T., … Janssens, V. (2021).

The broad phenotypic spectrum of PPP2R1A-related neurodevelopmental disorders correlates with the degree of biochemical dysfunction. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *23*(2), 352–362. PubMed. https://doi.org/10.1038/s41436-020-00981-2

Leonard, H., & Wen, X. (2002). The epidemiology of mental retardation: Challenges and opportunities in the new millennium. *Mental Retardation and Developmental Disabilities Research Reviews*, *8*(3), 117–134. https://doi.org/10.1002/mrdd.10031

Li, & Durbin, R. (2009a). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2009b). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Y., Fan, L., Luo, R., Yang, Z., Yuan, M., Zhang, J., & Gan, J. (2021). Case Report: De novo Variants of KMT2E Cause O'Donnell-Luria-Rodan Syndrome: Additional Cases and Literature Review. *Frontiers in Pediatrics*, *9*, 641841. https://doi.org/10.3389/fped.2021.641841

Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., Kumar, K., Khurana, E., Waszak, S., Korbel, J. O., Haber, J. E., Imielinski, M., Weischenfeldt, J., Beroukhim, R., & Campbell, P. J. (2020). Patterns of somatic structural variation in human cancer genomes. *Nature*, *578*(7793), Article 7793. https://doi.org/10.1038/s41586-019-1913-9

Lindstrand, A., Eisfeldt, J., Pettersson, M., Carvalho, C. M. B., Kvarnung, M., Grigelioniene, G., Anderlid, B.-M., Bjerin, O., Gustavsson, P., Hammarsjö, A., Georgii-Hemming, P., Iwarsson, E., Johansson-Soller, M., Lagerstedt-Robinson, K., Lieden, A., Magnusson, M., Martin, M., Malmgren, H., Nordenskjöld, M., … Nilsson, D. (2019). From cytogenetics to cytogenomics: Whole-genome sequencing as a first-line test comprehensively captures the

diverse spectrum of disease-causing genetic variation underlying intellectual disability. *Genome Medicine*, *11*(1), 68. https://doi.org/10.1186/s13073-019-0675-1

Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., Pellecchia, G., Sung, W. W. L., Wang, Z., Bikangaga, P., Boelman, C., Carter, M. T., Cordeiro, D., Cytrynbaum, C., Dell, S. D., … Marshall, C. R. (2018a). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *20*(4), 435–443. https://doi.org/10.1038/gim.2017.119

Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., Pellecchia, G., Sung, W. W. L., Wang, Z., Bikangaga, P., Boelman, C., Carter, M. T., Cordeiro, D., Cytrynbaum, C., Dell, S. D., … Marshall, C. R. (2018b). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *20*(4), 435–443. https://doi.org/10.1038/gim.2017.119

Lipton, J. O., & Sahin, M. (2014). The neurology of mTOR. *Neuron*, *84*(2), 275–291. https://doi.org/10.1016/j.neuron.2014.09.034

Liu, Liu, S., Guo, S., Lu, W., Zhang, Q., & Cheng, J. (2022). Evolutionary dynamics and conserved function of the Tudor domain-containing (TDRD) proteins in teleost fish. *Marine Life Science & Technology*, *4*(1), 18–30. https://doi.org/10.1007/s42995-021-00118-7

Liu, P., Erez, A., Nagamani, S. C. S., Dhar, S. U., Kołodziejska, K. E., Dharmadhikari, A. V., Cooper, M. L., Wiszniewska, J., Zhang, F., Withers, M. A., Bacino, C. A., Campos-Acevedo, L. D., Delgado, M. R., Freedenberg, D., Garnica, A., Grebe, T. A., Hernández-Almaguer, D., Immken, L., Lalani, S. R., … Bi, W. (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*, *146*(6), 889–903. PubMed. https://doi.org/10.1016/j.cell.2011.07.042

Liu, R., Sun, L., Wang, Y., Wang, Q., & Wu, J. (2023). New use for an old drug: Quinidine in KCNT1-related epilepsy therapy. *Neurological Sciences*, *44*(4), 1201–1206. https://doi.org/10.1007/s10072-022-06521-x

Liu, X., Huai, J., Endle, H., Schlüter, L., Fan, W., Li, Y., Richers, S., Yurugi, H., Rajalingam, K., Ji, H., Cheng, H., Rister, B., Horta, G., Baumgart, J., Berger, H., Laube, G., Schmitt, U., Schmeisser, M. J., Boeckers, T. M., … Vogt, J. (2016a). PRG-1 Regulates Synaptic Plasticity via Intracellular PP2A/β1-Integrin Signaling. *Developmental Cell*, *38*(3), 275–290. https://doi.org/10.1016/j.devcel.2016.06.019

Liu, X., Huai, J., Endle, H., Schlüter, L., Fan, W., Li, Y., Richers, S., Yurugi, H., Rajalingam, K., Ji, H., Cheng, H., Rister, B., Horta, G., Baumgart, J., Berger, H., Laube, G., Schmitt, U., Schmeisser, M. J., Boeckers, T. M., … Vogt, J. (2016b). PRG-1 Regulates Synaptic Plasticity via Intracellular PP2A/β1-Integrin Signaling. *Developmental Cell*, *38*(3), 275–290. https://doi.org/10.1016/j.devcel.2016.06.019

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews. Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., & Pallen, M. J. (2012). High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. *Nature Reviews. Microbiology*, *10*(9), 599–606. https://doi.org/10.1038/nrmicro2850

Loomes, R., Hull, L., & 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 and Adolescent Psychiatry*, *56*(6), 466–474. https://doi.org/10.1016/j.jaac.2017.03.013

Lord, J., & Baralle, D. (2021). Splicing in the Diagnosis of Rare Disease: Advances and Challenges. *Frontiers in Genetics*, *12*. https://www.frontiersin.org/articles/10.3389/fgene.2021.689892

Loveday, C., Tatton-Brown, K., Clarke, M., Westwood, I., Renwick, A., Ramsay, E., Nemeth, A., Campbell, J., Joss, S., Gardner, M., Zachariou, A., Elliott, A., Ruark, E., van Montfort, R., Childhood Overgrowth Collaboration, & Rahman, N. (2015). Mutations in the PP2A

regulatory subunit B family genes PPP2R5B, PPP2R5C and PPP2R5D cause human overgrowth. *Human Molecular Genetics*, *24*(17), 4775–4779. PubMed. https://doi.org/10.1093/hmg/ddv182

Lukmanji, S., Manji, S. A., Kadhim, S., Sauro, K. M., Wirrell, E. C., Kwon, C.-S., & Jetté, N. (2019). The co-occurrence of epilepsy and autism: A systematic review. *Epilepsy & Behavior*, *98*, 238–248. https://doi.org/10.1016/j.yebeh.2019.07.037

Lunter, G., & Goodson, M. (2011a). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. https://doi.org/10.1101/gr.111120.110

Lunter, G., & Goodson, M. (2011b). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. https://doi.org/10.1101/gr.111120.110

Lv, X., Jiang, H., Li, B., Liang, Q., Wang, S., Zhao, Q., & Jiao, J. (2014). The crucial role of Atg5 in cortical neurogenesis during early brain development. *Scientific Reports*, *4*, 6010. https://doi.org/10.1038/srep06010

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., … Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science (New York, N.Y.)*, *335*(6070), 823–828. https://doi.org/10.1126/science.1215040

Macdonald, R. L., & Kang, J.-Q. (2009). Molecular pathology of genetic epilepsies associated with GABAA receptor subunit mutations. *Epilepsy Currents*, *9*(1), 18–23. https://doi.org/10.1111/j.1535-7511.2008.01278.x

Madan, V., Madan, B., Brykczynska, U., Zilbermann, F., Hogeveen, K., Döhner, K., Döhner, H., Weber, O., Blum, C., Rodewald, H.-R., Sassone-Corsi, P., Peters, A. H. F. M., & Fehling, H. J. (2009). Impaired function of primitive hematopoietic cells in mice lacking the Mixed-Lineage-Leukemia homolog MLL5. *Blood*, *113*(7), 1444–1454. https://doi.org/10.1182/blood-2008-02-142638

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., & Jabado, N. (2011). What can exome sequencing do for you? *Journal of Medical Genetics*, *48*(9), 580–589. https://doi.org/10.1136/jmedgenet-2011-100223

Majewski, Ł., Sobczak, M., Havrylov, S., Jóźwiak, J., & Rędowicz, M. J. (2012). Dock7: A GEF for Rho-family GTPases and a novel myosin VI-binding partner in neuronal PC12 cells. *Biochemistry and Cell Biology = Biochimie Et Biologie Cellulaire*, *90*(4), 565–574. https://doi.org/10.1139/o2012-009

Makhnoon, S., Shirts, B. H., & Bowen, D. J. (2019). Patients' perspectives of variants of uncertain significance and strategies for uncertainty management. *Journal of Genetic Counseling*, *28*(2), 313–325. https://doi.org/10.1002/jgc4.1075

Malan, V., Martinovic, J., Sanlaville, D., Caillat, S., Perrier Waill, M.-C., Maurin Ganne, M.-L., Tantau, J., Attie-Bitach, T., Vekemans, M., & Morichon-Delvallez, N. (2006). Molecular characterisation of a prenatally diagnosed 5q15q21.3 deletion and review of the literature. *Prenatal Diagnosis*, *26*(3), 231–238. https://doi.org/10.1002/pd.1386

Mandahl N. (1992). Methods in solid tumour cytogenetics in human genetics, a practical approach. In: In Rooney DE (Ed.), *Human cytogenetics: Malignancy and acquired abnormalities* (pp. 165–203). Oxford University Press.

Marsh, E., Fulp, C., Gomez, E., Nasrallah, I., Minarcik, J., Sudi, J., Christian, S. L., Mancini, G., Labosky, P., Dobyns, W., Brooks-Kayal, A., & Golden, J. A. (2009). Targeted loss of Arx results in a developmental epilepsy mouse model and recapitulates the human phenotype in heterozygous females. *Brain: A Journal of Neurology*, *132*(Pt 6), 1563–1576. https://doi.org/10.1093/brain/awp107

Marsh, Heron, D., Edwards, T. J., Quartier, A., Galea, C., Nava, C., Rastetter, A., Moutard, M.-L., Anderson, V., Bitoun, P., Bunt, J., Faudet, A., Garel, C., Gillies, G., Gobius, I., Guegan, J., Heide, S., Keren, B., Lesne, F., … Depienne, C. (2017). Mutations in DCC cause isolated agenesis of the corpus callosum with incomplete penetrance. *Nature Genetics*, *49*(4), Article 4. https://doi.org/10.1038/ng.3794

Martínez, A. D., Acuña, R., Figueroa, V., Maripillan, J., & Nicholson, B. (2009). Gap-Junction Channels Dysfunction in Deafness and Hearing Loss. *Antioxidants & Redox Signaling*, *11*(2), 309–322. https://doi.org/10.1089/ars.2008.2138

Masset, H., Hestand, M. S., Van Esch, H., Kleinfinger, P., Plaisancié, J., Afenjar, A., Molignier, R., Schluth-Bolard, C., Sanlaville, D., & Vermeesch, J. R. (2016). A Distinct Class of Chromoanagenesis Events Characterized by Focal Copy Number Gains. *Human Mutation*, *37*(7), 661–668. https://doi.org/10.1002/humu.22984

Mas-Y-Mas, S., Barbon, M., Teyssier, C., Déméné, H., Carvalho, J. E., Bird, L. E., Lebedev, A., Fattori, J., Schubert, M., Dumas, C., Bourguet, W., & le Maire, A. (2016). The Human Mixed Lineage Leukemia 5 (MLL5), a Sequentially and Structurally Divergent SET Domain-Containing Protein with No Intrinsic Catalytic Activity. *PloS One*, *11*(11), e0165139. https://doi.org/10.1371/journal.pone.0165139

Matarazzo, M. R., Boyle, S., D'Esposito, M., & Bickmore, W. A. (2007). Chromosome territory reorganization in a human disease with altered DNA methylation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(42), 16546–16551. PubMed. https://doi.org/10.1073/pnas.0702924104

Matsuura, T., Sutcliffe, J. S., Fang, P., Galjaard, R.-J., Jiang, Y., Benton, C. S., Rommens, J. M., & Beaudet, A. L. (1997). De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (UBE3A) in Angelman syndrome. *Nature Genetics*, *15*(1), 74–77. https://doi.org/10.1038/ng0197-74

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J., Altshuler, D. M., & International HapMap Consortium. (2006). Common deletion polymorphisms in the human genome. *Nature Genetics*, *38*(1), 86–92. https://doi.org/10.1038/ng1696

McEntagart, M., Williamson, K. A., Rainger, J. K., Wheeler, A., Seawright, A., De Baere, E., Verdin, H., Bergendahl, L. T., Quigley, A., Rainger, J., Dixit, A., Sarkar, A., López Laso, E., Sanchez-Carpintero, R., Barrio, J., Bitoun, P., Prescott, T., Riise, R., McKee, S., … FitzPatrick, D. R. (2016). A Restricted Repertoire of De Novo Mutations in ITPR1 Cause

Gillespie Syndrome with Evidence for Dominant-Negative Effect. *American Journal of Human Genetics*, *98*(5), 981–992. https://doi.org/10.1016/j.ajhg.2016.03.018

McGillivray, G., Rosenfeld, J. A., McKinlay Gardner, R. J., & Gillam, L. H. (2012). Genetic counselling and ethical issues with chromosome microarray analysis in prenatal testing. *Prenatal Diagnosis*, *32*(4), 389–395. https://doi.org/10.1002/pd.3849

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Mefford, H. C., Batshaw, M. L., & Hoffman, E. P. (2012). Genomics, intellectual disability, and autism. *The New England Journal of Medicine*, *366*(8), 733–743. https://doi.org/10.1056/NEJMra1114194

Mencarelli, M. A., Kleefstra, T., Katzaki, E., Papa, F. T., Cohen, M., Pfundt, R., Ariani, F., Meloni, I., Mari, F., & Renieri, A. (2009). 14q12 Microdeletion syndrome and congenital variant of Rett syndrome. *European Journal of Medical Genetics*, *52*(2–3), 148–152. https://doi.org/10.1016/j.ejmg.2009.03.004

Menten, B., Maas, N., Thienpont, B., Buysse, K., Vandesompele, J., Melotte, C., de Ravel, T., Van Vooren, S., Balikova, I., Backx, L., Janssens, S., De Paepe, A., De Moor, B., Moreau, Y., Marynen, P., Fryns, J.-P., Mortier, G., Devriendt, K., Speleman, F., & Vermeesch, J. R. (2006). Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: A new series of 140 patients and review of published reports. *Journal of Medical Genetics*, *43*(8), 625–633. https://doi.org/10.1136/jmg.2005.039453

Menuz, K., & Nicoll, R. A. (2008). Loss of inhibitory neuron AMPA receptors contributes to ataxia and epilepsy in stargazer mice. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(42), 10599–10603. https://doi.org/10.1523/JNEUROSCI.2732-08.2008

Menzies, F. M., Fleming, A., & Rubinsztein, D. C. (2015). Compromised autophagy and neurodegenerative diseases. *Nature Reviews Neuroscience*, *16*(6), Article 6. https://doi.org/10.1038/nrn3961

Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews Genetics*, *11*(1), Article 1. https://doi.org/10.1038/nrg2626

Michelson, D. J., & Clark, R. D. (2020). Optimizing Genetic Diagnosis of Neurodevelopmental Disorders in the Clinical Setting. *Clinics in Laboratory Medicine*, *40*(3), 231–256. https://doi.org/10.1016/j.cll.2020.05.001

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., Church, D. M., Crolla, J. A., Eichler, E. E., Epstein, C. J., Faucett, W. A., Feuk, L., Friedman, J. M., Hamosh, A., Jackson, L., Kaminsky, E. B., Kok, K., Krantz, I. D., Kuhn, R. M., … Ledbetter, D. H. (2010). Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American Journal of Human Genetics*, *86*(5), 749–764. https://doi.org/10.1016/j.ajhg.2010.04.006

Miller, D. T., Lee, K., Abul-Husn, N. S., Amendola, L. M., Brothers, K., Chung, W. K., Gollob, M. H., Gordon, A. S., Harrison, S. M., Hershberger, R. E., Klein, T. E., Richards, C. S., Stewart, D. R., Martin, C. L., & ACMG Secondary Findings Working Group. Electronic address: documents@acmg.net. (2023). ACMG SF v3.2 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: Official Journal of the American College of Medical Genetics*, *25*(8), 100866. https://doi.org/10.1016/j.gim.2023.100866

Milligan, C. J., Li, M., Gazina, E. V., Heron, S. E., Nair, U., Trager, C., Reid, C. A., Venkat, A., Younkin, D. P., Dlugos, D. J., Petrovski, S., Goldstein, D. B., Dibbens, L. M., Scheffer, I. E., Berkovic, S. F., & Petrou, S. (2014). KCNT1 gain of function in 2 epilepsy phenotypes is reversed by quinidine. *Annals of Neurology*, *75*(4), 581–590. https://doi.org/10.1002/ana.24128

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Moeschler, J. B. (2008). Genetic evaluation of intellectual disabilities. *Seminars in Pediatric Neurology*, *15*(1), 2–9. https://doi.org/10.1016/j.spen.2008.01.002

Moeschler, J. B., Shevell, M., & Committee on Genetics. (2014). Comprehensive evaluation of the child with intellectual disability or global developmental delays. *Pediatrics*, *134*(3), e903-918. https://doi.org/10.1542/peds.2014-1839

Molina, Mateos, F., Merino, M, Epifanio, J. L, & Gorrono, M. (1989). *Aicardi syndrome in 2 sisters. 115*, 282–283.

Moog, U., Smeets, E. E., van Roozendaal, K. E., Schoenmakers, S., Herbergs, J., Schoonbrood-Lenssen, A. M., & Schrander-Stumpel, C. T. (2003). Neurodevelopmental disorders in males related to the gene causing Rett syndrome in females (MECP2). *European Journal of Paediatric Neurology*, *7*(1), 5–12. https://doi.org/10.1016/s1090-3798(02)00134-4

Morelli, P., Gaspari, M., Gabriele, C., Dastoli, S., Bennardo, L., Pavel, A. B., Patruno, C., Del Duca, E., & Nisticò, S. P. (2021). Proteomic analysis from skin swabs reveals a new set of proteins identifying skin impairment in atopic dermatitis. *Experimental Dermatology*, *30*(6), 811–819. https://doi.org/10.1111/exd.14276

Moreno-De-Luca, A., Myers, S. M., Challman, T. D., Moreno-De-Luca, D., Evans, D. W., & Ledbetter, D. H. (2013). Developmental brain dysfunction: Revival and expansion of old concepts based on new genetic evidence. *The Lancet. Neurology*, *12*(4), 406–414. https://doi.org/10.1016/S1474-4422(13)70011-5

Mulhern, M. S., Stumpel, C., Stong, N., Brunner, H. G., Bier, L., Lippa, N., Riviello, J., Rouhl, R. P. W., Kempers, M., Pfundt, R., Stegmann, A. P. A., Kukolich, M. K., Telegrafi, A., Lehman, A., CAUSES study, Lopez-Rangel, E., Houcinat, N., Barth, M., den Hollander, N., … Sands, T. T. (2018). NBEA: Developmental disease gene with early generalized epilepsy phenotypes. *Annals of Neurology*, *84*(5), 788–795. PubMed. https://doi.org/10.1002/ana.25350

Multani, P., Myers, R. H., Blume, H. W., Schomer, D. L., & Sotrel, A. (1994). Neocortical dendritic pathology in human partial epilepsy: A quantitative Golgi study. *Epilepsia*, *35*(4), 728–736. https://doi.org/10.1111/j.1528-1157.1994.tb02503.x

Myerowitz, R., Puertollano, R., & Raben, N. (2020). Impaired autophagy: The collateral damage of lysosomal storage disorders. *EBioMedicine*, *63*, 103166. https://doi.org/10.1016/j.ebiom.2020.103166

Nakamuta, S., Yang, Y.-T., Wang, C.-L., Gallo, N. B., Yu, J.-R., Tai, Y., & Van Aelst, L. (2017). Dual role for DOCK7 in tangential migration of interneuron precursors in the postnatal forebrain. *The Journal of Cell Biology*, *216*(12), 4313–4330. https://doi.org/10.1083/jcb.201704157

Nakazato, T., Ohta, T., & Bono, H. (2013). Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive. *PLOS ONE*, *8*(10), e77910. https://doi.org/10.1371/journal.pone.0077910

Nan, X., Meehan, R. R., & Bird, A. (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Research*, *21*(21), 4886–4892.

Nawy, T. (2013). Single-cell epigenetics. *Nature Methods*, *10*(11), Article 11. https://doi.org/10.1038/nmeth.2721

Neidich, J. A., Nussbaum, R. L., Packer, R. J., Emanuel, B. S., & Puck, J. M. (1990). Heterogeneity of clinical severity and molecular lesions in Aircardi syndrome. *The Journal of Pediatrics*, *116*(6), 911–917. https://doi.org/10.1016/S0022-3476(05)80649-3

Nemos, C., Lambert, L., Giuliano, F., Doray, B., Roubertie, A., Goldenberg, A., Delobel, B., Layet, V., N'guyen, M., Saunier, A., Verneau, F., Jonveaux, P., & Philippe, C. (2009). Mutational spectrum of CDKL5 in early-onset encephalopathies: A study of a large collection of French patients and review of the literature. *Clinical Genetics*, *76*(4), 357–371. https://doi.org/10.1111/j.1399-0004.2009.01194.x

Neul, J. L., Benke, T. A., Marsh, E. D., Skinner, S. A., Merritt, J., Lieberman, D. N., Standridge, S., Feyma, T., Heydemann, P., Peters, S., Ryther, R., Jones, M., Suter, B., Kaufmann, W. E., Glaze, D. G., & Percy, A. K. (2019). The array of clinical phenotypes of males with mutations in Methyl-CpG binding protein 2. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, *180*(1), 55–67. https://doi.org/10.1002/ajmg.b.32707

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

Ni, H., Kirschstein, T., Norwood, B. A., & Hsieh, C. L. (2019). Editorial: The Developmental Seizure-Induced Hippocampal Mossy Fiber Sprouting: Target for Epilepsy Therapies? *Frontiers in Neurology*, *10*. https://www.frontiersin.org/articles/10.3389/fneur.2019.01212

Ning, Y., Roschke, A., Smith, A. C. M., Macha, M., Precht, K., Riethman, H., Ledbetter, D. H., Flint, J., Horsley, S., Regan, R., Kearney, L., Knight, S., Kvaloy, K., & Brown, W. R. A. (1996). A complete set of human telomeric probes and their clinical application. *Nature Genetics*, *14*(1), Article 1. https://doi.org/10.1038/ng0996-86

Nishizuka, M., Okada, R., Seki, K., Arai, Y., & Iizuka, R. (1991). Loss of dendritic synapses in the medial amygdala associated with kindling. *Brain Research*, *552*(2), 351–355. https://doi.org/10.1016/0006-8993(91)90104-4

Nivón, L. G., Moretti, R., & Baker, D. (2013). A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLOS ONE*, *8*(4), e59004. https://doi.org/10.1371/journal.pone.0059004

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385. PubMed. https://doi.org/10.1038/nature11049

Nord, A. S., Pattabiraman, K., Visel, A., & Rubenstein, J. L. R. (2015). Genomic perspectives of transcriptional regulation in forebrain development. *Neuron*, *85*(1), 27–47. PubMed. https://doi.org/10.1016/j.neuron.2014.11.011

O'Donnell-Luria, A. H., Pais, L. S., Faundes, V., Wood, J. C., Sveden, A., Luria, V., Abou Jamra, R., Accogli, A., Amburgey, K., Anderlid, B. M., Azzarello-Burri, S., Basinger, A. A., Bianchini, C., Bird, L. M., Buchert, R., Carre, W., Ceulemans, S., Charles, P., Cox, H., … Rodan, L. H. (2019a). Heterozygous Variants in KMT2E Cause a Spectrum of Neurodevelopmental Disorders and Epilepsy. *American Journal of Human Genetics*, *104*(6), 1210–1222. https://doi.org/10.1016/j.ajhg.2019.03.021

O'Donnell-Luria, A. H., Pais, L. S., Faundes, V., Wood, J. C., Sveden, A., Luria, V., Abou Jamra, R., Accogli, A., Amburgey, K., Anderlid, B. M., Azzarello-Burri, S., Basinger, A. A., Bianchini, C., Bird, L. M., Buchert, R., Carre, W., Ceulemans, S., Charles, P., Cox, H., … Rodan, L. H. (2019b). Heterozygous Variants in KMT2E Cause a Spectrum of Neurodevelopmental

Disorders and Epilepsy. *American Journal of Human Genetics*, *104*(6), 1210–1222. PubMed. https://doi.org/10.1016/j.ajhg.2019.03.021

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

Olson, H., Shen, Y., Avallone, J., Sheidley, B. R., Pinsky, R., Bergin, A. M., Berry, G. T., Duffy, F. H., Eksioglu, Y., Harris, D. J., Hisama, F. M., Ho, E., Irons, M., Jacobsen, C. M., James, P., Kothare, S., Khwaja, O., Lipton, J., Loddenkemper, T., … Poduri, A. (2014). Copy number variation plays an important role in clinical epilepsy. *Annals of Neurology*, *75*(6), 943–958. PubMed. https://doi.org/10.1002/ana.24178

Oppermann, H., Marcos-Grañeda, E., Weiss, L. A., Gurnett, C. A., Jelsig, A. M., Vineke, S. H., Isidor, B., Mercier, S., Magnussen, K., Zacher, P., Hashim, M., Pagnamenta, A. T., Race, S., Srivastava, S., Frazier, Z., Maiwald, R., Pergande, M., Milani, D., Rinelli, M., … Platzer, K. (2023). CUX1-related neurodevelopmental disorder: Deep insights into phenotype-genotype spectrum and underlying pathology. *European Journal of Human Genetics*. https://doi.org/10.1038/s41431-023-01445-2

Ortiz-González, X. R., Tintos-Hernández, J. A., Keller, K., Li, X., Foley, A. R., Bharucha-Goebel, D. X., Kessler, S. K., Yum, S. W., Crino, P. B., He, M., Wallace, D. C., & Bönnemann, C. G. (2018). Homozygous Boricua TBCK Mutation Causes Neurodegeneration and Aberrant Autophagy. *Annals of Neurology*, *83*(1), 153–165. https://doi.org/10.1002/ana.25130

Pagnamenta, A. T., Lise, S., Harrison, V., Stewart, H., Jayawant, S., Quaghebeur, G., Deng, A. T., Murphy, V. E., Sadighi Akha, E., Rimmer, A., Mathieson, I., Knight, S. J. L., Kini, U., Taylor, J. C., & Keays, D. A. (2012). Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. *Journal of Human Genetics*, *57*(1), 70–72. https://doi.org/10.1038/jhg.2011.128

Panicker, N., Coutman, M., Lawlor-O'Neill, C., Kahl, R. G. S., Roselli, S., & Verrills, N. M. (2020). Ppp2r2a Knockout Mice Reveal That Protein Phosphatase 2A Regulatory Subunit, PP2A-B55α, Is an Essential Regulator of Neuronal and Epidermal Embryonic Development. *Frontiers in Cell and Developmental Biology*, *8*, 358–358. PubMed. https://doi.org/10.3389/fcell.2020.00358

Parenti, I., Mallozzi, M. B., Hüning, I., Gervasini, C., Kuechler, A., Agolini, E., Albrecht, B., Baquero-Montoya, C., Bohring, A., Bramswig, N. C., Busche, A., Dalski, A., Guo, Y., Hanker, B., Hellenbroich, Y., Horn, D., Innes, A. M., Leoni, C., Li, Y. R., … Kaiser, F. J. (2021). ANKRD11 variants: KBG syndrome and beyond. *Clinical Genetics*, *100*(2), 187–200. https://doi.org/10.1111/cge.13977

Park, S. M., Lim, J. S., Ramakrishina, S., Kim, S. H., Kim, W. K., Lee, J., Kang, H.-C., Reiter, J. F., Kim, D. S., Kim, H. H., & Lee, J. H. (2018). Brain Somatic Mutations in MTOR Disrupt Neuronal Ciliogenesis, Leading to Focal Cortical Dyslamination. *Neuron*, *99*(1), 83-97.e7. https://doi.org/10.1016/j.neuron.2018.05.039

Parker, E. M., Kindja, N. L., Cheetham, C. E. J., & Sweet, R. A. (2020). Sex differences in dendritic spine density and morphology in auditory and visual cortices in adolescence and adulthood. *Scientific Reports*, *10*(1), 9442. https://doi.org/10.1038/s41598-020-65942-w

Pasche, B., & Absher, D. (2011). Whole-genome sequencing: A step closer to personalized medicine. *JAMA*, *305*(15), 1596–1597. https://doi.org/10.1001/jama.2011.484

Patsalis, P. C., Evangelidou, P., Charalambous, S., & Sismani, C. (2004). Fluorescence in situ hybridization characterization of apparently balanced translocation reveals cryptic complex chromosomal rearrangements with unexpected level of complexity. *European Journal of Human Genetics*, *12*(8), Article 8. https://doi.org/10.1038/sj.ejhg.5201211

Pellestor, F., Anahory, T., Lefort, G., Puechberty, J., Liehr, T., Hedon, B., & Sarda, P. (2011). Complex chromosomal rearrangements: Origin and meiotic behavior. *Human Reproduction Update*, *17*(4), 476–494. https://doi.org/10.1093/humupd/dmr010

Pellestor, F., Gatinois, V., Puechberty, J., Geneviève, D., & Lefort, G. (2014). Chromothripsis: Potential origin in gametogenesis and preimplantation cell divisions. A review. *Fertility and Sterility*, *102*(6), 1785–1796. https://doi.org/10.1016/j.fertnstert.2014.09.006

Penzes, P., Cahill, M. E., Jones, K. A., VanLeeuwen, J.-E., & Woolfrey, K. M. (2011). Dendritic spine pathology in neuropsychiatric disorders. *Nature Neuroscience*, *14*(3), 285–293. https://doi.org/10.1038/nn.2741

Perche, O., Haddad, G., Menuet, A., Callier, P., Marcos, M., Briault, S., & Laudier, B. (2013). Dysregulation of FOXG1 pathway in a 14q12 microdeletion case. *American Journal of Medical Genetics Part A*, *161*(12), 3072–3077. https://doi.org/10.1002/ajmg.a.36170

Perrault, I., Hamdan, F. F., Rio, M., Capo-Chichi, J.-M., Boddaert, N., Décarie, J.-C., Maranda, B., Nabbout, R., Sylvain, M., Lortie, A., Roux, P. P., Rossignol, E., Gérard, X., Barcia, G., Berquin, P., Munnich, A., Rouleau, G. A., Kaplan, J., Rozet, J.-M., & Michaud, J. L. (2014). Mutations in DOCK7 in Individuals with Epileptic Encephalopathy and Cortical Blindness. *American Journal of Human Genetics*, *94*(6), 891–897. https://doi.org/10.1016/j.ajhg.2014.04.012

Perry, G. H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A. S., Hyland, C., Stone, A. C., Hurles, M. E., Tyler-Smith, C., Eichler, E. E., Carter, N. P., Lee, C., & Redon, R. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Research*, *18*(11), 1698–1710. PubMed. https://doi.org/10.1101/gr.082016.108

Petersen, B.-S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D., & Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics*, *18*(1), 14. https://doi.org/10.1186/s12863-017-0479-5

Petrovski, S., Gussow, A. B., Wang, Q., Halvorsen, M., Han, Y., Weir, W. H., Allen, A. S., & Goldstein, D. B. (2015). The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLOS Genetics*, *11*(9), e1005492. https://doi.org/10.1371/journal.pgen.1005492

Phan, L. K., Lin, F., LeDuc, C. A., Chung, W. K., & Leibel, R. L. (2002). The mouse mahoganoid coat color mutation disrupts a novel C3HC4 RING domain protein. *The Journal of Clinical Investigation*, *110*(10), 1449–1459. https://doi.org/10.1172/JCI16131

Pichini, A., & Bishop, M. (2022). A nationally agreed cross-professional competency framework to facilitate genomic testing. *Genetics in Medicine*, *24*(8), 1743–1752. https://doi.org/10.1016/j.gim.2022.04.023

Pickles, S., Vigié, P., & Youle, R. J. (2018). Mitophagy and Quality Control Mechanisms in Mitochondrial Maintenance. *Current Biology*, *28*(4), R170–R185. https://doi.org/10.1016/j.cub.2018.01.004

Pilarowski, G. O., Vernon, H. J., Applegate, C. D., Boukas, L., Cho, M. T., Gurnett, C. A., Benke, P. J., Beaver, E., Heeley, J. M., Medne, L., Krantz, I. D., Azage, M., Niyazov, D., Henderson, L. B., Wentzensen, I. M., Baskin, B., Guillen Sacoto, M. J., Bowman, G. D., & Bjornsson, H. T. (2018). Missense variants in the chromatin remodeler CHD1 are associated with neurodevelopmental disability. *Journal of Medical Genetics*, *55*(8), 561–566. https://doi.org/10.1136/jmedgenet-2017-104759

Pittman, A., & Hardy, J. (2013). Genetic analysis in neurology: The next 10 years. *JAMA Neurology*, *70*(6), 696–702. https://doi.org/10.1001/jamaneurol.2013.2068

Platzer, K., Cogné, B., Hague, J., Marcelis, C. L., Mitter, D., Oberndorff, K., Park, S.-M., Ploos van Amstel, H. K., Simonic, I., van der Smagt, J. J., Stegmann, A. P. A., Stevens, S. J. C., Stumpel, C. T. R. M., Vincent, M., Lemke, J. R., & Jamra, R. (2018a). Haploinsufficiency of CUX1 Causes Nonsyndromic Global Developmental Delay With Possible Catch-up Development. *Annals of Neurology*, *84*(2), 200–207. https://doi.org/10.1002/ana.25278

Platzer, K., Cogné, B., Hague, J., Marcelis, C. L., Mitter, D., Oberndorff, K., Park, S.-M., Ploos van Amstel, H. K., Simonic, I., van der Smagt, J. J., Stegmann, A. P. A., Stevens, S. J. C., Stumpel, C. T. R. M., Vincent, M., Lemke, J. R., & Jamra, R. (2018b). Haploinsufficiency of CUX1 Causes Nonsyndromic Global Developmental Delay With Possible Catch-up Development. *Annals of Neurology*, *84*(2), 200–207. https://doi.org/10.1002/ana.25278

Platzer, K., & Lemke, J. R. (2021). GRIN2B-Related Neurodevelopmental Disorder. In M. P. Adam, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, & A. Amemiya (Eds.), *GeneReviews®*. University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK501979/

Pockrandt, C., Alzamel, M., Iliopoulos, C. S., & Reinert, K. (2020). GenMap: Ultra-fast computation of genome mappability. *Bioinformatics (Oxford, England)*, *36*(12), 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222

Pons, M. E., & Garcia, C. A. (2008). Aicardi Syndrome in Monozygotic Twins. *Ophthalmic Genetics*, *29*(2), 87–88. https://doi.org/10.1080/13816810801968669

Poot, M., & Haaf, T. (2015). Mechanisms of Origin, Phenotypic Effects and Diagnostic Implications of Complex Chromosome Rearrangements. *Molecular Syndromology*, *6*(3), 110–134. PubMed. https://doi.org/10.1159/000438812

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), Article 10. https://doi.org/10.1038/nbt.4235

Posey, J. E., O'Donnell-Luria, A. H., Chong, J. X., Harel, T., Jhangiani, S. N., Coban Akdemir, Z. H., Buyske, S., Pehlivan, D., Carvalho, C. M. B., Baxter, S., Sobreira, N., Liu, P., Wu, N., Rosenfeld, J. A., Kumar, S., Avramopoulos, D., White, J. J., Doheny, K. F., Witmer, P. D., … Centers for Mendelian Genomics. (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *21*(4), 798–812. https://doi.org/10.1038/s41436-018-0408-7

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341. https://doi.org/10.1186/1471-2164-13-341

Quang, D., Chen, Y., & Xie, X. (2015). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, *31*(5), 761–763. https://doi.org/10.1093/bioinformatics/btu703

Raczy, C., Petrovski, R., Saunders, C. T., Chorny, I., Kruglyak, S., Margulies, E. H., Chuang, H.-Y., Källberg, M., Kumar, S. A., Liao, A., Little, K. M., Strömberg, M. P., & Tanner, S. W. (2013). Isaac: Ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*, *29*(16), 2041–2043. https://doi.org/10.1093/bioinformatics/btt314

Rao, Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D Map of the Human

Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Rao, R. C., & Dou, Y. (2015). Hijacked in cancer: The KMT2 (MLL) family of methyltransferases. *Nature Reviews. Cancer*, *15*(6), 334–346. https://doi.org/10.1038/nrc3929

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. PubMed. https://doi.org/10.1016/j.cell.2014.11.021

Raposo, A. A. S. F., Vasconcelos, F. F., Drechsel, D., Marie, C., Johnston, C., Dolle, D., Bithell, A., Gillotin, S., van den Berg, D. L. C., Ettwiller, L., Flicek, P., Crawford, G. E., Parras, C. M., Berninger, B., Buckley, N. J., Guillemot, F., & Castro, D. S. (2015). Ascl1 Coordinately Regulates Gene Expression and the Chromatin Landscape during Neurogenesis. *Cell Reports*, *10*(9), 1544–1556. https://doi.org/10.1016/j.celrep.2015.02.025

Redin, C., Brand, H., Collins, R. L., Kammin, T., Mitchell, E., Hodge, J. C., Hanscom, C., Pillalamarri, V., Seabra, C. M., Abbott, M.-A., Abdul-Rahman, O. A., Aberg, E., Adley, R., Alcaraz-Estrada, S. L., Alkuraya, F. S., An, Y., Anderson, M.-A., Antolik, C., Anyane-Yeboa, K., … Talkowski, M. E. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature Genetics*, *49*(1), Article 1. https://doi.org/10.1038/ng.3720

Reiff, M., Bernhardt, B. A., Mulchandani, S., Soucier, D., Cornell, D., Pyeritz, R. E., & Spinner, N. B. (2012). "What does it mean?": Uncertainties in understanding results of chromosomal microarray testing. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *14*(2), 250–258. https://doi.org/10.1038/gim.2011.52

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019a). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019b). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Reynhout, S., & Janssens, V. (2019). Physiologic functions of PP2A: Lessons from genetically modified mice. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1866*(1), 31–50. https://doi.org/10.1016/j.bbamcr.2018.07.010

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG Laboratory Quality Assurance Committee. (2015a). 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

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG Laboratory Quality Assurance Committee. (2015b). 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

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., WGS500 Consortium, Wilkie, A. O. M., McVean, G., & Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, *46*(8), 912–918. https://doi.org/10.1038/ng.3036

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: A tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, *83*(5), 610–615. https://doi.org/10.1016/j.ajhg.2008.09.017

Rodriguez, O. L., Sharp, A. J., & Watson, C. T. (2021). Limitations of lymphoblastoid cell lines for establishing genetic reference datasets in the immunoglobulin loci. *PloS One*, *16*(12), e0261374–e0261374. PubMed. https://doi.org/10.1371/journal.pone.0261374

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., & Campbell, C. (2018). FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. *Bioinformatics (Oxford, England)*, *34*(3), 511–513. https://doi.org/10.1093/bioinformatics/btx536

Roller, E., Ivakhno, S., Lee, S., Royce, T., & Tanner, S. (2016). Canvas: Versatile and scalable detection of copy number variants. *Bioinformatics*, *32*(15), 2375–2377. https://doi.org/10.1093/bioinformatics/btw163

Rosenfeld, J. A., Lacassie, Y., El-Khechen, D., Escobar, L. F., Reggin, J., Heuer, C., Chen, E., Jenkins, L. S., Collins, A. T., Zinner, S., Babcock, M., Morrow, B., Schultz, R. A., Torchia, B. S., Ballif, B. C., Tsuchiya, K. D., & Shaffer, L. G. (2011). New cases and refinement of the critical region in the 1q41q42 microdeletion syndrome. *European Journal of Medical Genetics*, *54*(1), 42–49. https://doi.org/10.1016/j.ejmg.2010.10.002

Rosenfeld, J. A., & Patel, A. (2017). Chromosomal Microarrays: Understanding Genetics of Neurodevelopmental Disorders and Congenital Anomalies. *Journal of Pediatric Genetics*, *6*(1), 42–50. https://doi.org/10.1055/s-0036-1584306

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5), R51. https://doi.org/10.1186/gb-2013-14-5-r51

Rossini, L., De Santis, D., Mauceri, R. R., Tesoriero, C., Bentivoglio, M., Maderna, E., Maiorana, A., Deleo, F., de Curtis, M., Tringali, G., Cossu, M., Tumminelli, G., Bramerio, M., Spreafico, R., Tassi, L., & Garbelli, R. (2021). Dendritic pathology, spine loss and synaptic reorganization in human cortex from epilepsy patients. *Brain*, *144*(1), 251–265. https://doi.org/10.1093/brain/awaa387

Ruark, E., Snape, K., Humburg, P., Loveday, C., Bajrami, I., Brough, R., Rodrigues, D. N., Renwick, A., Seal, S., Ramsay, E., Duarte, S. D. V., Rivas, M. A., Warren-Perry, M., Zachariou, A., Campion-Flora, A., Hanks, S., Murray, A., Ansari Pour, N., Douglas, J., … Rahman, N. (2013). Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature*, *493*(7432), 406–410. https://doi.org/10.1038/nature11725

Sadleir, L. G., Vears, D., Regan, B., Redshaw, N., Bleasel, A., & Scheffer, I. E. (2012). Family studies of individuals with eyelid myoclonia with absences. *Epilepsia*, *53*(12), 2141–2148. https://doi.org/10.1111/j.1528-1167.2012.03692.x

Sanganahalli, B. G., Herman, P., Hyder, F., & Kannurpatti, S. S. (2013). Mitochondrial calcium uptake capacity modulates neocortical excitability. *Journal of Cerebral Blood Flow & Metabolism*, *33*(7), 1115–1126. https://doi.org/10.1038/jcbfm.2013.61

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sansregret, L., & Nepveu, A. (2008). The multiple roles of CUX1: Insights from mouse models and cell-based assays. *Gene*, *412*(1–2), 84–94. https://doi.org/10.1016/j.gene.2008.01.017

Santini, E., & Klann, E. (2014). Reciprocal signaling between translational control pathways and synaptic proteins in autism spectrum disorders. *Science Signaling*, *7*(349), re10–re10. PubMed. https://doi.org/10.1126/scisignal.2005832

Sardiello, M., Palmieri, M., di Ronza, A., Medina, D. L., Valenza, M., Gennarino, V. A., Di Malta, C., Donaudy, F., Embrione, V., Polishchuk, R. S., Banfi, S., Parenti, G., Cattaneo, E., & Ballabio, A. (2009). A Gene Network Regulating Lysosomal Biogenesis and Function. *Science*, *325*(5939), 473–477. https://doi.org/10.1126/science.1174447

Sato, S., Yamamoto, K., Matsushita, T., Isobe, N., Kawano, Y., Iinuma, K., Niino, M., Fukazawa, T., Nakamura, Y., Watanabe, M., Yonekawa, T., Masaki, K., Yoshimura, S., Murai, H., Yamasaki, R., & Kira, J. (2015). Copy number variations in multiple sclerosis and neuromyelitis optica. *Annals of Neurology*, *78*(5), 762–774. https://doi.org/10.1002/ana.24511

Savaskan, N. E., Bräuer, A. U., & Nitsch, R. (2004). Molecular cloning and expression regulation of PRG-3, a new member of the plasticity-related gene family. *The European Journal of Neuroscience*, *19*(1), 212–220. https://doi.org/10.1046/j.1460-9568.2003.03078.x

Schaaf, C. P., Wiszniewska, J., & Beaudet, A. L. (2011). Copy number and SNP arrays in clinical diagnostics. *Annual Review of Genomics and Human Genetics*, *12*, 25–51. https://doi.org/10.1146/annurev-genom-092010-110715

Scheibel, M. E., Crandall, P. H., & Scheibel, A. B. (1974). The Hippocampal-Dentate Complex in Temporal Lobe Epilepsy. *Epilepsia*, *15*(1), 55–80. https://doi.org/10.1111/j.1528-1157.1974.tb03997.x

Schneider, P., Petzold, S., Sommer, A., Nitsch, R., Schwegler, H., Vogt, J., & Roskoden, T. (2018). Altered synaptic phospholipid signaling in PRG-1 deficient mice induces exploratory behavior and motor hyperactivity resembling psychiatric disorders. *Behavioural Brain Research*, *336*, 1–7. https://doi.org/10.1016/j.bbr.2017.08.032

Schoenfelder, S., & Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, *20*(8), 437–455. https://doi.org/10.1038/s41576-019-0128-0

Schönfeld, P., & Wojtczak, L. (2008). Fatty acids as modulators of the cellular production of reactive oxygen species. *Free Radical Biology & Medicine*, *45*(3), 231–241. https://doi.org/10.1016/j.freeradbiomed.2008.04.029

Schwessinger, R., Gosden, M., Downes, D., Brown, R. C., Oudelaar, A. M., Telenius, J., Teh, Y. W., Lunter, G., & Hughes, J. R. (2020). DeepC: Predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods*, *17*(11), 1118–1124. https://doi.org/10.1038/s41592-020-0960-3

Seaby, E. G., Smedley, D., Tavares, A. L. T., Brittain, H., Jaarsveld, R. H. van, Baralle, D., Rehm, H. L., O'Donnell-Luria, A., & Ennis, S. (2022). A gene-to-patient approach uplifts novel disease gene discovery and identifies 18 putative novel disease genes. *Genetics in Medicine*, *24*(8), 1697–1707. https://doi.org/10.1016/j.gim.2022.04.019

Shagin, D. A., Lukyanov, K. A., Vagner, L. L., & Matz, M. V. (n.d.). *Regulation of average length of complex PCR product*.

Sharawat, I. K., Panda, P. K., & Dawman, L. (2021). Clinical Characteristics and Genotype-Phenotype Correlation in Children with KMT2E Gene-Related Neurodevelopmental Disorders: Report of Two New Cases and Review of Published Literature. *Neuropediatrics*, *52*(2), 98–104. https://doi.org/10.1055/s-0040-1715629

Shariff, K., Ghosal, S., & Matouschek, A. (2004). The Force Exerted by the Membrane Potential during Protein Import into the Mitochondrial Matrix. *Biophysical Journal*, *86*(6), 3647–3652. https://doi.org/10.1529/biophysj.104.040865

Sharma, N., Sosnay, P. R., Ramalho, A. S., Douville, C., Franca, A., Gottschalk, L. B., Park, J., Lee, M., Vecchio-Pagan, B., Raraigh, K. S., Amaral, M. D., Karchin, R., & Cutting, G. R. (2014). Experimental Assessment of Splicing Variants Using Expression Minigenes and Comparison with In Silico Predictions. *Human Mutation*, *35*(10), 1249–1259. https://doi.org/10.1002/humu.22624

Sharma, S. R., Gonda, X., & Tarazi, F. I. (2018). Autism Spectrum Disorder: Classification, diagnosis and therapy. *Pharmacology & Therapeutics*, *190*, 91–104. https://doi.org/10.1016/j.pharmthera.2018.05.007

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), Article 10. https://doi.org/10.1038/nbt1486

Sheng, M., & Kim, E. (2000). The Shank family of scaffold proteins. *Journal of Cell Science*, *113*(11), 1851–1856. https://doi.org/10.1242/jcs.113.11.1851

Sheridan, M. B., Kato, T., Haldeman-Englert, C., Jalali, G. R., Milunsky, J. M., Zou, Y., Klaes, R., Gimelli, G., Gimelli, S., Gemmill, R. M., Drabkin, H. A., Hacker, A. M., Brown, J., Tomkins, D., Shaikh, T. H., Kurahashi, H., Zackai, E. H., & Emanuel, B. S. (2010). A palindrome-mediated recurrent translocation with 3:1 meiotic nondisjunction: The t(8;22)(q24.13;q11.21). *American Journal of Human Genetics*, *87*(2), 209–218. https://doi.org/10.1016/j.ajhg.2010.07.002

Shevell, M., Ashwal, S., Donley, D., Flint, J., Gingold, M., Hirtz, D., Majnemer, A., Noetzel, M., Sheth, R. D., Quality Standards Subcommittee of the American Academy of Neurology, & Practice Committee of the Child Neurology Society. (2003). Practice parameter: Evaluation of the child with global developmental delay: report of the Quality Standards Subcommittee of the American Academy of Neurology and The Practice Committee of the Child Neurology Society. *Neurology*, *60*(3), 367–380. https://doi.org/10.1212/01.wnl.0000031431.81555.16

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., Gaunt, T. R., & Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics (Oxford, England)*, *31*(10), 1536–1543. https://doi.org/10.1093/bioinformatics/btv009

Sica, V., Izzo, V., Bravo-San Pedro, J. M., Zamzami, N., & Maiuri, M. C. (2016). Mitophagy. In *Autophagy: Cancer, Other Pathologies, Inflammation, Immunity, Infection, and Aging* (pp. 91–104). Elsevier. https://doi.org/10.1016/B978-0-12-802936-7.00003-9

Singh, G., & Cooper, T. A. (2006). Minigene reporter for identification and analysis of cis elements and trans factors affecting pre-mRNA splicing. *BioTechniques*, *41*(2), 177–181. https://doi.org/10.2144/000112208

Sismani, C., Kitsiou-Tzeli, S., Ioannides, M., Christodoulou, C., Anastasiadou, V., Stylianidou, G., Papadopoulou, E., Kanavakis, E., Kosmaidou-Aravidou, Z., & Patsalis, P. C. (2008). Cryptic genomic imbalances in patients with de novo or familial apparently balanced translocations and abnormal phenotype. *Molecular Cytogenetics*, *1*, 15. https://doi.org/10.1186/1755-8166-1-15

Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., Huang, J., Du, Y., Lin, L., Dong, Z., Yang, L., Liu, X., Jiang, H., Xu, X., Wang, J., … Mulder, J. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science (New York, N.Y.)*, *367*(6482), eaay5947. https://doi.org/10.1126/science.aay5947

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/10.1002/cpmb.59

Smedley, D., Jacobsen, J. O. B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., Bone, W. P., Haendel, M. A., & Robinson, P. N. (2015a). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, *10*(12), 2004–2015. https://doi.org/10.1038/nprot.2015.124

Smedley, D., Jacobsen, J. O. B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., Bone, W. P., Haendel, M. A., & Robinson, P.

N. (2015b). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, *10*(12), 2004–2015. https://doi.org/10.1038/nprot.2015.124

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., Haendel, M. A., Mungall, C. J., Lewis, S. E., Groza, T., Valentini, G., & Robinson, P. N. (2016a). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *American Journal of Human Genetics*, *99*(3), 595–606. https://doi.org/10.1016/j.ajhg.2016.07.005

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., Haendel, M. A., Mungall, C. J., Lewis, S. E., Groza, T., Valentini, G., & Robinson, P. N. (2016b). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *American Journal of Human Genetics*, *99*(3), 595–606. https://doi.org/10.1016/j.ajhg.2016.07.005

Smith, K. M., Youssef, P. E., Wirrell, E. C., Nickels, K. C., Payne, E. T., Britton, J. W., Shin, C., Cascino, G. D., Patterson, M. C., & Wong-Kisiel, L. C. (2018). Jeavons Syndrome: Clinical Features and Response to Treatment. *Pediatric Neurology*, *86*, 46–51. https://doi.org/10.1016/j.pediatrneurol.2018.06.001

Sobreira, N., Schiettecatte, F., Valle, D., & Hamosh, A. (2015). GeneMatcher: A matching tool for connecting investigators with an interest in the same gene. *Human Mutation*, *36*(10), 928–930. https://doi.org/10.1002/humu.22844

Soden, S. E., Saunders, C. J., Willig, L. K., Farrow, E. G., Smith, L. D., Petrikin, J. E., LePichon, J.-B., Miller, N. A., Thiffault, I., Dinwiddie, D. L., Twist, G., Noll, A., Heese, B. A., Zellmer, L., Atherton, A. M., Abdelmoity, A. T., Safina, N., Nyp, S. S., Zuccarelli, B., … Kingsmore, S. F. (2014). Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Science Translational Medicine*, *6*(265), 265ra168. https://doi.org/10.1126/scitranslmed.3010076

Soto, C., & Pritzkow, S. (2018). Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nature Neuroscience*, *21*(10), 1332–1340. https://doi.org/10.1038/s41593-018-0235-9

Splinter, K., Adams, D. R., Bacino, C. A., Bellen, H. J., Bernstein, J. A., Cheatle-Jarvela, A. M., Eng, C. M., Esteves, C., Gahl, W. A., Hamid, R., Jacob, H. J., Kikani, B., Koeller, D. M., Kohane, I. S., Lee, B. H., Loscalzo, J., Luo, X., McCray, A. T., Metz, T. O., … Undiagnosed Diseases Network. (2018). Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *The New England Journal of Medicine*, *379*(22), 2131–2139. https://doi.org/10.1056/NEJMoa1714458

Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., Firth, H. V., Frazier, T., Hansen, R. L., Prock, L., Brunner, H., Hoang, N., Scherer, S. W., Sahin, M., Miller, D. T., & NDD Exome Scoping Review WorkGroup. (2019). Meta-analysis and multidisciplinary consensus statement: Exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *21*(11), 2413–2421. https://doi.org/10.1038/s41436-019-0554-6

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., McLaren, S., Lin, M.-L., McBride, D. J., Varela, I., Nik-Zainal, S., Leroy, C., Jia, M., Menzies, A., Butler, A. P., … Campbell, P. J. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, *144*(1), 27–40. https://doi.org/10.1016/j.cell.2010.11.055

Stevenson, R. E., Procopio-Allen, A. M., Schroer, R. J., & Collins, J. S. (2003). Genetic syndromes among individuals with mental retardation. *American Journal of Medical Genetics. Part A*, *123A*(1), 29–32. https://doi.org/10.1002/ajmg.a.20492

Strauss, U., & Bräuer, A. U. (2013). Current views on regulation and function of plasticity-related genes (PRGs/LPPRs) in the brain. *Biochimica Et Biophysica Acta*, *1831*(1), 133–138. https://doi.org/10.1016/j.bbalip.2012.08.010

Striano, P., Paravidino, R., Sicca, F., Chiurazzi, P., Gimelli, S., Coppola, A., Robbiano, A., Traverso, M., Pintaudi, M., Giovannini, S., Operto, F., Vigliano, P., Granata, T., Coppola, G., Romeo,

A., Specchio, N., Giordano, L., Osborne, L. R., Gimelli, G., … Zara, F. (2011). West syndrome associated with 14q12 duplications harboring FOXG1. *Neurology*, *76*(18), 1600–1602. https://doi.org/10.1212/WNL.0b013e3182194bbf

Striano, S., Striano, P., Nocerino, C., Boccella, P., Bilo, L., Meo, R., & Ruosi, P. (2002). Eyelid myoclonia with absences: An overlooked epileptic syndrome? *Neurophysiologie Clinique = Clinical Neurophysiology*, *32*(5), 287–296. https://doi.org/10.1016/s0987-7053(02)00343-x

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), Article 7571. https://doi.org/10.1038/nature15394

Sun, Y., Ma, L., Jin, M., Zheng, Y., Wang, D., & Ni, H. (2021). Effects of Melatonin on Neurobehavior and Cognition in a Cerebral Palsy Model of plppr5-/- Mice. *Frontiers in Endocrinology*, *12*, 598788. https://doi.org/10.3389/fendo.2021.598788

Sutton, V. R., & Van-Den-Vyver, I. B. (2020). Aicardi Syndrome. In M. P. Adam, D. B. Everman, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, & A. Amemiya (Eds.), *GeneReviews®*. University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK1381/

Swaminathan, G. J., Bragin, E., Chatzimichali, E. A., Corpas, M., Bevan, A. P., Wright, C. F., Carter, N. P., Hurles, M. E., & Firth, H. V. (2012). DECIPHER: Web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Human Molecular Genetics*, *21*(R1), R37-44. https://doi.org/10.1093/hmg/dds362

Takagi, M., Sasaki, G., Mitsui, T., Honda, M., Tanaka, Y., & Hasegawa, T. (2013). A 2.0Mb microdeletion in proximal chromosome 14q12, involving regulatory elements of FOXG1, with the coding region of FOXG1 being unaffected, results in severe developmental delay, microcephaly, and hypoplasia of the corpus callosum. *European Journal of Medical Genetics*, *56*(9), 526–528. https://doi.org/10.1016/j.ejmg.2013.05.012

Tanaka, M., Martinez-Juarez, I., Duron, R., Nguyen, V.-H., Ochoa, A., Jara-Prado, A., Alonso, M., Lopez-Ruiz, M., Medina, M., Guilhoto, L., Yacubian, M., Silva, R., Patterson, C., Bailey,

J., & Delgado-Escueta, A. (2018). Autophagy in Juvenile Myoclonic Epilepsy (JME) (P2.275). *Neurology*, *90*(15 Supplement). https://n.neurology.org/content/90/15_Supplement/P2.275

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-L., … Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, *163*(7), 1611–1627. https://doi.org/10.1016/j.cell.2015.11.024

Tangvarasittichai, S. (2015). Oxidative stress, insulin resistance, dyslipidemia and type 2 diabetes mellitus. *World Journal of Diabetes*, *6*(3), 456–480. https://doi.org/10.4239/wjd.v6.i3.456

Tankard, R. M., Bennett, M. F., Degorski, P., Delatycki, M. B., Lockhart, P. J., & Bahlo, M. (2018). Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *American Journal of Human Genetics*, *103*(6), 858–873. https://doi.org/10.1016/j.ajhg.2018.10.015

Tasdogan, A., Kumar, S., Allies, G., Bausinger, J., Beckel, F., Hofemeister, H., Mulaw, M., Madan, V., Scharffetter-Kochanek, K., Feuring-Buske, M., Doehner, K., Speit, G., Stewart, A. F., & Fehling, H. J. (2016). DNA Damage-Induced HSPC Malfunction Depends on ROS Accumulation Downstream of IFN-1 Signaling and Bid Mobilization. *Cell Stem Cell*, *19*(6), 752–767. https://doi.org/10.1016/j.stem.2016.08.007

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A. R., Attar, M., Babbs, C., Becq, J., Beeson, D., Bento, C., Bignell, P., Blair, E., Buckle, V. J., … McVean, G. (2015a). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, *47*(7), 717–726. https://doi.org/10.1038/ng.3304

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A. R., Attar, M., Babbs, C., Becq, J., Beeson, D., Bento, C., Bignell, P., Blair, E., Buckle, V. J., … McVean, G. (2015b). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, *47*(7), 717–726. https://doi.org/10.1038/ng.3304

Teinert, J., Behne, R., Wimmer, M., & Ebrahimi-Fakhari, D. (2020). Novel insights into the clinical and molecular spectrum of congenital disorders of autophagy. *Journal of Inherited Metabolic Disease*, *43*(1), 51–62. https://doi.org/10.1002/jimd.12084

Tommerup, N. (1993). Mendelian cytogenetics. Chromosome rearrangements associated with mendelian disorders. *Journal of Medical Genetics*, *30*(9), 713–727. https://doi.org/10.1136/jmg.30.9.713

Tønnesen, J., & Nägerl, U. V. (2016). Dendritic Spines as Tunable Regulators of Synaptic Signals. *Frontiers in Psychiatry*, *7*, 101. https://doi.org/10.3389/fpsyt.2016.00101

Traveling with Lennox – the petit mal triad | Beyond the Ion Channel. (n.d.). Retrieved 20 January 2023, from http://epilepsygenetics.net/2012/10/23/traveling-with-lennox-the-petit-mal-triad/

Trotti, D., Danbolt, N. C., & Volterra, A. (1998). Glutamate transporters are oxidant-vulnerable: A molecular link between oxidative and excitotoxic neurodegeneration? *Trends in Pharmacological Sciences*, *19*(8), 328–334. https://doi.org/10.1016/s0165-6147(98)01230-9

Truty, R., Paul, J., Kennemer, M., Lincoln, S. E., Olivares, E., Nussbaum, R. L., & Aradhya, S. (2019). Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *21*(1), 114–123. PubMed. https://doi.org/10.1038/s41436-018-0033-5

Turkdogan, D., Turkyilmaz, A., Gormez, Z., Sager, G., & Ekinci, G. (2019). A novel truncating mutation of DOCK7 gene with an early-onset non-encephalopathic epilepsy. *Seizure*, *66*, 12–14. https://doi.org/10.1016/j.seizure.2019.01.010

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., Halai, D., Baple, E., Craig, C., Hamblin, A., Henderson, S., Patch, C., O'Neill, A., Devereau, A., Smith, K., Martin, A. R., Sosinsky, A., McDonagh, E. M., Sultana, R., … 100 000 Genomes Project. (2018). The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *BMJ (Clinical Research Ed.)*, *361*, k1687. https://doi.org/10.1136/bmj.k1687

Upadhyay, A., Amanullah, A., Chhangani, D., Mishra, R., Prasad, A., & Mishra, A. (2016). Mahogunin Ring Finger-1 (MGRN1), a Multifaceted Ubiquitin Ligase: Recent Unraveling

of Neurobiological Mechanisms. *Molecular Neurobiology*, *53*(7), 4484–4496. https://doi.org/10.1007/s12035-015-9379-8

van Coevorden-Hameete, M. H., de Graaff, E., Titulaer, M. J., Hulsenboom, E., Sabater, L., Hoogenraad, C. C., & Sillevis Smitt, P. A. (2015). Plasticity-related gene 5: A novel surface autoantigen in paraneoplastic cerebellar degeneration. *Neurology® Neuroimmunology & Neuroinflammation*, *2*(5), e156. https://doi.org/10.1212/NXI.0000000000000156

van El, C. G., Cornel, M. C., Borry, P., Hastings, R. J., Fellmann, F., Hodgson, S. V., Howard, H. C., Cambon-Thomsen, A., Knoppers, B. M., Meijers-Heijboer, H., Scheffer, H., Tranebjaerg, L., Dondorp, W., & de Wert, G. M. W. R. (2013). Whole-genome sequencing in health care. *European Journal of Human Genetics*, *21*(6), Article 6. https://doi.org/10.1038/ejhg.2013.46

Van-Den-Vyver, I. B., Panichkul, P. P., Antalffy, B. A., Sun, Y., Hunter, J. V., & Armstrong, D. D. (2004). Presence of filamin in the astrocytic inclusions of Aicardi syndrome. *Pediatric Neurology*, *30*(1), 7–15. https://doi.org/10.1016/S0887-8994(03)00311-4

Van-Den-Vyver, I. B. van den. (2002). Microphthalmia with linear skin defects (MLS), Aicardi, and Goltz syndromes: Are they related X-linked dominant male-lethal disorders? *Cytogenetic and Genome Research*, *99*(1–4), 289–296. https://doi.org/10.1159/000071606

Van-Dyke, D. L., Weiss, L., Roberson, J. R., & Babu, V. R. (1983). The frequency and mutation rate of balanced autosomal rearrangements in man estimated from prenatal genetic studies for advanced maternal age. *American Journal of Human Genetics*, *35*(2), 301–308.

Vaz-Drago, R., Custódio, N., & Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Human Genetics*, *136*(9), 1093–1111. https://doi.org/10.1007/s00439-017-1809-4

Velmans, C., O'Donnell-Luria, A. H., Argilli, E., Tran Mau-them, F., Vitobello, A., Chan, M. C., Fung, J. L.-F., Rech, M., Abicht, A., Aubert Mucca, M., Carmichael, J., Chassaing, N., Clark, R., Coubes, C., Denommé-Pichon, A.-S., de Dios, J. K., England, E., Funalot, B., Gerard, M., … Erger, F. (2021). O'Donnell-Luria-Rodan syndrome: Description of a second multinational cohort and refinement of the phenotypic spectrum. *Journal of Medical Genetics*, *59*(7), 697–705. https://doi.org/10.1136/jmedgenet-2020-107470

Velmans, C., O'Donnell-Luria, A. H., Argilli, E., Tran Mau-Them, F., Vitobello, A., Chan, M. C., Fung, J. L.-F., Rech, M., Abicht, A., Aubert Mucca, M., Carmichael, J., Chassaing, N., Clark, R., Coubes, C., Denommé-Pichon, A.-S., de Dios, J. K., England, E., Funalot, B., Gerard, M., … Erger, F. (2022). O'Donnell-Luria-Rodan syndrome: Description of a second multinational cohort and refinement of the phenotypic spectrum. *Journal of Medical Genetics*, *59*(7), 697–705. https://doi.org/10.1136/jmedgenet-2020-107470

Velmans, T., Battefeld, A., Geist, B., Farrés, A. S., Strauss, U., & Bräuer, A. U. (2013). Plasticity-related gene 3 promotes neurite shaft protrusion. *BMC Neuroscience*, *14*(1), 36. https://doi.org/10.1186/1471-2202-14-36

Villard, L. (2007). MECP2 mutations in males. *Journal of Medical Genetics*, *44*(7), 417–423. https://doi.org/10.1136/jmg.2007.049452

Vissers, L. E. L. M., Gilissen, C., & Veltman, J. A. (2016). Genetic studies in intellectual disability and related disorders. *Nature Reviews. Genetics*, *17*(1), 9–18. https://doi.org/10.1038/nrg3999

Vissers, L. E. L. M., van Nimwegen, K. J. M., Schieving, J. H., Kamsteeg, E.-J., Kleefstra, T., Yntema, H. G., Pfundt, R., van der Wilt, G. J., Krabbenborg, L., Brunner, H. G., van der Burg, S., Grutters, J., Veltman, J. A., & Willemsen, M. A. A. P. (2017). A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *19*(9), 1055–1063. https://doi.org/10.1038/gim.2017.1

Vogt, J., Yang, J.-W., Mobascher, A., Cheng, J., Li, Y., Liu, X., Baumgart, J., Thalman, C., Kirischuk, S., Unichenko, P., Horta, G., Radyushkin, K., Stroh, A., Richers, S., Sahragard, N., Distler, U., Tenzer, S., Qiao, L., Lieb, K., … Nitsch, R. (2016). Molecular cause and functional impact of altered synaptic lipid signaling due to a prg-1 gene SNP. *EMBO Molecular Medicine*, *8*(1), 25–38. https://doi.org/10.15252/emmm.201505677

Volkmar, F. R., & McPartland, J. C. (2014). From Kanner to DSM-5: Autism as an evolving diagnostic concept. *Annual Review of Clinical Psychology*, *10*, 193–212. https://doi.org/10.1146/annurev-clinpsy-032813-153710

von Campe, G., Spencer, D. D., & de Lanerolle, N. C. (1997). Morphology of dentate granule cells in the human epileptogenic hippocampus. *Hippocampus*, *7*(5), 472–488. https://doi.org/10.1002/(SICI)1098-1063(1997)7:5<472::AID-HIPO4>3.0.CO;2-J

Vortkamp, A., Gessler, M., & Grzeschik, K.-H. (1991). GLI3 zinc-finger gene interrupted by translocations in Greig syndrome families. *Nature*, *352*(6335), 539–540. https://doi.org/10.1038/352539a0

Wain K.E. & Thorland E.C. (2011). *Clinical utility of chromosomal microarray testing*. *36*(2), 1–10.

Wallerstein, R., Sugalski, R., Cohn, L., Jawetz, R., & Friez, M. (2008). Expansion of the ARX spectrum. *Clinical Neurology and Neurosurgery*, *110*(6), 631–634. https://doi.org/10.1016/j.clineuro.2008.03.007

Wanders, R. J. A., & Waterham, H. R. (2006). Biochemistry of mammalian peroxisomes revisited. *Annual Review of Biochemistry*, *75*, 295–332. https://doi.org/10.1146/annurev.biochem.74.082803.133329

Wang, D., Jin, M.-F., Li, L., Liu, Y., Sun, Y., & Ni, H. (2021). PRG5 Knockout Precipitates Late-Onset Hypersusceptibility to Pilocarpine-Induced Juvenile Seizures by Exacerbating Hippocampal Zinc Signaling-Mediated Mitochondrial Damage. *Frontiers in Neuroscience*, *15*, 715555. https://doi.org/10.3389/fnins.2021.715555

Wang, D., Liu, Y., Zhao, D., Jin, M., Li, L., & Ni, H. (2022a). Plppr5 gene inactivation causes a more severe neurological phenotype and abnormal mitochondrial homeostasis in a mouse model of juvenile seizure. *Epilepsy Research*, *183*, 106944. https://doi.org/10.1016/j.eplepsyres.2022.106944

Wang, D., Liu, Y., Zhao, D., Jin, M., Li, L., & Ni, H. (2022b). Plppr5 gene inactivation causes a more severe neurological phenotype and abnormal mitochondrial homeostasis in a mouse model of juvenile seizure. *Epilepsy Research*, *183*, 106944. https://doi.org/10.1016/j.eplepsyres.2022.106944

Wang, T., Kim, C. N., Bakken, T. E., Gillentine, M. A., Henning, B., Mao, Y., Gilissen, C., Nowakowski, T. J., & Eichler, E. E. (2022). Integrated gene analyses of de novo variants from 46,612 trios with autism and developmental disorders. *Proceedings of the National*

Academy of Sciences of the United States of America, *119*(46), e2203491119. https://doi.org/10.1073/pnas.2203491119

Wang, Y., Qu, Z., Fang, Y., Chen, Y., Peng, J., Song, J., Li, J., Shi, J., Zhou, J.-Q., & Zhao, Y. (2022). *Chromosome territory reorganization through artificial chromosome fusion is negligible to cell fate determination and mouse development*. https://doi.org/10.1101/2022.10.09.511475

Wang, Y., Zhao, X., Ju, W., Flory, M., Zhong, J., Jiang, S., Wang, P., Dong, X., Tao, X., Chen, Q., Shen, C., Zhong, M., Yu, Y., Brown, W. T., & Zhong, N. (2015). Genome-wide differential expression of synaptic long noncoding RNAs in autism spectrum disorder. *Translational Psychiatry*, *5*(10), e660–e660. PubMed. https://doi.org/10.1038/tp.2015.144

Warburton, D. (1991a). De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: Clinical significance and distribution of breakpoints. *American Journal of Human Genetics*, *49*(5), 995–1013.

Warburton, D. (1991b). De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: Clinical significance and distribution of breakpoints. *American Journal of Human Genetics*, *49*(5), 995–1013.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. *G3: Genes|Genomes|Genetics*, *5*(8), 1543–1550. https://doi.org/10.1534/g3.115.018564

Watabe-Uchida, M., John, K. A., Janas, J. A., Newey, S. E., & Van Aelst, L. (2006). The Rac activator DOCK7 regulates neuronal polarity through local phosphorylation of stathmin/Op18. *Neuron*, *51*(6), 727–739. https://doi.org/10.1016/j.neuron.2006.07.020

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., & di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-13212-3

Westrate, L. M., Drocco, J. A., Martin, K. R., Hlavacek, W. S., & MacKeigan, J. P. (2014). Mitochondrial Morphological Features Are Associated with Fission and Fusion Events. *PLOS ONE*, *9*(4), e95265. https://doi.org/10.1371/journal.pone.0095265

Willmore, L. J., Ballinger, W. E., Boggs, W., Sypert, G. W., & Rubin, J. J. (1980). Dendritic alterations in rat isocortex within an iron-induced chronic epileptic focus. *Neurosurgery*, *7*(2), 142–146. https://doi.org/10.1227/00006123-198008000-00005

Wolfe, K., Strydom, A., Morrogh, D., Carter, J., Cutajar, P., Eyeoyibo, M., Hassiotis, A., McCarthy, J., Mukherjee, R., Paschos, D., Perumal, N., Read, S., Shankar, R., Sharif, S., Thirulokachandran, S., Thygesen, J. H., Patch, C., Ogilvie, C., Flinter, F., … Bass, N. (2017). Chromosomal microarray testing in adults with intellectual disability presenting with comorbid psychiatric disorders. *European Journal of Human Genetics*, *25*(1), Article 1. https://doi.org/10.1038/ejhg.2016.107

Wong, B. K. Y., & Sutton, V. R. (2018). Aicardi syndrome, an unsolved mystery: Review of diagnostic features, previous attempts, and future opportunities for genetic examination. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *178*(4), 423–431. https://doi.org/10.1002/ajmg.c.31658

Wright, C. F., Campbell, P., Eberhardt, R. Y., Aitken, S., Perrett, D., Brent, S., Danecek, P., Gardner, E. J., Chundru, V. K., Lindsay, S. J., Andrews, K., Hampstead, J., Kaplanis, J., Samocha, K. E., Middleton, A., Foreman, J., Hobson, R. J., Parker, M. J., Martin, H. C., … Study, the D. (2022). *Optimising diagnostic yield in highly penetrant genomic disease* (p. 2022.07.25.22278008). medRxiv. https://doi.org/10.1101/2022.07.25.22278008

Yamawaki, R., Thind, K., & Buckmaster, P. S. (2015). Blockade of excitatory synaptogenesis with proximal dendrites of dentate granule cells following rapamycin treatment in a mouse model of temporal lobe epilepsy. *Journal of Comparative Neurology*, *523*(2), 281–297. https://doi.org/10.1002/cne.23681

Yang, J., Zhang, Z., Roe, S. M., Marshall, C. J., & Barford, D. (2009). Activation of Rho GTPases by DOCK exchange factors is mediated by a nucleotide sensor. *Science (New York, N.Y.)*, *325*(5946), 1398–1402. https://doi.org/10.1126/science.1174468

Yang, S. Y., He, X. Y., & Schulz, H. (1987). Fatty acid oxidation in rat brain is limited by the low activity of 3-ketoacyl-coenzyme A thiolase. *The Journal of Biological Chemistry*, *262*(27), 13027–13032.

Yap, C. X., Alvares, G. A., Henders, A. K., Lin, T., Wallace, L., Farrelly, A., McLaren, T., Berry, J., Vinkhuyzen, A. A. E., Trzaskowski, M., Zeng, J., Yang, Y., Cleary, D., Grove, R., Hafekost, C., Harun, A., Holdsworth, H., Jellett, R., Khan, F., … Gratten, J. (2021). Analysis of common genetic variation and rare CNVs in the Australian Autism Biobank. *Molecular Autism*, *12*(1), 12–12. PubMed. https://doi.org/10.1186/s13229-020-00407-5

Yap, D. B., Walker, D. C., Prentice, L. M., McKinney, S., Turashvili, G., Mooslehner-Allen, K., de Algara, T. R., Fee, J., de Tassigny, X. d'Anglemont, Colledge, W. H., & Aparicio, S. (2011). Mll5 is required for normal spermatogenesis. *PloS One*, *6*(11), e27127. https://doi.org/10.1371/journal.pone.0027127

Yasui, D. H., Peddada, S., Bieda, M. C., Vallero, R. O., Hogart, A., Nagarajan, R. P., Thatcher, K. N., Farnham, P. J., & LaSalle, J. M. (2007). Integrated epigenomic analyses of neuronal MeCP2 reveal a role for long-range interaction with active genes. *Proceedings of the National Academy of Sciences*, *104*(49), 19416–19421. https://doi.org/10.1073/pnas.0707442104

Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *11*(2–3), 377–394. https://doi.org/10.1089/1066527041410418

Youle, R. J., & Narendra, D. P. (2011). Mechanisms of mitophagy. *Nature Reviews Molecular Cell Biology*, *12*(1), Article 1. https://doi.org/10.1038/nrm3028

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., & Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, *3*(4), e59. https://doi.org/10.1371/journal.pcbi.0030059

Yu, J., Szabo, A., Pagnamenta, A. T., Shalaby, A., Giacopuzzi, E., Taylor, J., Shears, D., Pontikos, N., Wright, G., Michaelides, M., Halford, S., & Downes, S. (2021). *SVRare: Discovering disease-causing structural variants in the 100K Genomes Project*. https://doi.org/10.1101/2021.10.15.21265069

Yu, P., Agbaegbu, C., Malide, D. A., Wu, X., Katagiri, Y., Hammer, J. A., & Geller, H. M. (2015). Cooperative interactions of LPPR family members in membrane localization and alteration

of cellular morphology. *Journal of Cell Science*, *128*(17), 3210–3222. https://doi.org/10.1242/jcs.169789

Zak, J., Vives, V., Szumska, D., Vernet, A., Schneider, J. E., Miller, P., Slee, E. A., Joss, S., Lacassie, Y., Chen, E., Escobar, L. F., Tucker, M., Aylsworth, A. S., Dubbs, H. A., Collins, A. T., Andrieux, J., Dieux-Coeslier, A., Haberlandt, E., Kotzot, D., … Lu, X. (2016). ASPP2 deficiency causes features of 1q41q42 microdeletion syndrome. *Cell Death & Differentiation*, *23*(12), Article 12. https://doi.org/10.1038/cdd.2016.76

Zepeda-Mendoza, C. J., Ibn-Salem, J., Kammin, T., Harris, D. J., Rita, D., Gripp, K. W., MacKenzie, J. J., Gropman, A., Graham, B., Shaheen, R., Alkuraya, F. S., Brasington, C. K., Spence, E. J., Masser-Frye, D., Bird, L. M., Spiegel, E., Sparkes, R. L., Ordulu, Z., Talkowski, M. E., … Morton, C. C. (2017). Computational Prediction of Position Effects of Apparently Balanced Human Chromosomal Rearrangements. *American Journal of Human Genetics*, *101*(2), 206–217. https://doi.org/10.1016/j.ajhg.2017.06.011

Zhang, C.-Z., Leibowitz, M. L., & Pellman, D. (2013). Chromothripsis and beyond: Rapid genome evolution from complex chromosomal rearrangements. *Genes & Development*, *27*(23), 2513–2530. PubMed. https://doi.org/10.1101/gad.229559.113

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., & Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, *41*(7), Article 7. https://doi.org/10.1038/ng.399

Zhang, X., Novera, W., Zhang, Y., & Deng, L.-W. (2017). MLL5 (KMT2E): Structure, function, and clinical relevance. *Cellular and Molecular Life Sciences: CMLS*, *74*(13), 2333–2344. https://doi.org/10.1007/s00018-017-2470-8

Zhang, Y., Wong, J., Klinger, M., Tran, M. T., Shannon, K. M., & Killeen, N. (2009). MLL5 contributes to hematopoietic stem cell fitness and homeostasis. *Blood*, *113*(7), 1455–1463. https://doi.org/10.1182/blood-2008-05-159905

Zhang, Yang, J., Liu, Y., Chen, X., Yu, T., Jia, J., & Liu, C. (2009). PR55 alpha, a regulatory subunit of PP2A, specifically regulates PP2A-mediated beta-catenin dephosphorylation. *The*

Journal of Biological Chemistry, *284*(34), 22649–22656. PubMed. https://doi.org/10.1074/jbc.M109.013698

Zhao, R.-Z., Jiang, S., Zhang, L., & Yu, Z.-B. (2019). Mitochondrial electron transport chain, ROS generation and uncoupling (Review). *International Journal of Molecular Medicine*, *44*(1), 3–15. https://doi.org/10.3892/ijmm.2019.4188

Zhou, W.-Z., Zhang, J., Li, Z., Lin, X., Li, J., Wang, S., Yang, C., Wu, Q., Ye, A. Y., Wang, M., Wang, D., Pu, T. Z., Wu, Y.-Y., & Wei, L. (2019). Targeted resequencing of 358 candidate genes for autism spectrum disorder in a Chinese cohort reveals diagnostic potential and genotype-phenotype correlations. *Human Mutation*, *40*(6), 801–815. PubMed. https://doi.org/10.1002/humu.23724

Žiburkus, J., Cressman, J. R., & Schiff, S. J. (2013). Seizures as imbalanced up states: Excitatory and inhibitory conductances during seizure-like events. *Journal of Neurophysiology*, *109*(5), 1296–1306. https://doi.org/10.1152/jn.00232.2012

Ziegler, A., Colin, E., Goudenège, D., & Bonneau, D. (2019). A snapshot of some pLI score pitfalls. *Human Mutation*, *40*(7), 839–841. https://doi.org/10.1002/humu.23763

Zorov, D. B., Vorobjev, I. A., Popkov, V. A., Babenko, V. A., Zorova, L. D., Pevzner, I. B., Silachev, D. N., Zorov, S. D., Andrianova, N. V., & Plotnikov, E. Y. (2019). Lessons from the Discovery of Mitochondrial Fragmentation (Fission): A Review and Update. *Cells*, *8*(2), 175. https://doi.org/10.3390/cells8020175

Zungu, M., Schisler, J., & Willis, M. S. (2011). All the little pieces. -Regulation of mitochondrial fusion and fission by ubiquitin and small ubiquitin-like modifer and their potential relevance in the heart.-. *Circulation Journal: Official Journal of the Japanese Circulation Society*, *75*(11), 2513–2521. https://doi.org/10.1253/circj.cj-11-0967

Zweier, M., Begemann, A., McWalter, K., Cho, M. T., Abela, L., Banka, S., Behring, B., Berger, A., Brown, C. W., Carneiro, M., Chen, J., Cooper, G. M., Deciphering Developmental Disorders (DDD) Study, Finnila, C. R., Guillen Sacoto, M. J., Henderson, A., Hüffmeier, U., Joset, P., Kerr, B., … Rauch, A. (2019). Spatially clustering de novo variants in CYFIP2, encoding the cytoplasmic FMRP interacting protein 2, cause intellectual disability and seizures.

European Journal of Human Genetics: EJHG, *27*(5), 747–759. https://doi.org/10.1038/s41431-018-0331-z

Appendix

Table S3.1: List of prioritised variants in DEE23 patient for further scrutinization. Text in colour "red" indicates the reason for exclusion from further analysis.

MOI	Gene	Variant	Pathogenicity predictors	gnomAD
***De novo* (het)**	*FCGR3A*	NC_000001.10(NM_000569.8): c.700G>A; p.(Arg233*)	CADD = 35 pLoF= Low-confidence (END_TRUNC)	0.00004784
	PCGF5	NC_000010.10(NM_032373.5):c.-57892G>A	15.97	-
	RNF135	NC_000017.10:g.29297137T>C	5.45	0.00006415
	RPS9	NC_000019.9(NM_001321706.2):c.-24C>T	CADD: 16.4 SpliceAI: 0.0200 (acceptor gain)	0.00001774
Recessive (simple)	***DOCK7***	NC_000001.10(NM_001271999.1):c.5691+1G>T	CADD = 35 SpliceAI _DS_DL= 1	-
	CACHD1	NC_000001.10(NM_020925.4):c.261+14609C>T	CADD = 15.5	0.00009569
	IFI44L	NC_000001.10:g.79086101T>A	CADD = 5.4	
	HES6	NC_000002.11:g.239149016_239149019del	-	-
	GABRA2	NC_000004.11(NM_000807.4): c.*1880G>A	CADD = 22.4	-
		NC_000005.9:g.4418261T>G	CADD = 18.1	0.001751 (1 hmz)
	RPS14	NC_000005.9(NM_005617.4): c.382G>A; p.Arg128Trp	CADD = 23.2 SIFT = deleterious PolyPhen-2 = benign	-
	TNIP1	NC_000005.9(NM_006058.5):c.-359C>T	CADD = 4.89	-
	MCF2L	NC_000013.10(NM_001112732.3):c.40G>A; p.Glu14Lys	CADD = 24.2 SIFT = deleterious PolyPhen-2 = benign	0.00003918
	SCAF1	NC_000019.9(NM_021228.3):c.1742_1747del; p.Arg581_Thr583delinsPro	-	-
Recessive (compound heterozygous)	*VWA5B1*	NC_000001.10(NM_001039500.3):c.1224C>G; p.(Ser408Arg)	CADD = 4.08 SIFT deleterious PolyPhen-2 = benign	0.0003937
		NC_000001.10(NM_001377531.1):c.2683G>T; p.(Ala895Ser)	CADD = 2.8 SIFT = benign PolyPhen-2 = tolerated	-
	PPFIA4	NC_000001.10:g.202995311C>G	3.23	0.006753 (1 hmz)
		NC_000001.10:g.202995589G>T	8.074	-
	PKHD1	NC_000006.11(NM_138694.4):c.11075G>A; p.(Arg3692Gln)	CADD = 17.36 SIFT = tolerated Polyphen-2 = benign	0.000007966
		NC_000006.11(NM_170724.3):c.7675G>C; p.(Val2559Leu)	CADD = 2.11 SIFT = tolerated Polyphen-2 = benign	
	TOX	NC_000008.10:g.60032211G>A	CADD = 7.547	-
		NC_000008.10:g.60032335G>T	CADD = 8.702	-
	COQ10A	NC_000012.11:g.56660759A>G	CADD = 17.1	0.00003188
		NC_000012.11(NM_144576.4):c.233C>T; p.(Ala78Val)	CADD = 22.1 SIFT = tolerated Polyphen-2 = benign	0.0003125

S3.2.3 Clinical phenotypic comparison between patients harboring biallelic variants in DOCK7

Eight patients have been reported in the literature (Perrault et al., 2014; Turkdogan et al., 2019; Bai et al., 2019; Haberlandt et al., 2021; Kapoor et al., 2022). These patients shared a core set of symptoms but showed significant heterogeneity. Epilepsy in these patients is poorly responsive to treatment, except for the patient reported by Turkdogan et al., who responded to clozapam and has since been seizure-free (Turkdogan et al., 2019). A patient reported by Kapoor et al. responded to valproate; however later died of seizure complications at 14 months (Kapoor et al., 2022). Similarly, all patients showed multifocal epileptic activity on EEG examination except the Turkdogan et al. patient with reportedly normal EEG for over 16 months. Development is severely impaired in all patients with motor, speech and social skills manifestations. However, two siblings with identical compound heterozygous *DOCK7* variants reported by Perrault et al. exhibit different degrees of impairment: whereas one sibling could not walk or speak independently at age 7, her sister is able to walk independently, can speak approx. 30 words and communicate in simple commands. Similarly, although one of the reported patients died at 14 months, two patients reached adulthood with multiple developmental milestones (Haberlandt et al., 2021; Kapoor et al., 2022). All reported patients (8/8) showed vision abnormalities associated with the diagnosis of cortical blindness, chiefly manifesting as lack of ocular contact and reaction to visual stimuli. However, one patient developed grossly normal visual pursuit over time. Additional shared features include low anterior hairline, reduced occipitofrontal circumference, abnormally marked pontobulbar sulcus and pontine hypoplasia. Other features show greater variability between patients. However, the full genetic spectrum and the associated clinical spectrum of pathogenic *DOCK7* variants are unknown. Moreover, the penetrance of each feature of the *DOCK7*-associated epileptic syndrome remains to be characterised.

Table S3.2: Clinical features of HICF2 case (section 3.2.3) and reported patients in the literature with biallelic DOCK7 mutations. Het, heterozygous; hom, homozygous; ROH, region of homozygosity

	Case in section 3.2.3	Perrault A1	Perrault A2	Perrault	Turkdogan2019	Bai2019	Haberlandt2021_P1	Haberlandt2021_P2	Kapoor2022
DOCK7 amino acid change	c.5691+1G>T	p.Asp837Alafs*48; p.Arg1237*	p.Asp837Alafs*48; p.Arg1237*	p.Glu2078*; p.Ser328*	p.Leu1117*	c.5291-1G>C, p.Arg827*	c.390_3936dup	c.390_3936dup	c.4897-1G>A
variant		NC_000001.10(NM_001271999.1):c.2510del	NC_000001.10(NM_001271999.1):c.3709C>T	NC_000001.10(NM_001271999.1):c.6232G>T; NC_000001.10(NM_001271999.1):c.983C>G	NC_000001.10(NM_001271999.1):c.3350T>A	NC_000001.10(NM_001271999.1):c.5929-1G>C; NC_000001.10(NM_001271999.1):c.2479C>T	NC_000001.10(NM_001271999.1):c.390-2468_3936+719dup	NC_000001.10(NM_001271999.1):c.390-2468_3936+719dup	NC_000001.10(NM_001271999.1):c.4897-1G>A
in gnomAD (v2.1.1: 4/12/19)	-	- missense only (Asp837Asn; Arg1237Lys; Arg1237Gly)	- missense only (Asp837Asn; Arg1237Lys; Arg1237Gly)	-	- synonymous only (Leu1086Leu)	- missense only (Arg827Leu, Arg827Gln)	-	-	-
Inheritance	homozygous	compound heterozygous	compound heterozygous	compound heterozygous	homozygous	compound heterozygous	homozygous	homozygous	homozygous
CNV analysis	dup:chr1; regions of homozygosity								
Sex	F	F	F	F	M	F	F	F	F
Age	?						27 y	23 y	14 m (deceased)
Pregnancy, birth									term delivery, 2.4 kg
Epileptic encephalopathy	+	+	+	+	+	+	+	+	+
Seizures	+	+	+	+	+	+	+	+	+
Seizure treatment response	presentation at 9m; controlled with valproate and levetiracetam combination	infantile spams disappeared on adenocorticotropic hormone; poorly controlled in general	poorly controlled myoclonic, partial complex, tonic seizures	poorly controlled repeated tonic-clonic seizures	generalised tonic-clonic seizures evolving to status epilepticus; frequent motor seizures responsive to clobazam (seizure free since)	head nodding and infantile spasms 20x a day starting 6m. Drug response unknown	Currently, seizure-free under antiepileptic therapy (levetiracetam, clobazam, zonisamide, midazolam)	Currently, short absences despite antiepileptic therapy	well controlled by sodium valproate
Abnormal EEG	+	+ multifocal epileptic activity; hypsarrythmia	+ multifocal epileptic activity	+ multifocal epileptic activity (2-3y)	+ initial generalised sharp waves, now normal EEG	+ multifocal epileptic activity; hypsarrythmia	+ Hypsarrhythmia, multifocal epileptic activity	+ Hypsarrhythmia, multifocal epileptic activity	normal background activity with multifocal spike and wave discharges
Cortical blindness	+	+ lack of attentiveness to visual stimuli; normal eye movements, pupillary rxn, fundus; cortical blindness but consistent improvement - grossly normal visual pursuit	+ lack of attentiveness to visual stimuli; normal eye movements, pupillary rxn, fundus; cortical blindness; cannot see well enough to play with toys	+ cortical blidness, complete absence of rxn to visual threat; wandering eye movements	+ cortical blindness (VEP diagnosis); lack of any visual contact	+ cortical blindness; lack of ocular contact and reaction to visual stimulus	+ FEVP: normal waveform of markedly decreased amplitude	+ nearly no vision	+ absent visual evoked potentials, did not follow light
Vision abnormalities	+	+	+	+	+	+	+	+	+
Heterochromia of irides						NA			
Nystagmus	?						+	+	
Strabismus	?								
Ptosis									
Dysmorphic features		+	+	+	+	+			+
Bitemporal narrowness	+			+					
Low anterior (= frontal) hairline		+	+	+	+	+	+	+	
Thick eyebrows				+					+
Synophrisis				+					
Synophrisis				+					
Periorbital fullness	+	+	+			+	+	+	+
Telecanthus		+	+				+	+	
Broad nasal tip		+	+	+	+		+	+	
Short philtrum				+	+		+		
Others					normo-brachycephaly; narrow forehead; prominent ears; full cheeks; long eyelashes; thin upper lip	protruding ears; highly arched palate; low-set ears; gingival maldevelopment	highly arched palate, long eyelashes, low set ears, thin upper lip		brachycephaly with prominent metopic suture, white forlock of hair on the forehead, synophrys, enopthalmia, small nose with an anteverted bulbous nasal tip, low set protruding cup-shaped ears, smooth philtrum with thin upper lip, fetal finger pads and fleshy hands
Microcephaly	+								+
OFC	-2 SD	45 cm (10 perc.)	42.2 cm (10 perc.)	-1 SD	NA	47.5 cm at 3y (30%)			-4.53 SD
Brain abnormalities		+	+	+	+	+			+
Abnormally marked pontobulbar sulcus		+	N/A	+	+	+	+	+	+
Pontine hypoplasia	+	+	N/A	+	+	+	+	+	+
Thin corpus callosum	-	+	+		+	+			+
Occipital atrophy		+	N/A	+	+	-	+	+	+

Appendices for chapter 4:

A supplementary video of COS7 cells expressing the WT PLPPR5 – S4.1: "PLPPR5 WT expression in COS7 cells.pptx".

PLPPR5 WT-mClover3 plasmid was transfected into COS7 cells, which were imaged live as described in section 4.2.2.9 using the Olympus system. The cells were imaged for 30 minutes.

Appendices for chapter 5:

Provided as .xlsx file including:

- Table S5.1- All SVs identified in the gene-based analysis.
- Table S5.2- molecular characterisation of translocations and their breakpoints.