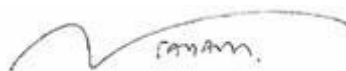


Clinical and molecular delineation of neurodevelopmental disorders within genetically isolated communities

Submitted by James Fasham
to the University of Exeter
as a thesis for the degree of
Doctor of Philosophy in Medical Studies
in October 2022

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.



Signature:

Acknowledgements and dedication

I would like to acknowledge and thank my tireless, dedicated supervisors and great supporters, Professors Emma Baple and Andrew Crosby. Without their interest and initial encouragement I would not have had the confidence or the platform to apply for my funding and take on this academic journey. Next, I thank my funders, GW4-CAT and Wellcome for believing in me and this project. Also, thanks to other supervisors and mentors, James Unay, Andrew Hattersley and Angus Jones, for their wisdom and guidance.

There would of course be no project if not for our patients and collaborators, I would especially like to thank Reham, who works tirelessly for her patients.

The lab team I have worked with have been wonderful and have become real friends. Joe's selflessness and calm make our lab run smoothly and a welcoming place to be; I continue to learn from Claire and Lettie, great writers, thinkers and clinicians. Nishanka is as helpful and productive a student as we could wish to have had. Finally, Adam, an adopted member of the group and friend. Thank you also members who have moved on and including Serene, Illaria, Barry and Gaurav who first introduced me to the lab. A special thank you goes to Matthew who kindled an interest in bioinformatics in me and continues to support me.

Last, but most importantly my family who have had to put up with less of me that I would have liked and have supported me throughout.

Abstract

The unique genetic make-up of genetically isolated communities, where otherwise rare genetic founder variants may become enriched, allows the clinical relevance of these variants to be more precisely determined where this may not otherwise be possible. This thesis details clinical and genomic studies of inherited neurodevelopmental disorders identified within Amish and Palestinian communities to advance understanding of the pathomolecular basis of these diseases.

Chapter 3 describes the discovery of a novel clinically recognisable syndromic microcephalic neuronal migration disorder with similarities to the “tubulinopathies”, resulting from biallelic variants in *CAMSAP1*, a microtubule-associated molecule. This finding, stemming from investigations in an extended Palestinian family, entailed collaborative clinical, genomic, cell and mouse studies. Four additional unrelated affected families were identified, highlighting the global relevance of such work. This chapter also documents preliminary studies identifying *CAMSAP2* and *CAMSAP3* variants as candidate causes of an overlapping phenotype.

Chapter 4 describes the use of whole-genome sequencing to identify a homozygous *SLC4A10* multi-exon deletion in two Palestinian children with microcephaly, abnormal slit-like lateral ventricles and features consistent with autistic spectrum disorder, closely mirroring findings in *Slc4a10*^{-/-} mice. Eight affected individuals from four unrelated families were subsequently identified. Collaborative mouse and functional data determined that presynaptic inhibitory

GABAergic transmission is compromised, identifying a potential therapeutic approach involving GABA_A receptor agonists.

Chapter 5 illustrates how the serendipitous accumulation in a community setting of otherwise rare gene variants enables their clinical relevance to be correctly elucidated. Clinical and genetic studies of *SCN9A* gene variants in the Amish and UK Biobank conclusively refute previous associations between *SCN9A* and epilepsy, leading ClinGen to re-evaluate this incorrect disease-gene association.

Together the work described in this thesis provides new insights into pathomolecular neurodevelopmental processes and improves scientific and clinical understanding of rare genetic variation, illustrating how community genomic research may expedite discovery of new rare diseases, improve clinical care and ultimately aid development of targeted treatments for these disorders.

Table of Contents

Acknowledgements and dedication.....	2
Abstract	3
Table of contents	5
List of Tables	7
List of Figures	9
Abbreviations.....	11
1. Introduction.....	15
1.1. Neurodevelopmental disorders	16
1.2. Human neurodevelopment.....	18
1.3. Neuronal migration	21
1.4. Neuronal migration disorders	23
1.5. Microtubules in cortical development	31
1.6. GABAergic neurodevelopment.....	34
1.7. Disorders of GABAergic dysfunction	36
1.8. Rare genetic disorders	37
1.9. Strategies to identify new monogenic causes of neurodevelopmental disorders	40
1.10. Autosomal recessive neurodevelopmental disorders.....	43
1.11. Genomic studies in genetically isolated communities empowers neurodevelopmental disorder disease gene identification	45
1.12. The North American Anabaptist communities	47
1.13. The Windows of Hope project	49
1.14. Palestinian communities.....	50
1.15. The need for increased ancestral diversity in genomic databases	53
1.16. Aims and objectives.....	56
1.17. Aims of the project.....	56
1.18. References	58
2. Materials and Methods.....	69
2.1. Buffers, reagents and stock solutions	70
2.2. Subjects and samples	71
2.3. Molecular methods	72
2.4. Next-generation sequencing	85
2.5. Bioinformatic methods.....	87
2.6. DDD complementary analysis project.....	96
2.7. Interrogation of the 100,000 Genomes Project dataset	97
2.8. References	100
3. Biallelic <i>CAMSAP1</i> variants cause a clinically recognizable neuronal migration disorder.....	102
3.1. Acknowledgements of co-authors and contributions to the paper	103
3.2. Manuscript.....	105
3.3. Supplemental Material.....	131
3.4. Further findings and future work.....	155
3.5. References	161
4. <i>SLC4A10</i> variants impair GABAergic transmission and CSF secretion causing a recognizable neurodevelopmental disorder in humans and mice.....	164
4.1. Acknowledgements of co-authors and contributions to the paper	165
4.2. Manuscript.....	167
4.3. Supplemental Material.....	207
4.4. Further findings and future work.....	231
4.5. References	235

5.	No association between <i>SCN9A</i> and monogenic human epilepsy disorders	242
5.1.	Acknowledgements of co-authors and contributions.....	243
5.2.	Abstract	244
5.3.	Introduction.....	245
5.4.	Discussion	252
5.5.	Supplemental material.....	256
5.6.	Further findings and future work.....	263
5.7.	References	265
6.	Concluding comments.....	272
6.1.	References	282
7.	Appendix	284
7.1.	Peer-reviewed manuscripts arising from this project	285
7.2.	Example QC metrics – Coverage @ 20X for Palestinian exome data.....	288
7.3.	QC.py – QC metric aggregator and plotter	289
7.4.	Additional scripts developed as part of this thesis	293
7.5.	VirtualPanel.py – A virtual panel creator	295
7.6.	SecondHit.sh - a tool to extract all unfiltered variants for a particular gene.....	297
7.7.	databaseQuery.sh - a tool to query a database of variants	298
7.8.	Research proposal: DDD (CAP330).....	300
7.9.	Research proposal: 100,000 Genomes project (RR349).....	301
7.10.	Biallelic.py – for extracting 100,000 Genomes project data	302
7.11.	Biallelic.py – for filtering 100,000 Genomes project data	303
7.12.	CAMSAP1 study data collection proforma	304
7.13.	Relative expression of CAMSAPs and MARK2 in human cells	305
7.14.	Identified individuals with variants in <i>CAMSAP3</i>	306
7.15.	Identified individuals with variants in MARK2.....	307
7.16.	<i>SLC4A10</i> study data collection proforma	308
7.17.	Additional individuals were identified with biallelic missense variants in <i>SLC4A10</i> ..	309
7.18.	Clinical features of all individuals identified with missense variants in <i>SLC4A10</i>	310
7.19.	Database frequency and <i>in silico</i> predictions of all missense variants identified in <i>SLC4A10</i>	311
7.20.	Distribution of <i>SLC4A10</i> missense variants with regard to protein domains and intra/extracellular location of the <i>SLC4A10</i> protein	312
7.21.	Additional protein modelling of missense variants in <i>SLC4A10</i>	313
7.22.	Functional studies of additional missense variants in <i>SLC4A10</i>	314

List of Tables

Chapter 1

Table 1.1: DSM-5 and ICD-11 subclassifications of neurodevelopmental disorders	17
Table 1.2: Tubulin subunits and complex proteins associated with neuronal migration disorders	25
Table 1.3: Microtubule associated gene families and associations with monogenic disease.....	30

Chapter 2

Table 2.1: Reagents used in this study	70
Table 2.2: Buffers and stock solutions used in this study.....	70
Table 2.3: Standard 10 µl PCR reaction mixture	78
Table 2.4: 10 µl PCR reaction mixture using an integrated PCR master mix	79
Table 2.5: Touchdown PCR protocol.....	79
Table 2.6: 22 µl ddPCR reaction mixture.....	85
Table 2.7: Default rules for filtering software	89-90
Table 2.8: Criteria for variant prioritisation.....	91-93

Chapter 3

Table 3.1: Summary of clinical and neurological features of individuals with CAMSAP1-related neuronal migration disorder.....	115
Table 3.S1: Pathogenic CAMSAP1 variants identified in this study with their frequency in population databases	150
Table 3.S2: Other variants identified through exome sequencing.....	151-153
Table 3.S3: Antibodies used in immunocytochemistry experiments.	154

Chapter 4

Table 4.1: Clinical findings in individuals with biallelic SLC4A10 variants.	184
Table 4.S1: Human and murine disorders associated with other SLC4A class molecules.	221
Table 4.S2: SLC4A10 variants identified in affected individuals in this study	223
Table 4.S3: Variants identified by exome / genome sequencing	224-226
Table 4.S4: Summary of electrophysiological recording from acute brain slices of Slc4a10 WT and knock-out mice	227
Table 4.S5: SLC4A10 GWAS associations	228-229

Chapter 5

Table 5.1: SCN9A variants proposed as a monogenic causes of epilepsy.	246
Table 5.S1: UK Biobank allele frequencies for the SCN9A variants in Table 5.1	260

Table 5.S2: Heterozygous <i>SCN9A</i> variants proposed as a monogenic cause of seizure disorders in subsequent publications, including the testing methodology employed.....	261
Table 5.S3: Rare variant burden analysis in UK Biobank	262

List of Figures

Chapter 1

Figure 1.1: A timeline of human neural development	18
Figure 1.2: Neurulation	19
Figure 1.3: Evolution of the neural tube	20
Figure 1.4: Human cortical development by radial migration	21
Figure 1.5: Network and phenotypic classifications of neuronal migration disorders	27
Figure 1.6: Molecular causes of lissencephaly in a cohort of 811 patients.....	28
Figure 1.7: The structure and properties of microtubules	31
Figure 1.8: The migration of GABAergic interneurons in the developing cortex.....	35
Figure 1.9: Number of gene-phenotype relationships in MIM database	38
Figure 1.10: The cycle of rare disease testing, discovery, and treatment.....	40
Figure 1.11: Cost per genome (2001-2021).....	41
Figure 1.12: The long tail of neurodevelopmental disorder disease gene discovery	44
Figure 1.13: Ancestral bottleneck events result in reduced genetic diversity with enrichment of rare alleles.....	46
Figure 1.14: Territory claimed by Palestine.....	51

Chapter 2

Figure 2.1: Bioinformatic pipeline for exome and genome sequencing	87
---	----

Chapter 3

Figure 3.1: Family pedigrees and biallelic CAMSAP1 variants associated with a syndromic neuronal migration disorder.....	110
Figure 3.2: Neuroimaging in 4 individuals with the CAMSAP1-related neuronal migration disorder	112
Figure 3.3: Patient iPSCs display decreased proliferation and differentiation and increased apoptosis of neural progenitor cells.	121
Figure 3.4: Expression of Camsap1 in the CNS and developing facial primordia	123
Figure 3.S1: Facial features of individuals with CAMSAP1-related disorder	141
Figure 3.S2: Additional MRI brain images from individuals with CAMSAP1-related disorder .	142
Figure 3.S3: Homozygous regions greater than 1Mb shared between individuals IV:10 and V:1, generated using AutoMap (Quinodoz et al. 2021)	143
Figure 3.S4: CAMSAP1 sequencing chromatograms from Family 1	144
Figure 3.S5: Recurrent CAMSAP1 deletion may result from homologous recombination.....	145
Figure 3.S6: Camsap1 genotyping and Mendelian survival.....	146
Figure 3.S7: Postnatal mutant mice do not survive in normal ratios but exhibit no skeletal abnormalities.	147
Figure 3.S8: Embryonic Camsap1 null mice exhibit no morphological phenotypes.	148

Figure 3.S9: Camsap1 target region for RNA scope probe (ACDBio Cat# 866521)	149
Figure 3.5: Proteins active at the minus end of the microtubule	155

Chapter 4

Figure 4.1: Family pedigrees and biallelic SLC4A10 variants	183
Figure 4.2: Neuroimaging from affected individuals with biallelic SLC4A10 variants	189
Figure 4.3: Slc4a10 ^{-/-} mice show behavioural abnormalities in the 2-object novel object recognition task and display grossly intact cortical architecture.	194
Figure 4.4: Localisation of SLC4A10 to GABAergic presynapses	196
Figure 4.5: SLC4A10 acts on presynaptic pH _i to promote GABA release in CA1 pyramidal neurons.	199
Figure 4.S1: Genome sequencing data of affected individuals in Family 1, reveals a biallelic, multi-exon, deletion of SLC4A10.	210
Figure 4.S2: ddPCR data confirms the presence of a multi-exon deletion of SLC4A10 in individuals III:1 and III:2 (Family 1).	211
Figure 4.S3: RT-PCR demonstrates mRNA splicing effect of NM_001178015:c.2863-2A>C p.(Gln954_Phe955ins*13) variant (Family 3).	212
Figure 4.S4: T2-weighted sagittal MRI scan of wild-type and Slc4a10 knockout mice.....	213
Figure 4.S5: Additional MRI images of individuals affected by SLC4A10-related disorder	214
Figure 4.S6: Heterologous expression of SLC4A10 wild-type and disease associated missense variants in N2a cells	215
Figure 4.S7: Acid extrusion by disease associated SLC4A10 missense variants is compromised.	216
Figure 4.S8: Disruption of Slc4a10 reduces the mIPSC frequency in CA3 pyramidal neurons.	217
Figure 4.S9: Disruption of Slc4a10 reduces the sIPSC frequency in CA1 pyramidal neurons.	218
Figure 4.S10: Intracellular acidification with sodium propionate impairs GABA release.	219
Figure 4.S11: Protein modelling of missense SLC4A10 variants	220

Chapter 5

Figure 5.1: Family pedigrees showing SCN9A NM_002977:c.1921A>T p.(Asn641Tyr) genotype data.....	250
Figure 5.S1: Sodium voltage-gated channel alpha subunit gene family expression data	258
Figure 5.S2: Variant clustering in SCN9A alongside other VGSC- α genes (SCN1A and SCN3A) associated with epilepsy.....	259

Chapter 6

Figure 6.1: Two isolated communities show different patterns of autozygosity	275
---	-----

Abbreviations

-TIPS	Microtubule minus-end-tracking proteins
+TIPS	Microtubule plus-end-tracking proteins
ABR	Auditory brainstem response
AC	Allele count
aCC	Agenesis of the corpus callosum
aCSF	Artificial cerebrospinal fluid
AD	Autosomal dominant
ADHD	Attention-Deficit/Hyperactivity Disorder
AF	Allele frequency
AMPA	α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ANOVA	Analysis of variance
API	Application Programming Interface
AR	Autosomal recessive
array CGH	Microarray-based comparative genomic hybridisation
AVS	Anabaptist Variant Server
BAM	Binary Alignment/Map
BAPTA	1,2-bis(o-aminophenoxy)ethane-N,N,N',N'-tetraacetic acid
BCECF	2',7'-Bis(2-carboxyethyl)-5(6)-carboxyfluorescein
BD	Twice daily
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like alignment tool
bp	base pairs
BR	Broad range
CADD	Combined Annotation Dependent Depletion
CASAVA	Consensus Assessment of Sequence And Variation
CC	Coiled-coil
CC BY	Creative Commons Attribution license
CC BY-SA	Creative Commons Attribution ShareAlike license
CCD	Charged-coupled device
CCHMC	Cincinnati Children's Hospital Medical Center
CKK	CAMSAP1, KIAA1078/CAMSAP2, KIAA1543/CAMSAP3 (domain)
CNS	Central nervous system
CNV	Copy number variant
co-IP	Co-immunoprecipitation
CRISPR	Clustered regularly interspaced short palindromic repeats
CSF	Cerebrospinal fluid
DAPI	4', 6-Diamidino-2-phenylindole dihydrochloride
DDD	Deciphering Developmental Disorders
ddH ₂ O	Double distilled water
ddPCR	Droplet digital PCR
dl-APV	DL-2-Amino-5-phosphonopentanoic acid
DMEM/F12	Dulbecco's Modified Eagle's Medium/Nutrient Mixture F-12
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide triphosphates
DSM-5	Diagnostic and Statistical Manual of Mental Disorders
E/I ratio	Excitatory-to-inhibitory ratio
EEG	Electroencephalogram

EGTA	Egtazic acid
EIPA	5-(N-ethyl-N-isopropyl) amiloride
ELIXIR	European Life-Science Infrastructure
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
ENST	Ensemble transcript
ERG	Electroretinogram
ESHG	European Society of Human Genetics
EU	European Union
FEP / FVEP	Flash (visual) evoked potential
FS+	Febrile seizures plus
GABA	Gamma-aminobutyric acid
GATK	Genome Analysis Toolkit
GC content	Guanine-cytosine content
GDD	Global developmental delay
GDPR	General Data Protection Regulation
GECIP	Genomics England Clinical Interpretation Partnership
GEFS+	Genetic epilepsy with febrile seizures plus
GenCC	the Gene Curation Coalition
GMFCS	Gross Motor Function Classification System
gnomAD	The Genome Aggregation Database
GRCh37	Genome Reference Consortium human genome build 37
GRCh38	Genome Reference Consortium human genome build 38
GTEX	Genotype-Tissue Expression
GTP	Guanosine-5'-triphosphate
gVCF	Genomic variant call format
GWAS	Genome-wide association studies
hCC	Hypogenesis of the corpus callosum
HCO ₃ ⁻	Bicarbonate
HE (staining)	Hematoxylin and eosin
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HES	Hospital episode statistics
HPC	High-performance computing cluster
HPO	Human phenotype ontology
HTA	Human Tissue Authority
ICD-9	International Classification of Diseases, Ninth Revision
ICD-10	International Statistical Classification of Diseases and Related Health Problems 10th Revision
ICD-11	International Statistical Classification of Diseases and Related Health Problems 11th Revision
ID	Intellectual disability
IGV	Integrative genome viewer
IHC	Immunohistochemistry
IM	Intramuscular
IMPC	International Mouse Phenotyping Consortium
InDel	Insertion and/or deletion
IPC	Intermediate progenitor cells
iPSCs	Induced Pluripotent Stem Cells
IQ	Intelligence quotient
IRB	Institutional review board
KO	Knockout
KS-test	Kolmogorov–Smirnov test
LAB buffer	Lithium acetate borate buffer

LINE	Long interspersed nuclear element
LMIC	Low / lower middle income country
LMS	Lambda-Mu-Sigma
LOEUF	Loss-of-function observed/expected upper bound fraction
MAF	Minor allele frequency
MANE select	Matched Annotation from NCBI and EMBL-EBI selected transcript
MAP	Microtubule-associated protein
mEPSC	Miniature excitatory postsynaptic currents
MIM	Mendelian Inheritance in Man
mIPSC	Miniature inhibitory postsynaptic currents
MMRRC	The Mutant Mouse Resource and Research Center
MOOC	Massive open online course
MPRF	Multiple potentially relevant findings
MR	Magnetic resonance
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MTOC	Microtubule organising centre
NCBE	Na ⁺ -coupled Cl ⁻ /HCO ₃ ⁻ exchanger
NCBI	National Center for Biotechnology Information
NFE	Non-Finnish European
NHS	(UK) National Health Service
NIHR	National Institute for Health and Care Research
nIPCs	Neuronal intermediate progenitor cells
NMDA	N-methyl-D-aspartate
NOR	Novel-object recognition
OFC	Occipitofrontal circumference
OMIM	Online Mendelian Diseases in Man
ORCID	Open Researcher and Contributor Identifier
oRG	Outer subventricular zone radial glia-like cells
P>A	Posterior more severe than anterior
PBMC	Peripheral blood mononuclear cell
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Parenteral gastrostomy
PFA	Paraformaldehyde
pHi	intracellular pH
PHRC	Palestinian Health Research Council
PMR	Plasma membrane region
PROMs	Patient reported outcome measures
PTZ	Pentylenetetrazole
QC	Quality control
REVEL	Rare Exome Variant Ensemble Learner
RFP	Red Fluorescent Protein
RNA	Ribonucleic acid
RT-PCR	Reverse transcription PCR
SDS	Standard deviation score
SEM	Standard error of the mean
sFTP	Secure file transfer protocol
SIFT	Sorting intolerant from tolerant
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism

SNV	Single nucleotide variant
SPB	Seal Point Balinese (pipeline)
SSVEP	Steady state visual evoked potential
TD-PCR	Touchdown PCR
T _m	Melting temperature
tRG	Truncated radial glia
TriMA	Trimethylamine chloride
TSA	Tyramide Signal Amplification
TTX	Tetrodotoxin
UCSC	University of California Santa Cruz
UCSD	University of California San Diego
UK	United Kingdom of Great Britain and Northern Ireland
UNICEF	United Nations International Children's Emergency Fund
US / USA	United States (of America)
UTR	Untranslated region
VCF	Variant call format
VDCCs	Voltage-gated calcium channels
VEP	Variant effect predictor
VGSC- α	voltage-gated sodium channel alpha
VUS	Variant of uncertain significance
VZ	Ventricular zone
WES	Whole-exome sequencing
WGA	Wheat germ agglutinin
WGS	Whole-genome sequencing
WHO	World health organisation
WT	Wild type
γ -TuRC	γ tubulin ring complex
1X	Working concentration
50X	50 times working concentration

1

Introduction

1.1. Neurodevelopmental disorders

The common involvement of the central nervous system (CNS) in many monogenic disorders is consistent with the observation that >80% of all human genes are expressed at some stage of brain development (Hawrylycz *et al.*, 2012). Disruption to gene function may result in a rare neurogenetic disorder, which include a vast spectrum of developmental or degenerative conditions that may affect any region of the brain or spinal cord. Despite the rapid progress of neurogenetic disease gene discovery, only ~25% of computationally annotated human genes currently have an established association with a disease phenotype (Mitani *et al.*, 2021) and more than half of patients do not receive a molecular diagnosis (Beaulieu *et al.*, 2014; Boycott *et al.*, 2017; Smedley *et al.*, 2021; Wright *et al.*, 2018). However, by providing fundamental insight into the disease process, neurogenetics has become established as an important discipline with potential to develop new therapeutic drugs to treat these conditions (Baple *et al.*, 2014; Harlalka *et al.*, 2013; Mitani *et al.*, 2021).

Neurodevelopmental disorders affect the formation and development of the nervous system. They typically manifest as a delay in achieving, or incomplete achievement of, psychomotor milestones (learning impairment), leading to difficulties in communication and cognitive impairment / intellectual disability (ID). Sometimes these features may exist in isolation, but they may also be associated with other symptoms or signs, such as epilepsy, and may form part of a recognised syndrome (Parenti *et al.*, 2020). The term neurodevelopmental disorder has broad usage. The classifications from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric

Association, 2017) and the International Statistical Classification of Diseases and Related Health Problems (“ICD-11”) (World Health Organization (WHO), 2019/2021), both widely used in psychiatry, education and healthcare billing, are probably the most established (**Table 1.1**).

DSM-5	ICD-11
Intellectual Disorders:	6A00 Disorders of Intellectual Development
- Intellectual Developmental Disorder	- 6A00.0 Disorders of Intellectual Development, mild
- Global Developmental Delay	- 6A00.1 Disorders of Intellectual Development, moderate
- Intellectual Disability	- 6A00.2 Disorders of Intellectual Development, severe
- Communication Disorders	- 6A00.3 Disorders of Intellectual Development, profound
Language Disorders	6A01 Developmental Speech or Language Disorders
Autism Spectrum Disorder	6A02 Autism Spectrum Disorder
Attention-Deficit/Hyperactivity Disorder (ADHD)	6A03 Developmental Learning Disorder
Specific Learning Disorders (in reading, writing or mathematics)	6A04 Developmental Motor Coordination Disorder
Motor Disorders	6A05 Attention-Deficit/Hyperactivity Disorder (ADHD)
Tic Disorders	6A06 Stereotyped Movement Disorder
	6A0Y Other Specified Neurodevelopmental Disorder

Table 1.1: DSM-5 and ICD-11 subclassifications of neurodevelopmental disorders

These classifications (**Table 1.1**) are broad; the prevalence of ID alone is above 1% (Vissers *et al.*, 2016). In this model the term neurodevelopmental disorder encompasses thousands of aetiological subtypes, most of these genetic

(Stessman *et al.*, 2014) [some ID gene panels contain >2,000 genes (Genomics England, 2022a)] with additional environmental causes (maternal alcohol abuse during pregnancy, infections and birth hypoxia). Defining the specific causes of neurodevelopmental disorders, by their molecular aetiologies enables precision medicine approaches for affected individuals.

1.2. Human neurodevelopment

Neurodevelopmental disorders are caused by abnormalities in the development of the human brain (**Figure 1.1**), which is potentially the most complex cellular arrangement seen in nature (Bear, 2020). The work detailed in this thesis focuses on the relationship between neurodevelopmental disorders and two specific processes integral to the normal formation and function of the cerebral cortex, neuronal migration and control of GABAergic signalling.

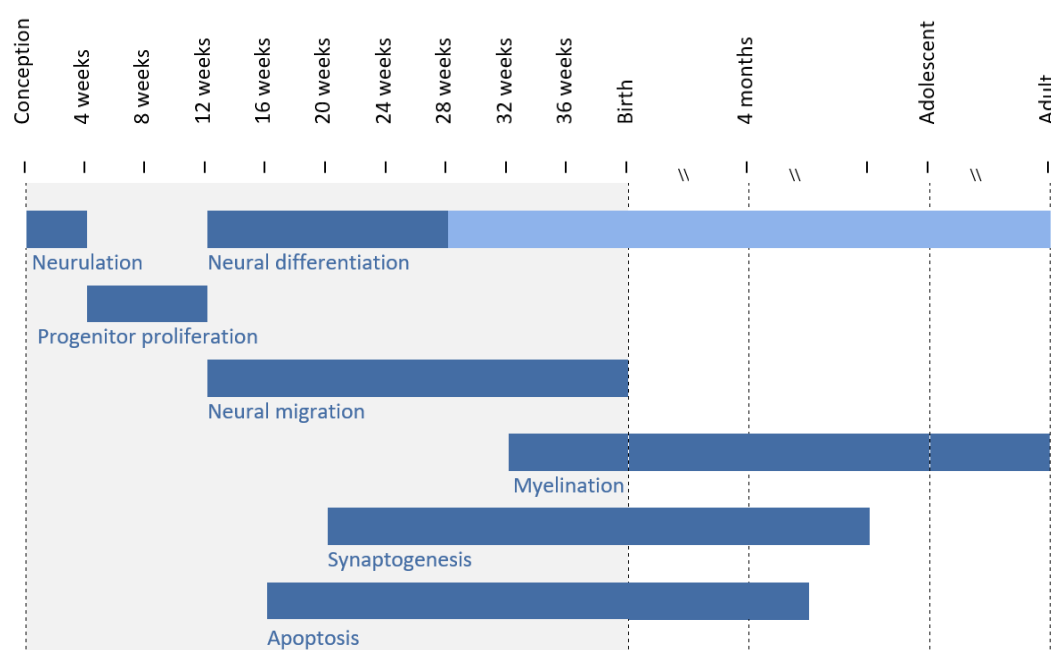


Figure 1.1: A timeline of human neural development

Grey shading represents prenatal period. Light blue shading represents ongoing differentiation occurring after the primary period (12-28w - dark blue). Adapted from (Ronan *et al.*, 2013)

Human neurodevelopment begins in the third week post gestation with the processes of neural plate induction and neurulation (**Figure 1.2**) (Sadler, 2003).

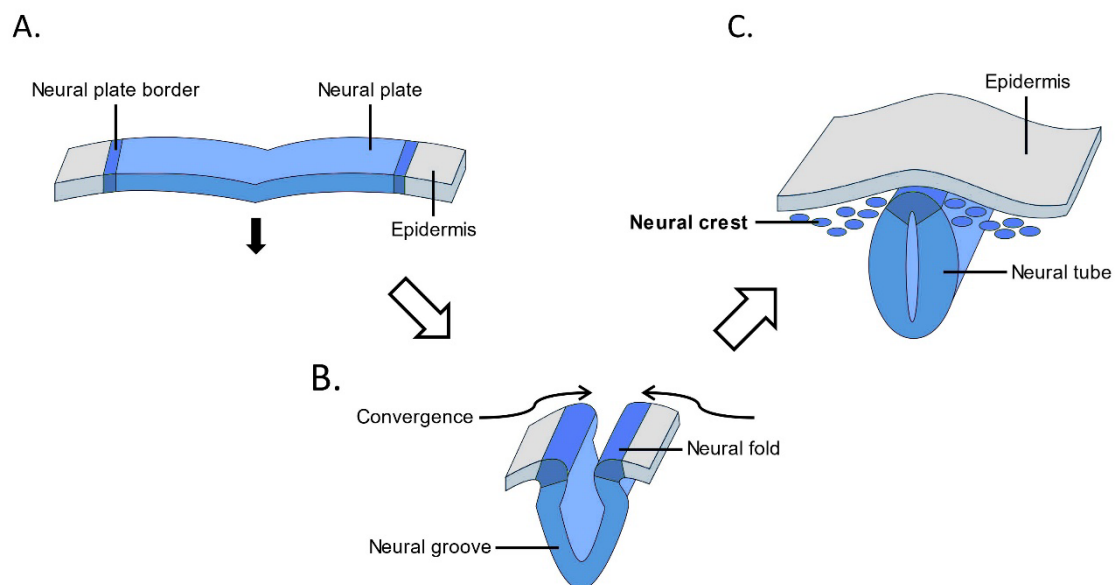


Figure 1.2: Neurulation

Adapted from commons.wikimedia.org, public domain

The trilaminar embryo (ectoderm, mesoderm and endoderm) develops a midline ectodermal thickening (the neural plate) and there is invagination of the midline of this structure (**A.**). Next the edges fold inwards (**B.**) and fuse above this invagination forming a closed neural tube by the end of the fourth week (**C.**). Disruption of this process (under the control of at least 80 genes in the mouse) (Copp *et al.*, 2003) is associated with neural tube defects in humans (e.g. *VANGL1*, MIM: 600145) (Kibar *et al.*, 2007). The rostral (“head”) portion of the neuronal tube progresses to form the brain with the caudal (“tail”) portion becoming the spinal cord.

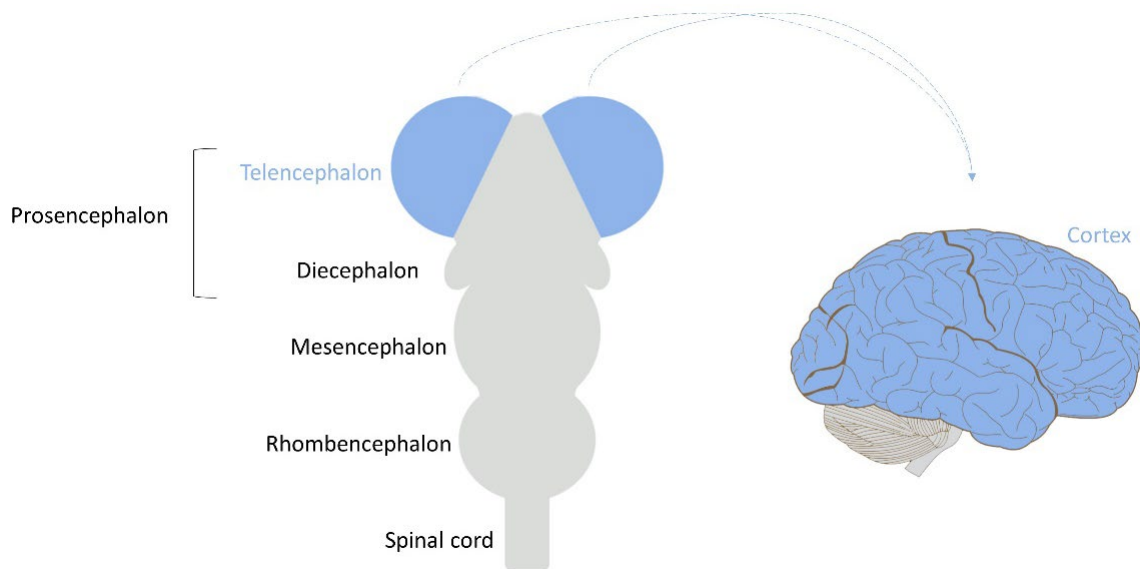


Figure 1.3: Evolution of the neural tube

Elements adapted from commons.wikimedia.org,
under Creative Commons Attribution ShareAlike (CC BY-SA) 3.0 license

The rostral portion develops three identifiable dilations (**Figure 1.3**): the prosencephalon (forebrain), which later derived the telencephalon; the mesencephalon (midbrain) and the rhombencephalon (hindbrain) (Sadler, 2003).

At the beginning of the fifth week, the cerebral hemispheres arise as the pallium, extensions of the lateral wall of the prosencephalon bulging into the lateral ventricle. Within an area of this is the neocortex (also called neopallium), the region that through extensive cellular proliferation and histogenesis including neuronal migration will become the laminated cerebral cortex.

1.3. Neuronal migration

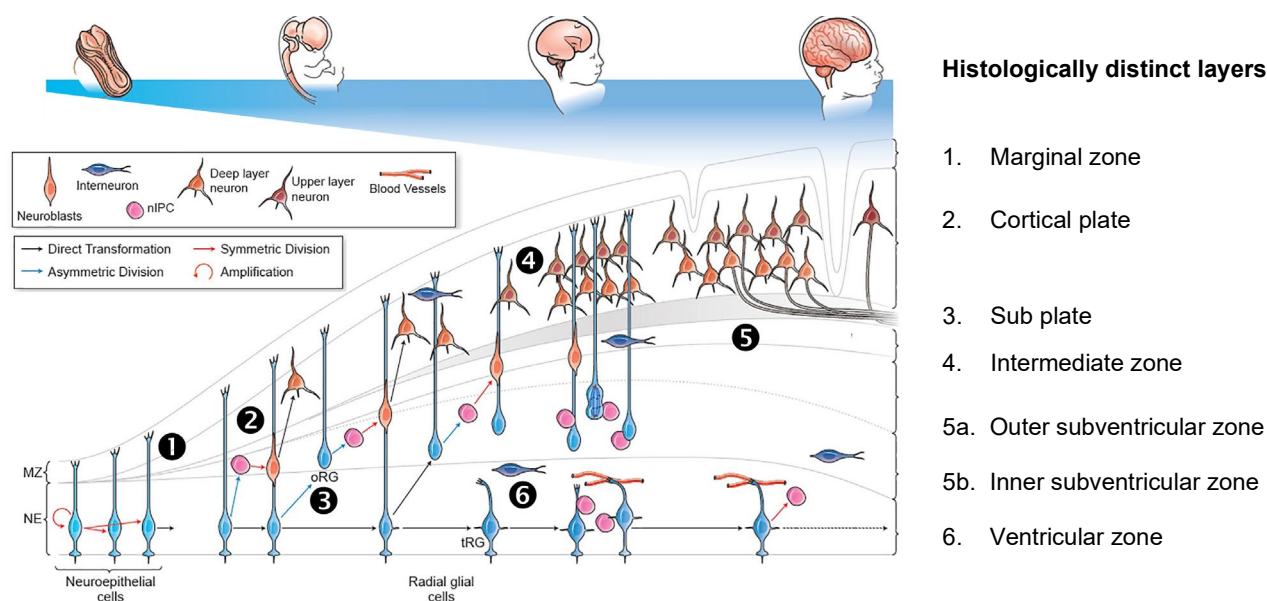


Figure 1.4: Human cortical development by radial migration

Adapted from (Subramanian *et al.*, 2019) under Creative Commons Attribution (CC BY) 4.0 license

Disruption of neuronal migration, a process occurring between 12 weeks gestational age and birth (**Figure 1.1**), can result in specific neurodevelopmental disorders with recognisable neuroradiology (**Section 1.4**). In this neurodevelopmental stage neuron progenitors leave their site of origin (for excitatory glutaminergic neurones this is the ventricular zone adjacent to the lateral ventricles) to take up final positions towards within a mature, layered cortex (**Figure 1.4**). Prior to the commencement of neuronal migration, in the first trimester, there is the symmetrical division of neuroepithelial cells (direct descendants of neural plate cells) in the neocortex, to expand the pool of progenitors, (❶ - **Figure 1.4**) (Bystron *et al.*, 2008).

From the twelfth week, asymmetric division of these cells also occurs (❷ - **Figure 1.4**) resulting in replacement of the progenitor and generation of a radial

glial cell. Radial glial cells, bipolar cells identified by their expression of glial fibrillary acidic protein, result from a transformation of neuroepithelial cells under the control of transcription factors *FOXP1*, *LHX2*, *PAX6*, *EMX2* (Bystron *et al.*, 2008; Mariani *et al.*, 2012). They are a source for neurons and several lineages of mature glia, including astrocytes and oligodendrocytes (Carlson, 2019). Neurone progenitors (neuroblasts - orange in **Figure 1.4**), are generated by radial glial cells from the end of the first trimester, either directly through asymmetric division or indirectly by generation of neuronal intermediate progenitor cells (nIPCs - pink in **Figure 1.4**). nIPCs also themselves divide symmetrically to amplify the clonal pool of neurons. In the second trimester radial glial cells lose contact with the outer subventricular zone (③ - **Figure 1.4**) giving rise to outer subventricular zone radial glia-like cells (oRG - **Figure 1.4**), which themselves generate neurons through IPCs. From approximately day 50 excitatory cortical pyramidal neurons, generated from radial glial cells and outer subventricular zone radial glia-like cells, migrate along the radial glial cells scaffold (④ - **Figure 1.4**) to establishing the cortical plate separated from the pia (the meningeal layer adherent to the brain) by the marginal zone. This process occurs in an “inside-out” manner, with the earliest generated neurons forming the deepest layers and the latest the most superficial. This arrangement is orchestrated by Cajal-Retzius cells in the marginal zone through the reelin signalling pathway (Bock & May, 2016). Toward the end of the second trimester the main elevated ridges (gyri) of the cortex form, resulting from the physical stresses associated with increased cortical cellularity, with further gyration continuing into the third trimester. In addition, as neurones migrate, they begin to associate locally and project long axons that form the cortical white matter (⑤

- **Figure 1.4**), including white matter structures such as the internal capsule and the corpus callosum. Finally, remaining radial glia transform into truncated radial glia (tRG - **Figure 1.4**) (6). The processes described above results in six histologically distinct layers present in the neocortex by the third trimester - the marginal zone, the cortical plate, sub plate, intermediate zone, subventricular zone and ventricular zone (**Figure 1.4**) (Subramanian *et al.*, 2019). The subventricular zone is sometimes further subdivided into outer and inner partitions. This organised layering is disrupted in some disorders, such as classic lissencephaly where it is replaced by a disordered four-layer arrangement.

1.4. Neuronal migration disorders

Neuronal migration disorders are a group of conditions arising from defects in the locomotion of neurones in the prenatal developing brain (Oegema *et al.*, 2020) that manifest with distinct neuroradiological abnormalities and result in early-onset developmental delay and seizures. These conditions are characterised neuroradiologically by abnormalities of cortical layering and an absence of normal folding (lissencephaly), with a paucity of gyral and sulcal development. The term lissencephaly, both a diagnosis and radiological and histological descriptor, encompasses distinct entities including classical lissencephaly, complete agyria (an absence of any gyri) and regional pachygyria (localised reduction in number of gyri). Further, there are other neuronal migration disorders including the clinically milder subcortical band heterotopias, anatomical division of the cortical plate into two layers (“double cortex syndrome”), and cobblestone complex, a form of over migration (Guerrini

& Dobyns, 2014). Neuronal migration disorder disease-gene discoveries have characterised key pathomechanistic pathways involved in normal neurodevelopment, enabling more precise diagnoses and benefits to prognostication, although no treatments to date.

The first neuronal migration disorder to have its molecular aetiology defined in 1997 was a form of autosomal dominant lissencephaly caused by variants in the platelet-activating factor acetylhydrolase, isoform 1b, alpha subunit (*PAFAH1B1*, previously *LIS1*) [MIM: 607432] (Lo Nigro *et al.*, 1997). A deletion of this gene with contiguous deletion of *YWHAE* on chromosome 17p13.3 is responsible for the Miller-Dieker lissencephaly syndrome (MIM: 247200), a condition also entailing facial dysmorphism and variable congenital malformations such as renal, gastrointestinal, and cardiac defects (Cardoso *et al.*, 2003). *PAFAH1B1* regulates the activity of the microtubule motor protein *DYNC1H1* (dynein, cytoplasmic 1, heavy chain 1) (Smith *et al.*, 2000), itself later shown to be associated with variable neuronal migration defects (MIM: 614563) (Poirier *et al.*, 2013; Vissers *et al.*, 2010). Shortly following the discovery of *PAFAH1B1*, the molecular causes of two X-linked recessive forms of lissencephaly were described: variants in *DCX* (MIM: 300067), a microtubule-associated protein (MAP) (Caspi *et al.*, 2000; des Portes *et al.*, 1998) and *ARX* (MIM: 300215) a transcription factor required for neuronal progenitor cell proliferation and GABAergic neuronal migration (Dobyns *et al.*, 1999; Friocourt *et al.*, 2008). Interestingly, whilst *DCX*-related lissencephaly is mainly transmitted as an X-linked recessive trait with males affected, some carrier females manifest the milder subcortical band heterotopia with this condition suggesting dosage-related expressivity of the *DCX*-related phenotype (Guerrini

& Dobyns, 2014). DCX is associated with ubiquitously expressed cytoplasmic actins, ACTB and ACTG1, later shown to cause Baraitser-Winter syndrome (MIM: 243310 and 614583), a dysmorphic developmental syndrome comprising of pachygyria, retinal coloboma, sensorineural deafness, reduced shoulder girdle muscle bulk and progressive joint stiffness (Rivière *et al.*, 2012). Later, variants in the kinesin (motor proteins) genes *KIF2A* and *KIF5C*, were also recognised as causes of complex cortical dysplasias (MIM: 615411, 615282) (Poirier *et al.*, 2013) as were microtubule-actin cross-linking factor 1 (MACF1) (MIM: 618325) (Dobyns *et al.*, 2018) and microtubule-associated protein 1b (MAP1B) (MIM: 618918) (Heinzen *et al.*, 2018).

The early molecular delineations of neuronal migration disorders already appear to converge around the microtubule and its associated proteins. This hypothesis was further confirmed by discoveries between 2007-2019 identifying genetic variants in the microtubule monomer subunits, tubulin alpha, and associated complex proteins as causes of, mostly autosomal dominant, neuronal migration disorders (**Table 1.2**).

Tubulin	Gene	OMIM	Inh.	Reference
Alpha	<i>TUBA1A</i>	611603	AD	(Keays <i>et al.</i> , 2007; Poirier <i>et al.</i> , 2007)
Beta	<i>TUBB</i>	615771	AD	(Breuss <i>et al.</i> , 2012)
-	<i>TUBB2A</i>	615763	AD	(Cushion <i>et al.</i> , 2014)
-	<i>TUBB2B</i>	610031	AD	(Jaglin <i>et al.</i> , 2009)
-	<i>TUBB3</i>	614039	AD	(Tischfield <i>et al.</i> , 2010)
Gamma	<i>TUBG1</i>	615412	AD	(Poirier <i>et al.</i> , 2013)
Gamma-tubulin complex	<i>TUBGCP2</i>	618737	AR	(Mitani <i>et al.</i> , 2019)

Table 1.2: Tubulin subunits and complex proteins associated with neuronal migration disorders

Abbreviations: Inh. Inheritance, OMIM. Online Mendelian Diseases in Man. A neuronal migration disorder associated with TUBA8 variants was proposed, but later disproved (Abdollahi *et al.*, 2009; Diggle *et al.*, 2017).

Tubulin-associated neuronal migration disorders are often termed “tubulinopathies” (Desikan & Barkovich, 2016), displaying characteristic neuroradiological features in addition to abnormalities of cortical layering, including dysmorphism of the basal ganglia, agenesis/hypogenesis of the corpus callosum (aCC / hCC), hypoplasia of the oculomotor and optic nerves, cerebellar hypodysplasia and dysmorphism of the hind-brain structures (Romaniello *et al.*, 2018).

Whilst the majority of neuronal migration disorders described so far have displayed autosomal dominant or X-linked inheritance, autosomal recessive causes do also exist, accounting for a lower proportion of affected individuals. These include the aforementioned *TUBGCP2* (Mitani *et al.*, 2019), *RELN* (Hong *et al.*, 2000) and *VLDLR* (Boycott *et al.*, 2005) (MIM: 257320 and 224050), both involved in reelin signalling, and *EML1* (Oegema *et al.*, 2019) (MIM: 600348).

As more affected individuals have been identified it has become possible to more precisely delineate the neuroradiological phenotypes associated with each specific genetic aetiology within the neuronal migration disorder subclasses. These can be classified by anatomical/histological appearance (lissencephaly, pachygyria, subcortical band heterotopia, microlissencephaly), degree (e.g. complete vs. partial) and location (anterior vs. posterior), and associated findings both neuroradiological (cerebellar hypoplasia, aCC) and extracerebral (genital anomalies, skeletal dysplasia). A number of these gene-phenotype relationships are summarised in **Figure 1.5**.

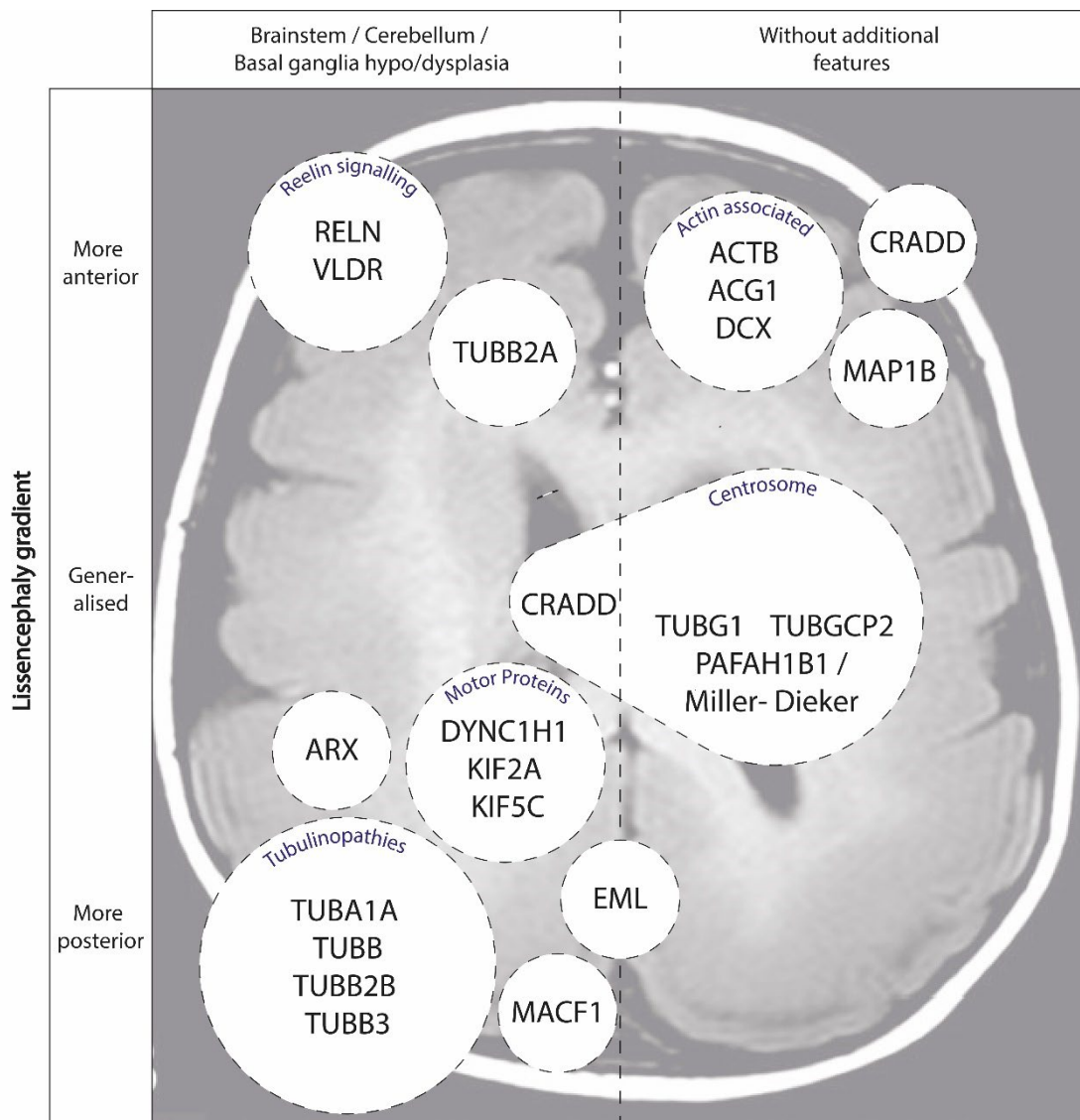


Figure 1.5: Network and phenotypic classifications of neuronal migration disorders

Derived from (Di Donato *et al.*, 2017; Di Donato *et al.*, 2018; Heinzen *et al.*, 2018; Kato, 2015; Schmidt *et al.*, 2021). Background image is a child with a *DYNC1H1* variant (Mutch *et al.*, 2016)

Cohort studies have attempted to quantify the contribution of each molecular aetiology to the overall prevalence of neuronal migration disorders (Accogli *et al.*, 2020; Di Donato *et al.*, 2018; Kolbjer *et al.*, 2021). The largest of these studies comprising of 811 patients with lissencephaly or subcortical band heterotopia, was assembled opportunistically by researchers at an international centre for lissencephaly research. Due to ascertainment bias it is unlikely to be

representative of all undiagnosed individuals but remains the most informative such study to date (Di Donato *et al.*, 2018) (**Figure 1.6**).

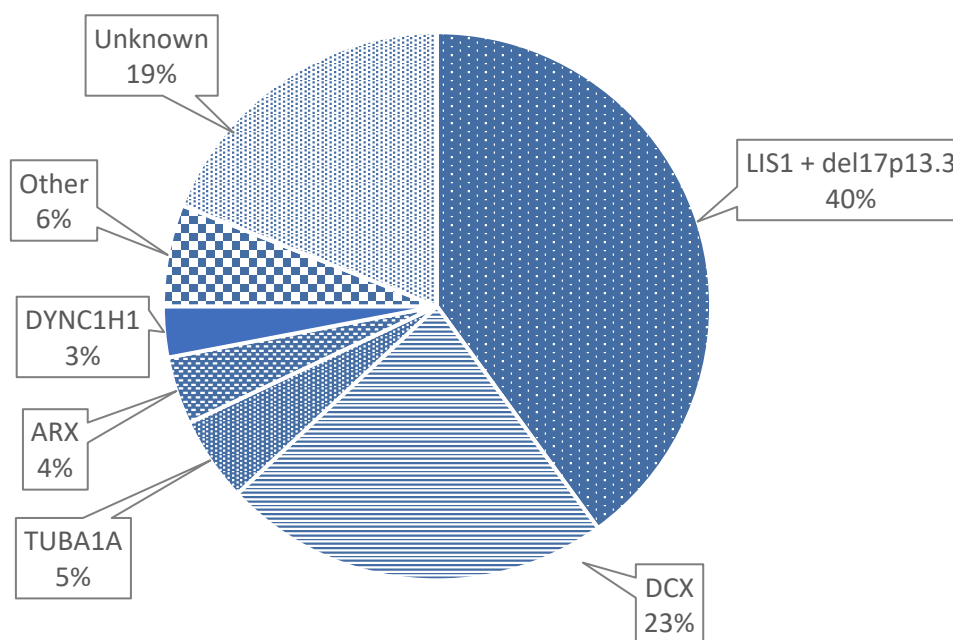


Figure 1.6: Molecular causes of lissencephaly in a cohort of 811 patients

Adapted from (Di Donato *et al.*, 2018)

The high proportion of individuals whose disease was caused by variants in *PAFAH1B1* (*LIS1*) is consistent with findings in other cohorts (Kolbjør *et al.*, 2021); the relatively low percentage of individuals accounted for by the tubulinopathies may result from strict recruitment criteria excluding individuals with non-cortical features. The diagnostic yield in this cohort, 81%, is higher than observed in comparable lissencephaly cohorts (60%) (Accogli *et al.*, 2020), possibly due to the extremely high diagnostic rate associated with their core phenotype, classical lissencephaly (Wiszniewski *et al.*, 2018).

Whilst more than two decades of research findings have clearly demonstrated the relationship between the genes encoding the microtubule and a number of its associated proteins with the pathogenesis of neuronal migration disorders (**Table 1.3**), more than 100 proteins known to be associated with microtubules have not yet been shown to be associated with human disease.

Class	NDD disorder	Other disorder	No monogenic disorder
Actins	<i>ACTB,ACTG (1-2)</i>	<i>ACTA (1-2), C1 G2</i>	<i>ACTBL2</i>
CAMSAP	-	-	<i>CAMSAP (1-3)</i>
CLASP	-	-	<i>CLASP (1-3)</i>
CLIP	-	-	<i>CLIP (1-4)</i>
Dynein	<i>DYNC1H1 DYNC1I2 DNAL4</i>	<i>DNAH (1, 2, 5, 8-11, 17) DNAI (1-2) DNAL1 DYNC2 (H1, I1, I2, LI1) DYNLT2B, NME8</i>	<i>DNAH (3, 6, 7, 12, 14) DNAI (3, 4, 7) DYNC (1I1, 1LI2) DYNL (L1, L2, RB1-2) DYNLT (1, 3)</i>
EML	<i>EML1</i>	-	<i>EML2-6</i>
MT-tethering	-	-	<i>HOOK(1-3)</i>
JPT	-	-	<i>JPT(1-2)</i>
Katanin	<i>KATNB1</i>	-	<i>KATNA1</i>
Kinesin	<i>KIF (1A, 2A, 4A, 5C, 7, 11, 14) CENPE</i>	<i>KIF (1B, 1C, 3B, 5A, 12, 20A, 21A, 22, 23) KIFBP1</i>	<i>KIF (2B, 2C, 3A, 3C, 4B, 5B, 6, 9, 13A, 13B, 15, 16B, 17, 18A, 18B, 19, 20B, 21B, 24, 25, 26A, 26B, 27) KIFC (1-3)</i>
MT-associated protein	<i>MAP1B, MAP11, MAPT</i>	-	<i>MAP (1A, 1S, 2-4, 6, 7, 7D1-D3, 9, 10)</i>
MT Cross-linking factor	<i>MACF1</i>	-	<i>MACF2</i>
MT-associated Ser/Thr kinase	<i>MAST1</i>	-	<i>MAST(2-3)</i>
MATCAP	-	-	<i>MATCAP</i>
MT-associated scaffold protein	-	-	<i>MTUS(1-2)</i>
End-binding	-	<i>MAPRE2 (EB2)</i>	<i>MAPRE1,3 (EB1,3)</i>
MARK	-	<i>MARK3</i>	<i>MARK1,2,4</i>
Regulator of MT dynamics	-	-	<i>RMDN(1-3)</i>
Alpha tubulin	<i>TUBA1A</i>	<i>TUBA4A, TUBA8</i>	<i>TUBA (1B, 1C, 3C, 3D, 3E, 4B)</i>
Beta tubulin	<i>TUBB, TUBB2A, TUBB (2B, 3, 4A)</i>	<i>TUBB (4B, 6)</i>	<i>TUBB (1, 8, 8B)</i>
Gamma tubulin	<i>TUBG1</i>	-	<i>TUBG2</i>
TUBGCP	<i>TUBGCP2</i>	-	-
Other tubulin	-	-	<i>TUBD1, TUBE1</i>

Table 1.3: Microtubule associated gene families and associations with monogenic disease

Data from (A. Akhmanova & Hoogenraad, 2015; A. Akhmanova & Steinmetz, 2019; Bodakuntla *et al.*, 2019) genenames.org and OMIM (accessed 19/07/22).

Abbreviations : **ACTG (1-2)**: *ACTG1* and *ACTG2* **CAMSAP (1-3)**: CAMSAP1, CAMSAP2, CAMSAP3 etc.

CAMSAP: Calmodulin-regulated spectrin-associated protein **CLASP**: cytoplasmic linker-associated protein

CLIP: CAP-GLY domain-containing linker protein

JPT: Jupiter MT-associated homolog;

MATCAP: MT-associated tyrosine carboxypeptidase

NDD: neurodevelopmental disorder

EML: Echinoderm MT-associated proteins

MARK: MAP/MT affinity-regulating kinase

MT: Microtubule

TUBGCP: tubulin gamma complex associated protein

1.5. Microtubules in cortical development

The formation of microtubules (**Figure 1.7**) involves tubulin heterodimers polymerising into linear protofilaments, 13 of which associate laterally with each other to create a rigid hollow polymer rod (Lasser *et al.*, 2018; Pollard & Goldman, 2018).

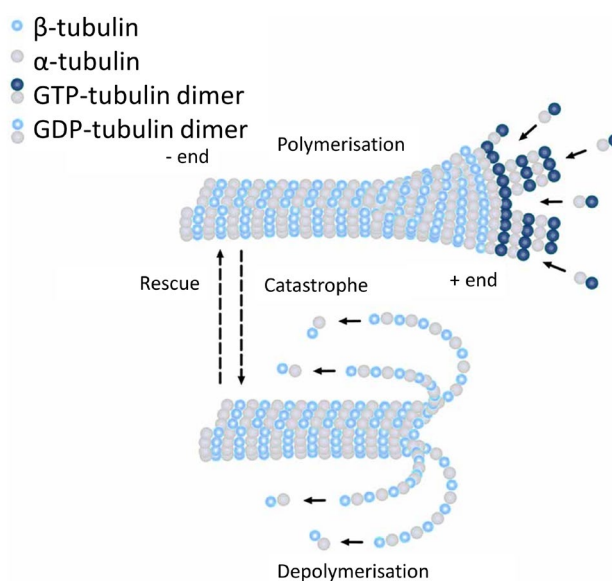


Figure 1.7: The structure and properties of microtubules

Adapted from (Lasser *et al.*, 2018) under CC BY 4.0 license

Microtubules are not symmetrical but polar, with specialised plus and minus ends. The plus end (at which β-tubulin is exposed) is dynamic and able to grow rapidly, but also prone to rapid “catastrophe” where large numbers of subunits rapidly depolymerise and microtubule length reduces substantially (Anna Akhmanova & Steinmetz, 2015; Mitchison & Kirschner, 1984). To prevent unintended catastrophe the plus end is stabilised by a Guanosine-5'-triphosphate (GTP) cap, a region where GTP hydrolysis has not yet occurred,

and microtubule plus-end-tracking proteins (+TIPs). The minus end of a microtubule (at which α -tubulin is exposed) grows more slowly, which makes it less prone to catastrophe despite the lack of a GTP cap (Akhmanova & Steinmetz, 2015). It is anchored in a microtubule organising centre (MTOC), which is usually the centrosome but can also be a spindle pole (during mitosis), a basal body (for cilia), the Golgi complex or the nucleus. Within the centrosome minus ends are nucleated and “capped” by a γ tubulin ring complex (γ -TuRC) comprising gamma tubulin, PCNT, ACAP450 and CDK5RAP2 among others (Tovey & Conduit, 2018). For non-centrosomal microtubules, whilst γ -TuRC is still present (Ori-McKenney *et al.*, 2012), the role of other factors, such as and microtubule minus-end-tracking proteins (-TIPs), is important. Less is known about this process, but the CAMSAP group of proteins, particularly CAMSAP2 and CAMSAP3, are thought to fulfil this stabilising role (Akhmanova & Hoogenraad, 2015; Jiang *et al.*, 2014).

Microtubules are extensively distributed throughout eukaryotic cells, providing a scaffold crucial for organellar positioning and for subcellular trafficking via the molecular motors dynein and kinesin (Hirokawa *et al.*, 2010). Additionally, they are focally organised into the mitotic spindle and the axonemes of cilia and flagella (Akhmanova & Steinmetz, 2019; Brouhard & Rice, 2018). Radially orientated microtubules attach to the centrosome; non-centrosomal microtubules, enable non-radial transport (e.g. apical-to-basal), support cellular protrusions, and form mitotic spindles (Hendershott & Vale, 2014; Keating & Borisy, 1999).

Microtubules are highly enriched in neurones and play a crucial role in their differentiation. Neurites, thin protrusions that will eventually become axons and dendrites, are created from the invasion of membrane protrusions (lamellipodia) by microtubules and actin (Götz & Huttner, 2005). Such neurites can be induced in *Drosophila* by simulating microtubule sliding using the microtubule motor protein kinesin-1 (human homologues include *KIF1A*) (Winding *et al.*, 2016). In neurones after the establishment of many neurites a polarity develops, with one neurite selected to become the axon, whilst the others become dendrites. This process again depends on the organisation of microtubules, with stabilisation of the microtubules in one neurite preceding its designation as the axon (Witte *et al.*, 2008). In early neuronal precursors microtubules are nucleated by the γ -TuRC from MTOCs, such as the centrosome. However, once neurons begin to establish polarity, the role of the centrosome as an MTOC diminishes and minus end stabilisation becomes increasingly important (Lasser *et al.*, 2018).

Following the specification of a single axon there is establishment of the growth cone, a dynamic structure at the axon tip responsible for elongating it by probing the extracellular environment (Bartolini & Gundersen, 2006; Lasser *et al.*, 2018). Within this structure are active microtubules and actin filaments which interact dynamically to advance, turn and eventually branch the tip of the axon (Dent *et al.*, 2011). These processes enable the intricate positioning of neurons within the developing mammalian brain during neuronal migration (Keays *et al.*, 2007).

Following the formation of axons and dendrites, an array of stable parallel microtubules are laid down providing “tracks” down which the molecular motors

dynein and kinesin convey payloads such as synaptic vesicles between the cell body and the synapse (Guedes-Dias & Holzbaur, 2019). In axons there is a uniform polarity, with minus ends orientated towards the cell body, whereas in dendrites the polarity is mixed (Tas *et al.*, 2017).

1.6. GABAergic neurodevelopment

The migration of excitatory glutamatergic cortical neurons is described above, and whilst these are the most numerous in the cortex, as many as 30% of cortical neurons are inhibitory GABAergic interneurons, which migrate differently. GABAergic interneurons, which express the primary inhibitory neurotransmitter gamma-aminobutyric acid (GABA), play an important role in controlling the propagation of cortical impulses and a key homeostatic role “maintaining network excitability and plasticity at optimal levels to facilitate the gating, processing and storage of information” (Tang *et al.*, 2021). GABA has also been shown to drive the formation of both inhibitory and excitatory synapses in early postnatal brains (Oh *et al.*, 2016), hence controlling not just the excitatory state of neuronal networks, but their development as well. The importance of intraneuronal connections cannot be overstated, indeed very recent work published in *Science* suggested that an interneuron-to-interneuron network an order of magnitude larger in humans compared to mice may be the most significant difference in network structure between these species (Loomba *et al.*, 2022).

There are subclassifications of GABAergic interneurons expressing different factors, each fulfilling a variety of roles. Fast-spiking interneurons expressing

the calcium-binding protein parvalbumin (PV⁺) directly target the neuronal soma of pyramidal neurons to reduce neuronal firing. Non-fast-spiking interneurons containing the somatostatin (SST⁺) target the distal dendrites, as do multipolar interneurons that express the glycoprotein reelin (RELN⁺). Finally, bipolar interneurons that express vasoactive intestinal peptide (VIP⁺) neurons target other inhibitory neurons thus indirectly reducing inhibition of pyramidal neurons (Tang *et al.*, 2021; Wamsley & Fishell, 2017). In addition to the synaptic role that GABA and specifically ligand-gated ionotropic GABA type A receptors (GABA_ARs) play in rapid phasic inhibition, GABA_ARs also exist in on extra synaptic neuronal membranes where it modulates a “slow and persistent” tonic GABAergic inhibition.

GABAergic interneurons originate in the medial and caudal ganglionic eminences, transient structures between the thalamus and caudate nucleus, and migrate tangentially along the marginal zone or in the subplate and SVZ (below CP and above VZ - **Figure 1.8**) before moving radially along the radial glial cell scaffold (see also **Figure 1.4**) (Bajaj *et al.*, 2021; Lui *et al.*, 2011).

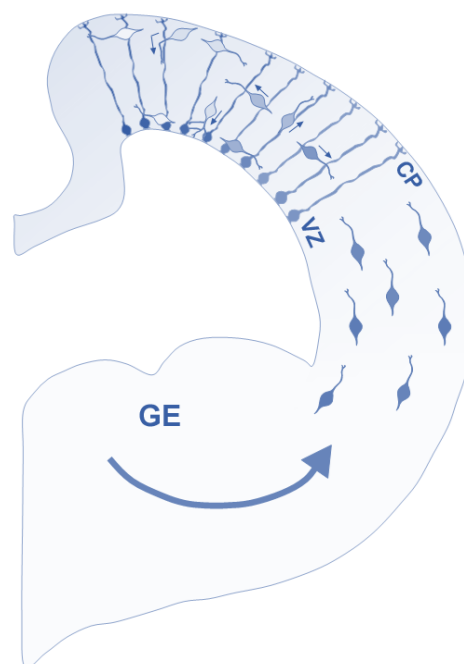


Figure 1.8: The migration of GABAergic interneurons in the developing cortex

Adapted from (Yokota *et al.*, 2007)(CC BY) CP: cortical plate, GE: ganglionic eminences, VZ: ventricular zone

Cellular migration is eventually terminated by a cellular response to local GABA levels (Bortone & Polleux, 2009). Migration of GABAergic neurons, commencing in the second trimester, continues until early infancy in humans.

1.7. Disorders of GABAergic dysfunction

There are a number of well described human neurodevelopmental disorders known to involve defects in GABAergic signalling, with common features of developmental impairment, ID and a reduced seizure threshold or epilepsy (Tang *et al.*, 2021). Clear examples include disease-causing variants in the 19 gene subunits of the GABA receptor known to cause autosomal dominant developmental and epileptic encephalopathy – a severe form of epilepsy with associated developmental impairment (e.g. *GABRA1* [MIM: 615744], *GABRB1* [MIM: 617153], *GABRB3* [MIM: 617113]) (Hernandez & MacDonald, 2019). However, aberrant GABAergic signalling also has a role in a more diverse group of highly penetrant monogenic human neurodevelopmental disorders. One example is Rett syndrome (MIM: 312750), a severe condition mostly affecting females, where mouse models show derangement of the balance of excitatory and inhibitory (GABAergic) synaptic membrane currents (E-I balance or E/I ratio) received by cortical pyramidal neurons (Banerjee *et al.*, 2016; Zhou & Yu, 2018). Additionally, for both fragile X syndrome (MIM: 300624) and Angelman syndrome (MIM: 105830) a reduction in GABA-related tonic inhibition on cortical pyramidal neurons is observed in murine models (Curia *et al.*, 2009; Egawa *et al.*, 2012; Tang *et al.*, 2021). Further, disordered GABAergic transmission has been postulated as having some role in a predisposition to polygenic disorders such as autism spectrum disorders (Tang *et al.*, 2021), with

one hypothesis, supported by an excess of seizures in individuals with autism spectrum disorder, suggesting increased E/I ratio may contribute to the pathomechanism in autism (Rubenstein & Merzenich, 2003).

1.8. Rare genetic disorders

Rare genetic diseases affect ~6% of the UK population and are over-represented in ethnic minority communities (Davies, 2016; Ferreira, 2019). It is currently estimated that there are somewhere between 4,000-7,000 rare disorders, with rare neurological disorders representing the single largest category of monogenic disease (Department of Health and Social Care, 2019, 2021; Ehrhart *et al.*, 2021) (**Figure 1.9**). Despite recent advances in the field of genomics, more than half of individuals with rare disease remain undiagnosed (Boycott *et al.*, 2017; Wright *et al.*, 2018). Additionally, progress in rare disease gene discovery has been uneven, with more rapid discovery of genetic disorders with a *de novo* dominant mechanism through trio exome-based projects (Beaulieu *et al.*, 2014; Wright *et al.*, 2015), and much slower progress for rare autosomal recessive conditions. This is particularly the case for neurological diseases, which represent by far the largest group of unexplained rare diseases (Mitani *et al.*, 2021).

The number of individuals recognised to be living with a rare disease has been growing rapidly; this can be attributed to three factors (Gorini *et al.*, 2021):

1. The ongoing discovery of new rare diseases (**Figure 1.9**)
2. A greater number of new diagnoses, resulting from a wider availability of genetic/genomic testing and increased public awareness of rare disease (Nguengang Wakap *et al.*, 2020).
3. Increasing life expectancy for rare disease patients, resulting from advances in therapies and potentially earlier diagnosis (lead-time bias) (Alonso-Ferreira *et al.*, 2018).

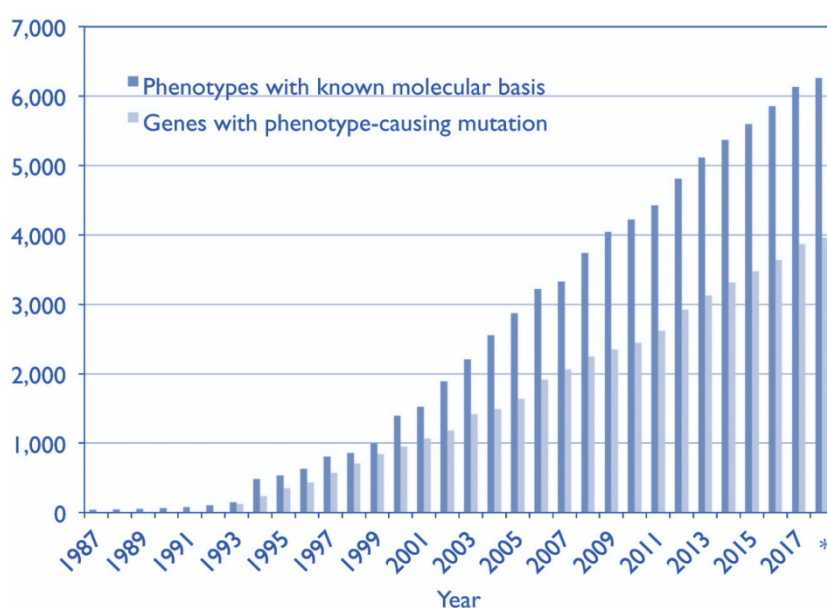


Figure 1.9: Number of gene-phenotype relationships in MIM database

Reproduced from (Amberger *et al.*, 2019) under CC BY 4.0 license

MIM: Mendelian Diseases in Man (<https://www.omim.org>)

Rare genetic diseases present an immense healthcare burden worldwide costing the UK National Health Service (NHS) >£300 million/year (Davies, 2016). Approximately 75% of rare genetic diseases affect children, a disproportionately high percentage compared to common disorders such as ischaemic heart disease, cancer and stroke (Department of Health and Social Care, 2021). Many children have undergone prolonged diagnostic odysseys of

>7 years with an average of three misdiagnoses during the course of their investigations (Davies, 2016). Rare disease in a child also affects the wider family, with additional caring responsibilities, strain on relationships and impacts on decisions to have children. The process of obtaining a diagnosis can be transformative for patients and their families, even if the clinical course of the condition cannot currently be modified (Bick *et al.*, 2021).

In the UK and internationally, many new rare disease initiatives and genomic programmes are currently being developed. The European Union (EU) has invested more than €2.4 billion and in the UK rare disorders have been highlighted as an NHS and Department of Health and Social Care priority area. In January 2021, a new UK Rare Disease Framework was published detailing the commitment of UK governments to improve lives of people with rare conditions (Department of Health and Social Care, 2021). The commitment to rapid genomic sequencing and the Genomic Medicine Service, is central to the new national genomics strategy in England (Lord Bethell of Romford, 2020) and the NHS Long Term Plan (<https://www.longtermplan.nhs.uk>). The 2018 House of Commons report on Genetics and Genome Editing in the NHS also emphasised the importance of embedding research in the new NHS Genomic Medicine Service (House of Commons, 2018). The UK Rare Diseases Framework (Department of Health and Social Care, 2021) highlighted four key priorities for progress in order to improve the lives of those living with rare disorders:

- 1 *Helping patients get a final diagnosis faster*
- 2 *Increasing awareness of rare diseases among healthcare professionals*
- 3 *Better coordination of care*
- 4: *Improving access to specialist care, treatments and drugs*

The role of research in achieving these goals is highlighted in **Figure 1.10**.

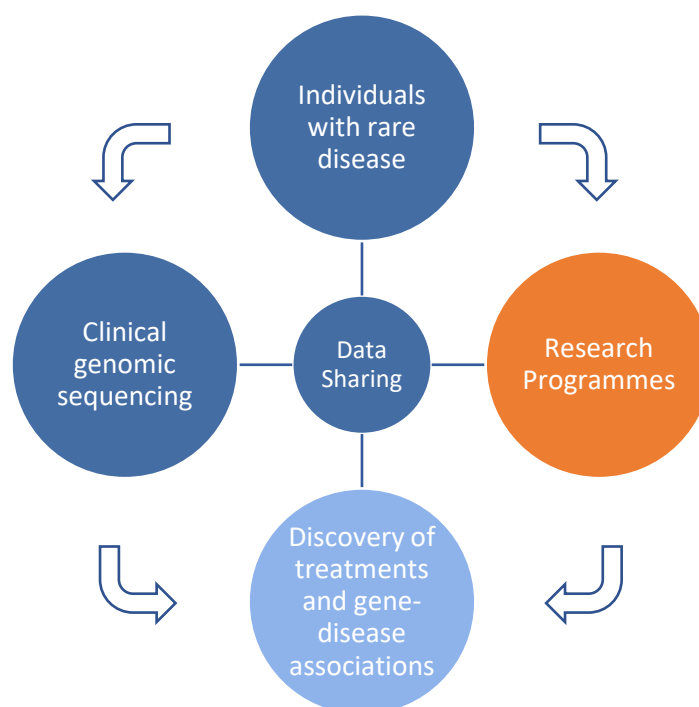


Figure 1.10: The cycle of rare disease testing, discovery, and treatment.

Adapted from (Rehm, 2022)

1.9. Strategies to identify new monogenic causes of neurodevelopmental disorders

Early studies to identify the genetic causes of rare diseases, carried out between the 1980s and early 2000s, involved genetic linkage or genome-wide single nucleotide polymorphism (SNP) mapping in large families to identify a locus likely containing the disease gene and small enough to perform gene-by-gene dideoxy (Sanger) sequencing (Claussnitzer *et al.*, 2020; Quinodoz *et al.*, 2021). Next-generation sequencing was introduced in the mid-2000s. This involves fragmentation of genomic DNA into pieces, sequencing these fragments in parallel and the reassembly of resulting sequence information

using bioinformatics approaches and comparison with a reference sequence. This enormously increased the speed of sequencing, reducing the per-base cost (**Figure 1.11**) and enabled genome-wide sequencing strategies to be considered.

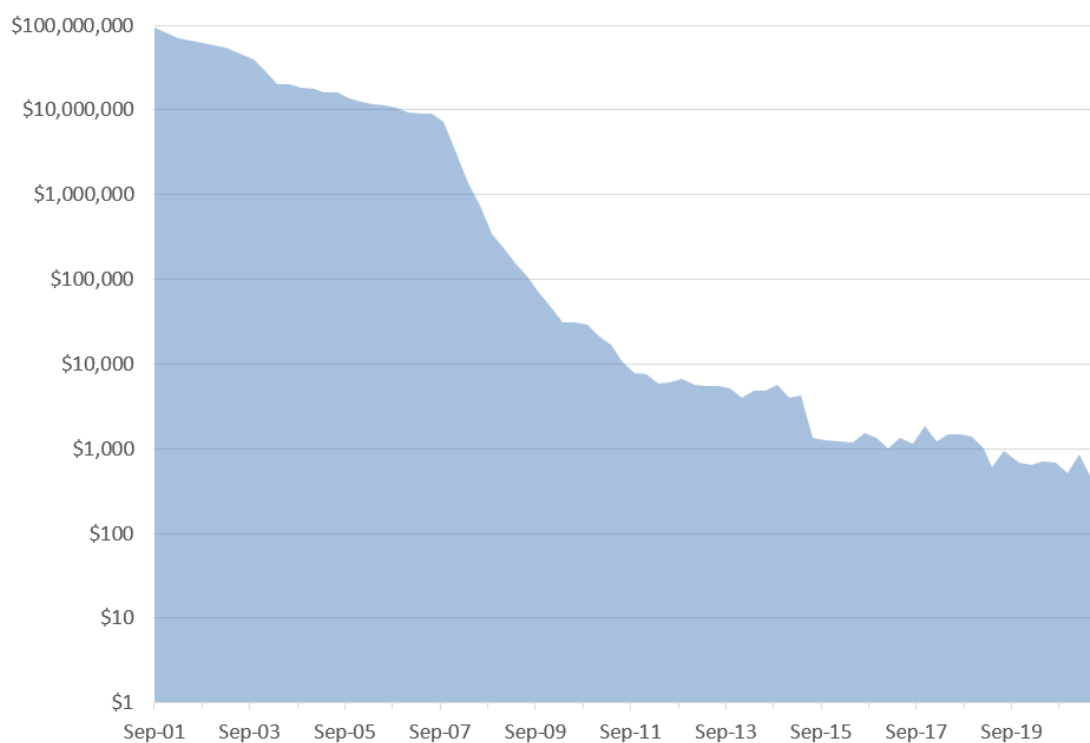


Figure 1.11: Cost per genome (2001-2021)

Data from: www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Note that cost (in dollars) is plotted on a logarithmic axis

This has had a profound impact on disease both gene discovery and clinical practice, with exome or genome now recommended early in the diagnostic pathway for ID (Manickam *et al.*, 2021; Smedley *et al.*, 2021; Splinter *et al.*, 2018).

The diagnostic yield of exome sequencing in neurodevelopmental disorders has been estimated at 30–43% by a 2019 meta-analysis [range 13-90%] (Srivastava *et al.*, 2019). The additive diagnosis achievable with genome sequencing,

particularly within clinical diagnostic practice, are controversial. Recent work suggested that up to 13% of diagnoses from genome sequencing are attributable to variants not reliably detected by exome sequencing (non-coding, mitochondrial, tandem repeat and structural variants). However, potential uplifts in diagnostic yield may still be practically limited by technical challenges in interpreting non-coding and structural variants (Rehm, 2022; Smedley *et al.*, 2021).

Revolutions of scale and cost in sequencing technologies have also enabled innovation in research study design. Since the 2010s very large relatively unselected cohorts have been recruited to identify and delineate new developmental disorders using exome or genome sequencing. The Deciphering Developmental Disorders (DDD) study, performed trio exome sequencing and microarray-based comparative genomic hybridisation (array CGH) on ~13,500 UK families, >70% of whom were affected with ID / developmental delay or learning disability, achieving a diagnostic yield of 40% (The Deciphering Developmental Disorders Study, 2015; Wright *et al.*, 2018). More recently, the 100,000 Genomes Project, performed genome sequencing on >33,000 individuals with rare disease, 39% of whom were affected by neurological or neurodevelopmental disorders (Best *et al.*, 2021; Genomics England, 2017; Whewey *et al.*, 2019). Results from the pilot stage of the Rare Diseases Program, reported a diagnostic yield of 29%, possibly reflecting a high number of individuals having previously undergone exome sequencing. In addition to the clinical benefits, both the 100,000 Genomes Project (251 publications, [genomicsengland.co.uk/research/publications](https://www.genomicsengland.co.uk/research/publications) accessed 18/10/2022) and the DDD study (275 publications, <https://www.ddduk.org/publications.html>; accessed

18/10/2022) have reported a large number of research findings to-date, identifying novel candidate disease-gene associations (e.g. *COL4A3BP*, *PCGF2*, *PPP2R5D*, *PPP2R1A*, *LRRC45*) as well as improving variant curation and disease phenotype delineation (Best *et al.*, 2021; DDD project, 2022; Genomics England, 2022b; Wright *et al.*, 2015). However, the successes have not been uniform across genetic diseases. Specifically, the DDD study, with its focus on a gene-agnostic trio exome approach, has excelled at the discovery of candidate autosomal dominant disease genes with a *de novo* mechanism (>28 genes) (Kaplanis *et al.*, 2020) but has reported relatively few candidate autosomal recessive disease genes (Wright *et al.*, 2018), possibly due to the low likelihood (~1%) of children born to European couples being affected with an autosomal recessive disorder (Fridman *et al.*, 2021).

1.10. Autosomal recessive neurodevelopmental disorders

Autosomal recessive disorders are the primary cause of ID in countries where close marriages of close family members are common, accounting for over 1 billion people worldwide (Hu *et al.*, 2018). It is widely believed that many gene-disease associations underlying such disorders may still remain undiscovered (Fumagalli *et al.*, 2018; Hu *et al.*, 2018; McInnes *et al.*, 2021; Shen *et al.*, 2015), with one model suggesting that such discoveries follow a Pareto model or “80-20” rule (20% of the work produces 80% of the output with the remaining 20% of the output coming from 80% of the work) (**Figure 1.12**). The graph below potentially demonstrates the current situation under that model, with a smaller number of known diseases contributing 50% of diagnoses in green, and a much larger number of rare disease still undiscovered (Shen *et al.*, 2015).

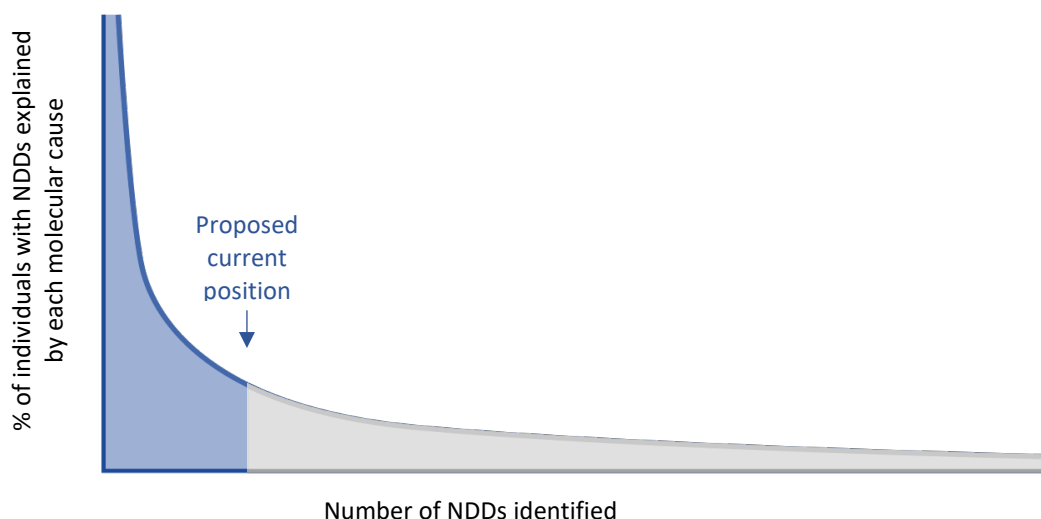


Figure 1.12: The long tail of neurodevelopmental disorder disease gene discovery

Adapted from (Shen *et al.*, 2015) Abbreviations: NDD = neurodevelopmental disorder

This hypothesis is supported by both the observation that new disease-gene associations are not declining (**Figure 1.9**) and a recent modelling analysis suggesting that “known recessive genes only contribute 50% of detectable recessive contribution to developmental disorders” (Balick *et al.*, 2022). An analysis of 6,040 European parent-child trio exomes from the DDD study, estimated the disease contribution of autosomal recessive coding variants at 3.6%, compared to 50% explained by *de novo* variants. However, the contribution was notably higher (31%) in patients with Pakistani ancestry, due to elevated autozygosity, with only half the recessive burden being attributable to known genes (Martin *et al.*, 2018).

The experience of our research group whilst working with datasets from both the DDD study and the 100,000 Genomes Project, is that despite the scale of these projects (over 10,000 UK developmental disorder families in the DDD study, over 33,000 individuals with rare disease in the 100,000 Genomes

project) many families will represent the only affected individual in a project with a particular rare autosomal recessive neurodevelopmental disorder. This underscores the profound heterogeneity and diverse disease mechanisms underlying neurogenetic disorders.

Other large-scale programmes have investigated individual nuclear consanguineous families typically with single affected siblings or sib pairs (Monies *et al.*, 2017; Najmabadi *et al.*, 2011). While this approach is invaluable for identifying potential candidate new disease genes/causative variants, it lacks the inherent power of community-based studies to definitively define causes of autosomal recessive disease. Evidencing this, a notable number of candidate pathogenic variants identified in such large-scale programmes are subsequently excluded as causative of disease as genomic knowledge of rare variants has increased (Monies *et al.*, 2019).

1.11. Genomic studies in genetically isolated communities empowers neurodevelopmental disorder disease gene identification

An alternative approach to those described above involves studies of neurodevelopmental disorders which occur with increased frequency in certain genetically isolated populations. In such communities an ancestral bottleneck may enrich certain disease-causing alleles. In communities of Northern European ancestry geographic mobility and cultural practices mean that the majority of conceptions do not involve closely related parents. However, this is not true of many communities worldwide (e.g. Pakistan, Arab Palestinians) in which there may be historic cultural preferences for marriages within the same

community (endogamy), or to close relatives often cousins (consanguinity) (Bittles, 2005). In other communities, low geographic mobility combined with a genetic bottleneck (**Figure 1.13**) results in increased degrees of autozygosity even where familial marriages are less commonly practiced but endogamy is the norm (the Amish). Both situations lead to enrichment of founder variants and an increased prevalence of autosomal recessive diseases. These communities are highly enriched for a unique combination of rare variants, enabling highly successful studies that have identified multiple novel candidate disease genes (Alazami *et al.*, 2015; Lopes *et al.*, 2016).

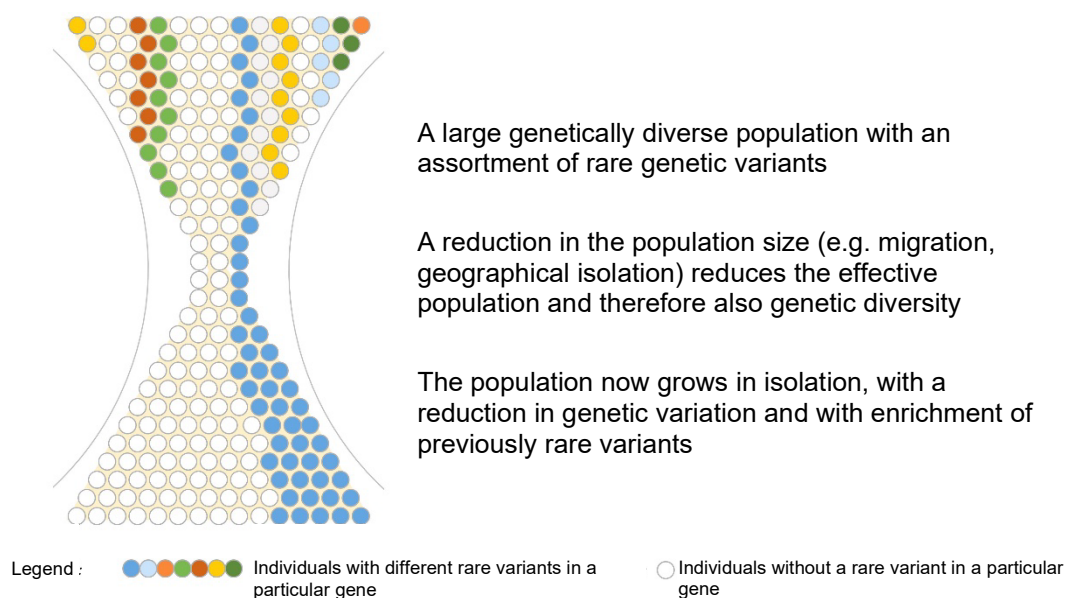


Figure 1.13: Ancestral bottleneck events result in reduced genetic diversity with enrichment of rare alleles.

Original work

In community-based studies, it is important that researchers engage with patients, families and the wider community at every stage of the process (project design, implementation and publication), so that the work brings benefits to the community. A consensus definition of community genetics was

reached by a group of scientists led by Leo Ten Kate. This group emphasised the benefit of genetic service provision to the individuals in a community as the primary aim.

“...the application of health and disease-related genetics and genomics knowledge and technologies in human populations and communities to the benefit of individuals therein.” (ten Kate et al., 2010)

Our Rare Disease Genomics research group has established a number of international translational community genomic research programmes that have been co-developed with the communities and their local and specialist healthcare providers. This thesis focusses on inherited neurodevelopmental disorders identified in affected individuals in families from Anabaptist (Amish / Mennonite) and Palestinian communities.

1.12. The North American Anabaptist communities

The Amish are an Anabaptist Christian church who arose under the leadership of Jakob Amman during the Protestant Reformation in Central Europe during the 16th century (Nolt, 2016). Their religious beliefs and their rejection of violence and infant baptism (Anabaptist), lead to their persecution in Europe resulting in their exodus to the “New World” in two waves in the 18th and 19th centuries. Around 500 Amish families migrated to the US and settled in the county of Pennsylvania between 1736 and 1770. A second wave of ~3,000 Amish immigrants travelled to Pennsylvania, Ohio, Indiana, Iowa and Illinois between 1818 and 1860. Migrations also occurred at a similar time from other Anabaptist groups often considered to be part of the Amish but with a distinct history and customs, such as the Mennonites and the Hutterites (Nolt, 2016).

The original emigree families in these communities were already extensively related, having separated themselves from the mainstream society whilst living in Europe. This gave rise to a gene pool among the immigrant Anabaptist peoples that was already more restricted than typical Northern European communities at the time.

The Amish are now one of the fastest growing populations in the world, doubling approximately every 20 years, due to most Amish couples having five or more children (in previous generations this number was greater). The Amish population in North America was estimated to be 373,620 in June 2022, more than half of whom are resident in Ohio, Pennsylvania and Indiana. (Elizabethtown College, 2022).

Amish communities are known for their traditional plain dress and simple rural lifestyle largely separated from modern communities. They use horse and buggy transportation, speak “Pennsylvanian Dutch” (a German originating dialect) and their children receive basic education only until the eighth grade. They hold church services within their homes rather than formal religious buildings, and practice non-violence and “shunning” (social exclusion of community members who have transgressed the rules of the community). Historical divisions among the Amish have resulted in a number of separate communities existing today (Old Order Amish, New Order Amish, Beachy Amish, Amish Mennonites) each with their own specific practices and ordinances (Nolt, 2016).

Many Amish live below the federal poverty threshold and have no health insurance. The standard clinical evaluation to diagnose a child with neurodevelopmental disability costs an average of \$19,000 (ranging typically from \$9,000 to \$35,000), excluding professional fees, indirect institutional expenses and genome sequencing (Soden *et al.*, 2014). Despite this, and an abundance of caution regarding new technologies, the Amish are often remarkably receptive to genetic/genomic testing and genomic research, given the potential to provide a diagnosis which enables targeted clinical management. A number of specialist not-for-profit clinics have been set up to serve Amish communities including the aforementioned New Leaf Clinic in which the Exeter-led translation community genomics project “Windows of Hope” is integrated.

1.13. The Windows of Hope project

The University of Exeter led *Windows of Hope* (WoH) Anabaptist translational genomic medicine study (<https://wohproject.com>) led by Professors Crosby and Baple, was established in 2000. The project has investigated many hundreds of families primarily with inherited neurological diseases from the Anabaptist (Amish/Mennonite) communities in North America. Our research team work closely with a number of healthcare services, in particular New Leaf Center Clinic for Special Children, Akron Children’s hospital and the Centre for Special Children in La Farge. This approach supports local healthcare providers, enabling translation of research findings into healthcare benefits for the community through improved understanding and recognition of the spectrum and causes of genetic disorders and development of diagnostic testing,

including a bespoke chip-based panel test for Amish variants now integrated into the genetic diagnostic laboratory in Wisconsin. The WoH study has discovered 25 new genetic disorders amongst the Amish since inception. Highlighting the global relevance of such disease gene discoveries, almost all disease genes initially identified in the Amish have subsequently been shown to cause similar diseases in other populations worldwide, (Fasham *et al.*, 2021). The successful approaches and good practices of the Windows of Hope project have been shared with other clinicians and researchers worldwide through a Massive Open Online Courses (MOOC) (<https://www.futurelearn.com/courses/community-genetics>)

Adopting the same approach as The Windows of Hope project in North America, The *Stories of Hope* project, which commenced 2019, works closely with local clinicians to characterise the clinical and molecular basis of rare diseases in Palestinian communities in order to translate research findings into improved diagnostic provision and clinical care for affected patients, families, and their communities. Further the study also aims to increase genetic testing capacity locally by recruiting, supporting, and training local scientists in genomics

1.14. Palestinian communities

The Middle East is the land-bridge joining Europe, Asia and Africa and has long been a crossroads for civilisation. Previously part of the Ottoman Empire, at its dissolution in 1917 the region of Palestine was placed under the administration of the British by the League of Nations. Large-scale Jewish immigration

occurred between 1922 and 1947 and in 1948, following the expiry of the British protectorate, the state of Israel declared independence. Wars between Israel, Palestinians and Arab neighbours in 1948, 1967 and 1973 resulted the occupation of the Gaza Strip and the West Bank by Israel, which continues to this day (**Figure 1.14**).



Figure 1.14: Territory claimed by Palestine

Marked in dark green, the Gaza strip to the left and the West Bank to the right.
Source: Natural Earth (Open Source)

Despite the Oslo accord of 1993 establishing Palestinian autonomy these regions remain under Israeli occupation and conflicts continue to arise. The majority of the Palestinian population has been displaced (7 million overseas, 5.3 million in the Palestinian Occupied Territories and ~1.7 million in Israel). A number of those who have remained live in temporary camps, unemployment is high (26%) wages low (£25/day) and quality of healthcare is very variable, but

typically low (Palestinian Central Bureau of Statistics, 2022). Access to highly specialist clinics, investigations and specifically genetic testing is limited within Palestinian hospitals.

Consanguineous marriage practices, generally highly prevalent in neighbouring Arab populations (25-43%), are particularly so among Palestinians (~45%), with the majority of marriages occurring between first cousins (Rahim *et al.*, 2009). The resulting degree of consanguinity (mean coefficient of inbreeding = 0.0198) has been estimated to lead result in 13.9 deaths per thousand live births in Palestine (Bittles, 2005). These figures may be higher in the 23% of the population living rurally (<https://data.worldbank.org>, 2021) where the risk of recessive genetic disease may be further compounded by a small number of community founders and endogamous practices in small towns and villages. Like the Amish, family sizes are large (averaging 4.5 births/woman) (Rahim *et al.*, 2009).

The Palestinian people are predominantly Muslims of Arab ancestry and frequently share genetic variants with Arab populations residing in other countries for example Turkey [*GAMT* NM_138924.2:c.491del: p.(Gly164Alafs*14) (Hengel *et al.*, 2020; Item *et al.*, 2004)], Jordan [*PRICKLE1* NM_001144881.1:c.311G>A: p.(Arg104Gln) (Bassuk *et al.*, 2008; Hengel *et al.*, 2020)] and Saudi Arabia, Tunisia and Bedouin communities [*TBCE* NM_003193.5: c.155_166delGCCACGAAGGGA p.(Ser52_Gly55del) (Padidela *et al.*, 2009; Parvari *et al.*, 2002; Touati *et al.*, 2019)] amongst others.

Despite the paucity of population level genetic data for Palestinian communities, there is a single recent published study that aimed to investigate the feasibility of first-line exome sequencing in Palestinians and Israeli Arabs affected by a neurological disorder (Hengel *et al.*, 2020). The researchers recruited 83 families achieving a definitive or likely diagnosis in 42 (51%). The study eligibility required that two family members be affected by a likely-genetic neurological disorder. It is thus unsurprising that 39/42 (93%) of families diagnosed were affected by an autosomal recessive condition, given the study recruitment bias towards this mode of inheritance. Whilst in most cases causative variants were apparently private to that family (not previously reported), four variants were shared between more than one family, suggesting the possibility of founder variants (*GAMT* NM_138924.2:c.491del: p.(Gly164Alafs*14) - three families, also recorded in Turkey (Item *et al.*, 2004), *GPT2* NM_133443.4:c.70C>T p.(Gln24*); *MTMR2* NM_016156.5:c.766_767del, p.(Lys256Glufs*20); *PRICKLE1* NM_001144881.1:c.311G>A: p.(Arg104Gln) - two families each, also recorded in Jordan (Bassuk *et al.*, 2008)).

1.15. The need for increased ancestral diversity in genomic databases

Currently, the “precision” element of precision genetic medicine is based on a foundation of genetic data derived predominantly from populations of European ancestry (Popejoy *et al.*, 2018). However, genomic reference data have a diversity problem (Sirugo *et al.*, 2019). For example, whilst almost a quarter of the world’s population are of South Asian origin, only 12% (15,308/125,748) of exomes in gnomAD v2.1.1 are from individuals of this ancestry. Similarly,

despite the Middle East comprising approximately 6% of world population (and ~1% of the UK population) only 0.2% (158/76,000) of genomes in gnomAD v3.1.1 are from individuals with Arab ancestry (Grace Tiao, 2020; Laurent Francioli, 2018; Office for National Statistics, 2011; UNICEF, 2019). GnomAD [the Genome Aggregation Database (Karczewski *et al.*, 2020)] is the publicly available genomic database most widely used in rare disease gene discovery and clinical diagnostics and the absence of this information results in an inability to exclude potentially benign variants of uncertain significance (VUSes) and a resultant inability to prioritise potentially causative variants (Gudmundsson *et al.*, 2021). A recent study provided evidence of this, maintaining the diagnostic rate achieved in ethnic minorities only at a cost of significantly increased analysis time (Bowling *et al.*, 2022). This striking genomic data disparity and its consequences underscores the importance of continued efforts to diversify the genomic evidence-base (Koch, 2020).

One example of the clinical impact of this genomic inequality, is provided by a recent study undertaken in Bradford, England, which reported that two thirds of child deaths involved infants under one year old and that families of South Asian origin, where consanguineous relationships are common, were over-represented. The authors presented data showing that >40% of deaths were attributable to genetic and/or congenital anomalies. As a consequence, Bradford Council put in place plans to cooperate with the NHS to increase access to genetic testing and counselling for families at most risk (Bradford Safeguarding Children Board, 2018; Merten, 2019).

To address the genomic data disparity, a number of positive developments are being made. In 2019, GenomeAsia published the results of a pilot study analysing the genomes of 1,739 people, representing the widest coverage of genetic diversity in Asia to date (Wall *et al.*, 2019). In the Middle East a number of groups have sought to establish databases of Middle Eastern individuals. At a regional level these include The Greater Middle East (GME) Variome Project (<http://igm.ucsd.edu/gme/>) and the Dalia database (CSIR-Institute of Genomics and Integrative Biology - <https://clingen.igib.res.in/dalia>) and at a national level population genetic projects such as Iranome database, Qatar genome project and Qatar Biobank (Elfatih *et al.*, 2021; Fattahi *et al.*, 2019; Razali *et al.*, 2021). While this shows that steps are being taken, they do not yet have the same broad utility as gnomAD much more still needs to be done to address inequalities in genomic research studies, and the consequent transferability of findings and benefits (Darr *et al.*, 2016).

For these reasons, and in order to support this goal we established an in-house database of Anabaptist exomes, which was later incorporated into the Anabaptist Variant Server (collaborators include Regeneron Genetics Center, University of Maryland School of Medicine Amish Program, Clinic for Special Children, Das Deutsche Clinic and National Institute of Mental Health Amish Program) comprising a database of more than 10,000 Amish and Mennonite individuals from various research centres working with the North American Anabaptist communities.

1.16. Aims and objectives

This study falls within an established international multicentre translational rare disorders research programme, based at the University of Exeter. The overarching objective of this PhD involves the delineation and phenotypic, genetic and pathomolecular characterisation of inherited neurodevelopmental disorders within the Palestinian and Anabaptist (Amish and Mennonite) communities. In order to achieve this, the following specific aims were pursued:

1.17. Aims of the project

1. To perform detailed clinical phenotyping in combination with molecular studies to identify the genetic cause of disease in Palestinian and Anabaptist families affected by likely monogenic neurodevelopmental disorders or unknown aetiology.
2. In the case of novel candidate genetic causes of neurodevelopmental disorders additional aims were:
 - a. To explore GeneMatcher and international collaborative networks to identify additional families with these conditions so as to more completely characterise the genetic and clinical spectrum of each disorder.
 - b. To undertake collaborative molecular, cell and model organism studies in order to functionally characterise each disorder, and investigate the pathogenic variants identified.
3. To gain new insights into the clinical relevance of otherwise rare genetic variation enriched within Anabaptist and Palestinian communities, aiding genomic variant interpretation internationally.
4. To translate research findings into improved diagnostic strategies and precision medicine approaches for patients and families affected by rare neurodevelopmental disorders worldwide.

Rationale for presentation of the thesis

We investigated 113 families with neurodevelopmental disorders (93 Palestinian and 20 Amish) identifying 59 diagnoses or likely diagnoses. We found important novel candidate disease-gene associations in 10 families. Chapters 3 and 4 focus on the two exemplar studies describing novel neurodevelopmental disorders (*CAMSAP1*-related neuronal migration disorder and *SLC4A10*-related neurodevelopmental disorder). Chapter 5 illustrates how the serendipitous accumulation in the Amish of otherwise rare gene variants enables their clinical relevance to be correctly elucidated.

1.18. References

- Abdollahi MR, Morrison E, Sirey T, Molnar Z, Hayward BE, Carr IM, Springell K, Woods CG, Ahmed M, Hattingh L, *et al.* Mutation of the variant alpha-tubulin TUBA8 results in polymicrogyria with optic nerve hypoplasia. *Am J Hum Genet* 2009;85(5):737-744.
- Accogli A, Severino M, Riva A, Madia F, Balagura G, Iacomino M, Carlini B, Baldassari S, Giacomini T, Croci C, *et al.* Targeted re-sequencing in malformations of cortical development: genotype-phenotype correlations. *Seizure* 2020;80:145-152.
- Akhmanova A, & Hoogenraad CC. Microtubule minus-end-targeting proteins. *Curr Biol* 2015;25(4):R162-171.
- Akhmanova A, & Steinmetz MO. Control of microtubule organization and dynamics: two ends in the limelight. *Nat Rev Mol Cell Biol.* 2015;16(12):711-726.
- Akhmanova A, & Steinmetz MO. Microtubule minus-end regulation at a glance. *J Cell Sci* 2019;132(11).
- Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, Hijazi H, Alshammari M, Aldahmesh MA, Salih MA, *et al.* Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Rep* 2015;10(2):148-161.
- Alonso-Ferreira V, Sánchez-Díaz G, Villaverde-Hueso A, Posada de la Paz M, & Bermejo-Sánchez E. A Nationwide Registry-Based Study on Mortality Due to Rare Congenital Anomalies. *Int J Environ Res Public Health.* 2018;15(8):1715.
- Amberger JS, Bocchini CA, Scott AF, & Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* 2019;47(D1):D1038-d1043.
- American Psychiatric Association APADSMF. (2017). *Diagnostic and statistical manual of mental disorders : DSM-5.* Arlington, VA: American Psychiatric Association.
- Bajaj S, Bagley JA, Sommer C, Vertesy A, Nagumo Wong S, Krenn V, Lévi-Strauss J, & Knoblich JA. Neurotransmitter signaling regulates distinct phases of multimodal human interneuron migration. *The EMBO Journal* 2021;40(23):e108714.
- Balick DJ, Jordan DM, Sunyaev S, & Do R. Overcoming constraints on the detection of recessive selection in human genes from population frequency data. *Am J Hum Genet.* 2022;109(1):33-49.
- Banerjee A, Rikhye RV, Breton-Provencher V, Tang X, Li C, Li K, Runyan CA, Fu Z, Jaenisch R, & Sur M. Jointly reduced inhibition and excitation underlies circuit-wide changes in cortical processing in Rett syndrome. *Proc Natl Acad Sci U S A* 2016;113(46):E7287-e7296.
- Baple E, Maroofian R, Chioza B, Izadi M, Cross H, Al-Turki S, Barwick K, Skrzypiec A, Pawlak R, Wagner K, *et al.* Mutations in KPTN Cause Macrocephaly, Neurodevelopmental Delay, and Seizures. *Am J Hum Genet* Vol. 2014;94:87-94.
- Bartolini F, & Gundersen GG. Generation of noncentrosomal microtubule arrays. *Journal of Cell Science* 2006;119(20):4155-4163.
- Bassuk AG, Wallace RH, Buhr A, Buller AR, Afawi Z, Shimojo M, Miyata S, Chen S, Gonzalez-Alegre P, Griesbach HL, *et al.* A Homozygous Mutation in Human PRICKLE1 Causes an Autosomal-Recessive Progressive Myoclonus Epilepsy-Ataxia Syndrome. *Am J Hum Genet.* 2008;83(5):572-581.

- Bear MF, Connors BW & Paradiso MA. (2020). *Neuroscience : exploring the brain*.
- Beaulieu CL, Majewski J, Schwartzentruber J, Samuels ME, Fernandez BA, Bernier FP, Brudno M, Knoppers B, Marcadier J, Dymont D, *et al*. FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *Am J Hum Genet* 2014;94(6):809-817.
- Best S, Lord J, Roche M, Watson CM, Poulter JA, Bevers RPJ, Stuckey A, Szymanska K, Ellingford JM, Carmichael J, *et al*. Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project. *J Med Genet* 2022;59(8):737-747.
- Bick D, Bick SL, Dimmock DP, Fowler TA, Caulfield MJ, & Scott RH. An online compendium of treatable genetic disorders. *Am J Med Genet C Semin Med Genet* 2021;187(1):48-54.
- Bittles AH. Endogamy, Consanguinity and Community Disease Profiles. *Public Health Genomics* 2005;8(1):17-20.
- Bock HH, & May P. Canonical and Non-canonical Reelin Signaling. *Front Cell Neurosci*. 2016;10.
- Bodakuntla S, Jijumon AS, Villablanca C, Gonzalez-Billault C, & Janke C. Microtubule-Associated Proteins: Structuring the Cytoskeleton. *Trends in Cell Biology* 2019;29(10):804-819.
- Bortone D, & Polleux F. KCC2 expression promotes the termination of cortical interneuron migration in a voltage-sensitive calcium-dependent manner. *Neuron* 2009;62(1):53-71.
- Bowling KM, Thompson ML, Finnila CR, Hiatt SM, Latner DR, Amaral MD, Lawlor JMJ, East KM, Cochran ME, Greve V, *et al*. Genome sequencing as a first-line diagnostic test for hospitalized infants. *Genet Med*. 2022;24(4):851-861.
- Boycott KM, Flavelle S, Bureau A, Glass HC, Fujiwara TM, Wirrell E, Davey K, Chudley AE, Scott JN, McLeod DR, *et al*. Homozygous deletion of the very low density lipoprotein receptor gene causes autosomal recessive cerebellar hypoplasia with cerebral gyral simplification. *Am J Hum Genet* 2005;77(3):477-483.
- Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, Brookes AJ, Brudno M, Carracedo A, den Dunnen JT, *et al*. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am J Hum Genet* 2017;100(5):695-705.
- Bradford Safeguarding Children Board. (2018). *Child Death Overview Panel (CDOP) Annual report 2017-18*. Retrieved from <https://saferbradford.co.uk/media/jjdcwbbz/cdop-annual-report-2017-2018-july-final-for-publication.pdf>
- Breuss M, Heng Julian I-T, Poirier K, Tian G, Jaglin Xavier H, Qu Z, Braun A, Gstrein T, Ngo L, Haas M, *et al*. Mutations in the β -Tubulin Gene TUBB5 Cause Microcephaly with Structural Brain Abnormalities. *Cell Reports* 2012;2(6):1554-1562.
- Brouhard GJ, & Rice LM. Microtubule dynamics: an interplay of biochemistry and mechanics. *Nat Rev Mol Cell Biol* 2018;19(7):451-463.
- Bystron I, Blakemore C, & Rakic P. Development of the human cerebral cortex: Boulder Committee revisited. *Nat Rev Neurosci* 2008;9(2):110-122.
- Cardoso C, Leventer RJ, Ward HL, Toyo-Oka K, Chung J, Gross A, Martin CL, Allanson J, Pilz DT, Olney AH, *et al*. Refinement of a 400-kb critical region allows genotypic differentiation between isolated lissencephaly, Miller-Dieker syndrome, and other phenotypes secondary to deletions of 17p13.3. *Am J Hum Genet* 2003;72(4):918-930.

- Carlson BMKPN. (2019). *Human embryology and developmental biology*.
- Caspi M, Atlas R, Kantor A, Sapir T, & Reiner O. Interaction between LIS1 and doublecortin, two lissencephaly gene products. *Hum Mol Genet* 2000;9(15):2205-2213.
- Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM, MacArthur DG, *et al*. A brief history of human disease genetics. *Nature* 2020;577(7789):179-189.
- Copp AJ, Greene ND, & Murdoch JN. The genetic basis of mammalian neurulation. *Nat Rev Genet* 2003;4(10):784-793.
- Curia G, Papouin T, Séguéla P, & Avoli M. Downregulation of tonic GABAergic inhibition in a mouse model of fragile X syndrome. *Cereb Cortex* 2009;19(7):1515-1520.
- Cushion TD, Paciorkowski AR, Pilz DT, Mullins JG, Seltzer LE, Marion RW, Tuttle E, Ghoneim D, Christian SL, Chung SK, *et al*. De novo mutations in the beta-tubulin gene TUBB2A cause simplified gyral patterning and infantile-onset epilepsy. *Am J Hum Genet* 2014;94(4):634-641.
- Darr A, Small N, Ahmad WI, Atkin K, Corry P, & Modell B. Addressing key issues in the consanguinity-related risk of autosomal recessive disorders in consanguineous communities: lessons from a qualitative study of British Pakistanis. *J Community Genet* 2016;7(1):65-79.
- Davies S. Generation genome: Annual report of the chief medical officer. Retrieved August 2016;11:2018.
- DDD project. (2022). DDD project: Publications. Retrieved from <https://www.ddduk.org/publications.html>
- Dent EW, Gupton SL, & Gertler FB. The growth cone cytoskeleton in axon outgrowth and guidance. *Cold Spring Harbor perspectives in biology* 2011;3(3):a001800.
- Department of Health and Social Care. (2019). *The UK strategy for rare diseases -2019 update to the Implementation Plan for England*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/781472/2019-update-to-the-rare-diseases-implementation-plan-for-england.pdf
- Department of Health and Social Care. (2021). *The UK Rare Diseases Framework*. Retrieved from <https://www.gov.uk/government/publications/uk-rare-diseases-framework>
- des Portes V, Pinard JM, Billuart P, Vinet MC, Koulakoff A, Carrié A, Gelot A, Dupuis E, Motte J, Berwald-Netter Y, *et al*. A novel CNS gene required for neuronal migration and involved in X-linked subcortical laminar heterotopia and lissencephaly syndrome. *Cell* 1998;92(1):51-61.
- Desikan RS, & Barkovich AJ. Malformations of cortical development. *Ann Neurol* 2016;80(6):797-810.
- Di Donato N, Chiari S, Mirzaa GM, Aldinger K, Parrini E, Olds C, Barkovich AJ, Guerrini R, & Dobyns WB. Lissencephaly: Expanded imaging and clinical classification. *Am J Med Genet A* 2017;173(6):1473-1488.
- Di Donato N, Timms AE, Aldinger KA, Mirzaa GM, Bennett JT, Collins S, Olds C, Mei D, Chiari S, Carvill G, *et al*. Analysis of 17 genes detects mutations in 81% of 811 patients with lissencephaly. *Genet Med* 2018;20(11):1354-1364.

- Diggle CP, Martinez-Garay I, Molnar Z, Brinkworth MH, White E, Fowler E, Hughes R, Hayward BE, Carr IM, Watson CM, *et al.* A tubulin alpha 8 mouse knockout model indicates a likely role in spermatogenesis but not in brain development. *PLoS One* 2017;12(4):e0174264.
- Dobyns WB, Aldinger KA, Ishak GE, Mirzaa GM, Timms AE, Grout ME, Dremmen MHG, Schot R, Vandervore L, van Slegtenhorst MA, *et al.* MACF1 Mutations Encoding Highly Conserved Zinc-Binding Residues of the GAR Domain Cause Defects in Neuronal Migration and Axon Guidance. *Am J Hum Genet.* 2018;103(6):1009-1021.
- Dobyns WB, Berry-Kravis E, Havernick NJ, Holden KR, & Viskochil D. X-linked lissencephaly with absent corpus callosum and ambiguous genitalia. *Am J Med Genet* 1999;86(4):331-337.
- Egawa K, Kitagawa K, Inoue K, Takayama M, Takayama C, Saitoh S, Kishino T, Kitagawa M, & Fukuda A. Decreased tonic inhibition in cerebellar granule cells causes motor dysfunction in a mouse model of Angelman syndrome. *Sci Transl Med* 2012;4(163):163ra157.
- Ehrhart F, Willighagen EL, Kutmon M, van Hoften M, Curfs LMG, & Evelo CT. A resource to explore the discovery of rare diseases and their causative genes. *Scientific Data* 2021;8(1):124.
- Elfatih A, Mifsud B, Syed N, Badii R, Mbarek H, Abbaszadeh F, & Estivill X. Actionable genomic variants in 6045 participants from the Qatar Genome Program. *Hum Mutat* 2021.
- Elizabethtown College. (2022). Amish Population Profile. Retrieved from <https://groups.etown.edu/amishstudies/statistics/amish-population-profile-2022/>
- Fasham J, Lin S, Ghosh P, Radio FC, Farrow EG, Thiffault I, Kussman J, Zhou D, Hemming R, Zahka K, *et al.* Elucidating the clinical spectrum and molecular basis of HYAL2 deficiency. *Genet Med* 2021.
- Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, Amini A, Arzhanghi S, Jalalvand K, Jamali P, *et al.* Iranome: A catalog of genomic variations in the Iranian population. *Hum Mutat* 2019;40(11):1968-1984.
- Ferreira CR. The burden of rare diseases. *Am J Med Genet A.* 2019;179(6):885-892.
- Fridman H, Yntema HG, Mägi R, Andreson R, Metspalu A, Mezzavila M, Tyler-Smith C, Xue Y, Carmi S, Levy-Lahad E, *et al.* The landscape of autosomal-recessive pathogenic variants in European populations reveals phenotype-specific effects. *Am J Hum Genet* 2021;108(4):608-619.
- Friocourt G, Kanatani S, Tabata H, Yozu M, Takahashi T, Antypa M, Raguénès O, Chelly J, Férec C, Nakajima K, *et al.* Cell-autonomous roles of ARX in cell proliferation and neuronal migration during corticogenesis. *J Neurosci* 2008;28(22):5794-5805.
- Fumagalli C, Vacirca D, Rappa A, Passaro A, Guarize J, Rafaniello Raviele P, de Marinis F, Spaggiari L, Casadio C, Viale G, *et al.* The long tail of molecular alterations in non-small cell lung cancer: a single-institution experience of next-generation sequencing in clinical molecular diagnostics. *J Clin Pathol* 2018;71(9):767-773.
- Genomics England. The 100,000 Genomes Project Protocol. *Genomics England Ltd: London* 2017.
- Genomics England. (2022a). Intellectual disability (Version 3.1500). Retrieved from <https://panelapp.genomicsengland.co.uk/panels/285/>

- Genomics England. (2022b). Research and Partnerships > Publications. Retrieved from <https://www.genomicsengland.co.uk/research/publications>
- Gorini F, Coi A, Mezzasalma L, Baldacci S, Pierini A, & Santoro M. Survival of patients with rare diseases: a population-based study in Tuscany (Italy). *Orphanet Journal of Rare Diseases* 2021;16(1):275.
- Götz M, & Huttner WB. The cell biology of neurogenesis. *Nat Rev Mol Cell Biol.* 2005;6(10):777-788.
- Grace Tiao JG. (2020). gnomAD v3.1 New Content, Methods, Annotations, and Data Availability. Retrieved from <https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1-new-content-methods-annotations-and-data-availability/>
- Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, Rehm HL, MacArthur DG, & O'Donnell-Luria A. Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat* 2021.
- Guedes-Dias P, & Holzbaur ELF. Axonal transport: Driving synaptic function. *Science* 2019;366(6462):eaaw9997.
- Guerrini R, & Dobyns WB. Malformations of cortical development: clinical features and genetic causes. *Lancet Neurol* 2014;13(7):710-726.
- Harlalka GV, Lehman A, Chioza B, Baple EL, Maroofian R, Cross H, Sreekantan-Nair A, Priestman DA, Al-Turki S, McEntagart ME, *et al.* Mutations in B4GALNT1 (GM2 synthase) underlie a new disorder of ganglioside biosynthesis. *Brain* 2013;136(Pt 12):3618-3624.
- Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 2012;489(7416):391-399.
- Heinzen EL, O'Neill AC, Zhu X, Allen AS, Bahlo M, Chelly J, Chen MH, Dobyns WB, Freytag S, Guerrini R, *et al.* De novo and inherited private variants in MAP1B in periventricular nodular heterotopia. *PLoS Genet* 2018;14(5):e1007281.
- Hendershott MC, & Vale RD. Regulation of microtubule minus-end dynamics by CAMSAPs and Patronin. *Proc Natl Acad Sci U S A* 2014;111(16):5860-5865.
- Hengel H, Buchert R, Sturm M, Haack TB, Schelling Y, Mahajnah M, Sharkia R, Azem A, Balousha G, Ghanem Z, *et al.* First-line exome sequencing in Palestinian and Israeli Arabs with neurological disorders is efficient and facilitates disease gene discovery. *Eur J Hum Genet* 2020;28(8):1034-1043.
- Hernandez CC, & Macdonald RL. A structural look at GABA(A) receptor mutations linked to epilepsy syndromes. *Brain Res* 2019;1714:234-247.
- Hirokawa N, Niwa S, & Tanaka Y. Molecular motors in neurons: transport mechanisms and roles in brain function, development, and disease. *Neuron* 2010;68(4):610-638.
- Hong SE, Shugart YY, Huang DT, Shahwan SA, Grant PE, Hourihane JO, Martin ND, & Walsh CA. Autosomal recessive lissencephaly with cerebellar hypoplasia is associated with human RELN mutations. *Nat Genet* 2000;26(1):93-96.
- House of Commons. (2018). *Genomics and genome editing in the NHS*. <https://publications.parliament.uk/> Retrieved from <https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/349/349.pdf>

- Hu H, Kahrizi K, Musante L, Fattahi Z, Herwig R, Hosseini M, Oppitz C, Abedini SS, Suckow V, Larti F, *et al.* Genetics of intellectual disability in consanguineous families. *Mol Psychiatry* 2018.
- Item CB, Mercimek-Mahmutoglu S, Battini R, Edlinger-Horvat C, Stromberger C, Bodamer O, Mühl A, Vilaseca MA, Korall H, & Stöckler-Ipsiroglu S. Characterization of seven novel mutations in seven patients with GAMT deficiency. *Hum Mutat* 2004;23(5):524.
- Jaglin XH, Poirier K, Saillour Y, Buhler E, Tian G, Bahi-Buisson N, Fallet-Bianco C, Phan-Dinh-Tuy F, Kong XP, Bomont P, *et al.* Mutations in the β -tubulin gene TUBB2B result in asymmetrical polymicrogyria. *Nat Genet* 2009;41(6):746-752.
- Jiang K, Hua S, Mohan R, Grigoriev I, Yau KW, Liu Q, Katrukha EA, Altelaar AF, Heck AJ, Hoogenraad CC, *et al.* Microtubule minus-end stabilization by polymerization-driven CAMSAP deposition. *Dev Cell* 2014;28(3):295-309.
- Kaplanis J, Samocha KE, Wiel L, Zhang Z, Arvai KJ, Eberhardt RY, Gallone G, Lelieveld SH, Martin HC, McRae JF, *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* 2020;586(7831):757-762.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581(7809):434-443.
- Kato M. Genotype-phenotype correlation in neuronal migration disorders and cortical dysplasias. *Front Neurosci.* 2015;9.
- Keating TJ, & Borisy GG. Centrosomal and non-centrosomal microtubules. *Biol Cell* 1999;91(4-5):321-329.
- Keays DA, Tian G, Poirier K, Huang GJ, Siebold C, Cleak J, Oliver PL, Fray M, Harvey RJ, Molnar Z, *et al.* Mutations in alpha-tubulin cause abnormal neuronal migration in mice and lissencephaly in humans. *Cell* 2007;128(1):45-57.
- Kibar Z, Torban E, McDearmid JR, Reynolds A, Berghout J, Mathieu M, Kirillova I, De Marco P, Merello E, Hayes JM, *et al.* Mutations in VANG1 associated with neural-tube defects. *N Engl J Med* 2007;356(14):1432-1437.
- Koch L. Exploring human genomic diversity with gnomAD. *Nat Rev Genet* 2020;21(8):448-448.
- Kolbjørn S, Martin DA, Pettersson M, Dahlin M, & Anderlid BM. Lissencephaly in an epilepsy cohort: Molecular, radiological and clinical aspects. *Eur J Paediatr Neurol* 2021;30:71-81.
- Lasser M, Tiber J, & Lowery LA. The Role of the Microtubule Cytoskeleton in Neurodevelopmental Disorders. *Front Cell Neurosci* 2018;12.
- Laurent Francioli GT, Konrad Karczewski, Matthew Solomonson, Nick Watts. (2018). gnomAD v2.1. Retrieved from <https://gnomad.broadinstitute.org/news/2018-10-gnomad-v2-1/>
- Lo Nigro C, Chong CS, Smith AC, Dobyns WB, Carrozzo R, & Ledbetter DH. Point mutations and an intragenic deletion in LIS1, the lissencephaly causative gene in isolated lissencephaly sequence and Miller-Dieker syndrome. *Hum Mol Genet* 1997;6(2):157-164.
- Looma S, Straehle J, Gangadharan V, Heike N, Khalifa A, Motta A, Ju N, Sievers M, Gempt J, Meyer HS, *et al.* Connectomic comparison of mouse and human cortex. *Science* 2022;0(0):eabo0924.

- Lopes FL, Hou L, Boldt AB, Kassem L, Alves VM, Nardi AE, & McMahon FJ. Finding Rare, Disease-Associated Variants in Isolated Groups: Potential Advantages of Mennonite Populations. *Hum Biol* 2016;88(2):109-120.
- Lord Bethell of Romford. (2020). *Genome UK: the future of healthcare*. <https://www.gov.uk/government/publications/genome-uk-the-future-of-healthcare>
Retrieved from <https://www.gov.uk/government/publications/genome-uk-the-future-of-healthcare>
- Lui JH, Hansen DV, & Kriegstein AR. Development and evolution of the human neocortex. *Cell* 2011;146(1):18-36.
- Manickam K, McClain MR, Demmer LA, Biswas S, Kearney HM, Malinowski J, Massingham LJ, Miller D, Yu TW, & Hisama FM. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). *Genet Med* 2021;23(11):2029-2037.
- Mariani J, Simonini MV, Palejev D, Tomasini L, Coppola G, Szekely AM, Horvath TL, & Vaccarino FM. Modeling human cortical development in vitro using induced pluripotent stem cells. *Proc Natl Acad Sci U S A* 2012;109(31):12770-12775.
- Martin HC, Jones WD, McIntyre R, Sanchez-Andrade G, Sanderson M, Stephenson JD, Jones CP, Handsaker J, Gallone G, Bruntraeger M, *et al*. Quantifying the contribution of recessive coding variation to developmental disorders. *Science* 2018;362(6419):1161-1164.
- McInnes G, Sharo AG, Koleske ML, Brown JEH, Norstad M, Adhikari AN, Wang S, Brenner SE, Halpern J, Koenig BA, *et al*. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am J Hum Genet* 2021;108(4):535-548.
- Merten M. Keeping it in the family: consanguineous marriage and genetic disorders, from Islamabad to Bradford. *BMJ* 2019;365:l1851.
- Mitani T, Isikay S, Gezdirici A, Gulec EY, Punetha J, Fatih JM, Herman I, Akay G, Du H, Calame DG, *et al*. High prevalence of multilocus pathogenic variation in neurodevelopmental disorders in the Turkish population. *Am J Hum Genet* 2021;108(10):1981-2005.
- Mitani T, Punetha J, Akalin I, Pehlivan D, Dawidziuk M, Coban Akdemir Z, Yilmaz S, Aslan E, Hunter JV, Hijazi H, *et al*. Bi-allelic Pathogenic Variants in TUBGCP2 Cause Microcephaly and Lissencephaly Spectrum Disorders. *Am J Hum Genet* 2019;105(5):1005-1015.
- Mitchison T, & Kirschner M. Dynamic instability of microtubule growth. *Nature* 1984;312(5991):237-242.
- Monies D, Abouelhoda M, AlSayed M, Alhassnan Z, Alotaibi M, Kayyali H, Al-Owain M, Shah A, Rahbeeni Z, Al-Muhaizea MA, *et al*. The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes. *Hum Genet* 2017;136(8):921-939.
- Monies D, Abouelhoda M, Assoum M, Moghrabi N, Rafiullah R, Almontashiri N, Alowain M, Alzaidan H, Alsayed M, Subhani S, *et al*. Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a Highly Consanguineous Population. *Am J Hum Genet*. 2019;104(6):1182-1201.
- Mutch CA, Poduri A, Sahin M, Barry B, Walsh CA, & Barkovich AJ. Disorders of Microtubule Function in Neurons: Imaging Correlates. *AJNR Am J Neuroradiol* 2016;37(3):528-535.

- Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, Chen W, Hosseini M, Behjati F, Haas S, Jamali P, *et al.* Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* 2011;478(7367):57-63.
- Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, & Rath A. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 2020;28(2):165-173.
- Nolt SM. (2016). *A history of the Amish*.
- Oegema R, Barakat TS, Wilke M, Stouffs K, Amrom D, Aronica E, Bahi-Buisson N, Conti V, Fry AE, Geis T, *et al.* International consensus recommendations on the diagnostic work-up for malformations of cortical development. *Nat Rev Neurol* 2020;16(11):618-635.
- Oegema R, McGillivray G, Leventer R, Le Moing A-G, Bahi-Buisson N, Barnicoat A, Mandelstam S, Francis D, Francis F, Mancini GMS, *et al.* EML1-associated brain overgrowth syndrome with ribbon-like heterotopia. *Am J Med Genet C Semin Med Genet*. 2019;181(4):627-637.
- Office for National Statistics. (2011). *2011 Census*. Retrieved from <https://www.ons.gov.uk/census/2011census>
- Oh WC, Lutz S, Castillo PE, & Kwon H-B. De novo synaptogenesis induced by GABA in the developing mouse cortex. *Science* 2016;353(6303):1037-1040.
- Ori-McKenney KM, Jan LY, & Jan Y-N. Golgi outposts shape dendrite morphology by functioning as sites of acentrosomal microtubule nucleation in neurons. *Neuron* 2012;76(5):921-930.
- Padidela R, Kelberman D, Press M, Al-Khawari M, Hindmarsh PC, & Dattani MT. Mutation in the TBCE Gene Is Associated with Hypoparathyroidism-Retardation-Dysmorphism Syndrome Featuring Pituitary Hormone Deficiencies and Hypoplasia of the Anterior Pituitary and the Corpus Callosum. *J Clin Endocrinol Metab* 2009;94(8):2686-2691.
- Palestinian Central Bureau of Statistics. (2022). *Palestine in figures 2021*. Ramallah, Palestine
- Parenti I, Rabaneda LG, Schoen H, & Novarino G. Neurodevelopmental Disorders: From Genetics to Functional Pathways. *Trends Neurosci* 2020;43(8):608-621.
- Parvari R, Hershkovitz E, Grossman N, Gorodischer R, Loeys B, Zecic A, Mortier G, Gregory S, Sharony R, Kambouris M, *et al.* Mutation of TBCE causes hypoparathyroidism-retardation-dysmorphism and autosomal recessive Kenny-Caffey syndrome. *Nat Genet* 2002;32(3):448-452.
- Poirier K, Keays DA, Francis F, Saillour Y, Bahi N, Manouvrier S, Fallet-Bianco C, Pasquier L, Toutain A, Tuy FP, *et al.* Large spectrum of lissencephaly and pachygyria phenotypes resulting from de novo missense mutations in tubulin alpha 1A (TUBA1A). *Hum Mutat* 2007;28(11):1055-1064.
- Poirier K, Lebrun N, Broix L, Tian G, Saillour Y, Boscheron C, Parrini E, Valence S, Pierre BS, Oger M, *et al.* Mutations in TUBG1, DYNC1H1, KIF5C and KIF2A cause malformations of cortical development and microcephaly. *Nat Genet* 2013;45(6):639-647.
- Pollard TD, & Goldman RD. Overview of the Cytoskeleton from an Evolutionary Perspective. *Cold Spring Harb Perspect Biol* 2018;10(7).
- Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorff LA, Koenig B, Ramos EM, Sorokin EP, Wand H, *et al.* The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Hum Mutat* 2018;39(11):1713-1720.

- Quinodoz M, Peter VG, Bedoni N, Royer Bertrand B, Cisarova K, Salmaninejad A, Sepahi N, Rodrigues R, Piran M, Mojarrad M, *et al.* AutoMap is a high performance homozygosity mapping tool using next-generation sequencing data. *Nat Commun* 2021;12(1):518.
- Rahim HF, Wick L, Halileh S, Hassan-Bitar S, Chekir H, Watt G, & Khawaja M. Maternal and child health in the occupied Palestinian territory. *Lancet* 2009;373(9667):967-977.
- Razali RM, Rodriguez-Flores J, Ghorbani M, Naeem H, Aamer W, Aliyev E, Jubran A, Ismail SI, Al-Muftah W, Badji R, *et al.* Thousands of Qatari genomes inform human migration history and improve imputation of Arab haplotypes. *Nat Commun* 2021;12(1):5929.
- Rehm HL. Time to make rare disease diagnosis accessible to all. *Nature Medicine* 2022.
- Rivière JB, van Bon BW, Hoischen A, Kholmanskikh SS, O'Roak BJ, Gilissen C, Gijsen S, Sullivan CT, Christian SL, Abdul-Rahman OA, *et al.* De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. *Nat Genet* 2012;44(4):440-444, s441-442.
- Romaniello R, Arrigoni F, Fry AE, Bassi MT, Rees MI, Borgatti R, Pilz DT, & Cushion TD. Tubulin genes and malformations of cortical development. *Eur J Med Genet* 2018;61(12):744-754.
- Ronan JL, Wu W, & Crabtree GR. From neural development to cognition: unexpected roles for chromatin. *Nat Rev Genet* 2013;14(5):347-359.
- Rubenstein JL, & Merzenich MM. Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes Brain Behav* 2003;2(5):255-267.
- Sadler TW. (2003). *Langman's Medical Embryology. With CD-ROM*. Philadelphia: Lippincott Williams & Wilkins.
- Schmidt L, Wain KE, Hajek C, Estrada-Veras JI, Guillen Sacoto MJ, Wentzensen IM, Malhotra A, Clause A, Perry D, Moreno-De-Luca A, *et al.* Expanding the Phenotype of TUBB2A Related Tubulinopathy: Three Cases of a Novel, Heterozygous TUBB2A Pathogenic Variant p.Gly98Arg. *Molecular Syndromology* 2021;12(1):33-40.
- Shen T, Lee A, Shen C, & Lin CJ. The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. *Genetics Research* 2015;97:e15.
- Sirugo G, Williams SM, & Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell* 2019;177(1):26-31.
- Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, Ellingford JM, Arno G, Tucci A, Vandrovcova J, *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med* 2021;385(20):1868-1880.
- Smith DS, Niethammer M, Ayala R, Zhou Y, Gambello MJ, Wynshaw-Boris A, & Tsai LH. Regulation of cytoplasmic dynein behaviour and microtubule organization by mammalian Lis1. *Nat Cell Biol* 2000;2(11):767-775.
- Soden SE, Saunders CJ, Willig LK, Farrow EG, Smith LD, Petrikin JE, LePichon JB, Miller NA, Thiffault I, Dinwiddie DL, *et al.* Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Sci Transl Med* 2014;6(265):265ra168.
- Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheatle-Jarvela AM, Eng CM, Esteves C, Gahl WA, Hamid R, *et al.* Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N Engl J Med* 2018;379(22):2131-2139.

- Srivastava S, Love-Nichols JA, Dies KA, Ledbetter DH, Martin CL, Chung WK, Firth HV, Frazier T, Hansen RL, Prock L, *et al.* Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med* 2019;21(11):2413-2421.
- Stessman HA, Bernier R, & Eichler EE. A genotype-first approach to defining the subtypes of a complex disease. *Cell* 2014;156(5):872-877.
- Subramanian L, Calcagnotto ME, & Paredes MF. Cortical Malformations: Lessons in Human Brain Development. *Front Cell Neurosci* 2019;13:576.
- Tang X, Jaenisch R, & Sur M. The role of GABAergic signalling in neurodevelopmental disorders. *Nat Rev Neurosci* 2021;22(5):290-307.
- Tas RP, Chazeau A, Cloin BM, Lambers ML, Hoogenraad CC, & Kapitein LC. Differentiation between oppositely oriented microtubules controls polarized neuronal transport. *Neuron* 2017;96(6):1264-1271. e1265.
- ten Kate LP, Al-Gazali L, Anand S, Bittles A, Cassiman J-J, Christianson A, Cornel MC, Hamamy H, Kääriäinen H, Kristoffersson U, *et al.* Community genetics. Its definition 2010. *Journal Comm Genet* 2010;1(1):19-22.
- The Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 2015;519(7542):223-228.
- Tischfield MA, Baris HN, Wu C, Rudolph G, Van Maldergem L, He W, Chan WM, Andrews C, Demer JL, Robertson RL, *et al.* Human TUBB3 mutations perturb microtubule dynamics, kinesin interactions, and axon guidance. *Cell* 2010;140(1):74-87.
- Touati A, Nouri S, Halleb Y, Kmiha S, Mathlouthi J, Tej A, Mahdhaoui N, Ben Ahmed A, Saad A, Bensignor C, *et al.* Additional Tunisian patients with Sanjad-Sakati syndrome: A review toward a consensus on diagnostic criteria. *Arch Pediatr* 2019;26(2):102-107.
- Tovey Corinne A, & Conduit Paul T. Microtubule nucleation by γ -tubulin complexes and beyond. *Essays Biochem* 2018;62(6):765-780.
- UNICEF. (2019). *MENA Generation 2030*. Retrieved from <https://data.unicef.org/resources/middle-east-north-africa-generation-2030/>
- Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, *et al.* A de novo paradigm for mental retardation. *Nat Genet* 2010;42(12):1109-1112.
- Vissers LELM, Gilissen C, & Veltman JA. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet* 2016;17(1):9-18.
- Wall JD, Stawiski EW, Ratan A, Kim HL, Kim C, Gupta R, Suryamohan K, Gusareva ES, Purbojati RW, Bhangale T, *et al.* The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 2019;576(7785):106-111.
- Wamsley B, & Fishell G. Genetic and activity-dependent mechanisms underlying interneuron diversity. *Nat Rev Neurosci* 2017;18(5):299-309.
- Wheway G, Consortium GER, Mitchison HM, Ambrose JC, Baple EL, Bleda M, Boardman-Pretty F, Boissiere JM, Boustred CR, Caulfield MJ, *et al.* Opportunities and Challenges for Molecular Understanding of Ciliopathies—The 100,000 Genomes Project. *Front Genet.* 2019;10:127.

- Winding M, Kelliher MT, Lu W, Wildonger J, & Gelfand VI. Role of kinesin-based microtubule sliding in *Drosophila* nervous system development. *Proceedings of the National Academy of Sciences* 2016;113(34):E4985-E4994.
- Windows of Hope Project. Welcome to Windows of Hope. Retrieved from <https://wohproject.com/>
- Wiszniewski W, Gawlinski P, Gambin T, Bekiesinska-Figatowska M, Obersztyn E, Antczak-Marach D, Akdemir ZHC, Harel T, Karaca E, Jurek M, *et al.* Comprehensive genomic analysis of patients with disorders of cerebral cortical development. *Eur J Hum Genet* 2018;26(8):1121-1131.
- Witte H, Neukirchen D, & Bradke F. Microtubule stabilization specifies initial neuronal polarization. *Journal of Cell Biology* 2008;180(3):619-632.
- World Health Organization (WHO). (2019/2021). International Classification of Diseases, Eleventh Revision (ICD-11). Retrieved from <https://icd.who.int/browse11>. Retrieved 26/07/2022 <https://icd.who.int/browse11>
- Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, King DA, Ambridge K, Barrett DM, Bayzatinova T, *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* 2015;385(9975):1305-1314.
- Wright CF, McRae JF, Clayton S, Gallone G, Aitken S, FitzGerald TW, Jones P, Prigmore E, Rajan D, Lord J, *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med* 2018;20(10):1216-1223.
- Yokota Y, Gashghaei HT, Han C, Watson H, Campbell KJ, & Anton ES. Radial glial dependent and independent dynamics of interneuronal migration in the developing cerebral cortex. *PLoS One* 2007;2(8):e794.
- Zhou S, & Yu Y. Synaptic E-I Balance Underlies Efficient Neural Coding. *Front Neurosci.* 2018;12.

2

Materials and Methods

2.1. Buffers, reagents and stock solutions

Reagents, buffers and stock solutions used in this study are listed in **Table 2.1** and **Table 2.2**. Laboratory consumables (pipette tips, plastic vessels) were purchased from Alpha Laboratories (Eastleigh, UK), STARLAB Group (Milton Keynes, UK) and Mettler Toledo (Columbus, Ohio, USA) for Rainin™ products.

Reagents	Manufacturer
Agarose (molecular grade)	Thermo Fisher Scientific (Waltham, MA, USA)
Automated Droplet Generation Oil for EvaGreen	Bio-Rad (Hercules, CA, USA)
Boric Acid	Thermo Fisher Scientific (Waltham, MA, USA)
Deoxyribonucleoside triphosphate (dNTP) solution mix	Solis BioDyne (Tartu, Estonia)
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich (St. Louis, MI, USA)
DreamTaq™ DNA polymerase	Thermo Fisher Scientific (Waltham, MA, USA)
DreamTaq™ green buffer (10x)	Thermo Fisher Scientific (Waltham, MA, USA)
DreamTaq™ Green PCR Master Mix (2x)	Thermo Fisher Scientific (Waltham, MA, USA)
Ethanol	Thermo Fisher Scientific (Waltham, MA, USA)
Ethidium bromide	Thermo Fisher Scientific (Waltham, MA, USA)
Exonuclease I	New England Biolabs (Ipswich, MA, USA)
Invitrogen™ Qubit™ dsDNA HS and BR Assay Kit	Thermo Fisher Scientific (Waltham, MA, USA)
Isohelix™ Xtreme DNA Kit	Isohelix (Harrietsham, Kent, UK)
Lithium acetate dehydrate	Alfa Aesar (Ward Hill, MA, USA)
GeneRuler™ 1kb DNA ladder	Thermo Fisher Scientific (Waltham, MA, USA)
GeneRuler™ 1kb Plus DNA ladder	Thermo Fisher Scientific (Waltham, MA, USA)
ORACollect® for Pediatrics OC-175 buccal swabs	DNA Genotek (Ottawa, Canada)
QX200 EvaGreen ddPCR Supermix	Bio-Rad (Hercules, CA, USA)
QX200 ddPCR EvaGreen Buffer Control Kit	Bio-Rad (Hercules, CA, USA)
ReliaPrep™ Miniprep system	Promega Corporation (Madison, WI, USA)
Shrimp alkaline phosphatase	New England Biolabs (Ipswich, MA, USA)

Table 2.1: Reagents used in this study

Buffer/Stock solution	Composition
Lithium acetate borate (LAB) buffer	For 1 L 50x stock solution: 51 g lithium acetate dehydrate, 31 g boric acid in ddH ₂ O to final volume
ExoSAP enzyme mix	For 1 ml: 2.5 µl Exonuclease I 20,000 U/ml), 25 µl shrimp alkaline phosphatase (1000U/ml) in ddH ₂ O to final volume

Table 2.2: Buffers and stock solutions used in this study

2.2. Subjects and samples

Recruitment

The Palestinian “Stories of Hope” and Anabaptist “Windows of Hope” translational genomic research programmes were reviewed and approved by Palestinian Health Research Council (PHRC/51819) and Akron Children’s Hospital Institutional Review Board (IRB), Ohio, USA (#986876) respectively. All the studies described in this thesis were conducted in accordance with the principles of the Declaration of Helsinki. Recruitment to both research programmes required submission of the appropriate signed consent from all participants and/or parent/legal guardian, detailed phenotypic information and a blood or buccal sample.

Phenotyping of affected individuals

Affected individuals were identified by their clinician or using GeneMatcher (Sobreira *et al.*, 2015). Phenotypic information was obtained through the clinical care provider using a targeted questionnaire. Where possible, a full medical and developmental history and systemic examination for the purposes of the research study was obtained and photographic records of pertinent examination findings were taken with informed consent. Where appropriate, additional clinical records (including results of any investigations) and educational reports were requested using the appropriate information release form.

Height and weight were measured barefoot. Occipitofrontal circumference (OFC) was measured following standard procedures. Height, weight and OFC

standard deviations and Z-scores were calculated using a Microsoft Excel add-in to access growth references based on the Lambda-Mu-Sigma (LMS) method (Cole *et al.*, 1998; Pan & Cole, 2011). Whenever possible, original neuroimaging was requested for review locally by Professor William Dobyns, University of Minnesota for the *CAMSAP1* study or Dr Lucy Lees Consultant Neuroradiologist, University Hospitals Plymouth NHS Trust for the *SLC4A10* study.

Data Management

On collection of a blood/buccal/DNA sample, each family was assigned a pedigree ID, and each individual an anonymised individual ID. Family pedigrees were constructed using HaploPainter v1.043 (Thiele & Nürnberg, 2005) using information provided by the family and in the case of Anabaptist families the online Swiss Anabaptist Geneological Association database (www.saga-omii.org). Clinical and molecular information was recorded in a password protected database. All data storage and management was compliant with UK General Data Protection Regulation (GDPR).

2.3. Molecular methods

Sample collection, DNA extraction and storage

The studies described in this thesis were carried out in compliance with the Human Tissue Authority (HTA) Codes of Practice and Standards (Code E: Research). The Human Tissue Act 2004 defines as any material 'which consists of or includes human cells' to be 'relevant material' and this includes blood

samples. Therefore, all blood and buccal samples (and subsequent DNA extractions) were stored in HTA-licensed premises. Peripheral venous blood samples were kept cool during transport and stored in Exeter at -20°C. Buccal samples were collected using ORAcollect® for Pediatrics OC-175 buccal swabs (DNA Genotek, Ottawa, Canada) and stored in Exeter at 4°C.

DNA extraction from whole blood samples

Lymphocyte DNA was extracted and purified using the ReliaPrep™ Miniprep system (Promega Corporation, Madison, Wisconsin, USA) according to the manufacturer's protocol. This is a 4-stage ethanol-free process designed to maximise the purity of extracted DNA. Filter pipette tips were used at all stages to reduce the risk of contamination of the pipette with blood and DNA.

1. *Preparation of blood samples*

Frozen blood samples were thawed then thoroughly mixed by hand for at least 2 mins.

2. *Cell lysis*

200 µl whole blood was added to a 1.5 ml microcentrifuge tube containing 20 µl of proteinase K and briefly mixed. To this 200 µl of cell lysis buffer was added followed by mixing using a vortex device (Topmix FB15024 Vortex Mixer, Fisher Scientific, Waltham, MA, USA) and incubation at 56°C in a water bath (Mixer HC Thermoblock, STARLAB, Milton Keynes, UK) for 10 minutes.

3. *DNA bound to binding column*

The lysate was removed from the water bath and 250 µl of binding buffer was added before further vortex mixing. At this stage the lysate was visually inspected to ensure a dark green colour, as specified in the protocol. The contents of the tube were then added to the binding column and centrifuged at maximum speed (13,000rpm) for 1 minute with the flowthrough discarded as hazardous waste. The binding column was then visually inspected to ensure that the lysate had completely passed

through the membrane. If any lysate remained above the membrane the column was centrifuged for a further 1 min.

4. *DNA purification*

Three identical washes were performed. In each, 500 µl of column wash was added to the column and centrifuged at maximum speed for 3 minutes with the flowthrough discarded.

5. *DNA elution*

The column was placed in a new 1.5 ml microcentrifuge tube and 200 µl of nuclease-free water was added to the column. This was centrifuged at maximum speed for 1 minute resulting in elution of the DNA. Desired concentration was >30 ng/µl, resulting in a desired yield >6 mg.

6. *DNA storage*

Extracted DNA was stored at -20°C.

DNA extraction from buccal swabs

The Isohelix™ Xtreme DNA Kit (Isohelix. Harrietsham, Kent, UK) was used for extracting DNA following the manufacturers protocol. This is an alternative column-based DNA extraction protocol. It is optimised for the extraction and purification of DNA from salivary and buccal samples. It uses ethanol, which improves the precipitation of DNA and aids the removal of some impurities. Otherwise, the principles were very similar to those described in the blood extraction step above.

1. *Cell lysis*

The buccal swab arrives in a collection tube also containing transport medium and saliva. To this tube 20 µl of proteinase K and 500 µl of lysis buffer were added. They were mixed using a vortex device and incubated at 60°C in a water bath for 10 minutes.

2. *DNA bound to binding column*

To the same tube 750 µl of column binding buffer and 1.25 ml of ethanol were added with further vortex mixing. A 700 µl aliquot from the sample tube was added to the column and centrifuged at maximum speed for 1 minute with the flowthrough discarded until no sample remained. This

step was repeated until no sample remained in the sample tube, usually requiring 4 - 5 aliquots (2.8 - 3.5 ml volume in total).

3. *DNA purification*

Two identical washes were performed. In each 750 μ l of wash buffer was added to the column and centrifuged at maximum speed for 1 minute with the flowthrough discarded. Finally, the empty column was centrifuged for 3 minutes to remove all traces of ethanol.

4. *DNA elution*

The column was placed into a new 1.5 ml microcentrifuge tube, 100 μ l of elution buffer (pre-heated to 70°C) was added and it was allowed to stand for 3 minutes. It was then centrifuged for 1 minute at maximum speed resulting in elution of the DNA. Desired concentration was >30 ng/ μ l resulting in a desired yield >3 mg.

5. *DNA storage*

Extracted DNA was stored at -20°C.

Quantification of DNA

Extracted DNA concentration was routinely quantified using spectrophotometry. Samples undergoing or ddPCR or next-generation sequencing were further quantified using fluorometry.

DNA quantification using spectrophotometry

Concentration and purity of eluted DNA was assessed using a the NanoDrop™ 2000 and associated software (Thermo Fisher Scientific, Waltham, MA, USA), which uses a modified Beer-Lambert equation to calculate nucleic acid concentration based on the sample's absorption of ultraviolet (UV) light at 260 nm (A₂₆₀), the peak absorbance wavelength of nucleic acids. DNA purity was also assessed by comparing with absorbance at 280 nm (the peak absorbance wavelength for proteins/phenolic compounds) and 230 nm (the peak

absorbance wavelength for organic compounds). A260/A280 and A260/A230 ratios represent DNA purity, with an A260/A280 ratio or A260/A230 ratio <1.8 indicating the contamination of the sample with non-nucleotide molecules.

Quantification of double-stranded DNA using fluorometry

The Qubit™ system was used to quantify double-stranded DNA following extraction. The Invitrogen™ Qubit™3 fluorometer (Thermo Fisher Scientific, MA, USA) accurately measures DNA, RNA and protein quantity using a fluorescent dye that emits a signal only when bound to the target. The Invitrogen™ Qubit™ dsDNA BR (broad-range) Assay Kit (Thermo Fisher Scientific, MA) was utilised according to the manufacturer's protocol.

1. *Preparation of working solution*

A working solution was prepared by diluting Qubit™ dsDNA BR Reagent 1:200 in Qubit™ dsDNA BR Buffer (1 µl reagent, 199 µl buffer per sample + 2) in one or more 15 ml Falcon tubes.

2. *Preparation of DNA samples*

For each DNA sample requiring quantification 1 µl sample DNA was added to 199 µl working solution in a separate Qubit™ Assay Tube.

3. *Preparation of BR standards*

Qubit™ dsDNA BR standards #1 and #2 were each added to 190 µl of working solution in separate Qubit™ Assay Tubes.

4. *Calibrating the fluorometer*

Standards #1 and #2 (whose DNA concentrations are known) were measured.

5. *Measurement of DNA samples*

Samples were measured. A curve-fitting algorithm performed by the fluorometer allows DNA concentration to be calculated based on the measured fluorescence and the relationship between the two standards used in the calibration.

Polymerase chain reaction (PCR), Agarose gel electrophoresis and dideoxy sequencing

Primer design, dilution and storage

The UCSC (University of California Santa Cruz) Genome Browser (genome.ucsc.edu) was used to identify a region of Genome Reference Consortium human genome build 37 (GRCh37) sequence surrounding the variant (including 500 bp upstream and downstream). This locus was checked for known repetitive elements (short interspersed nuclear elements [SINEs], long interspersed nuclear elements [LINEs], low complexity sequences, di/trinucleotide repeats (microsatellites) and homopolymer repeats), common SNPs and regions of high GC content (guanine-cytosine content). If any of these were present the region would be redefined (made larger or smaller, moved upstream or downstream) to exclude them where possible. Within this region bespoke primers used for PCR amplification were designed for each variant using Primer3Plus software (<https://www.bioinformatics.nl/>) following the criteria below:

- Amplicon length: 200 bp (100-1000 bp)
- Primer Size: 20 bp (20 - 27 bp)
- Primer GC% - (20 - 80%)
- Melting temperature: 60°C (58 - 62°C)
- Primer melting temperatures to be within 5°C of each other

It was important that primers are uniquely complementary to a single locus, to prevent non-specific binding and amplification of multiple regions. This was

confirmed using BLAT, the BLAST-like alignment tool (Kent, 2002), integrated within the UCSC Genome Browser. Occasionally conditions outside of those stated above (higher/lower melting temperature, larger amplicon size) were required to obtain unique primers or overcome regions of high GC content.

Preparation of primers

Dry single-stranded DNA primers were ordered from Integrated DNA Technologies, BVBA (Leuven, Belgium). These were reconstituted to create a 100 μ M stock concentration, by adding 10 μ l of double distilled water (ddH₂O) for each nanogram of DNA, and then stored frozen (-20°C). The stock was further diluted 1:20 (10 μ l stock primer solution, 190 μ l ddH₂O) to make a working solution at 5 μ M, which was used for polymerase chain reaction (PCR). This working solution was stored between 0 and 4°C.

Polymerase chain reaction (PCR)

A standard protocol for DNA amplification was followed. PCR was performed using a standard 10 μ l reaction (**Table 2.3**) or an equivalent mastermix (**Table 2.4**). This included 0.8 μ l of DNA at a concentration of 10-30 ng/ μ l.

Component (concentration)	Volume
Primer forward (5 μ M)	0.4 μ l
Primer reverse (5 μ M)	0.4 μ l
DreamTaq™ green buffer (10x)	1.0 μ l
DreamTaq™ DNA polymerase (5 units/ μ l)	0.1 μ l
dNTP mix solution (10 mM)	0.4 μ l
ddH ₂ O	6.9 μ l
Total	9.2 μ l

Table 2.3: Standard 10 μ l PCR reaction mixture

Component (concentration)	Volume
Primer forward (5 μ M)	0.4 μ l
Primer reverse (5 μ M)	0.4 μ l
DreamTaq™ Green PCR Master Mix (2X)	5.0 μ l
ddH ₂ O	3.4 μ l
Total	9.2 μ l

Table 2.4: 10 μ l PCR reaction mixture using an integrated PCR master mix

PCR was undertaken in a thermal cycler (Mastercycler® ep gradient S, Eppendorf, Hamburg, Germany) using a touchdown PCR (TD-PCR) protocol, which employs initial annealing temperatures above the projected melting temperature (T_m). This approach helps to avoid amplification of non-specific products by incrementally lowering the initial annealing temperature (by 2°C) every 2 cycles until the required melting temperature was reached, improving specificity, sensitivity and yield over standard PCR techniques (Korbie & Mattick, 2008).

Step	Action	Temperature	Time
1	Initial denaturation	95°C	5 mins
2	Denature	95°C	30 sec
3	Anneal	$T_a + 4^\circ\text{C}$	30 sec
4	Extension	72°C	30-60s
<i>Repeat 1-4 for a total of 2 cycles</i>			
5	Denature	95°C	30 sec
6	Anneal	$T_a + 2^\circ\text{C}$	30 sec
7	Extension	72°C	30-60 sec
<i>Repeat 5-7 for a total of 2 cycles</i>			
8	Denature	95°C	30 sec
9	Anneal	T_a	30 sec
10	Extension	72°C	30-60 sec
<i>Repeat 8-10 for a total of 35-45 cycles</i>			
11	Final extension	72°C	5 mins

Table 2.5: Touchdown PCR protocol

The annealing temperature (T_a) was calculated as 2°C below the expected melting temperature (for the primer with the lowest melting temperature), as

predicted by Primer3Plus. Primer pairs were trialled on control DNA samples that were known to amplify well in previous reactions. If the PCR reaction produced weak or no product then the contents of the mix and amplification conditions were altered (see *Optimisation of PCR reaction composition and conditions*). In order to ensure that the desired DNA template was being amplified and not a contaminant, a water control was included in each reaction.

Optimisation of PCR reaction composition and conditions

Large target region

Where the size of the target region for amplification exceeded 501 base pairs (bp) in length the extension time was increased for a further 30 seconds for each for further 1-500 bp (e.g. total extension time 60 seconds for 501-1000 bp, 90 seconds for 1001-1500 bp, 120 seconds for 1501-2000 bp).

High GC content

Where template GC content was high (typically 60-80%) there was increased difficulty denaturing and separating double-stranded DNA due to the increased number of hydrogen bonds present. To overcome these adversities, the PCR reaction was supplemented with the addition of Dimethyl sulfoxide (DMSO), which is thought to improve denaturation and prevent the formation of secondary structures by binding to the major and minor grooves of the template DNA (Hardjasa *et al.*, 2010). DMSO was typically used at 10% (1 μ l DMSO replacing 1 μ l ddH₂O in the above reaction mixture), but DMSO concentrations of 5% (0.5 μ l DMSO replacing 0.5 μ l ddH₂O) and 15% (1.5 μ l DMSO replacing 1.5 μ l DMSO) were also occasionally required.

Large difference in primer melting temperatures

Where the difference in primer melting temperatures was large (>2°C difference between the two primers) then a temperature gradient optimisation was performed prior to PCR. This was also required if the initial reaction yielded no product. Briefly this involves performing PCR using positive control DNA across a range of temperatures over which the reaction was expected to work with the highest temperature that yields a single clear band being selected.

Agarose gel electrophoresis

To determine the adequacy of PCR amplification of the DNA, the resulting products underwent agarose gel electrophoresis. This is a process used to separate DNA molecules of differing molecular weights. The agarose gel forms a highly cross-linked matrix through which differently sized negatively charged DNA molecules move at different speeds towards the positively charged anode in response to an electric current. Differing concentrations of agarose produce gels with different properties: low concentration gels (0.8% to 1%) enable DNA molecules to migrate rapidly which is of benefit when amplicon sizes were large. High concentration gels (1.5 to 1.8%) provide great impedance to DNA, which allows the differentiation of small products from each other. In general, a 1% gel was used in this study unless products were very large or the differentiation of two small products of similar size was required.

1% agarose gels were made by dissolving 1.0 g molecular grade agarose powder in 100 mL 1X Lithium acetate borate (LAB) buffer (consisting of 10 mM lithium acetate and 10 mM boric acid). 1X LAB buffer (working solution) was

obtained by diluting 50X LAB buffer (stock solution) 1:50 (mixing 1 part LAB stock solution with 49 parts water). To promote the dissolution of the agarose the mixture was heated in a 700 W domestic microwave until the agarose had completely dissolved on visual inspection (45 to 55 seconds). 5 µl of 1% ethidium bromide (a DNA-binding fluorophore) was added to the solution, which was then poured into a gel tray with between one and two well combs in place. As this solution cooled at room temperature it solidified, the process taking approximately 30 to 60 minutes.

The solidified agarose gel was then placed in an electrophoresis tank containing 1X LAB buffer solution and 2 µl of PCR product were pipetted directly into each well. Into an adjacent well (one per comb row) was pipetted an appropriate molecular weight marker ("ladder") such as the GeneRuler™ 1kb DNA ladder (Thermo Fisher Scientific) corresponding to the size of expected product to allow estimation of the molecular weight of the PCR product.

A constant voltage of 150 V was passed across the gel electrophoresis tank for a period of between 20 and 60 minutes depending on the expected size of the PCR product. The gel was removed from the gel electrophoresis tank and excess buffer was removed using a paper towel. The gel was then trans-illuminated using an ultraviolet light causing fluorescence of ethidium bromide, where this was bound to DNA. A photograph was taken using a static camera mounted within a gel imaging and analysis system (InGenius, Syngene) (GeneSnap image acquisition software, Syngene).

Purification of PCR products and dideoxy sequencing

If the agarose gel showed a single clear band the remaining sample (8 µl) would be prepared for dideoxy sequencing. Prior to this it was necessary to remove unincorporated primers and deoxynucleotide triphosphates (dNTPs). This was achieved using ExoSAP, a combination of hydrolytic enzymes exonuclease I (Exo I), degrading single-stranded DNA, and shrimp alkaline phosphatase (rSAP), dephosphorylating nucleotides (Werle *et al.*, 1994). 5 µl of PCR product (of 8 µl total) was pipetted into a new 0.2 mL Eppendorf tube and 2 µl of ExoSAP was added. This was incubated using a thermocycler at 37°C for 30 minutes. Enzymes were then deactivated by a second incubation at 95°C for five minutes.

Dideoxy sequencing of PCR product which had undergone ExoSAP clean-up was performed by Source BioScience (Cambridge, UK), with resultant chromatograms visualised using an FinchTV© (Geospiza, Inc, Seattle, WA)

Droplet digital PCR (ddPCR)

ddPCR, a water-oil emulsion droplet method of performing digital PCR, was used as an orthogonal method to confirm copy number variants (CNVs - deletions and/or duplications) identified by exome or genome sequencing. It can sensitively quantify the amount of target DNA in a sample without the use of standard curves.

Primers were designed using a similar strategy to that described above [see *Polymerase Chain reaction (PCR)*] with the following specific considerations:

1. The desired amplicon size was adjusted (minimum 70 bp; optimum between 100 and 120 bp; maximum 200 bp).
2. Primers were designed to anneal at as close to 58°C as possible by setting Primer3Plus to allow melting temperatures of 59 to 61°C.
3. At least two (ideally three) primers were designed within the suspected CNV with at least one primer (ideally two primers) outside the region on either side (total 4 - 7 primer pairs).
4. Regions of high GC content were avoided wherever possible as variation in temperature and DMSO concentrations were not possible.

Stock solutions of each primer (100 µM) were created as described above (*Preparation of primers*). From these, a working solution (1 µM) of the primer pair mix was created by diluting each 1:100 in the same tube (adding 1 µl of each primer to 98 µl of water). Additionally, previously validated control primers targeting the gene *RPP30* were used as a positive control.

Sample selection

Control DNA was used in each experiment. Where variants were present on the X-chromosome, sex matched controls were selected. Additionally, for homozygous deletions ideally both parental DNA samples (expected to be heterozygous) were analysed alongside DNA from the affected individual to confirm segregation and accurate detection across the range of gene dosage abnormalities.

DNA sample preparation

Double-stranded DNA concentration was quantified using the *Quantification of double-stranded DNA using Qubit™* method described above. 200 ng of DNA was diluted with ddH₂O to make up a total volume of 50 µl, sufficient for eight 22 µl reactions (**Table 2.6**).

Component (concentration)	Volume
Primer pair mix (2pmol/ μ l)	1.1 μ l
EvaGreen Supermix	11 μ l
ddH ₂ O	4.4 μ l
Template DNA	5.5 μ l
Total	22 μ l

Table 2.6: 22 μ l ddPCR reaction mixture

Each 22 μ l reaction mixture was pipetted into one well of a 96 well plate. The plate was sealed using a PX1 plate sealer (Bio-Rad, Hercules, CA, USA) and agitated using a vortex machine before being centrifuged using a plate spinner. Droplets were generated by an automated droplet generator (Bio-Rad) using DG32 Automated Droplet Generator Cartridges (Bio-Rad) and Automated Droplet Generation Oil for EvaGreen (Bio-Rad). Following droplet generation, PCR was performed in before droplets were quantified using the QX200 Droplet Digital PCR System (Bio-Rad). Resultant data were analysed using QX manager 1.2 Standard Edition (Bio-Rad).

2.4. Next-generation sequencing

Exome sequencing

Exome sequencing was performed using either (i) an Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) at the Royal Devon and Exeter NHS Foundation Trust, (ii) an Illumina NovaSeq 6000 at University of Exeter Sequencing Service or (iii) using a DNBSEQ (BGI Tech Solutions, BGI Genomics, Hong Kong). Exon targeting used either (a) Twist Human Core Exome (Twist Bioscience, CA, USA) [Exeter Only] (b) Agilent SureSelect Whole Exome v6 (Agilent, Santa Clara, CA, USA) [Exeter and BGI]. Data was received

as fastq files either (i) via direct file transfer to the University of Exeter server [Exeter] (ii) via secure file transfer protocol [BGI] (iii) on an external hard-drive [BGI]. For all platforms a minimum read depth and coverage of 20x was achieved for 90-95% of the exome (**Appendix 7.2**).

Genome sequencing

Genome sequencing was performed by BGI on a DNBSEQ (BGI Tech Solutions, BGI Genomics, Hong Kong). Data was transferred as fastq files on an external hard drive.

2.5. Bioinformatic methods

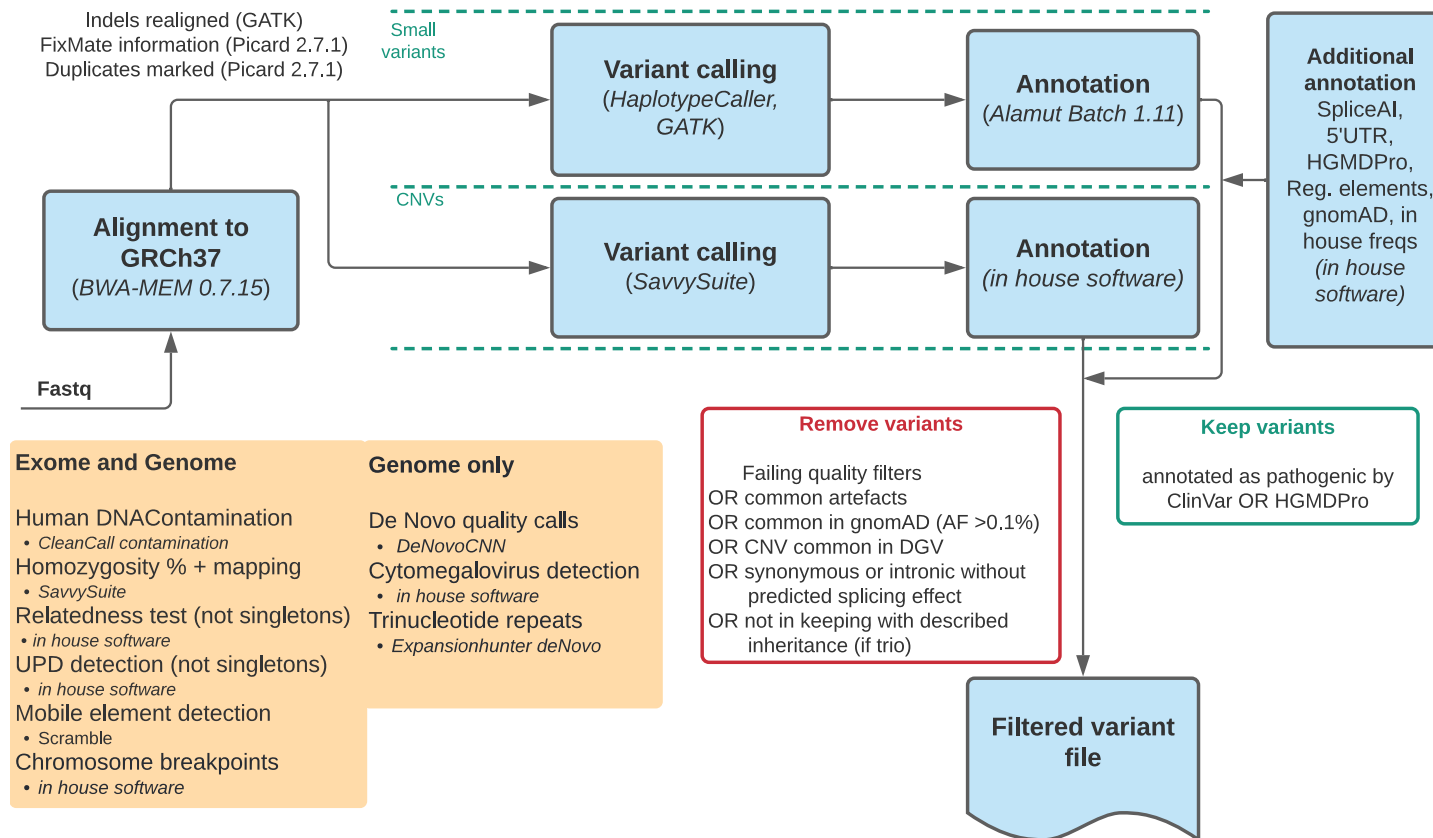


Figure 2.1: Bioinformatic pipeline for exome and genome sequencing

Filtering parameters: Maximum frequency cutoffs: 0.001, HGMD/Clinvar: 0.01, homozygous HGMD/Clinvar variants: 0.01, de novo variants: 1.0E-4
 Minimum GQ cutoff: Heterozygous variants: 49.0, homozygous variants: 15.0. Minimum CNV Bayes Factor cutoff: Recessive: 8.0, De novo: 16.0
 Minimum mapping quality: No minimum. Variants filtered by VCF filter column - SNVs must have "PASS" or "MQ40", indels may also have "QD2"

Bioinformatic pipeline for exome data

All exome data were processed using the same bioinformatic pipeline (**Figure 2.1**).

1. **Read alignment/mapping:** Reads, received in compressed fastq format (fq.gz), were aligned to GRCh37 using Burrow-Wheeler Alignment minimal exact matches (BWA-MEM) v0.7.17 (Li & Durbin, 2010) to generate Binary Alignment/Map (BAM) files.
2. **Mate information verified:** Information regarding the mate coordinates and insert size was checked using FixMateInformation (Picard Tools v2.15.0).
3. **Duplicate reads removed:** Duplicate reads were marked and removed using MarkDuplicates (Picard Tools v2.15.0).
4. **Insertion and/or deletion (InDel) realignment:** InDel realignment was performed using IndelRealigner (Genome Analysis Toolkit [GATK] v3.7.0).
5. **Calling small variants:** Single nucleotide variants (SNVs) and small InDels were called using HaplotypeCaller (GATK v3.7.0) resulting in a variant call format (VCF) file.
6. **Initial filtering:** The VCF file was filtered, excluding a list of known VCF calls containing common population variants and known artefacts to create a filtered VCF file. This initial filtering step reduces the size of the file before annotation.
7. **Annotation:** The filtered VCF file was annotated with information from:
 - Alamut® Batch (v1.10 - v1.11) including population frequencies (from gnomAD and 1000 Genomes Project), predicted protein outcomes and some *in silico* tool predictions (PolyPhen, SIFT, AlignGVGD, REVEL, MaxEntScan).
 - OMIM gene-disease associations
 - UTRannotator on 5' untranslated region (UTR) disruption (Zhang *et al.*, 2020)
 - HGMD Pro (a database of mostly disease associated variants)
 - in-house database of Exeter exomes (mostly unwell children and their unaffected parents).

8. **Calling CNVs:** CNV calling was performed using SavvyCNV (Laver *et al.*, 2022) and/or ExomeDepth (Plagnol *et al.*, 2012):
- ExomeDepth alone before mid-2018
 - SavvyCNV and ExomeDepth together from mid-2019 to mid-2018
 - SavvyCNV alone since mid-2019

ExomeDepth is a R package that relies solely on read depth data to detect CNVs (Plagnol *et al.*, 2012). SavvyCNV, available since 2018, outperforms ExomeDepth and a number of comparable tools by utilising off-target reads to detect CNVs which may or may not fall outside of exome targets (Laver *et al.*, 2022). CNVs were formatted in VCF format and annotated with the same information as the SNVs where applicable (some tools, e.g. SpliceAI, are not applicable to CNVs).

9. Additional steps

- **Detecting runs of homozygosity:** Runs of homozygosity and total homozygosity were detected from using SavvyHomozygosity (<https://github.com/rdemolgen/SavvySuite>).
- **Detecting mobile element insertion:** Mobile element insertion was detected using Scramble (Torene *et al.*, 2020)
- **Detecting contamination with human DNA:** Sample contamination with human DNA was quantified by CleanCall contamination (Flickinger *et al.*, 2015).

10. Further filtering and final output

Short variants, called by HaplotypeCaller, and CNVs, called by SavvyCNV, were combined using in-house Java-based software (Trio) and further filtered according to the following rules (**Table 2.7**).

Inheritance filters	
Family structure is a singleton	
- All variants	Retain
Family structure is a duo	
- Shared variants	Retain
- Else	Exclude
Family structure is a trio	
- Homozygous variants	Retain
- Hemizygous variants in the proband where the mother is a carrier (<i>X-linked recessive</i>)	Retain

- Multiple heterozygous variants in one gene in the proband where ≥ 1 variant has been inherited from each parent (<i>compound heterozygous</i>)	Retain
- Heterozygous or mosaic variants in the proband not present in either parent (<i>de novo</i>)	Retain
- Inherited heterozygous variants in imprinted genes	Retain
- Else	Exclude

Positive filters - retain variant if ≥ 1 apply regardless of negative filters	
Present in the HGMD Pro database Marked as 'DM' or 'DM?' and not retired	Retain
Present in ClinVar database Marked as 'Pathogenic' or 'Likely Pathogenic'	Retain

Negative filters - remove variant if ≥ 1 apply, unless positive filter applies	
Gene	
Not associated with a gene	Exclude
Gene starts with <i>MUC</i> , <i>HLA</i> , <i>LINC</i>	Exclude
Maximum allele frequency (gnomAD)	
HGMD / ClinVar variants	Exclude if >0.01
<i>De novo</i> variants	Exclude if >0.0001
All other variants	Exclude if >0.001
Predicted protein impact	
Protein change (missense, nonsense, frameshift, InDel)	Retain
Significant splicing impact predicted (MaxEntScan)	Retain
Disrupting 5' UTR	Retain
No protein or splicing impact (e.g. synonymous)	Exclude
Genotype quality	
Heterozygous variants	Exclude if <49
Homozygous variants	Exclude if <15
CNV Bayes factor	
Recessive	Exclude if <8.0
<i>De novo</i>	Exclude if <16.0
Alternate reads	
<i>De novo</i>	Exclude if <4
Flags	
Insertions or deletions	Exclude if NOT "QD2", "PASS", "MQ40"
SNVs	Exclude if NOT "PASS", "MQ40"

Table 2.7: Default rules for filtering software

Bioinformatic pipeline for genome data

The pipeline for genome data closely mirrors the bioinformatics pipeline used to process the exome data (**Figure 2.1**). Specifically, the same filtering criteria were employed. There are the following notable differences:

- Genome data were uploaded to the University of Exeter high-performance computer cluster (HPC due to the requirement for a greater number of more powerful nodes and greater storage.
- The Exeter genomes dataset (>1,000 individuals, enriched for rare and unusual forms of diabetes, sequenced by BGI and analysed using the Exeter bioinformatics pipeline) was used instead of the Exeter exomes dataset.
- Additional tools included:
 - DeNovoCNN (Khazeeva *et al.*, 2021) – used to call *de novo* variants with high confidence using a Bayesian framework
 - ExpansionHunter deNovo (Dolzhenko *et al.*, 2020) – used to detect known and novel repeat expansions genome-wide

Variant prioritisation

Manual variant prioritisation of the resultant Excel format file was carried out (**Table 2.8**).

Parameter	Criteria	Source
Inheritance pattern	Retain only those compatible with the inheritance pattern and pedigree structure.	Pedigree and clinical records
Frequency in population databases	Retain only those with a mean allele frequency <0.01 in all databases and an absence of homozygous unaffected individuals	gnomAD v2.1.1, gnomAD v3.1.2 (Karczewski <i>et al.</i> , 2020) Exeter Exome or Genome databases (<i>see below</i>) Internal control datasets for Amish and

		Palestinian individuals (see below)
Likely artefacts	<p>Exclude those with read depth <5</p> <p>Exclude multiple (>4) variant calls in the same gene with implausible allele depths.</p> <p>Exclude trinucleotide repeat expansions / contractions with allele depths not suggestive of either heterozygous or homozygous change</p> <p>Exclude variants confirmed to be artefacts by visual inspection of aligned reads</p> <p>Exclude CNV calls with read-depth ratios outside the following biologically plausible values:</p> <ul style="list-style-type: none"> - 0.0-0.1 (homozygous deletion) - 0.4-0.6 (heterozygous deletion) - 1.4-1.6 (heterozygous duplication) - 1.9-2.1 (homozygous duplication) 	Integrative Genome Viewer (IGV) (Broad institute)
Potentially compound heterozygous variants <i>in cis</i>	<p>Exclude if <i>in cis</i> (and no other variant) using the following approaches:</p> <p>1) both variants are within ~100 base pairs of each other and confirmed to be <i>in cis</i> by identifying both variants within the same read using IGV.</p> <p>2) both variants are present and exonic in gnomAD 2.1.1 and a common haplotype (implying <i>in cis</i>) was identified using the gnomAD variant-cooccurrence tool.</p>	IGV gnomAD v2.1.1, (Karczewski <i>et al.</i> , 2020)
Previously reported variants	<p>Prioritise variants annotated:</p> <ul style="list-style-type: none"> - “DM” or “DM?” - “pathogenic” or “likely pathogenic” 	HGMDPro (Stenson <i>et al.</i> , 2014) ClinVar
Disease-gene association	Prioritise variants in established/previously published candidate disease genes	OMIM (Amberger <i>et al.</i> , 2019) PubMed HGMDPro (Stenson <i>et al.</i> , 2014)
Protein effect	Prioritise the predicted loss-of-function variants (nonsense, frameshift, canonical splice site, CNV deletion)	-

<i>In silico</i> missense prediction tools	Prioritise if Rare Exome Variant Ensemble Learner (REVEL) score >0.7	REVEL (Ioannidis <i>et al.</i> , 2016)
Gene Constraint scores	<p>Prioritise if:</p> <p>Variant is a missense variant and missense Z score is >3.09</p> <p>Variant is a predicted loss-of-function variant and "loss-of-function observed/expected upper bound fraction" or (LOEUF) score < 0.35</p>	gnomAD (Lek <i>et al.</i> , 2016)
<i>In silico</i> splicing prediction tools	Prioritise if score >0.15 (D. Barelle Personal communication)	SpliceAI (Broad Institute) (Jaganathan <i>et al.</i> , 2019)

Table 2.8: Criteria for variant prioritisation

The following additional criteria were used for prioritising variants in candidate disease-genes.

- Genes with high expression in tissue of interest in the Genotype-Tissue Expression (GTEx) database (Carithers *et al.*, 2015)
- Genes with a knockout mouse model phenotype compatible with the human phenotype. Databases interrogated include the International Mouse Phenotyping Consortium (IMPC) database (Dickinson *et al.*, 2016) and the Mouse Genome Database (Blake *et al.*, 2021)
- Genes known to be within a gene-pathway associated with the human phenotype [<https://string-db.org>] (a European Life-Science Infrastructure (ELIXIR) Core Data Resource)]

All variants meeting criteria for final consideration were visualised in IGV.

Quality control (QC)

To ensure that all exome and genome sequencing data were of sufficient quality regardless of sequencing provider, QC metrics were obtained routinely during both bioinformatics pipeline (HsMetrics – Picard Tools v2.15.0) and reviewed with each new sample. At intervals it was also necessary to appraise the overall quality of sequencing obtained by each provider and for each batch, and for this a Python script was designed to analyse the data by batch (**Appendix 7.3**).

Briefly, this crawls all our available exome data, extracting QC metrics and appending batch information, eventually presenting the data in bar plots separated by batch. Example data are provided (**Appendix 7.2**). Additional custom scripts to allow the generation of virtual gene panels and to search for cryptic variants in were also developed as part of this thesis (**Appendices 7.4 - 7.7**).

Generation of an internal variant database

The next-generation sequencing data amassed by the “Windows of Hope” and “Stories of Hope” research programmes includes more than 440 processed genome and exome samples, entailing a unique resource defining the genetic variation present in the Anabaptist and Palestinian communities. An in-house database of variant frequencies for the Windows of Hope project was created.

The steps used for this included:

1. Combining all unfiltered VCF format files into a multi-sample VCF file using GATK (v3.8.0) CombineVariants (with -genotypeMergeOptions UNIQUIFY)
2. Splitting this large multi-sample VCF by chromosome

3. Importing these into Excel

Steps 2. and 3. aimed to maximise utility of this resource to those working exclusively in a Microsoft Windows environment. However, Microsoft Excel (which has a file size limit of 1,048,576 rows) placed significant limitations on this process, such that each chromosome needed to be divided into greater number of subdivisions as the number of samples grew. Over time it became clear that a new approach was needed.

The current approach is to utilise a bash script (**Appendix 7.7**) that leverages Unix grep to search across the files in real-time. This has the advantage of automatically updating as new samples are added and being flexible to allow for searches that include gene, position, reference SNP (rs), coding effect, HGMD annotated phenotype. Resultant data can also be filtered for gnomAD frequency, REVEL score or SpliceAI score using Unix awk.

Anabaptist variant server

Work performed on the in-house Anabaptist variant database was assimilated within a collaborative Anabaptist variant database called the “Anabaptist variant server”. This database contains genomic variant data from The University of Maryland School Amish Program (Maryland, USA), Clinic for Special Children (Strasburg, Pennsylvania, USA), Das Deutsch Clinic (Middlefield, Ohio, USA), National Institute of Mental Health Amish Program (Bethesda, Maryland, USA), University of Exeter Windows of Hope Project, and Regeneron Genetics Center LLC (New York, USA). This resource provides access to summary level data from thousands of exome samples performed on individuals of Anabaptist

heritage, greatly empowering the ability to interpret variants from these communities.

Structural analyses of proteins

X-ray diffraction or nuclear magnetic resonance–derived structures of human proteins of interest were sought in UniProt and the Protein Data Bank (PDB) to identify the 3-dimensional geometry surrounding pathogenic missense variants and protein/substrate interactions. Where no structures could be identified homologues with high sequencing identity (>40%) were identified using the Basic Local Alignment Search Tool-Protein. In some cases, homology models were used to create structures for proteins of interest where no crystallographic structure was available (<https://swissmodel.expasy.org>, accessed June 18, 2021). In other cases, variants were visualised on the homologue. Residues were visualised and annotated using Pymol 2.3 (Schrödinger LLC, 2019).

2.6. DDD complementary analysis project

Genomic variants in the *CAMSAP1* gene and the paralogues *CAMSAP2*, *CAMSAP3* and known interactor *MARK2* identified through the DDD project were obtained as part of a complementary analysis project application (CAP330), alongside Human phenotype ontology (HPO) terms submitted for each affected individual (**Appendix 7.8**). Between 2011 and 2015 the DDD study recruited ~13,500 children from the UK affected by severe, undiagnosed developmental disorders, >70% of whom were affected with ID / developmental delay or learning disability (The Deciphering Developmental Disorders Study,

2015; Wright *et al.*, 2018). Anonymised data including (1) a list of all variants in the above genes detected through exome sequencing in the cohort, (2) VCF files of exome sequencing for all individuals with a variant in the above genes, (3) HPO terms for all affected individuals with a variant in the above genes, was downloaded from Wellcome Sanger Genome campus servers using sFTP and stored on a controlled-access server at the University of Exeter. Individuals with either biallelic (homozygous or compound heterozygous, determined by parental inheritance) or *de novo* heterozygous potentially disease-causing variants were retained. These were defined as (1) rare, with a mean allele frequency <0.01 in gnomAD v2.1.1 and v3.1.2 with an absence of homozygous unaffected individuals, (2) protein altering (missense / stop gain / stop loss / InDel / frameshift / splice donor or acceptor). Phenotypic information was retrieved from a file containing HPO terms and manually inspected alongside the variants.

2.7. Interrogation of the 100,000 Genomes Project dataset

The 100,000 Genomes Project rare disease dataset v13 (30/09/2021 research-help.genomicsengland.co.uk/pages/viewpage.action?pagelId=45024632) was interrogated to identify additional individuals affected by the novel disorders identified as part of this PhD project, in order to robustly confirm the disease-gene association and more fully explore the genetic and phenotypic spectrum of each disorder. These data include genome sequence data (in VCF format) for all rare disease patients and their recruited family members (73,700 genomes in total). In addition, there is anonymised primary clinical (phenotype) data coded in HPO format and secondary data from Hospital episode statistics (HES) and

Patient reported outcome measures (PROMs) for all rare disease probands.

These data are stored on the Genomics England HPC “Helix”. They are available to authorised researchers through the Genomics England research portal, accessed via the virtual desktop provided by Inuvika (Toronto, Canada).

Authorisation to access this portal was obtained through the Enhanced Interpretation GECIP (Genomics England Clinical Interpretation Partnership), and our approved project “Novel insights into rare inherited disorders” (RR349)

(**Appendix 7.9**). Routine datasets provided by Genomics England do not support searches to identify all individuals across the dataset with biallelic variants in a specified gene. To facilitate the identification of such patients a new pipeline was required. This comprised two scripts run sequentially “biallelic.sh” and “biallelic.py” (**Appendices 7.10 & 7.11**). biallelic.sh is written in bash and awk to prioritise speed in an enormous dataset. The pipeline utilises an aggregated VCF file of all Genome Reference Consortium human genome build 38 (GRCh38) genomes generated by Genomics England “aggV2”, which is split into more than a hundred partitions. Additionally, the associated variant effect predictor (VEP) annotations and an Ensemble gene coordinates file are used as inputs to perform the following tasks in sequence:

1. Define the start and finish positions (“the locus”) of the given gene using the gene coordinates file provided by Genomics England
2. Identify the aggV2 partition containing the gene
3. Use BCFtools (SAMtools, Genome Research Ltd) to extract the locus
4. Merging the aggV2 variant calls with their VEP annotations
5. Filters the resulting calls to retain variants with an expected significant effect on the protein (missense / stop gain / stop loss / InDel / frameshift / splice donor or acceptor)

The second script, `biallelic.py` is written in Python, specifically utilises the pandas library for more nuanced filtering and annotation of variants. The script takes the output files from `biallelic.sh` and performs the following tasks in sequence:

1. Genotypes (e.g. 0/0, 0/1, 1/1) stored as strings are converted to float variables (“scores”) where 1 represents a heterozygous genotype and 2 represents a homozygous variant genotype.
2. Variants are filtered based on frequency in the 100,000 Genomes database excluding those with 400 or more alleles (Minor allele frequency [MAF] ~ 0.003). This threshold was arrived at after iterative testing.
3. Genotype scores are summed across each individual (an individual with two different heterozygous variants and an individual with a single homozygous variant would both receive scores of 2)
4. Individuals with a score < 2 are removed
5. Variants that no longer have any heterozygous or homozygous individuals are removed
6. Demographic and phenotype information relevant to the individual is imported from LabKey and appended

2.8. References

- Amberger JS, Bocchini CA, Scott AF, & Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* 2019;47(D1):D1038-d1043.
- Blake JA, Baldarelli R, Kadin JA, Richardson JE, Smith CL, & Bult CJ. Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res* 2021;49(D1):D981-d987.
- Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank* 2015;13(5):311-319.
- Cole TJ, Freeman JV, & Preece MA. British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Stat Med.* 1998;17(4):407-429.
- Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, Meehan TF, Weninger WJ, Westerberg H, Adissu H, *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* 2016;537(7621):508-514.
- Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, van Vugt JJFA, Nguyen C, Narzisi G, Gainullin VG, Gross AM, *et al.* ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol* 2020;21(1):102.
- Flickinger M, Jun G, Abecasis GR, Boehnke M, & Kang HM. Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. *Am J Hum Genet* 2015;97(2):284-290.
- Hardjasa A, Ling M, Ma K, & Yu H. (2010). *Investigating the Effects of DMSO on PCR Fidelity Using a Restriction Digest-Based Method.*
- Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* 2016;99(4):877-885.
- Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 2019;176(3):535-548.e524.
- Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* 2002;12(4):656-664.
- Khazeeva G, Sablauskas K, van der Sanden B, Steyaert W, Kwint M, Rots D, Hinne M, van Gerven M, Yntema H, Vissers L, *et al.* DeNovoCNN: A deep learning approach to de novo variant calling in next generation sequencing data. *bioRxiv* 2021:2021.2009.2020.461072.
- Korbie DJ, & Mattick JS. Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nature Protocols* 2008;3(9):1452-1456.
- Laver TW, De Franco E, Johnson MB, Patel KA, Ellard S, Weedon MN, Flanagan SE, & Wakeling MN. SavvyCNV: Genome-wide CNV calling from off-target reads. *PLOS Computational Biology* 2022;18(3):e1009940.
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536(7616):285-291.

- Li H, & Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26(5):589-595.
- Pan H, & Cole TJ. (2011). LMS growth, a Microsoft Excel add-in to access growth references based on the LMSmethod. Version 2.77. Retrieved from <http://www.healthforallchildren.co.uk/>
- Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 2012;28(21):2747-2754.
- Sobreira N, Schiettecatte F, Valle D, & Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat* 2015;36(10):928-930.
- Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, & Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 2014;133:1-9.
- The Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 2015;519(7542):223-228.
- Thiele H, & Nürnberg P. HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics* 2005;21(8):1730-1732.
- Torene RI, Galens K, Liu S, Arvai K, Borroto C, Scuffins J, Zhang Z, Friedman B, Sroka H, Heeley J, *et al.* Mobile element insertion detection in 89,874 clinical exomes. *Genet Med* 2020;22(5):974-978.
- Werle E, Schneider C, Renner M, Völker M, & Fiehn W. Convenient single-step, one tube purification of PCR products for direct sequencing. *Nucleic Acids Res* 1994;22(20):4354-4355.
- Wright CF, McRae JF, Clayton S, Gallone G, Aitken S, FitzGerald TW, Jones P, Prigmore E, Rajan D, Lord J, *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med* 2018;20(10):1216-1223.
- Zhang X, Wakeling M, Ware J, & Whiffin N. Annotating high-impact 5'untranslated region variants with the UTRannotator. *Bioinformatics* 2020;37(8):1171-1173.

3

Biallelic *CAMSAP1* variants cause a clinically recognizable neuronal migration disorder.

Reham Khalaf-Nazzal*, James Fasham*, Katherine A. Inskeep*, Lauren E. Blizzard, Joseph S. Leslie, Matthew N. Wakeling, Nishanka Ubeyratna, Tadahiro Mitani, Jennifer L. Griffith, Wisam Baker, Fida' Al-Hijawi, Karen C. Keough, Alper Gezdirici, Loren Pena, Christine G. Spaeth, Peter D. Turnpenny, Joseph R. Walsh, Randal Ray, Amber Neilson, Evguenia Kouranova, Xiaoxia Cui, David T. Curiel, Davut Pehlivan, Zeynep Coban Akdemir, Jennifer E. Posey, James R. Lupski, William B. Dobyns, Rolf W. Stottmann#, Andrew H. Crosby#, Emma L. Baple#

Am J Hum Genet 2022;109(11):2068-2079

3.1. Acknowledgements of co-authors and contributions to the paper

This study was undertaken as part of the “Stories of Hope” Palestinian translational genomic research programme, conceived, designed, and led by Dr Reham Khalaf-Nazzal (Arab American University), Prof Peter Turnpenny, Prof Emma Baple and Prof Andrew Crosby. Where specific experiments or analyses were performed by study collaborators or members of my supervisors’ group other than myself, then these are detailed below.

Clinical and genomic studies

Clinical data were obtained by local clinical care providers using a standardised proforma that I designed. Phenotype data for Family 1 was provided by Dr Khalaf-Nazzal. Support with interpretation of clinical data was provided by Prof Baple. Prof William Dobyns (University of Minnesota) reviewed available neuroradiological data and assisted me with interpretation of the findings.

I performed analysis of exome data from Family 1. This was also undertaken in parallel by Dr Khalaf-Nazzal. Technical assistance with validation and cosegregation studies for the *CAMSAP1* variants identified was provided by Mr Joseph Leslie and Mr Nishanka Ubeyratna (University of Exeter).

Exome sequencing data from Family 2 was initially analysed by GeneDx as part of diagnostic testing, I reanalysed the data using the same bioinformatics pipeline used to analyse the data from Family 1.

Analysis of exome data from families 3,4 and 5 was undertaken by our collaborators. I reviewed all of the *CAMSAP1* variants identified and any genomic variants that could not be excluded through those analyses.

Cell and mouse studies

iPSCs were generated by Evguenia Kouranova (Cincinnati Children's Pluripotent Stem Cell Facility).

Cell and mouse studies were carried out in collaboration with Prof Rolf Stottmann (Cincinnati Children's Hospital Medical Center). Katherine Inskeep performed all of the iPSC studies and Lauren Blizzard performed the mouse and RNAScope studies. I assisted with analysis of the results of these studies.

Manuscript draft and revision

I wrote and revised the manuscript with Prof Baple, Prof Crosby and Prof Stottmann. All the co-authors provided comments and feedback prior to submission. Figure 3.3 was produced by Katherine Inskeep, Figure 3.4 and figures 3.S6 - 3.S8 were produced by Lauren Blizzard.

Further work

I extended the work described in the accepted manuscript by identifying candidate genomic variants in *MARK2* and other *CAMSAP* genes as potential causes of neuronal migration disorders in data sets from the 100,000 Genomes Project, DDD study, GeneMatcher and other collaborating genomic research groups.

3.2. Manuscript

Abstract

Non-centrosomal microtubules are essential cytoskeletal filaments, important for neurite formation, axonal transport and migration. They require stabilisation by microtubule minus-end-targeting proteins including the CAMSAP family of molecules. Using exome sequencing on samples from five unrelated families, we demonstrate that biallelic *CAMSAP1* loss-of-function variants cause a clinically recognizable, syndromic neuronal migration disorder. The cardinal clinical features of the syndrome include a characteristic craniofacial appearance, primary microcephaly, severe neurodevelopmental delay, cortical visual impairment and seizures. The neuroradiological phenotype comprises a highly recognizable combination of classic lissencephaly with a posterior more severe than anterior gradient similar to *PAFAH1B1(LIS1)*-related lissencephaly and severe hypoplasia or absence of the corpus callosum, dysplasia of the basal ganglia, hippocampus and midbrain, and cerebellar hypodysplasia, similar to the tubulinopathies, a group of monogenic tubulin-associated disorders of cortical dysgenesis. Neural cell rosette lineages derived from affected individuals displayed findings consistent with these phenotypes including abnormal morphology, decreased cell proliferation and neuronal differentiation. *Camsap1* null mice displayed increased perinatal mortality and RNAScope studies identified high expression levels in the brain throughout neurogenesis and in facial structures, consistent with the mouse and human neurodevelopmental and craniofacial phenotypes. Together our findings confirm a fundamental role of CAMSAP1 in neuronal migration and brain development

and define it as the cause of an autosomal recessive neurodevelopmental disorder in humans and mice.

Neuronal migration disorders arise from defects in the locomotion of neurons in the prenatal developing brain, resulting in early onset developmental impairment and seizures (Oegema *et al.* 2020; Severino *et al.* 2020). These conditions are characterised by neuroradiological and histopathological abnormalities of cortical layering, absence of normal folding (lissencephaly), aCC and hypo/dysgenesis of the cerebellum. A number of monogenic causes have been described, almost all of which impact the formation or functioning of microtubules (Di Donato *et al.* 2017; Fry, Cushion and Pilz 2014). Owing to the severity of the neurological phenotype most cases are sporadic, resulting from *de novo* heterozygous loss-of-function variants. In published cohorts, heterozygous pathogenic variants in *PAFAH1B1* (platelet-activating factor acetylhydrolase, isoform 1b, alpha subunit, formerly *LIS1*), which interacts with the microtubule motor cytoplasmic dynein (Smith *et al.* 2000), account for more than a third of affected individuals (Di Donato *et al.* 2018) and hemizygous or heterozygous pathogenic variants in doublecortin (*DCX*), important for microtubule stabilisation and inhibition of neurite outgrowth (Caspi *et al.* 2000), a further quarter (Di Donato *et al.* 2018). Variants in several other microtubule interacting molecules including dynein, cytoplasmic 1, heavy chain 1 (*DYNC1H1*) as well as the alpha (*TUBA1A*) [MIM: 611603], beta (*TUBB2A*, *TUBB2B*, *TUBB3*, *TUBB4A*, *TUBB*) [MIM: 615763, 610031, 614039, 615771] and gamma-tubulin (*TUBG1*) [MIM: 615412] subunits are also well recognised genetic causes of neuronal migration and brain malformations disorders (Fry, Cushion & Pilz 2014; Kumar *et al.* 2010). Those cases caused by pathogenic variants in tubulin subunits are collectively termed the “tubulinopathies” (Desikan & Barkovich 2016) and display distinctive neuroradiological features,

which in addition to the cortical layering abnormalities include hypoplasia/aplasia of the corpus callosum, hypoplasia of the oculomotor and optic nerves, cerebellar hypodysplasia (including foliar dysplasia) and dysmorphism of the basal ganglia and hind-brain structures (Romaniello *et al.* 2018).

The CAMSAP (calmodulin-regulated spectrin-associated protein) family contains three human proteins (CAMSAP1, CAMSAP2 [CAMSAP1L1 / KIAA1078] and CAMSAP3 [KIAA1543]) essential for the formation and maintenance of the minus-end of non-centrosomal microtubules (Hendershott *et al.* 2018). At least one orthologue exists in all eumetazoans, including the well-described *Patronin* in *Drosophila*, indicating a fundamental requirement for at least one CAMSAP molecule in mitotic processes (Baines *et al.* 2009; Pavlova *et al.* 2019). CAMSAP proteins are defined by a highly conserved 'CKK' (CAMSAP1, KIAA1078/CAMSAP2, KIAA1543/CAMSAP3) domain at the C-terminus (Baines *et al.* 2009) with an additional 5' calponin homology (CH) domain and intermediate coiled-coil (CC) motif also invariably present (Akhmanova & Hoogenraad 2015). The CKK domain, predicted to adopt a β -barrel conformation with a single invariant tryptophan residue within its core, directs each CAMSAP protein to the microtubule minus end. CAMSAP1 binds transiently to the outermost microtubule ends (King *et al.* 2014), stabilizing them without affecting tubulin incorporation rate *in vitro* (Atherton *et al.* 2017; Hendershott and Vale 2014; Jiang *et al.* 2014). Conversely, CAMSAP2 and CAMSAP3 remain bound to and decorate the microtubule lattice, exerting a stabilizing effect and reducing tubulin incorporation rates (Hendershott & Vale 2014; Jiang *et al.* 2014; Yau *et al.* 2014). A recently described *Camsap1*

knockout mouse model (*Camsap1*^{-/-}) manifested preweaning lethality with epileptic seizures and abnormal cortical lamination, highly suggestive of impaired neuronal migration (Zhou *et al.* 2020). Neurons from *Camsap1*^{-/-} mice displayed abnormal polarisation resulting in a multi-axon phenotype. To date however, no human monogenic disorder has been associated with variants in any of the *CAMSAP* genes. Here we define biallelic *CAMSAP1* gene variants as a cause of a clinically and radiologically distinct syndromic neuronal migration disorder.

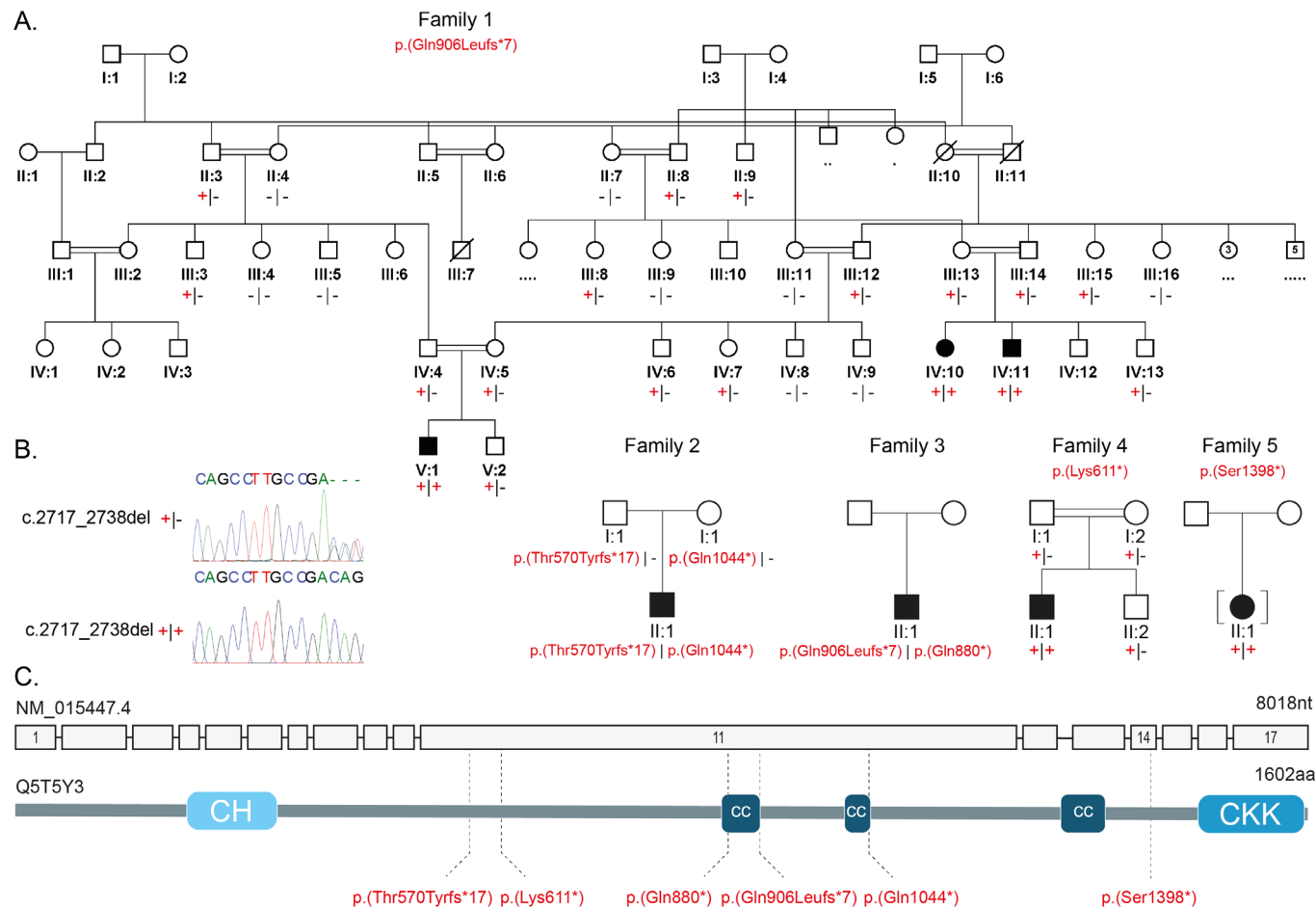


Figure 3.1: Family pedigrees and biallelic *CAMSAP1* variants associated with a syndromic neuronal migration disorder

A: Simplified pedigrees of families in the study showing cosegregation of the variants identified ('-' wild type allele; '+' familial variant), with reference to transcript NM_015447.4. **B:** chromatogram for the c.2717_2738 deletion is shown with heterozygous (top) and homozygous variant (bottom) individuals shown. **C:** Intron/exon genomic organisation of *CAMSAP1* (top) and protein domain architecture of *CAMSAP1* (bottom) illustrating the calponin homology (CH), coiled coil (CC) and calmodulin-regulated spectrin-associated CKK (*CAMSAP1*, *KIAA1078/CAMSAP2*, *KIAA1543/CAMSAP3*) domain) domains alongside the location of each of the identified pathogenic variants (dotted line).

We initially identified an extended Palestinian kinship comprising of two interlinking nuclear families with three children aged between 4 months to 3 years 9 months (Family 1, V:1, 1V:10 and IV:11, pedigree is shown in **Fig. 3.1**), affected by a syndromic neuronal migration disorder (recruited with informed consent and Palestinian Health Research Council PHRC/51819 ethical approval). The three children presented with severe primary microcephaly (-4.8 to -6.4 SDS), profound global developmental delay (GDD) and craniofacial dysmorphism including large ears, high palate, metopic ridging, a flat wide nasal bridge (**Fig. 3.S1A-B** Individual IV:10; **Fig. 3.S1C-D** Individual IV:11). Neurological examination findings included central hypotonia and peripheral hypertonia with brisk reflexes and positive Babinski sign bilaterally. All three affected individuals developed seizures at an early age, which have been refractory to treatment, and the two older children (V:1, IV:10) have a diagnosis of cortical visual impairment. Magnetic resonance imaging (MRI) neuroimaging findings in all three children were consistent and include agyria/severe pachygyria with a posterior greater than anterior gradient, dysmorphic basal ganglia and aCC (**Fig. 3.2A-D** for Individual V:1, **Fig. 3.S2A-D** for Individual IV:11).

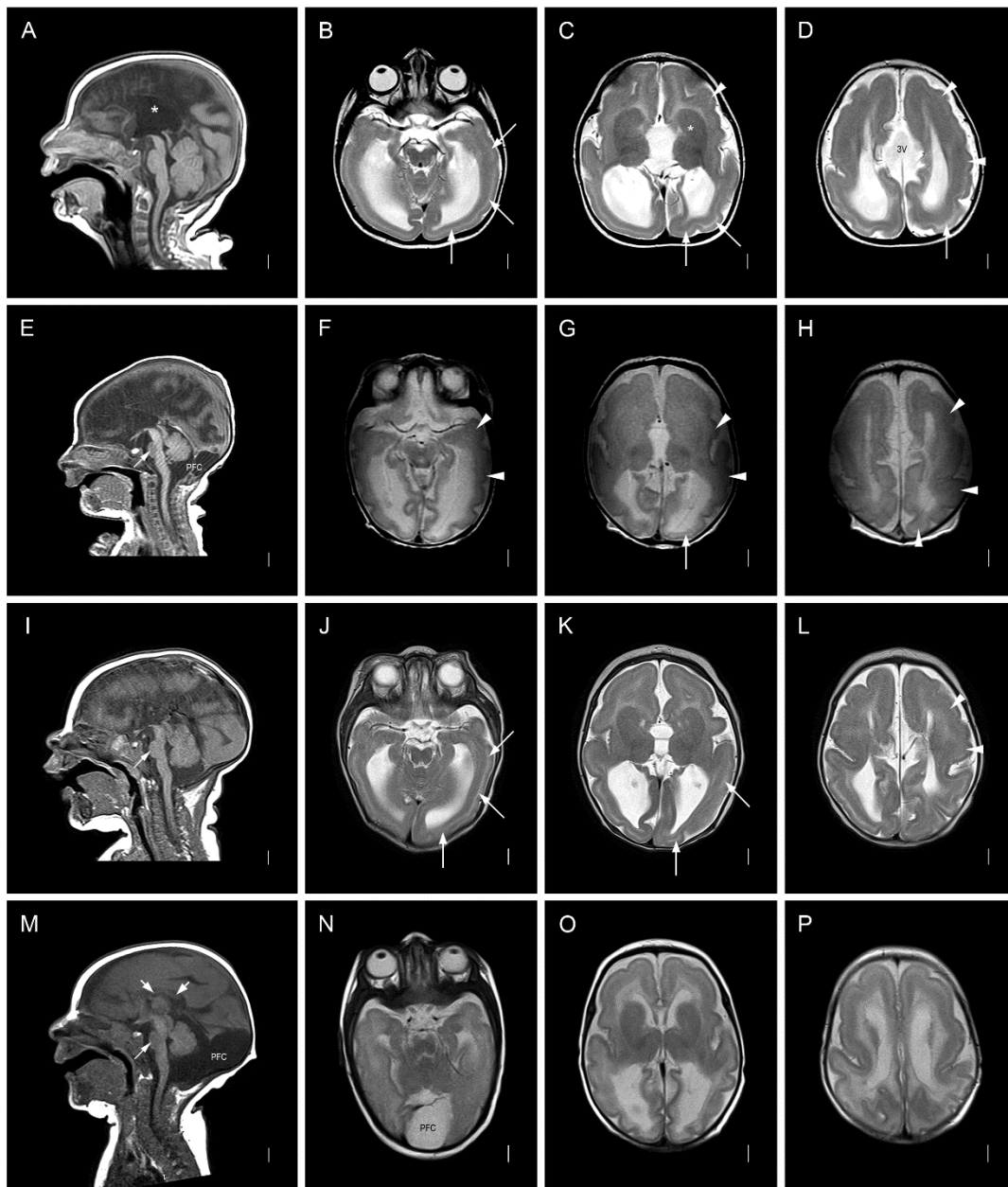


Figure 3.2: Neuroimaging in 4 individuals with the *CAMSAP1*-related neuronal migration disorder

Row 1 (A-D) is **Family 1, V:1** aged seven months Row 2 (E-H) is **Family 2, II:1** aged two days
 Row 3 (I-L) is **Family 3, II:1** aged three months Row 4 (M-P) is **Family 4, II:1** aged three months.
 T1-weighted midline sagittal images show absent (asterisk in A, also E, I) or short and thin (short white arrows in M) corpus callosum and small base of the pons (thin white arrow in E, I, M). Enlarged posterior fossa or "mega-cisterna magna" (PFC in E, M) was seen in 2/4 subjects. T2-weighted axial images show posterior more severe than anterior gradient with areas of agyria or severe pachygyria with prominent cell sparse zones and reduced thickness of the cerebral mantle/wall in posterior regions (white arrows in B-D, G, J-K) and areas of less severe pachygyria with thicker cerebral mantle/wall in anterior regions (white arrowheads in C-D, F-H, L). The gradient in Family 4, II:1 (N-P) was less clear with lower resolution images. The boundaries of the basal ganglia and thalami were difficult to see and the internal capsule are not seen (* in C, also in G, K, O). The 3rd ventricle was enlarged in all, and dramatically enlarged into a midline interhemispheric cyst in Family 1, V:1 (3V in D).

To define the genetic cause of disease, exome sequencing was undertaken using DNA from individuals IV:10 and V:1 (Illumina HiSeq and Twist Human Core Exome Kit), assuming homozygosity for a founder variant, although also considering other inheritance mechanisms. SNVs and InDels were detected using GATK HaplotypeCaller and annotated using Alamut batch (v1.8). CNVs were detected using both SavvyCNV (Laver *et al.* 2022) and ExomeDepth (<https://github.com/vplagnol/ExomeDepth>). Variants failing quality filters or present at a frequency of >0.1% or with >1 homozygous individual in gnomAD (v2.1.1 or v3.1.1) or in our in-house database were excluded. Homozygous and compound heterozygous variants common to both affected individuals and present in exons or within ± 6 nucleotides in the intron were then evaluated (a full description of the exome bioinformatic pipeline is included in the **Supplemental Text**). A single standout candidate cause of the phenotype was identified, a homozygous variant common to both individuals (Chr9(GRCh38): g.135821923_135821944del NM_015447.4(CAMSAP1): c.2717_2738del p.(Gln906Leufs*7), located within a ~8Mb region of homozygosity shared between IV:10 and V:1 (Chr9(GRCh38):g.130299324 to the 9q terminus) (**Fig. 3.S3**). The variant was confirmed by dideoxy sequencing (**Fig. 3.S4**) and found to cosegregate as expected for an autosomal recessive disease in the third affected child (IV:11), parents and three unaffected siblings. The variant, located in exon 11/17, is predicted to cause a frameshift and premature termination, likely resulting in nonsense-mediated decay and biallelic loss of function.

Through international collaboration (including GeneMatcher) (Sobreira *et al.* 2015), we identified four additional affected children from four unrelated

families, in whom exome sequencing undertaken on Illumina platforms identified biallelic predicted loss-of-function *CAMSAP1* variants. These individuals (aged one – six years) presented with clinical and neuroradiological features overlapping those of the Palestinian children. Clinical data were obtained with informed consent by local clinicians using a standardised proforma. Clinical findings on the seven affected individuals are summarised in **Table 3.1**. The pedigrees and clinical photos depicting morphological features are shown in **Fig. 3.1** and **Fig. 3.S1**. Detailed clinical descriptions of all individuals are provided in the **Supplemental Data**. Retrospective analysis of neuroimaging findings (5/5 where complete data were available) was undertaken by author WBD and is shown in **Fig. 3.2** and **Fig. 3.S2**. The *CAMSAP1* variants and exome variants identified in each family in this study are listed in **Table 3.S1** and **Table 3.S2** respectively.

INDIVIDUAL	Family 1, V:1	Family 1, IV:10	Family 1, IV:11	Family 2, II:1	Family 3, II:1	Family 4, II:1	Family 5, II:1
CAMSAP1 variants (NM_015447.4)	Homozygous c.2717_2738del p.(Gln906Leufs*7)	Homozygous c.2717_2738del p.(Gln906Leufs*7)	Homozygous c.2717_2738del p.(Gln906Leufs*7)	c.1707dupT p.(Thr570Tyrfs*17) c.3130C>T p.(Gln1044*)	c.2717_2738del p.(Gln906Leufs*7) c.2638C>T p.(Gln880*)	Homozygous c.1831A>T p.(Lys611*)	Homozygous c.4193C>G p.(Ser1398*)
Region	Palestine	Palestine	Palestine	North America	North America	Turkey	North America
Sex	M	F	M	M	M	M	F
GROWTH							
Measurement age (y)	3y	3.8y	0.3y	3.6y	1.1y	6.4y	4.8y
Birth OFC cm (SDS)	Microcephaly	Microcephaly (1m)	Microcephaly	Microcephaly	Microcephaly (4m)	32 (-2.5)	31.8 (-2.2)
Height cm (SDS)	NK	NK	NK	99.1cm (-0.2)	66cm (-4.2)	125cm (+1.3)	106.5cm (-0.1)
Weight kg (SDS)	NK	NK	NK	13.5kg (-1.4)	7.7kg (-14.5)	25kg (+1.0)	17.3kg (-0.2)
OFC cm (SDS)	42.7cm (-4.8)	40.7cm (-6.4)	Microcephaly	45cm (-4.8)	42cm (-5.0)	53cm (-0.2)	42.5cm (-7.1)*
NEUROLOGY							
Age at assessment (y)	3y	3.8y	0.3y	5y	1.7y	5.5y	4.8y
Global Dev. Delay	Profound	Profound	Profound	Severe	Severe	Severe	Severe
Central tone	↓	↓	↓	↓	↓	↓ (severe)	↓
Peripheral tone	↑	↑	NK	↑	↓	↑	↑
Deep tendon reflexes	+++	+++	NK	++; brisk	++	+++	+++
Plantar reflexes	↑	↑	NK	NK	absent	NK	NK
Seizures/age of onset	from 1m	-	from 2m	from 4-5m	from 5m	from 5m	from 2m
EEG findings	NK	NK	NK	Mod. hypsarrhythmia	hypsarrhythmia	burst suppression	Multifocal discharges
Cortical visual impairment	+	+	NK	+	+	+	+
Feeding difficulties	+	+	+	+	gastrostomy	+	gastrostomy
FACIAL FEATURES							
Prominent metopic suture	+	+	+	+	+	-	-
Wide nasal bridge	+	+	+	+	+	+	+
Pronounced cupid's bow	+	+	+	+	+	-	+
Large prominent ears	+	+	+	+	+	+	-
High arched palate	+	+	NK	-	NK	+	+
NEUROIMAGING							
Lissencephaly/Pachygyria	+	+	+	+	+	+	+
aCC / severe hCC	+	+	+	+	+	+	+
Dysplastic basal ganglia	+	+	+	+	+	+	+
Enlarged posterior fossa	-	-	-	+	-	+	-
Cerebellar hypoplasia	mild	mild	+	+	+	+	mild
OTHER CLINICAL FEATURES	Hyperopia & astigmatism	-	-	Cryptorchidism	-	Cryptorchidism, Femoral hernia	Deceased age 5.5y

Table 3.1: Summary of clinical and neurological features of individuals with CAMSAP1-related neuronal migration disorder

Abbreviations: +, feature is present; -, feature is absent; ↑, increased; ↓, decreased; +++, hyperactive; aCC, Agenesis of the corpus callosum; cm, centimetres; Dev., Developmental; DWS, Dandy-Walker syndrome; F, female; hCC, hypogenesis of the corpus callosum; m, months; M, male; NK, not known; OFC, Occipitofrontal circumference; SDS, Standard deviations; y, years. * OFC at 3.8y

Family 2, II:2 is the eldest child of unaffected, unrelated North American parents of North European ancestry. Antenatal brain imaging showed lissencephaly, aCC and a small cerebellum. He was hypotonic at birth and exhibited early feeding difficulties. From four months he was affected by infantile spasms and tonic seizures and his EEG showed a modified hypsarrhythmia pattern. At age 5 years, he has severe GDD, is unable to sit unsupported and non-verbal but uses a gaze tracking device to indicate his needs. Neurological findings include axial hypotonia, peripheral hypertonia and cortical visual impairment. He has a similar facial gestalt (prominent metopic suture, wide nasal bridge and prominent cupid's bow) to the affected Palestinian children (**Fig. 3.S1E-F**). MRI revealed diffuse severe pachygyria with a "posterior more severe than anterior" (P>A) gradient, dysmorphic basal ganglia, absent corpus callosum and enlarged posterior fossa or "mega-cisterna magna" (**Fig. 3.2E-H**). He underwent diagnostic trio exome sequencing through GeneDx (U.S.A), which identified *in trans* compound heterozygous predicted loss-of-function *CAMSAP1* variants [Chr9(GRCh38):g.135822954dupA NM_015447.4:c.1707dupT p.(Thr570Tyrfs*17) and Chr9(GRCh38):g.135821531G>A NM_015447.4:c.3130C>T p.(Gln1044*)]. To exclude other potential causes of disease, the exome data were then

reanalysed using the same bioinformatic pipeline and filtering strategy applied to the exome data from Family 1 (variant list – **Table 3.S2**).

Family 3, II:1, a 20-month-old male child born to North American parents of Northern European ancestry, was noted to be microcephalic and display abnormal movements in early infancy. His EEG showed hypsarrhythmia and he was diagnosed with infantile spasms at 5 months of age. His seizures have progressed, requiring polytherapy for effective control. He has severe GDD (crawling, non-verbal, reaching for objects), cortical visual impairment and generalised hypotonia. He has a history of feeding difficulties requiring nasogastric feeding and parenteral gastrostomy (PEG) placement for an unsafe swallow. His craniofacial features are similar to those of the other affected children. (**Fig. 3.S1F**). Brain MRI (**Fig. 3.2I-L**) revealed pachygyria with thicker cerebral mantle anteriorly, enlarged 3rd ventricle and dysmorphic basal ganglia and thalami with internal capsule not seen. Singleton diagnostic exome sequencing undertaken through GeneDx (U.S.A), identified compound heterozygous *in trans* predicted loss-of-function variants in *CAMSAP1*, including the same frameshift variant identified in Family 1 [Chr9(GRCh38): g.135821923_135821944del NM_015447.4:c.2717_2738del p.(Gln906Leufs*7)] and a novel nonsense variant Chr9(GRCh38):g.135822023G>A NM_015447.4:c.2638C>T p.(Gln880*).

Family 4, II:1 is a six-year-old child of related Turkish parents, investigated as part of a large cohort study to define candidate genetic causes of neurodevelopmental disorders (Mitani *et al.* 2021). He was found to be microcephalic at birth (-2.5 SDS) and was diagnosed with infantile spasms aged

5 months, before going on to develop multiple other seizure types. His EEG showed a burst-suppression pattern. He has mild craniofacial dysmorphism, profound GDD, central hypotonia, limb spasticity, epilepsy and cortical visual impairment. Limb movements are described as dyskinetic with varying spasticity of his limbs and intermittent guarded rigidity. MRI revealed diffuse lissencephaly, dysmorphic basal ganglia, a thin corpus callosum and enlarged posterior fossa or “mega cisterna magna” (**Fig. 3.2M-P**). Trio exome sequencing performed at Baylor College of Medicine (U.S.A), as previously described (Mitani *et al.* 2021), identified a novel homozygous candidate nonsense variant, Chr9(GRCh38):g.135822830T>A; NM_015447.4:c.1831A>T; p.(Lys611*), located within a 4.2Mb region of homozygosity.

Family 5, II:1 is a four-year-old adopted North American child, with profound GDD, generalised seizures and severe microcephaly (-7.1 SDS). Neurological examination revealed central hypotonia with bilateral lower limb spasticity and dystonic movements, with craniofacial features including a wide nasal bridge and pronounced cupid's bow (**Fig. 3.S1H**). EEG findings of multifocal epileptiform discharges were consistent with electroclinical seizures that appeared to lateralise to either hemisphere. MRI findings include holohemispheric bilateral lissencephaly and grey matter band heterotopia with notable white matter volume loss, a prominent cisterna magna and diffusely small brainstem with decreased volume of the dorsal pons (**Fig. 3.S2E-F**). Proband-only exome sequencing performed at Cincinnati Children's Hospital Medical Center (CCHMC) using previously described methodology (Liegel *et al.* 2019) identified a homozygous *CAMSAP1* variant [Chr9:g.135818055G>C NM_015447.4:c.4193C>G p.(Ser1398*)].

All the *CAMSAP1* variants identified as part of this study were predicted to result in nonsense mediated mRNA decay and loss of function. The variants were predominantly (5/6) located in the largest exon of the *CAMSAP1* gene (exon 11/17, **Fig. 3.1C**) and are absent from gnomAD v2.1.1 and v3.1.1, with the exception of p.(Gln906Leufs*7), present in two unrelated families in this study and one additional heterozygous Finnish individual in gnomAD (**Table 3.S1**). Further inspection of the genomic architecture surrounding this variant suggests its recurrent nature may result from a homologous recombination event (**Fig. 3.S5**). Furthermore, there are no homozygous loss-of-function *CAMSAP1* variants listed in publicly accessible genomic databases. The exome filtering steps followed in Families 3-5 were similar to those described for Families 1 and 2 and are detailed in the **Supplemental Data**. In all families the *CAMSAP1* variants cosegregated as expected for an autosomal recessive trait (**Fig. 3.1**). In Family 3, the closely collocated compound heterozygous variants in *CAMSAP1* p.(Gln906Leufs*7) and p.(Gln880*) could be determined to be *in trans* by phasing the short read exome sequencing data.

We next investigated the functional consequences of the *CAMSAP1* variants in iPSCs. Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood from an affected individual (Family 2, II:1), cultured to enrich erythroid progenitor cells, which were transduced with a Sendai viral cocktail that expresses the Yamanaka factors, Klf4, cMyc, Oct4 and Sox2, to obtain iPSCs. The *CAMSAP1* genotype of the iPSC line was confirmed by next-generation sequencing. These iPSCs were used alongside an iPSC wildtype control cell line (iPSC72.3, CCHMC Pluripotent Stem Cell Facility) to generate neural rosettes: radially organised two-dimensional structures of neural progenitor cells

(NPCs) and differentiating neurons. We analysed the neural rosettes at both eight days *in vitro*, when they are entirely composed of progenitor cells, and eleven days *in vitro*, when the progenitors begin to produce differentiating neurons along the outer edge of each rosette. The morphology of rosettes derived from the affected child (Family 2, II:2) were abnormal, showing large clusters of cells improperly collecting in the centre of each rosette (**Fig. 3.3A-B**). We stained rosettes for PAX6, a marker of neuronal progenitors, and TUJ1 which marks differentiated neurons. These data demonstrated significantly fewer differentiated neurons present at 11 days *in vitro* in rosettes from the affected individual (**Fig. 3.3C-H**). Additionally, a proliferation defect was evident by the reduced number of pHH3+ cells present at both 8 and 11 days (**Fig. 3.3I-J**). Finally, cleaved caspase-3 (CC3) is upregulated in disease-associated rosettes, indicating an increased level of cell death which may explain the presence of the dense cell clusters previously noted (**Fig. 3.3K-N**). Overall, iPSCs and neural rosettes derived from the affected child with the *CAMSAP1*-related neuronal migration disorder show increased apoptosis, and decreased proliferation and differentiation of neuronal progenitor cells, consistent with the neuronal migration defects observed in affected individuals.

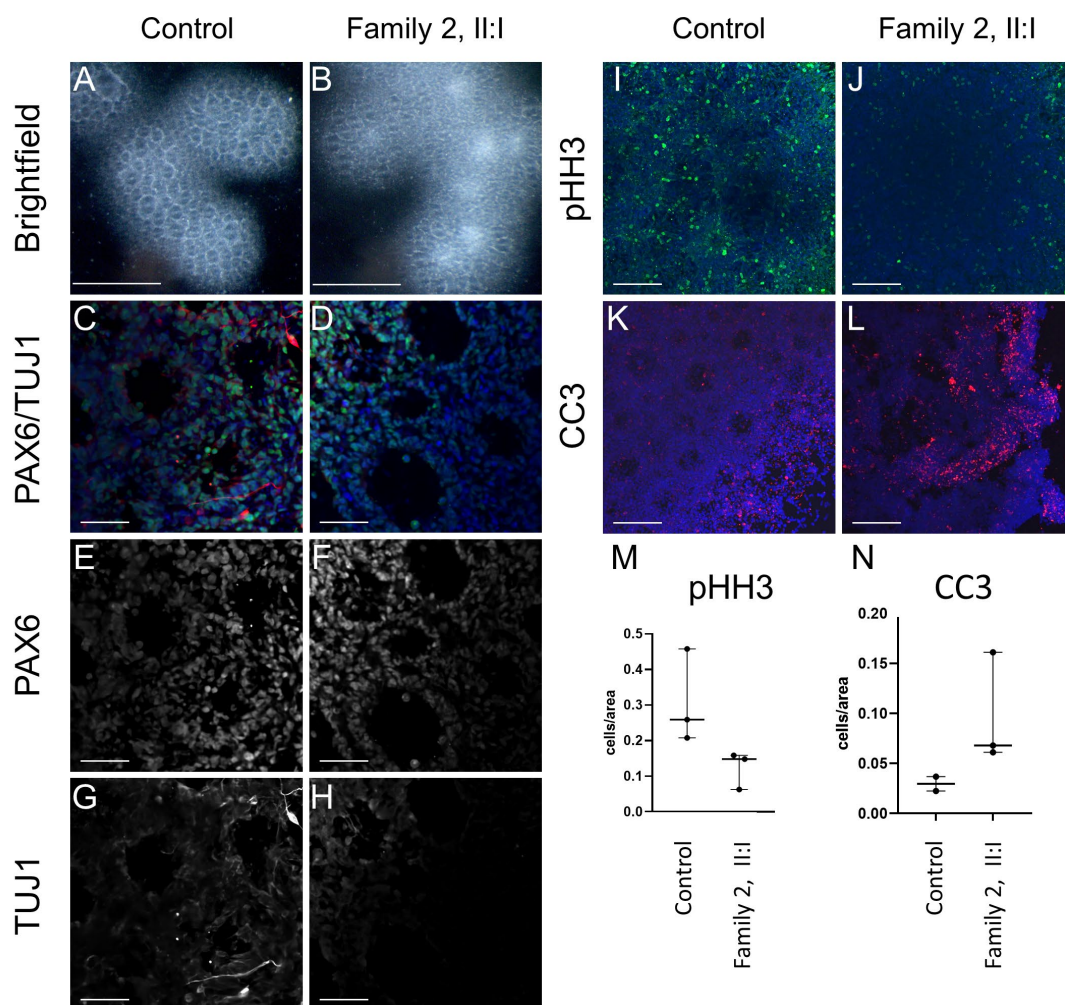


Figure 3.3: Patient iPSCs display decreased proliferation and differentiation and increased apoptosis of neural progenitor cells.

A-B - Brightfield images showing abnormal clustering of cells in rosettes from an affected individual at 11 days *in vitro*. Control cell rosettes have a clearly visible interior region of reduced cell density as compared to more dense cells from the affected individual. Scale bar = 500 μm

[C-L] - Immunohistochemistry (IHC) analysis highlights molecular features of affected individual-derived rosettes. Scale bars = 50 μm

C-H - IHC for PAX6 (green, a marker of neuronal progenitors), and TUJ1/TUBB3 (red, marks differentiated neurons) demonstrated significantly fewer differentiated neurons at 11 days in rosettes derived from affected individual iPSCs compared to control iPSCs. (**E-H**) PAX6 (**E,F**) and TUJ1 (**G,H**) shown independently.

I,J - IHC for phosphohistone H3 (pHH3) shows reduced staining suggesting a proliferation defect in rosettes derived from iPSCs obtained from an affected individual.

K,L - IHC for cleaved caspase-3 (CC3) demonstrating increased apoptosis in rosettes derived from iPSCs obtained from an affected individual.

M,N - Quantification of counts for pHH3+ cells (n=3 images x 3 replicates) and CC3+ cells (n=3 images x 3 replicates). iPSCs from the affected individual were derived from Family 2, II:1, one clone of which was received from the Genome Engineering & Stem Cell Center, Department of Genetics, School of Medicine, Washington University in Saint Louis. Control cells are from iPSC line iPSC72.3.

Despite the availability of RNASeq datasets describing general expression of the *CAMSAPs* in mouse and human tissues including during brain development, no detailed temporospatial expression studies of *CAMSAP1* in the developing embryo have been performed. This is relevant given the craniofacial and specific CNS abnormalities associated with this condition. We thus performed RNAScope *in situ* RNA hybridisation in the developing mouse at several stages from embryonic day 10.5 (E10.5) to adulthood (postnatal day 27). At E10.5, *Camsap1* is highly and ubiquitously expressed in the head, particularly in the brain and throughout the first pharyngeal arch, and neural tube (**Fig. 3.4A-B**). At E14.5 or mid-neurogenesis, expression in the brain becomes slightly more localised to the ganglionic eminences and cortex (**Fig. 3.4C**) and present (albeit somewhat reduced) in the caudal neural tube (**Fig. 3.4D**). At late stages of neurogenesis (E18.5), *Camsap1* in the brain is expressed most highly in the cortex, particularly in upper layers of differentiated neurons and the ventricular zone (**Fig. 3.4E-F**). Postnatally, *Camsap1* is particularly evident in upper cortical layers and in the hippocampus (**Fig. 3.4G-J**). High embryonic expression in the developing face tissues and cortex from E10.5-18.5 would be consistent with craniofacial abnormalities and CNS malformations such as microcephaly and lissencephaly which we observe in individuals with *CAMSAP1*-related neuronal migration disorder. Postnatal expression in upper cortical layers correlates with previous mouse and rat expression data indicating that *Camsap1* is expressed in specific populations of neurons and astrocytes (Yamamoto *et al.* 2009).

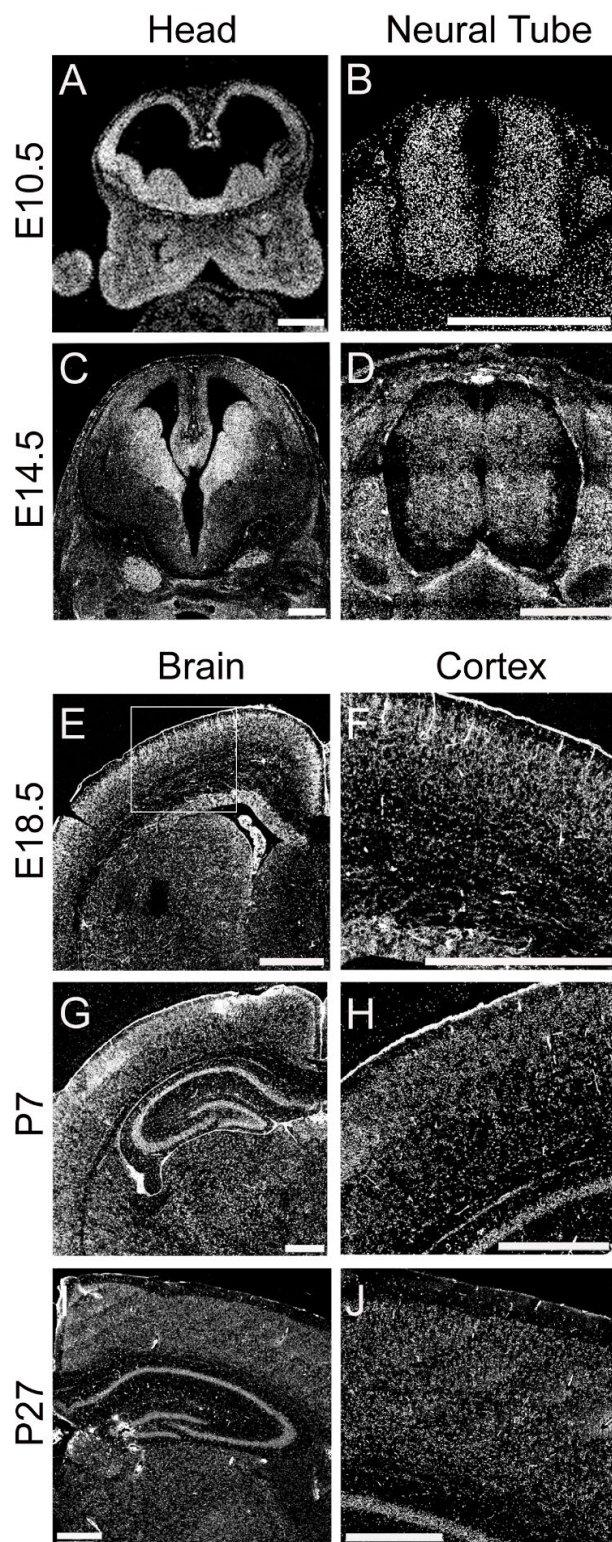


Figure 3.4: Expression of Camsap1 in the CNS and developing facial primordia

RNAScope probe for Camsap1 demonstrate robust mRNA expression at E10.5 (A-B) and E14.5 (C-D) in the developing head (A,C), and neural tube (B,D) with a clear enrichment in the neural tissues (A-J) in addition to

the developing pharyngeal arches (A). Expression at E18.5 (E-F), P7 (G-H) and P27 (I-J) remains high in the brain with slightly enriched expression in upper layers (F,H,J are higher magnification views of E,G,I, respectively). All scale bars= 500µm.

Whilst previous studies of *Camsap1* knockout mice were robust (Zhou *et al.* 2020), they focused on the laminar organisation of the cortex and neuronal polarity, and not craniofacial phenotypes. We thus obtained a null allele of *Camsap1* (*Camsap1^{em1(IMPC)J}*; hereafter referred to as the “null” allele) in order to investigate the wider effects of CAMSAP1 loss in the mouse. The null allele entails an 826 bp deletion encompassing exon 2 and 563 bp of flanking intronic sequence including the splice donor and acceptor site (www.informatics.jax.org/allele/MGI:6361950 - accessed 23/11/2022). It is predicted to result in early truncation and nonsense-mediated decay. Homozygous null mice were examined at birth (P0) and during weaning (P21) (see. **Fig. 3.S6A** for conclusive genotyping) revealing that whilst we observe normal Mendelian survival of *Camsap1^{null/null}* animals at embryonic stages E14.5-E18.5 (**Fig. 3.S6B**), *Camsap1^{null/null}* animals do not survive in Mendelian ratios postnatally, with only ~1/3rd of expected homozygous null animals present at P1 (**Fig. 3.S6C**) and zero at P21 (**Fig. 3.S6D**). Embryos display no gross morphological or histological abnormalities of the brain or facial structures and whilst a single heterozygous animal was affected by hydrocephaly, surviving *Camsap1^{null/null}* animals manifest no overt morphological or skeletal abnormalities (**Fig. 3.S7**).

Together our clinical, genetic, murine and molecular findings define biallelic likely loss-of-function variants in *CAMSAP1* as a cause of a novel recognizable

syndromic microcephalic neuronal migration disorder. The cardinal clinical features of the *CAMSAP1*-related neuronal migration disorder include craniofacial dysmorphism comprising of large ears, a prominent metopic suture, wide nasal bridge and pronounced cupid's bow (7/7 individuals, **Fig. 3.S1**), primary microcephaly (7/7 individuals, defined as >2 standard deviations below the mean for age and sex; i.e. z-score = - 2) (Centers for Disease Control and Prevention 2020), severe to profound GDD (7/7 individuals), feeding difficulties (7/7), sometimes requiring gastrostomy (2/7), cortical visual impairment (6/6 for which data were available) and seizures (6/7), typically infantile spasms with onset before one year of age, progressing to other generalised seizures refractory to antiepileptic treatment. Neurological findings although variable, include central hypotonia, with peripheral hypertonia, brisk reflexes and positive Babinski sign.

The neuroradiological abnormalities (**Fig. 3.2**) are strikingly consistent across all of the affected individuals in this study. Affected individuals display a classic (thick) lissencephaly with P>A gradient. In posterior regions, areas of agyria or severe pachygyria with prominent cell-sparse zones and reduced thickness of the cerebral mantle/wall are seen, with areas of less severe pachygyria and thicker cerebral mantle/wall in anterior regions. Extra-cortical features include dysplasia of the hippocampus (short, globular and under-rotated), basal ganglia and thalami, alongside absence of internal capsule, an absent or extremely short and thin corpus callosum, mild-moderate brainstem hypoplasia, small base of the pons, severe underdeveloped frontal horn (likely due to the basal ganglia abnormalities), enlarged third ventricle, mild borderline enlarged tectum and cerebellar hypoplasia,

although the cerebellar folia pattern remains normal (within the limits of resolution). An enlarged posterior fossa or “mega-cisterna magna” was also observed in 2/5 affected individuals.

These observed abnormalities combine cortical findings characteristic of *PAFAH1B1(LIS1)*-related classic lissencephaly with wider brain findings more in keeping with the tubulinopathies. The classical thick lissencephaly appearance with P>A gradient and prominent cell-sparse zone seen in the *CAMSAP1*-related neuronal migration disorder is analogous to the findings seen in *PAFAH1B1(LIS1)*-related lissencephaly (Di Donato *et al.* 2017), and thus predictive of a 4-layer cortical histopathology. However, the extra-cortical malformations are more in keeping with a severe tubulinopathy disorder, most similar to classical *TUBA1A*-related disease (Bahi-Buisson *et al.* 2014; Di Donato *et al.* 2017; Hebebrand *et al.* 2019; Romaniello *et al.* 2018). Notable neuroradiological differences between the *CAMSAP1*-related neuronal migration disorder and the tubulinopathies involve the cerebellar and corpus callosum phenotypes. At the milder end of the tubulinopathies spectrum cerebellar hypoplasia with an abnormal folia pattern may be the only abnormality present, while folia patterning is apparently preserved in the *CAMSAP1*-related neuronal migration disorder. Conversely, hypogenesis of the corpus callosum is typically seen in the tubulinopathies, whereas agenesis is very rarely described but was universally seen in individuals affected by the *CAMSAP1*-related neuronal migration disorder (Hebebrand *et al.* 2019). Taken together, the *CAMSAP1*-related neuroradiology may best be described as resembling that of an atypical tubulinopathy, but with a distinct and unusual pattern of abnormalities

(classic lissencephaly with a P>A gradient and complete aCC) which appear pathognomonic of the disorder.

The close alignment of neuroradiological features between the tubulinopathies, *PAFAH1B1(LIS1)*-related lissencephaly and the *CAMSAP1*-related neuronal migration disorder is consistent with a shared pathomolecular mechanism underlying these diseases and highlights the role of the minus end of the microtubule in neuronal migration disorders. *PAFAH1B1(LIS1)* facilitates the function of the minus-end-directed microtubule motor dynein, and variants affecting the heavy chain of cytoplasmic dynein (*DYNC1H1*), are associated with a neuronal migration disorder which may also show P>A gradient of lissencephaly in addition to extra-cortical malformations (Di Donato *et al.* 2017; Splinter *et al.* 2012). Dynein is a molecular motor that traffics substrates and organelles to this microtubule minus-end, whereas *CAMSAP* proteins play a stabilizing role and tubulins are the core structural component (Atherton *et al.* 2017; Hendershott & Vale 2014; Jiang *et al.* 2014).

There are, however, distinct clinical, neuroradiological and pathophysiological features only associated with the *CAMSAP1*-related neuronal migration disorder. We thus investigated the temporospatial expression of *Camsap1*, to better understand these phenotypes. Previous studies have identified high *Camsap1* expression in the cortex, subventricular zone and hippocampus of mice and rats (Yamamoto *et al.* 2009; Zhou *et al.* 2020), and in neurons, astrocytes and their precursor neuronal stem cells (Yamamoto *et al.* 2009; Yoshioka *et al.* 2012). Our

murine brain expression studies determined that *Camsap1* is widely expressed early in neurogenesis in the developing brain and neural tube, in keeping with the wide range of brain malformations observed in the disorder. In addition, the fine spatial resolution that our methods enable highlight CAMSAP1 specificity to the ganglionic eminences, ventricular zone, and outer cortical layers throughout embryonic neurogenesis, in keeping with a role in neuronal migration. Supporting this, postnatal *Camsap1* expression remains evident in the outer cortex and hippocampus, the endpoints of these migration routes. The normal development of these structures is critically dependent on precisely controlled molecular mechanisms, which govern the highly intricate neuronal migration events ongoing during these developmental stages (Ayala, Shu & Tsai 2007). Our findings thus implicate *Camsap1* in these neuronal migration processes, potentially explaining the clinical findings in affected individuals with the *CAMSAP1*-related neuronal migration disorder.

Previously studied *Camsap1* knockout mice generated by Zhou and colleagues identified radial migration defects of cortical neurons leading to cortical laminar disorganisation, and a significant increase of neurons in the intermediate zone after the completion of neuronal migration (Zhou *et al.* 2020). Whilst mice do not have sufficiently folded brains to adequately model human lissencephaly, the identification of impaired neuronal migration by Zhou and colleagues is highly consistent with the lissencephaly phenotype identified here. The mice were born in Mendelian ratios, although they exhibited reduced brain and body size and a high level of early postnatal mortality resulting from seizures in the immediate postnatal

period (Zhou *et al.* 2020). This phenotype is analogous to the early and severe forms of epilepsy invariably observed in individuals with the *CAMSAP1*-related neuronal migration disorder (**Table 3.1**), with the lack of perinatal mortality observed in humans potentially explained by the efficacy of pharmacological anti-epileptic therapies. Our studies of *Camsap1*^{null/null} mice, indicate a significantly increased mortality occurring in the third trimester or at birth for null animals, with survival into the second trimester (E14.5-E18.5) unaffected. This may reflect a variation in phenotype severity, leading to prenatal seizures or CNS abnormalities affecting fundamental physiological processes and related to differences in the genomic knockout strategies and/or potentially different genetic strains used in generating each murine model. Alternatively, stochastic variation of expression of a dosage sensitive gene may play a role. While we did not observe gross anatomical or histological changes in the second trimester (E14.5-E18.5), postnatal histological findings in the mice generated by Zhou and colleagues suggest that the absence of CAMSAP1 leads to disorganisation of cortical layering (Zhou *et al.* 2020).

Histological studies in mammalian neurons have established that CAMSAP1 inhibits neurite extension (Baines *et al.* 2009) and is critical for the differentiation of neurites into mature axons and dendrites (Zhou *et al.* 2020). A proposed mechanism for ensuring the development of (typically) a single neuronal axon involves the accumulation of CAMSAP1-associated microtubule clusters in the longest neurite, given the finding that depletion of CAMSAP1 results in an abnormal multi-axon phenotype (Zhou *et al.* 2020). The kinase MARK2, which

phosphorylates CAMSAP1 controlling its ability to bind microtubules, appears to regulate this process (Zhou *et al.* 2020). Additionally a role in neuronal polarisation and axonal / dendritic differentiation may not be unique to CAMSAP1, with other studies also implicating a similar role for CAMSAP2 and CAMSAP3 (Toya *et al.* 2016) in this process (Jiang *et al.* 2014; Yau *et al.* 2014). Other previous *in vitro* and animal studies have suggested roles for CAMSAP1 in mammalian neuronal polarisation, axon / dendrite differentiation and cytotaxis comprising important elements of neuronal migration, findings which are now consolidated by our studies. The discovery of a non-tubulin cause of a tubulinopathy-like disorder highlights other microtubule-associated proteins, including microtubule minus-end targeting proteins, as candidate genetic causes of neuronal migration disorder.

3.3. Supplemental Material

Supplemental Text

Detailed clinical summaries

Family 1, IV:10 is a three-year-old girl born to first-cousin parents at full term following an uncomplicated pregnancy. At one month of age, she was noted to be microcephalic with absent movement on her right side and increased tone. At three years and nine months of age she had made no developmental progress and made no eye contact. She startles to noise but has no verbal or non-verbal communication. She was severely microcephalic (-6.4 SDS) with large ears, thick gums, high palate, metopic ridging, a flat wide nasal bridge (**Fig. 3.S1A-B**) and bilateral fifth finger clinodactyly. Neurological examination revealed generally increased peripheral tone with upgoing planter reflexes; her right arm adopts a rigid extensor posture.

Her younger male sibling (**IV:11**) was found to have microcephaly with aCC on an antenatal ultrasound scan at 22 weeks gestation. He developed intractable epilepsy at nine weeks and was noted to have similar craniofacial dysmorphism to his sister.

A male second cousin (**V:1**) now aged three years, was born at 36 weeks gestation weighing 2kg (-1.8 SDS), microcephaly and shortened long bones were identified on antenatal scans. He had his first generalised seizure at six weeks of age, at three years these continue to be refractory to anti-epileptic treatment. Like his cousins, he has profound GDD. He displayed extreme irritability and has suffered repeated, complicated respiratory tract infections. He is severely microcephalic (-4.8 SDS) with a prominent synophrys, metopic ridging, a flat wide nasal bridge, large ears, thick gums, high palate (**Fig. 3.S1C-D**) and fifth finger clinodactyly

bilaterally. Neurological findings include generalised hypertonia, hyperreflexia and positive Babinski sign bilaterally.

MRI neuroimaging findings in all three children (**IV:10**, **IV:11** and **V:1**) are consistent and include agyria/severe pachygyria with a P>A gradient, dysmorphic basal ganglia and aCC (V-1: **Fig. 3.2A-D**, IV-11: **S2A-D**).

Family 2, III:2 is a 5-year-old male, the eldest of two siblings born to unaffected, unrelated North American parents of North European ancestry (**Fig. 3.1**). Antenatal brain imaging showed lissencephaly, aCC and a small cerebellum. His birth at 38+2 weeks was uncomplicated, although he was subsequently noted to be hypotonic and exhibited early feeding difficulties.

Probable tonic seizure activity began at four-to-five months old as ocular roving/deviation, arm extension, and truncal extension in addition to infantile spasms. His EEG showed diffuse beta frequencies, poor organisation and right parietal interictal discharges. Seizure activity was refractory to initial levetiracetam monotherapy which was discontinued. Repeat EEG was performed that demonstrated a “modified hypsarrhythmia pattern” classically associated with West Syndrome. A 30-day prednisolone taper stabilised this pattern; seizure activity resolved and was maintained with vigabatrin and clobazam dual-therapy for approximately 18 months. Thereafter there was good control on clobazam monotherapy (0.5-0.6 mg/kg/day). Recently breakthrough seizure activity has occurred requiring supplementation of the treatment regimen with Topiramate (~2 mg/Kg BD). At 3 years 7 months of age growth parameters were: Height 99.1 cm (-0.2 SDS), 13.5 kg (-1.4 SDS), 45 cm (-4.8 SDS).

Aged 5 years his neurodevelopment is severely, globally, delayed despite extensive physical therapy and without regression. He can roll and support his head and trunk for short periods, although cannot sit unsupported or crawl. He tolerates 30-45 minutes of upright weight-bearing daily using a standing aid and demonstrates slow gait/locomotion while supported in a gait trainer. His axial tone is hypotonic with variable spasticity of his limbs with intermittent guarded rigidity.

He has been treated with botulinum toxin IM injections, which caused unacceptable side effects, and is currently trialling Carbidopa/Levodopa (Sinemet). He has 15° thoracic rotoscoliosis deformity of his spine and severe left hip dysplasia, with corrective surgical intervention is planned for the latter at around five and a half years of age. He can sometimes reach for toys, recognises a few words (“book”) and uses a gaze tracking device to make choices between two or three options on a screen, but his productive speech is limited to babbles and consonant sounds without purposeful phonation of words.

There have been longstanding feeding difficulties and per oral intake consists of thickened liquids from a “sippy cup” and pureed diet via flat-plastic spoon. Steady state and flash visual evoked potential (SSVEP and FVEP) demonstrate mild cortical vision impairment and ophthalmoscopy demonstrates normal appearance of optic discs bilaterally. He also has hyperopia with astigmatism for which he wears corrective lens. There have been no concerns regarding his hearing. He was born with left unilateral cryptorchidism and has been affected by chronic constipation. Examination findings demonstrated central hypotonia with peripheral spasticity and brisk reflexes without ankle clonus.

MRI (**Fig. 3.2E-H**) performed on day two revealed diffuse severe pachygyria with a P>A gradient, dysmorphic basal ganglia, aCC and enlarged posterior fossa or “mega cisterna magna” (**Fig. 3.2E-H**)

Diagnostic trio exome was performed at GeneDx (U.S.A) using a proprietary targeting system and a custom developed analysis tool (Xome analyzer), with raw data later reanalysed in Exeter for robustness using the same pipeline as for Family 1. This identified no plausible variants in known disease genes but did reveal novel, *in trans* compound heterozygous *CAMSAP1* variants in exon 11/17: a paternally inherited Chr9(GRCh38):g.135822954dupA NM_015447.4:c.1707dupT p.(Thr570TyrFs) variant, and a maternal Chr9(GRCh38):g.135821531G>A NM_015447.4:c.3130 C>T p.(Gln1044*) variant also in exon 11.

Family 3, II:1 is the only child of North American parents of Northern European origin, born at 41 weeks gestation with a normal birthweight (3.27kg). Abnormal movements were noticed in early infancy and diagnosed as infantile spasms at 5 months of age. Spasms continues and additional seizure semiologies were noted over time including focal episodes of arm posturing and head turning with chaotic eye movements, asymmetric tonic seizures, asymmetric epileptic spasms and occasional epileptic status. Seizures eventually responded to a combination of vigabatrin, clobazam, levetiracetam and zonisamide, following unsuccessful trials of steroids and topiramate. Cannabidiol was also trialled with unclear benefit. Ketogenic diet was not attempted due to chronic intolerance of other formulas. EEG showed hypsarrhythmia. Microcephaly was noted at an early age with severe delay to neurodevelopment subsequently apparent without evidence of neurodevelopmental regression – at age one there was very poor trunk/head control and intermittent extensor arm posturing, but no purposeful movements and no communication. At 1 year 1 month his growth parameters were: Height 66cm (-4.2 SDS), weight 7.7kg (-14.5 SDS), OFC 42cm (-5.0 SDS).

At 1 year 8 months he has started turning his head purposefully, occasionally reaching for and picking up objects with his hands and scooting in crawling position. He has been diagnosed with cortical visual impairment with poor eye tracking and deprivation nystagmus. He was affected by severe feeding difficulties and there were concerns about aspiration risk requiring nasogastric feeding with a PEG placed at 15 months of age. On examination he has decreased central and peripheral tone with symmetrical, normal deep tendon reflexes and absent Babinski sign.

MRI neuroimaging (**Fig. 3.2I-L**) revealed pachygyria with thicker cerebral mantle anteriorly, enlarged 3rd ventricle and dysmorphic basal ganglia and thalami with internal capsule not seen.

Proband-only exome was performed at GeneDx (U.S.A), identified compound heterozygous predicted loss-of-function variants in *CAMSAP1*, including the same

frameshift variant identified in Family 1

[Chr9(GRCh38):g.135821923_135821944del NM_015447.4:c.2717_2738del p.(Gln906Leufs*7)] and a novel nonsense variant

Chr9(GRCh38):g.135822023G>A NM_015447.4: c.2638C>T p.(Gln880*). These were proven to be *in trans* using the short-read exome sequencing data and parental segregation was not required. A previously reported pathogenic heterozygous variant in *POLG* [NM_002693.2:c.202C>T; p.(Gln68*), inheritance unknown] was also identified. This variant has been published twice previously, in both cases associated with mitochondrial DNA depletion syndrome 4A (MIM: 203700), also known as Alpers–Huttenlocher syndrome, when *in trans* with another pathogenic *POLG* variant and with transmitting parents unaffected (Wong *et al.* 2008, Saneto *et al.* 2010). Since no second pathogenic *POLG* variant could be identified in this individual this is unlikely to be the cause of their neurodevelopmental disorder.

Family 4, II:1 is a male child of related Turkish parents. He was born at full term weighing 3.26kg (-0.3 SDS), length 49cm (-0.6 SDS) and with a head circumference of 32 cm (-2.5 SDS). He presented with severe developmental delay, central hypotonia, limb spasticity, epilepsy and relative microcephaly. Seizures were first observed at 5 months of age as infantile spasms then progressing to other semiology, occurring most frequently on waking and in the early morning and requiring Sodium Valproate, Vigabatrin, Clobazam for control. EEG showed a burst suppression pattern.

At 5 years 6 months he has no head control or other gross motor development, no fine motor development, cortical visual impairment and an unsafe swallow to both liquids and solids. He makes sounds, but these are without clear meaning and has no communicative language. Limb movements are dyskinetic with varying spasticity of his limbs and intermittent guarded rigidity. Central tone is severely reduced, with peripheral spasticity and hyperactive deep tendon reflexes. He has a prominent and wide nasal root, relatively large ears, an open mouth with high arched palate and left sided unilateral ptosis.

At 6 years 5 months growth parameters were: Height 125cm (+1.3 SDS), weight 25kg (+1.0 SDS) OFC 53cm (-0.2 SDS).

Brain MRI (**Fig. 3.2M-P**) revealed diffuse lissencephaly, dysmorphic basal ganglia, a thin corpus callosum and an enlarged posterior fossa or "mega-cistern magna".

Trio exome sequencing at Baylor College of Medicine, using previously described methods (Pehlivan *et al.* 2019), identified a novel homozygous candidate nonsense variant, Chr9(GRCh38):g.135822830T>A; NM_015447.4:c.1831A>T; p.(Lys611*), located within a 4.2Mb region of homozygosity.

Family 5, II:1 is the adopted child of North American parents, only limited information was available regarding her biological parents. She was born at 41+1 weeks gestation following a high-risk pregnancy with a birthweight within the normal range, and head circumference of 31.8cm (-2.2 SDS). Neurodevelopment was reported as severely delayed, rolling first at 3y10m, no speech at 4y9m and dependent for all activities of daily living (Gross Motor Function Classification System (GMFCS) level V. Dystonia, presenting as neck extension and back arching, was diagnosed at 11m of age and managed with diazepam, gabapentin, and baclofen. Seizures semiologies include complex partial seizures involving the right side and unresponsive episodes with associated eye rolling, likely generalised seizures. These are associated with EEG findings of multifocal epileptiform discharges suggesting electroclinical seizures that appeared to lateralise to either hemisphere and have been treated with vigabatrin and diazepam for extended seizures. Neurological examination and revealed central hypotonia with bilateral lower extremity spasticity, and investigations were consistent with cortical visual impairment. Her hearing was normal when assessed at 11m of age. At 4 years and 9 months she was no longer able to roll her growth parameters were: Height 106.5cm (-0.1 SDS), weight 17.3kg (-0.2 SDS) and OFC 42.5cm (-7.1 SDS). During her 5th year of life she developed episodes of "emesis with dark fluid", the family opted for hospice care after an extensive evaluation for her symptoms and she died at 5.5 years of age.

Brain MRI findings (**Fig. 3.S2E-F**) include holohemispheric bilateral lissencephaly and grey matter band heterotopia with notable white matter volume loss, a prominent cisterna magna and diffusely small brainstem with decreased volume of the dorsal pons.

Proband-only exome sequencing, performed at Cincinnati Children's Hospital Medical Center and analysed using VarSeq 2.2.3 (Golden Helix, Bozeman, MT) identified a homozygous *CAMSAP1* variant [Chr9:g.135818055G>C NM_015447.3: c.4193C>G p.(Ser1398*)].

Supplemental Methods

iPSC culture

Human iPSCs from a control line (iPSC72.3) and from an affected individual (Family 2, II:1) were cultured in MTeSR media (STEMCELL) in Nunc plates (Fisher) on a matrix of Matrigel (CCHMC PSCF) dissolved in Dulbecco's Modified Eagle's Medium/Nutrient Mixture F-12 (DMEM/F12). One clone of each line was received from the CCHMC PSCF and the Genome Engineering & Stem Cell Center, Department of Genetics, School of Medicine, Washington University in Saint Louis respectively. Cells were passaged every 7 days using Gentle Cell Dissociation Reagent (GCDR, STEMCELL) and fed daily. The STEMdiff™ SMADi Neural Induction Kit (STEMCELL) was used to generate neural rosettes from high-quality iPSC colonies. Briefly, iPSCs were dissociated into single cells and plated into an Aggrewell 800 well at 10,000 cells per well, forming embryoid bodies. These were fed daily, then replated on day 5 by filtering through a 37µM reversible strainer into a 24-well plate containing coverslips coated in Matrigel/DMEM/F12. On day 8, percent neural induction was visually estimated for each well and confirmed to be 75% or above for neural rosettes which were harvested on day 8 or day 11.

Clinical and genetic methods

DNA was extracted from blood/buccal samples using standard techniques. Exome sequencing was performed using DNA from individuals V:1 and IV:10 (Family 1) using either Agilent SureSelect Whole Exome v6 (Agilent Technologies, Santa Clara, CA) or Twist Human Core Exome Kit (Twist Bioscience, San Francisco, CA) exon targeting respectively. Reads were aligned (BWA-MEM v0.7.17), mate-pairs fixed and duplicates removed (Picard v2.15.0), InDel realignment/base quality recalibration (GATK v3.7.0), SNV / InDel detection (GATK HaplotypeCaller), annotation (Alamut v1.8), and read depth was determined for the whole exome through our in-house pipeline. This conforms to GATK best practices. Variants were filtered based on call quality, segregation with disease, impact on gene function and allele frequency in population databases. Homozygous or compound heterozygous variants present in exons or adjacent intronic regions were evaluated and assessed for clinical correlation with phenotype.

Diagnostic exome was performed for Individual II:1 (Family 2; trio of both parents and proband) and Individual II:1 (Family 3; proband-only) at GeneDx (U.S.A) using a proprietary targeting system and a custom developed analysis tool (Xome analyzer). Raw data from Individual II:1 (Family 2) were also re-analysed using the same bioinformatic pipeline and filtering strategy as in Family 1, with the inclusion of the analysis of *de novo* variants. Trio exome was performed for Individual II:1 (Family 4) at Baylor College of Medicine as described previously (Mitani *et al.* 2021). Individual II:1 (Family 5) was analysed using - VarSeq 2.2.3 from Golden Helix. Protein coding variants with plausible variant allele fraction (VAF - 0.3-0.7 for heterozygous variants) were selected and filtered by frequency (MAF < 0.0001, <2 individuals in gnomAD for heterozygous variants) and Combined Annotation Dependent Depletion (CADD) score prediction (>25).

Immunocytochemistry

Cells were plated onto coverslips in a 24-well tissue culture plate and fixed in 4% paraformaldehyde for 15 minutes. To permeabilise cell membranes, coverslips were immersed in 0.1% Triton-X 100 for 5 minutes prior to blocking. Coverslips

were blocked in 4% normal goat serum for 30 mins before addition of primary antibody(ies) at 4°C overnight. Secondary antibody was applied for 1 hour, then coverslips were co-stained with DAPI (4', 6-Diamidino-2-phenylindole dihydrochloride) for 15 min. They were sealed to glass slides with ProLong Gold Antifade Mountant. Images were acquired on a Nikon C2 Confocal Microscope. Tuj1+ (n=2), PHH3+ (n=3), and CC3+ (n=3) cells were quantified using the Brightspot Detection automated measurement function in NIS-Elements AR software. Three images were captured per coverslip per experiment, with n=2 or n=3 experimental replicates as listed above. The averages of each set of three images are shown in **Fig. 3.3**. Antibodies and concentrations for immunocytochemistry are listed in **Table 3.S3**.

RNAScope

Wild-type mouse embryos maintained on a CD1 genetic background were dissected at ages E10.5, E14.5, E18.5, P7, and P27 and fixed in formalin for 16-24 h; brains were sub-dissected for ages E18.5-P27. The tissue was washed in phosphate-buffered saline (PBS), then dehydrated and paraffin embedded by the Cincinnati Children's Hospital Medical Center (CCHMC) Pathology Core. Paraffin blocks were sectioned at 5 μ m, placed on SuperFrost slides, and baked at 60°C for 1 hour. Target retrieval steps outlined in the manual assay protocol were followed based on recommendations for brain tissue, then slides were dried at room temperature overnight. Hybridisation and amplification steps were performed using the HybEZ oven set at 40C. Manual assay protocol from ACDBio was performed using RNAScope Multiplex Fluorescent Reagent Kit V2 (323100), Tyramide Signal Amplification (TSA) Cyanine 3 Fluorophores (NEL744001KT) at 1:750, and Camsap1 probe made to order by ACDBio (Cat. # 866521) (**Fig. 3.3.S9**).

Mouse husbandry

All animals were maintained through a protocol approved by the Cincinnati Children's Hospital Medical Center Institutional Animal Care and Use Committee (IACUC2019-0068). C57BL/6NJ-*Camsap1*^{em1(IMPC)J} / Mmjax mice (Jackson Labs,

Mutant Mouse Resource and Research Center (MMRRC) Stock No. 65662-JAX) were housed in a vivarium with a 12-h light cycle with food and water *ad libitum*. Mice were maintained by intercross from the stock commercially obtained for up to three generations. Mice for dissection were euthanised with isoflurane and cervical dislocation. Whole-brain and skeletal images were taken on a Zeiss Discovery V8 microscope.

Histology

Embryos were dissected, fixed in Bouin's fixative for 48h, washed in 70% ethanol, and dehydrated and paraffin embedded by the CCHMC Pathology Core. Blocks were sectioned by microtome at 10um, then sections were placed on SuperFrost slides, baked >1 hour, and stained with hematoxylin and eosin using standard methods.

Skeletal Preparations

Pups were collected at P0-P1 and frozen. Skin and fat were removed from the embryos prior to fixation in 95% ethanol for 2-5d. The skeletons were stained with Alizarin red and Alcian blue and cleared with potassium hydroxide using standard procedures. Bone measurements were taken with Zen software and assessed for statistical significance with one-way analysis of variance (ANOVA) (n=6-12 animals per genotype).

Study approvals

Studies were conducted in accordance the declaration of Helsinki. Written informed consent was received from participants prior to inclusion in the study.

- Palestinian Health Research Council - PHRC/HC/518/19
- Cincinnati Children's Hospital Medical Center - 2014-3789
- Baylor College of Medicine - H-29697

Figure 3.S1: Facial features of individuals with CAMSAP1-related disorder

Clinical photographs demonstrating the cardinal features of the CAMSAP1-related disorder including microcephaly, large ears, prominent metopic suture, wide nasal bridge and pronounced cupid's bow. (A,B) Family 1, IV:10; (C,D) Family 1, IV:11; (E) Family 2, II:1; (F) Family 2, II:1 (5y8m) (G) Family 3, II:1; (H) Family 5, II:1.



Figure 3.S2: Additional MRI brain images from individuals with CAMSAP1-related disorder

Row 1 (A-D) **Family 1, IV:11** aged 4 months, showing agenesis of the corpus callosum (aCC), dysmorphic basal ganglia, posterior-more-severe-than-anterior gradient pachygyria and cerebellar hypoplasia.

Row 2 (E-F) **Family 5, II:1**. In this case full imaging was not available for re-review, but report and static images were reviewed. There is holohemispheric bilateral lissencephaly with notable white matter volume loss and a prominent cisterna magna.

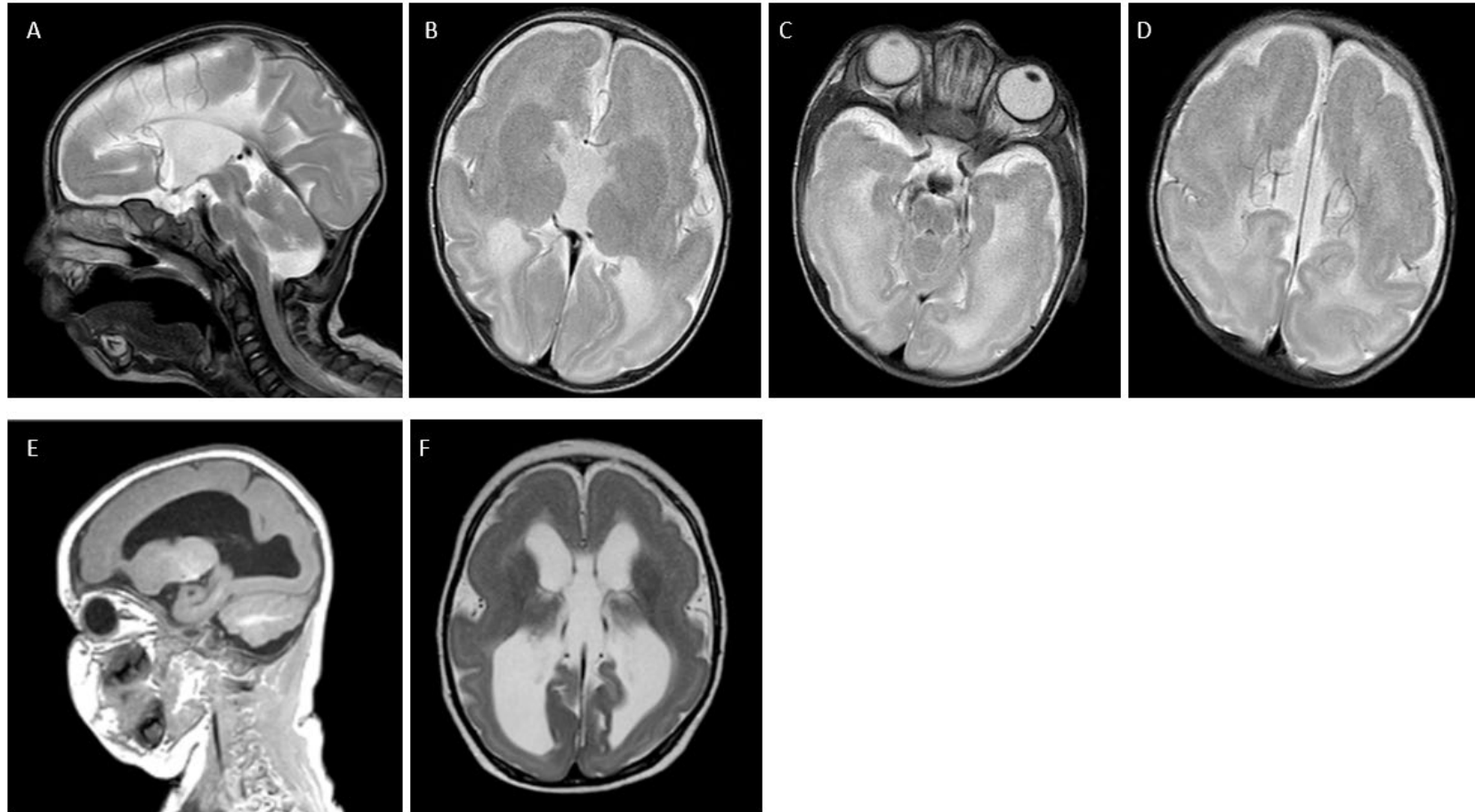
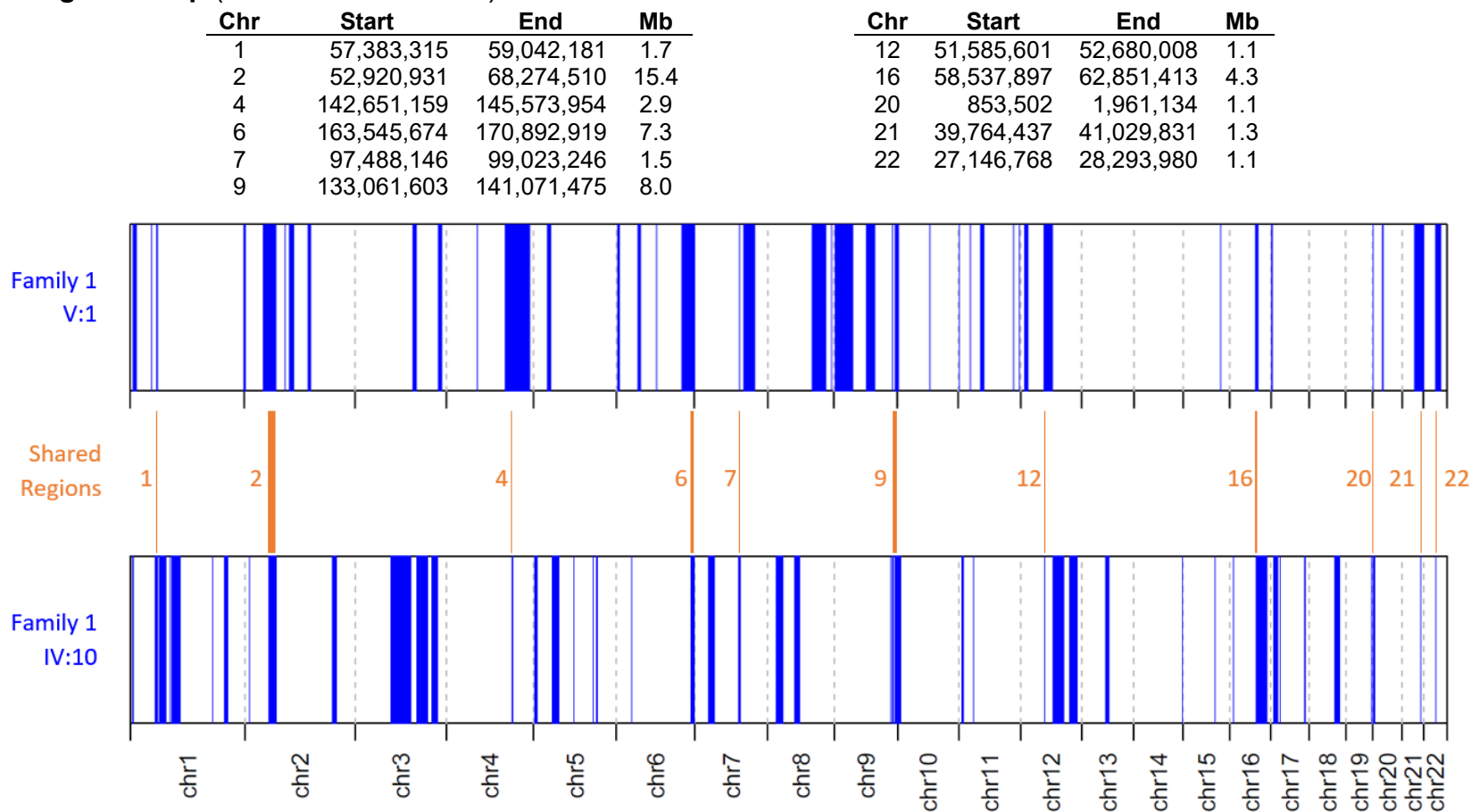


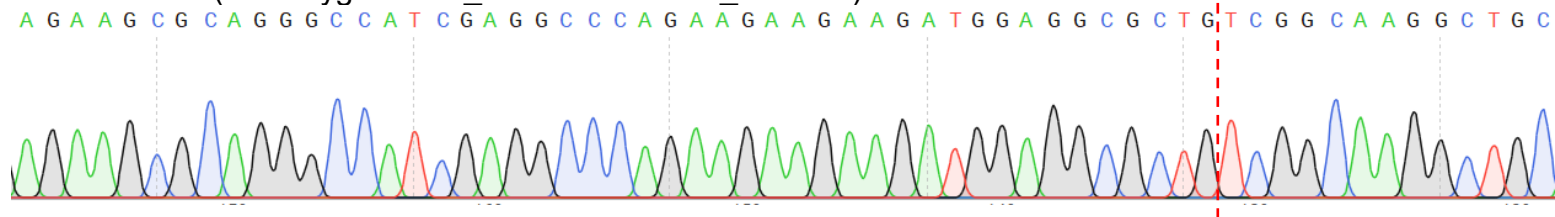
Figure 3.S3: Homozygous regions greater than 1Mb shared between individuals IV:10 and V:1, generated using AutoMap (Quinodoz *et al.* 2021)



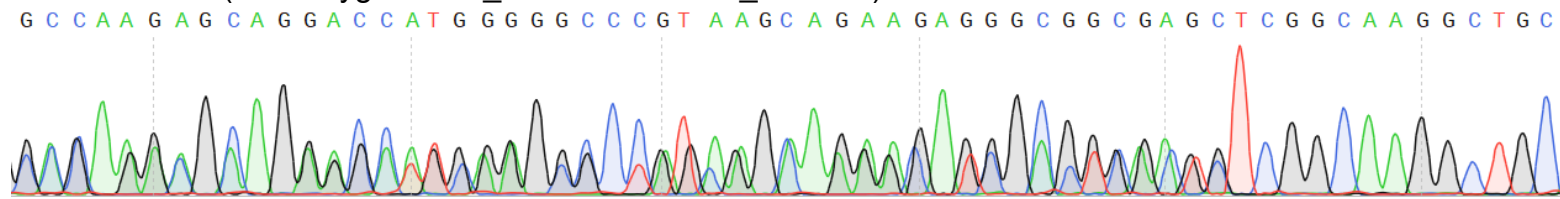
Quinodoz, M., V. G. Peter, N. Bedoni, B. Royer Bertrand, K. Cisarova, A. Salmaninejad, N. Sepahi, R. Rodrigues, M. Piran, M. Mojarrad, A. Pasdar, A. Ghanbari Asad, A. B. Sousa, L. Coutinho Santos, A. Superti-Furga and C. Rivolta (2021). "AutoMap is a high performance homozygosity mapping tool using next-generation sequencing data." *Nature Communications* **12**(1): 518.

Figure 3.S4: CAMSAP1 sequencing chromatograms from Family 1

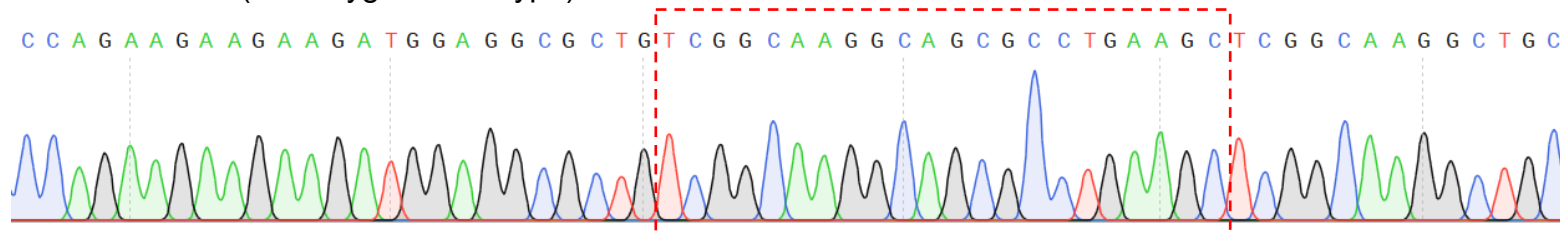
Individual V:1 (homozygous NM_015447.4 c.2717_2738del)



Individual IV:4 (heterozygous NM_015447.4 c.2717_2738del)



Individual IV:12 (homozygous wild type)



Positions of the 22 deleted bases are shown on the homozygous individual with a red line and the wild-type individual with a red box

Figure 3.S5: Recurrent CAMSAP1 deletion may result from homologous recombination

Alignment of reads generated through exome sequencing for Family 1, IV:10 and Family 1, V:1, visualised using the Integrative Genomics Viewer (version 2.9.2, Broad Institute). Position on Chromosome 9 (GRCh37) is shown at the top. Read depth is shown above a number of representative reads. The reference sequence is shown below the 22-base-pair deletion NM_015447.4 c.2717_2738del; p.(Gln906Leufs*7) is shown with the homologous sequence GCCTTGCCGA highlighted in red and yellow. Recombination of these adjacent regions may be an explanation for the recurrent nature of this variant.

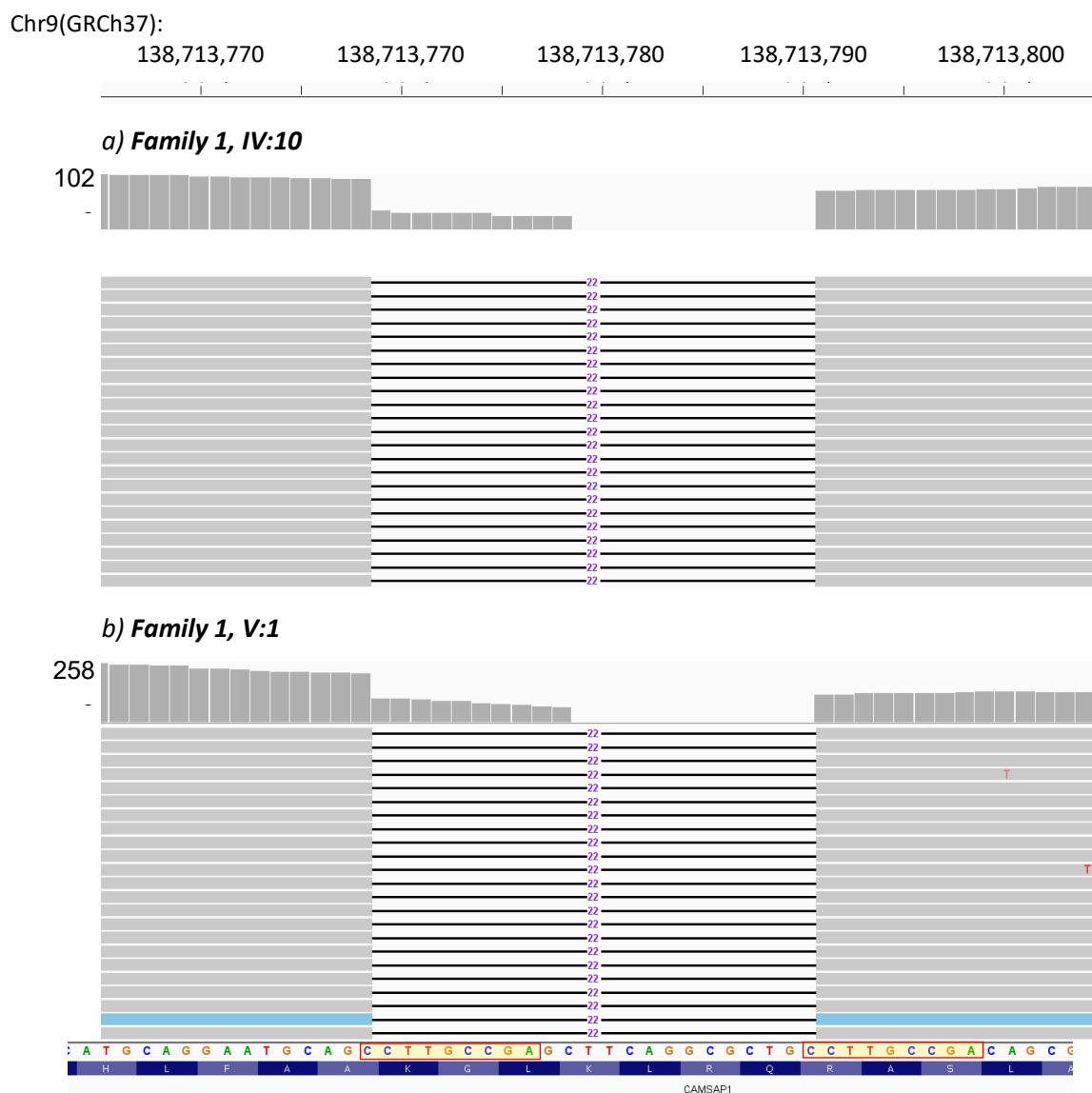
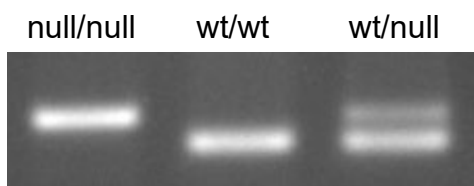


Figure 3.S6: *Camsap1* genotyping and Mendelian survival

A: Conclusive genotyping example for *Camsap1* homozygous (null/null), control (wt/wt), and heterozygous (wt/null) animals. The null allele is *Camsap1*^{em1(IMPC)J}.

**Mendelian survival tables for mice**

B: E14-E18.5: Embryonic survival follows Mendelian ratios.

	Total	wt/wt	wt/null	null/null
Expected		12	24	12
Observed	48	9	28	11
% of total surviving	100%	19%	58%	23%

p=0.47

C: P0-P1: Approximately one-third of expected homozygous null animals survive birth.

	Total	wt/wt	wt/null	null/null
Expected		9	18	9
Observed	36	16	17	3
% of total surviving	100%	44%	47%	8%

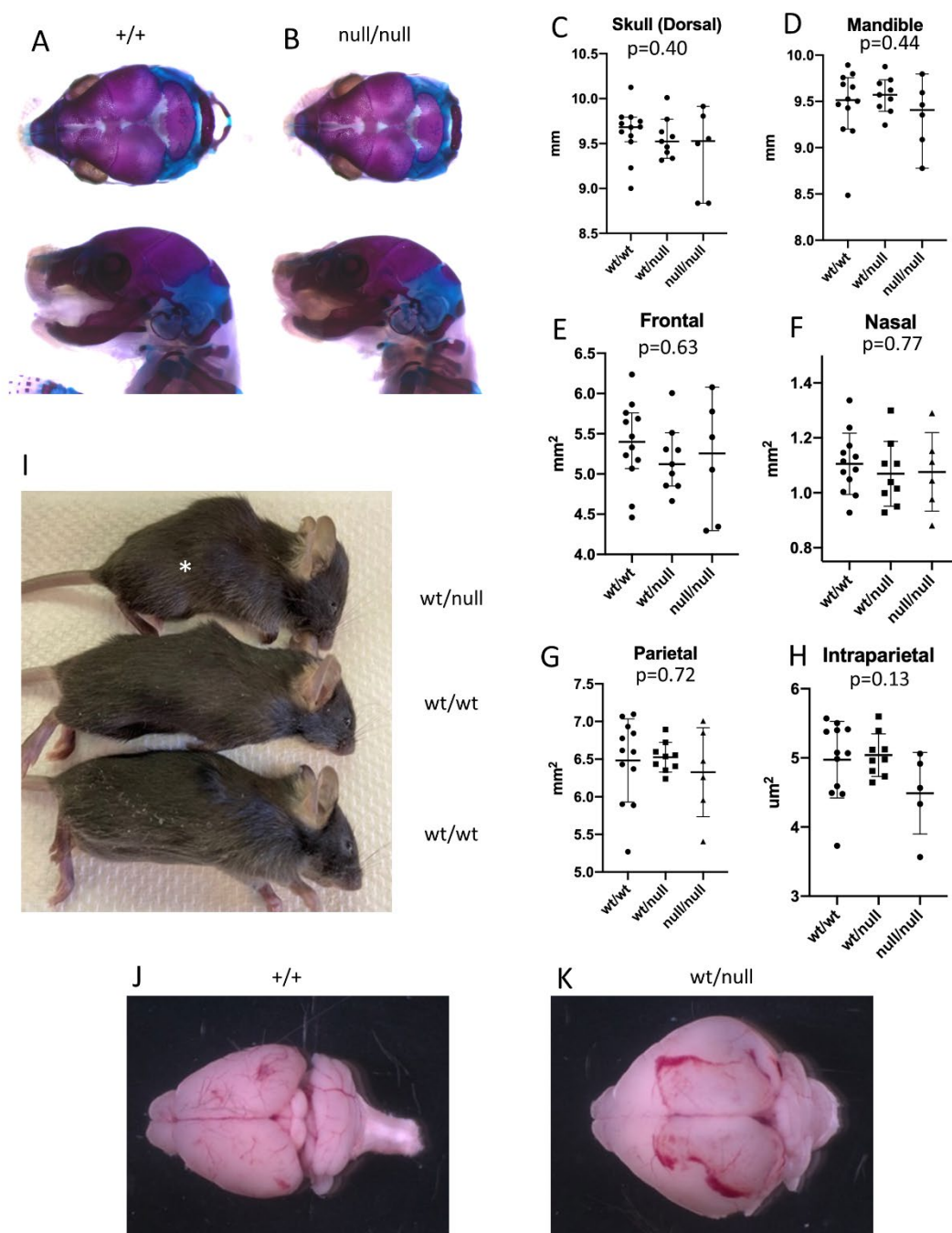
p=0.087

D: P21: Homozygous null animals do not survive to weaning

	Total	wt/wt	wt/null	null/null
Expected		9.25	18.5	9.25
Observed	37	10	27	0
% of total surviving	100%	26%	69%	0%

p=0.001

Figure 3.S7: Postnatal mutant mice do not survive in normal ratios but exhibit no skeletal abnormalities.



A,B - P0 skeletal preparation example images exhibiting no evident abnormalities in homozygous null animals. **C-H** - Quantification of bone length measurements from skeletal preparations ($n=6-12$ animals per genotype as shown in plots). **I** - P21 littermates, *Camsap1*^{null/wt} individual with hydrocephaly marked with asterisk. **J-K** - Brains dissected from littermate control and hydrocephalic *Camsap1*^{null/wt} individual. The null allele is *Camsap1*^{em1(IMPC)J}.

Figure 3.S8: Embryonic *Camsap1* null mice exhibit no morphological phenotypes.

Whole-mount control and *Camsap1*^{null/null} embryos at stages E14.5 (A-B) and E16.5 (C-D). Hematoxylin and eosin-stained histological images of control and *Camsap1*^{null/null} embryos at E14.5 (E-F) and E16.5 (G-H). No gross morphological or histological abnormalities were noted. The null allele is *Camsap1*^{em1(IMPC)J}.

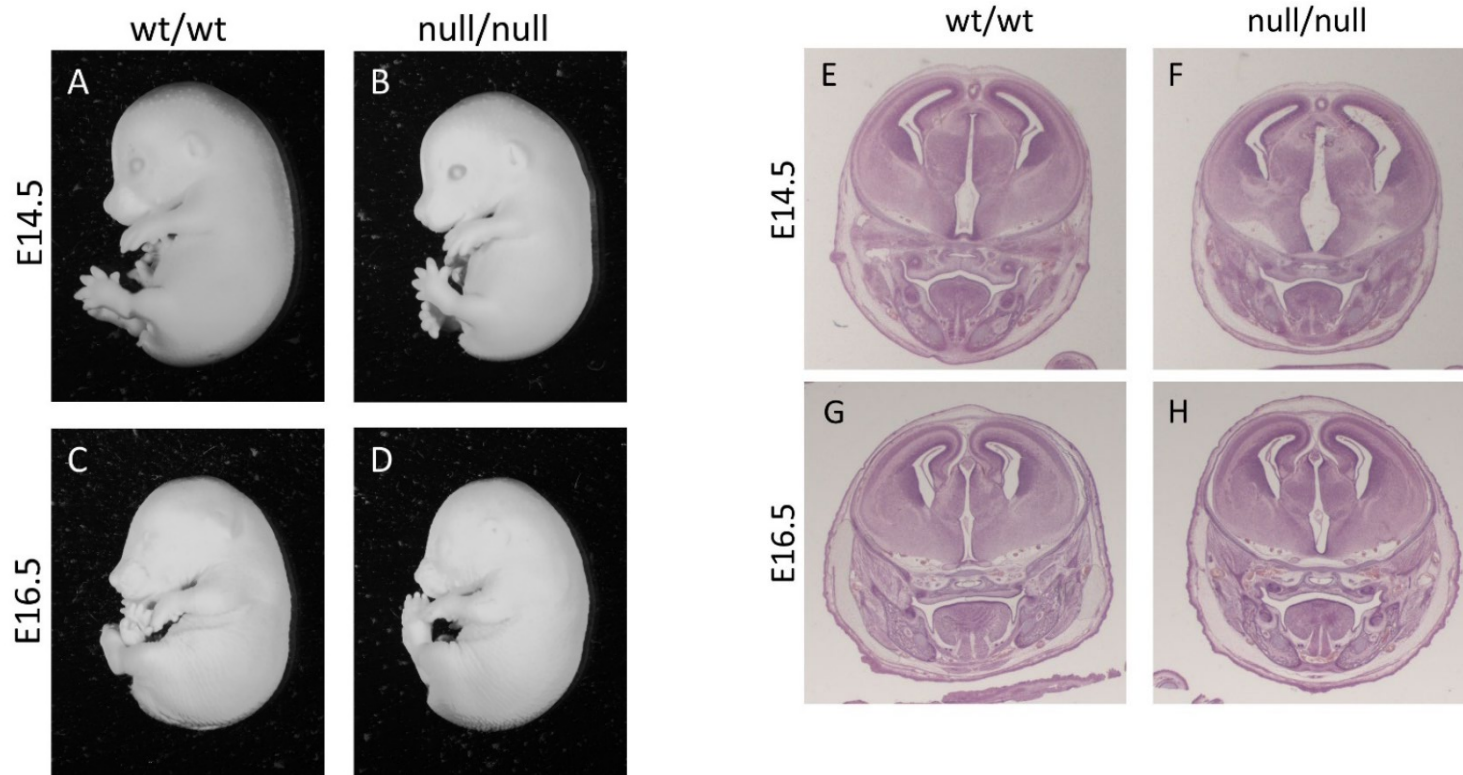


Figure 3.S9: Camsap1 target region for RNA scope probe (ACDBio Cat# 866521)

Species* :	Mouse
Species (common):	House Mouse
Entrez Gene ID :	227634
Gene Alias :	9530003A05Rik
Accession No:	NM_001276359.1
Target Region :	6606 - 7920
No. of Pairs :	20

6606 acagt gtccacatc ttctggcttc agggcagctg cgcatgactt ttgggttacc
 6661 taggaattta ttttcatga agagttgaa gacctgggtg agagcaggaa gttgcatttc
 6721 tgctcgactt ttaatcggg gacttggcat gtggcctgtc ctgaaggctg ttcagaacg
 6781 cagggctctg atgtgaagta cagtgccagt gtctcaaggc cgggctgtgg aggccatgtg
 6841 gcttgatggc tcagaagcac tcgtgccttc tgtgtccat gacacagaag tagttttgat
 6901 tttttttta attaggaagt ttccggta ca ggattttgtg gtgggtgatct ggagcctcca
 6961 ggggtgtggg acaactgtct gaccaactgt gctgggaagg ctactcagct ttcgggcag
 7021 aaagtgccaa gaaattgaat acatgacgac tcatgcagct caticcttag taacaacaga
 7081 cccccccgag gagcccaggc cggttggcat taaaaatatt tctcagccgg gcgggtgggg
 7141 cgcacgcctt taatccaagc acttgggagg cagagacagg cggatttctg agtttgaggc
 7201 cagcctgac tacaaagtga gttccaggac agccaggact acacagagaa acctgtctc
 7261 gaaaaaataa aaaaaaaaaa ttcagaact agatgtcag gtctacatgt ctgtgtcgt
 7321 gtgggatgta cctgtgtgtg cgtatatgag tatgtgcaca caaacatgtg catatggcct
 7381 gtgtacacgt gtgtatgcat gccaaagata gcgtttccag tgtgacctt tgacctgga
 7441 ttctcgggtg gttcctgca tacaatctcc aatcagactt ttaaggcca gactgctca
 7501 ggcaacattg aaaagtggca taatacaaaa ttacttcta gattgttga aactatggt
 7561 gttacttga agaagaaagt gtaaaagtc cattttctt gttgaaagt aatcaactg
 7621 agtaaacctt tatagattgg tgctggttta cctagtgaaa tggctttgat ctgggtacc
 7681 tgagcctatt ggtgattca ttctatctgt gtccagtacc acatgtgtaa agccagttct
 7741 aactcctgt ttgtgactat ggccaagtca caagcccaa ctgggaacga gccatgcca
 7801 gcatcctctt gtttctact gattcctgg caataaact gtacctgct gactcagggt
 7861 ctctcgtg tctgaggtgc actctggagg tgctttggt atctggtatc actggttgc

Supplemental Tables

Table 3.S1: Pathogenic *CAMSAP1* variants identified in this study with their frequency in population databases

<i>CAMSAP1</i> variant (NM_015447.4)	Population	Chr9(GRCh38): g.	Chr9(GRCh37): g.	Nucleotide change	Exon	gnomAD v2.1.1	gnomAD v3.1.1
c.1707dupT p.(Thr570Tyrfs*17)	White European	135,822,954	138,714,800	insA	11/17	absent	absent
c.1831A>T p.(Lys611*)	Turkey	135,822,830	138,714,676	T>A	11/17	absent	absent
c.2638C>T p.(Gln880*)	White European	135,822,023	138,713,869	G>A	11/17	absent	absent
c.2717_2738del p.(Gln906Leufs*7)	Arab / White European	135,821,923_ 135,821,944	138,713,769_ 138,713,790	AGCCTTGCCG AGCTTCAGGC GCT > A	11/17	1 het	absent
c.3130 C>T p.(Gln1044*)	North America	135,821,531	138,713,377	G>A	11/17	absent	absent
c.4193C>G p.(Ser1398*)	North America	135,818,055	138,709,901	G>C	14/17	absent	absent

Abbreviations: het, heterozygous individual

Table 3.S2: Other variants identified through exome sequencing

Individual(s)	Variant	Zygosity, Inheritance	Gene	OMIM phenotype	Gene expression	gnomAD v2.1.1	ClinVar	SIFT	Polyphen2	Interpretation
Family 1 V:1, IV:10	NM_024757.4:c.623C>T; p.(Pro208Leu)	Homozygous	<i>EHMT1</i>	(AD) Kleefstra Syndrome 1; 610253	Widespread	4 het	-	Damaging 0.001	Probably damaging 0.998	Phenotype absent from parents
	NM_153710.4:c.1819G>A; p.(Asp607Asn)	Homozygous	<i>STKLD1</i>	-	Testes only	10 het	-	Tolerated 0.257	Possibly damaging 0.855	No phenotype, Expression only in testes
	NM_004269.3:c.107A>G; p.(Lys36Arg)	Homozygous	<i>MED27</i>	-	Oesophagus, v low in brain	27 het	-	Tolerated 0.240	Possibly damaging 0.899	No phenotype, Low brain expression
	NM_020695.3:c.3547G>A; p.(Asp1183Asn)	Heterozygous	<i>REXO1</i>	-	Widespread	3 het	-	Tolerated 0.141	Benign 0.001	No phenotype, Predicted benign
	NM_020695.3:c.1493G>A; p.(Arg498His)	Heterozygous	<i>REXO1</i>	-	Widespread	13 het	-	Damaging 0.022	Probably damaging 0.955	No phenotype, Second REXO1 variant predicted benign
	NM_016252.3:c.11032A>G; p.(Ile3678Val)	Heterozygous	<i>BIRC6</i>	-	Widespread	174 het	-	Tolerated 0.353	Benign 0	No phenotype, predicted benign
	NM_016252.3:c.14294A>G; p.Glu4765Gly)	Heterozygous	<i>BIRC6</i>	-	Widespread	132 het	-	Damaging 0.000	Probably damaging 0.96	No phenotype, Second BIRC6 variant predicted benign
Family 2 II:1	NM_013358.2:c.592A>G; p.(Lys198Glu)	Heterozygous <i>De novo</i>	<i>PADI1</i>	-	Widespread	absent	-	Tolerated 0.624	Possibly damaging 0.844	No phenotype
	NM_001369.2:c.7246C>T; p.(Arg2416Cys)	Heterozygous Paternal	<i>DNAH5</i>	Ciliary dyskinesia, primary, 3, with or without situs inversus	Low in brain	2 hets	-	Damaging 0.000	Benign 0.055	Phenotype absent, Second <i>DNAH5</i> variant predicted benign
	NM_001369.2:c.4687G>A; p.(Gly1563Ser)	Heterozygous Maternal	<i>DNAH5</i>	Ciliary dyskinesia, primary, 3, with or without situs inversus	Low in brain	24 hets	Conflicting interpret- ations	Tolerated 0.54	Benign 0.002	Phenotype absent, Predicted benign
	NM_002850.3:c.1828C>G; p.(Arg610Gly)	Heterozygous Maternal	<i>PTPRS</i>	-	Widespread	1 het	-	Tolerated 0.36	Benign 0.081	No phenotype, Predicted benign
	NM_002850.3:c.1438G>T; p.(Val480Leu)	Heterozygous Paternal	<i>PTPRS</i>	-	Widespread	2 hets	-	Damaging 0.000	Probably damaging 0.969	No phenotype, Second <i>PTPRS</i> variant predicted benign

Individual(s)	Variant	Zygoty, Inheritance	Gene	OMIM phenotype	Gene expression	gnomAD v2.1.1	ClinVar	SIFT	Polyphen2	Interpretation
Family 3 II:1	NM:002693.2:c.202C>T; p.(Gln68*)	Heterozygous	<i>POLG</i>	Mitochondrial DNA depletion syndrome 4A; 203700, 4B; 613622 & others	Widespread	absent	Pathogenic	-	-	Mitochondrial phenotypes are recessive, no second variant
<i>Full variant list not available</i>										
Family 4 II:1	NM_004817.4:c.2065C>T; p.(Arg689Cys)	Homozygous	<i>TJP2</i>	Cholestasis, progressive familial intrahepatic 4; 615878	Widespread	1 het	-	Damaging 0.001	Damaging 1	Phenotype absent in the proband
	NM_001100112.2:c.2654C>T; p.(Thr885Met)	Homozygous	<i>MYH2</i>	Proximal myopathy and ophthalmoplegia; 605637	Low in brain, high in muscle	16 het	-	Damaging 0.029	Benign 0.001	Phenotype absent in the proband
	NM_014733.6:c.1670C>G; p.(Ser557Cyst)	Heterozygous <i>de novo</i>	<i>ZFYVE16</i>	-	Widespread	17 het	-	Damaging 0.037	Damaging 0.995	No phenotype
Family 5 II:1	NM_001099439.1:c.712G>A; p.Gly238Arg	Heterozygous	<i>EPHA10</i>	-	Low expression in cortex	3 hets for same amino acid	-	Damaging 0.000	-	No phenotype, Low brain expression
	NM_001134479.1:c.442T>C; p.Phe148Leu	Heterozygous	<i>LRRC8D</i>	-	Widespread	absent	-	Damaging 0.003	Probably damaging 0.975	No phenotype
	NM_001531.2:c.602_603delCA; p.Thr201Argfs*47	Heterozygous	<i>MR1</i>	-	Widespread	1, multiple examples of fs* in earlier amino acids	-	-	-	No phenotype
	NM_018263.4:c.1952G>C; p.Arg651Thr	Heterozygous	<i>ASXL2</i>	Shashi-Pena syndrome; 617190	Widespread	absent	-	Damaging 0.000	Probably damaging 0.942	Does not match established phenotypes
	NM_020381.3:c.154G>A; p.Val52Ile	Heterozygous	<i>PDSS2</i>	Coenzyme q10 deficiency, primary, 3; 614652	Widespread	1	-	Tolerated 0.258	Possibly damaging 0.76	Does not match established phenotypes

Individual(s)	Variant	Zygoty, Inheritance	Gene	OMIM phenotype	Gene expression	gnomAD v2.1.1	ClinVar	SIFT	Polyphen2	Interpretation
Family 5 II:1	NM_001277115.1:c.2966G>A; p.Arg989Gln	Heterozygous	<i>DNAH11</i>	Ciliary dyskinesia, primary, 7; 611884	Low expression in cortex	1 het for same amino acid	conflicting	Damaging 0.004	Probably damaging 0.994	Does not match established phenotypes
	NM_006955.2:c.1910A>G; p.Tyr637Cys	Heterozygous	<i>ZNF33B</i>	-	Widespread	1 het	-	Damaging 0.001	Probably damaging 0.991	No phenotype
	NM_019006.3:c.579C>G; p.Ile193Met	Heterozygous	<i>ZFAND6</i>	-	Widespread	absent	-	Damaging 0.002	Probably damaging 0.991	No phenotype
	NM_018146.3:c.175delC; p.Arg59Alafs*46	Heterozygous	<i>MRM3</i>	-	Widespread	2 het for same fs*	-	-	-	No phenotype
	NM_030962.3:c.191C>T; p.Thr64Met	Homozygous	<i>SBF2</i>	Charcot-Marie- Tooth disease, type 4b2; 604563	Widespread	absent	-	Damaging 0.011	Probably damaging 0.996	Does not match established phenotypes
	NM_001039958.1: c.546_558delAGGGCAGGGGCGAG; p.Gln184Argfs	Homozygous	<i>MESP2</i>	Spondylocostal dysostosis 2, autosomal recessive; 608681	Low expression in cortex	absent	-	-	-	Does not match established phenotypes
	NM_001145402.1:c.2344delG; p.Glu782Argfs*2	Homozygous	<i>FAM71E2</i>	-	Low expression in cortex	absent	-	-	-	No phenotype, Low brain expression
	NM_001145402.1:c.2333delA; p.Gln778Argfs*6	Homozygous	<i>FAM71E2</i>	-	Low expression in cortex	absent	-	-	-	No phenotype, Low brain expression
	NM_152503.5: c.93- 1_93insCTTATAGACAGGGCC CCGCGCCGGCACT; p.Asn31Lysfs*10,	Homozygous	<i>MROH8</i> , <i>RPN2</i>	-	Widespread	absent	-	-	-	No phenotype
	NM_032796.3:c.455G>A; p.Arg152His	Homozygous	<i>SYAP1</i>	-	Widespread	1 het	-	Damaging 0.000	Probably damaging 0.994	No phenotype

Table 3.S3: Antibodies used in immunocytochemistry experiments.

Antibody Name	Concentration	Source	Catalogue #
Cleaved Caspase-3	1:1000	Cell Signaling	9661
Pax6	1:500	MBL	PD022
PHH3	1:500	Sigma	369A
TUJ1	1:500	Abcam	ab18207

3.4. Further findings and future work

This study, by combining clinical analysis, genomic investigations and molecular studies carried out on patient iPSCs, has confirmed that biallelic variants in *CAMSAP1* are associated with a novel syndromic neuronal migration disorder resembling a tubulinopathy. In so doing, it highlights the functional importance of stabilisation of the negative end of the microtubule in neurodevelopment, a process which currently contributes few molecular causes of neurodevelopmental disorders.

Specifically, our findings raise the likelihood that other CAMSAP molecules and their known binding partners (**Figure 3.5**) may contribute to the currently undiagnosed portion of neuronal migration disorders.

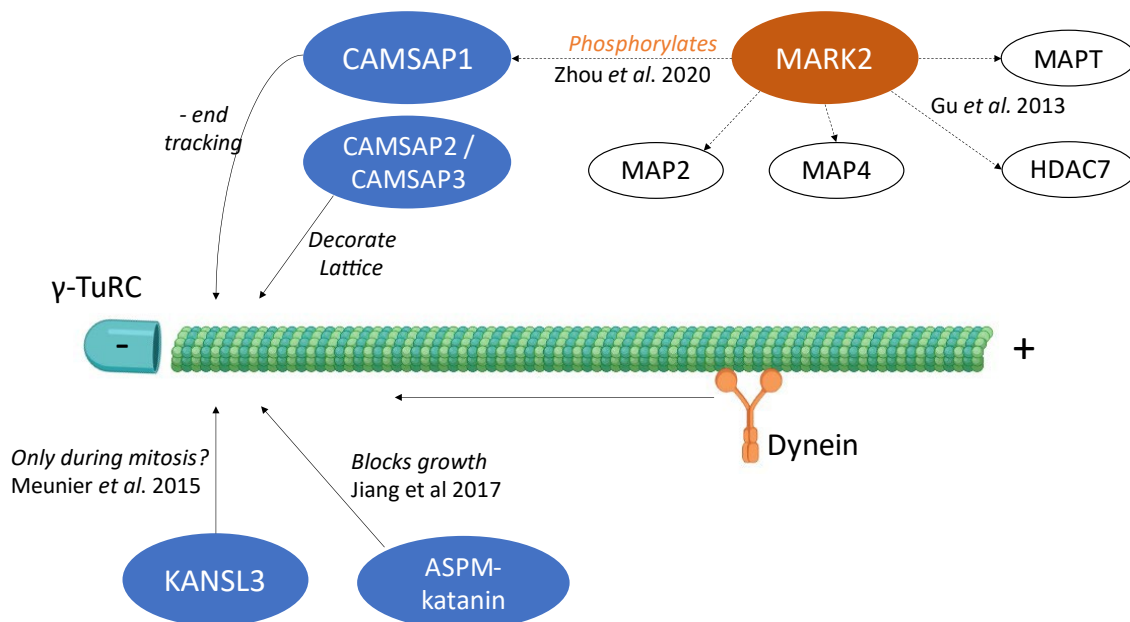


Figure 3.5: Proteins active at the minus end of the microtubule

Shown in shaded blue ovals. MARK2, known to phosphorylate CAMSAP1 is shown in an orange oval with its other targets (not minus end targeted) unshaded. This image is original work presented in this thesis. The microtubule and dynein pictograms were created in BioRender.

CAMSAP1 has two paralogues, *CAMSAP2* and *CAMSAP3*, both of which share a high degree of sequence similarity with each other and *CAMSAP1*, particularly within the family-defining microtubule-targeting CKK domain. Both *CAMSAP2/3* molecules have been documented to possess similar, but distinct roles in the stabilisation of the minus end of the microtubule (Hendershott & Vale, 2014; Jiang *et al.*, 2014; Yau *et al.*, 2014). All homologues show high expression in brain tissues, as does a key *CAMSAP*-related kinase MARK2. Whilst *CAMSAP1* and *CAMSAP2* have widespread expression, *CAMSAP3* expression is more pronounced in skin (**Appendix 7.13**).

Preliminary investigations, performed as part of this study, identified a number of individuals with undiagnosed rare diseases and mono- or bi-allelic variants in *CAMSAP3* (**Appendix 7.14**). The most convincing of these entails an individual with severe GDD, cortical atrophy and microcephaly and a *CAMSAP3* NM_020902:c.1050-2A>G canonical splicing variant *in trans* with a NM_020902:p.(Pro774Alafs*95) frameshift variant identified in the 100,000 Genomes Project Rare Disease patient cohort. The phenotype of the affected child closely mirrors the phenotypical findings of microcephaly and thin corpus callosum in a *CAMSAP3*^{-/-} mouse model (Collins *et al.*, 2019), and so may represent the first potential family identified with *CAMSAP3*-associated disease. Follow-up studies, confirming the genetic segregation of these variants and investigating their functional impact are planned, alongside a search for further affected individuals using Genematcher.

Additionally, within the DDD study, the 100,000 genomes study and ClinVar, we identified ten individuals with neurodevelopmental disorders and *de novo* predicted loss-of-function and missense variants in *MARK2* (**Appendix 7.15**). These findings suggest that heterozygous variants in *MARK2* may be a significantly under recognized cause of monogenic neurodevelopmental disorders, and our findings are contributing to a manuscript currently being prepared for submission (work led by our collaborators group).

In the future further individuals will undoubtedly be identified with the *CAMSAP1*-related neuronal migration disorder, for example via GeneMatcher (Sobreira *et al.*, 2015), through projects such as SOLVE-RD (Schüle *et al.*, 2021), in lissencephaly cohorts (Di Donato *et al.*, 2018) and through diagnostic testing internationally. An expanded cohort will enable more detailed studies into the molecular basis of this disorder alongside a clearer delineation of the genetic and phenotypic spectrum of *CAMSAP1*-related disorder. Further, the presence or absence and distribution of disease-associated missense variants, not yet observed with the *CAMSAP1*-related neuronal migration disorder, may further inform our understanding of molecular function of CAMSAP1 and the pathomolecular basis of the disorder. Functional investigations, either using iPSCs derived from affected individuals or transfection of constructs containing pathogenic variants into established, immortalised cell lines (e.g. HEK293 cells), will more precisely determine outcomes on protein levels, intracellular localisation and molecular interactions with microtubules and partner molecules such as spectrin, calmodulin and MACF1 (Akhmanova & Steinmetz, 2019). The results of these studies will enable links to be drawn between phenotype severity and the location of the variant in relation to CAMSAP1 polypeptide

domains and post translational modifications, improving understanding of the physiological role that CAMSAP1 plays during neurodevelopment.

The clinical similarities between the *CAMSAP1*-related neuronal migration disorder and the tubulinopathy disorders indicates a possible underlying shared pathomechanism. Further investigation of this may aid understanding of the role of CAMSAP1 and its interactors in normal physiology, which are currently limited. Co-immunoprecipitation (co-IP) studies in an established wild-type neuronal cell line (e.g. SH-SY5Y), with a *CAMSAP1*^{-/-} knockout achieved using CRISPR-Cas9 providing a negative control, may highlight other binding partners that could be contributing to the pathology of rare neuronal migration disorders, as is the case for *MARK2*. The role of CAMSAP1 on tubulin assembly could also be investigated using kymography (Smal *et al.*, 2010), a method that captures microtubule dynamics over time using time-lapse fluorescence microscopy in living cells. These studies would enable comparisons to those already performed as part of studies of the tubulinopathy group of disorders with which the *CAMSAP1*-related neuronal migration disorder displays phenotypic overlap (Oegema *et al.*, 2015).

A further avenue for investigation derives from the distribution of truncating variants seen in our studies to date. 5/6 pathogenic variants that we report occur in the same exon (exon 11/17), with the remaining variant identified occurring downstream, in exon 14. The unusually large size of exon 11 (encoding 807/1603 aa), coupled with stochastic variation in a relatively small cohort, may explain its overrepresentation among the currently defined disease variants. However other explanations, such as varied phenotypic severity (both

increased and decreased) associated with truncating variants in other exons or an additional autosomal dominant mechanism for some variants, may also be possible. This gene exhibits extremely high loss-of-function constraint (loss-of-function observed/expected upper bound fraction [LOEUF] = 0.24) that would be more typical for a gene showing haploinsufficiency with an early onset phenotype affecting reproductive fitness. However, there is currently no evidence of any phenotype among the parents of affected individuals. Variable splicing of *CAMSAP1*, potentially allowing some isoforms to be expressed despite the presence of a nonsense variant in one alternatively spliced exon, may be one explanation for the observation of variable phenotypes with loss-of-function variants. Whilst *CAMSAP1* does have at least three recognised transcripts (ENST00000389532.4 [MANE select - Matched Annotation from NCBI and EMBL-EBI, equivalent to NM_015447.4], ENST00000312405.6 and ENST00000409386.3) there is limited evidence in the GTEx database for brain expression of any apart from the MANE select transcript in adults. Important developmental transcripts cannot currently be excluded by GTEx, but some are expected to be added in future releases. Studies of iPSCs from an affected proband (Family 2, Individual II:1) aiming to confirm the absence of *CAMSAP1* protein were unfortunately not possible due to inadequate specificity of the *CAMSAP1* antibody. Additionally, transcriptomic studies were delayed by the loss of sample material in transit and could not be completed before the submission of this thesis. Therefore, while the pathogenicity of the *CAMSAP1* variants currently identified is in little doubt, future work aiming to measure transcript or protein levels in iPSCs derived from patient cells would be of value

and involve either design of a novel antibody specific to the CAMSAP1 N-terminus or transcriptomic studies.

3.5. References

- Akhmanova A, & Hoogenraad CC. Microtubule minus-end-targeting proteins. *Curr Biol* 2015;25(4):R162-171.
- Akhmanova, A., and Steinmetz, M. O. Microtubule minus-end regulation at a glance. *J Cell Sci*. 2019;132(11).
- Atherton J, Jiang K, Stangier MM, Luo Y, Hua S, Houben K, van Hooff JJE, Joseph A-P, Scarabelli G, Grant BJ, *et al*. A structural model for microtubule minus-end recognition and protection by CAMSAP proteins. *Nat Struct Mol Biol*. 2017;24(11):931-943.
- Ayala R, Shu T, & Tsai L-H. Trekking across the Brain: The Journey of Neuronal Migration. *Cell* 2007;128(1):29-43.
- Bahi-Buisson N, Poirier K, Fourniol F, Saillour Y, Valence S, Lebrun N, Hully M, Bianco CF, Boddaert N, Elie C, *et al*. The wide spectrum of tubulinopathies: what are the key features for the diagnosis? *Brain* 2014;137(Pt 6):1676-1700.
- Baines AJ, Bignone PA, King MD, Maggs AM, Bennett PM, Pinder JC, & Phillips GW. The CKK domain (DUF1781) binds microtubules and defines the CAMSAP/ssp4 family of animal proteins. *Mol Biol Evol* 2009;26(9):2005-2014.
- Caspi M, Atlas R, Kantor A, Sapir T, & Reiner O. Interaction between LIS1 and doublecortin, two lissencephaly gene products. *Hum Mol Genet* 2000;9(15):2205-2213.
- Centers for Disease Control and Prevention. (2020). Facts about Microcephaly. Retrieved from <https://www.cdc.gov/ncbddd/birthdefects/microcephaly.html>
- Desikan RS, & Barkovich AJ. Malformations of cortical development. *Ann Neurol* 2016;80(6):797-810.
- Di Donato N, Chiari S, Mirzaa GM, Aldinger K, Parrini E, Olds C, Barkovich AJ, Guerrini R, & Dobyns WB. Lissencephaly: Expanded imaging and clinical classification. *Am J Med Genet A* 2017;173(6):1473-1488.
- Di Donato N, Timms AE, Aldinger KA, Mirzaa GM, Bennett JT, Collins S, Olds C, Mei D, Chiari S, Carvill G, *et al*. Analysis of 17 genes detects mutations in 81% of 811 patients with lissencephaly. *Genet Med* 2018;20(11):1354-1364.
- Fry AE, Cushion TD, & Pilz DT. The genetics of lissencephaly. *Am J Med Genet C Semin Med Genet* 2014;166(2):198-210.
- Gu, GJ, Lund H, Wu D, Blokzijl A, Classon C, von Euler G, Landegren U, Sunnemark D, & Kamali-Moghaddam M. Role of individual MARK isoforms in phosphorylation of tau at Ser²⁶² in Alzheimer's disease. *Neuromolecular Med*. 2013;15(3): 458-469.
- Hebebrand M, Hüffmeier U, Trollmann R, Hehr U, Uebe S, Ekici AB, Kraus C, Krumbiegel M, Reis A, Thiel CT, *et al*. The mutational and phenotypic spectrum of TUBA1A-associated tubulinopathy. *Orphanet Journal of Rare Diseases* 2019;14(1):38.
- Hendershott MC, & Vale RD. Regulation of microtubule minus-end dynamics by CAMSAPs and Patronin. *Proc Natl Acad Sci U S A* 2014;111(16):5860-5865.
- Jiang K, Hua S, Mohan R, Grigoriev I, Yau KW, Liu Q, Katrukha EA, Altelaar AF, Heck AJ, Hoogenraad CC, *et al*. Microtubule minus-end stabilization by polymerization-driven CAMSAP deposition. *Dev Cell* 2014;28(3):295-309.

- Jiang, K., Rezabkova, L., Hua, S., Liu, Q., Capitani, G., Altelaar, A. F. M., Heck, A. J. R., Kammerer, R. A., Steinmetz, M. O., & Akhmanova, A. Microtubule minus-end regulation at spindle poles by an ASPM-katanin complex. *Nat Cell Biol.* 2017;19(5): 480-492.
- King MD, Phillips GW, Bignone PA, Hayes NV, Pinder JC, & Baines AJ. A conserved sequence in calmodulin regulated spectrin-associated protein 1 links its interaction with spectrin and calmodulin to neurite outgrowth. *J Neurochem* 2014;128(3):391-402.
- Kumar RA, Pilz DT, Babatz TD, Cushion TD, Harvey K, Topf M, Yates L, Robb S, Uyanik G, Mancini GM, *et al.* (2010). TUBA1A mutations cause wide spectrum lissencephaly (smooth brain) and suggest that multiple neuronal migration pathways converge on alpha tubulins. In *Hum Mol Genet* (Vol. 19, pp. 2817-2827).
- Laver TW, De Franco E, Johnson MB, Patel KA, Ellard S, Weedon MN, Flanagan SE, & Wakeling MN. SavvyCNV: Genome-wide CNV calling from off-target reads. *PLOS Comp Biol* 2022;18(3):e1009940.
- Liegel RP, Finnerty E, Blizzard L, DiStasio A, Hufnagel RB, Saal HM, Sund KL, Prows CA, & Stottmann RW. Using human sequencing to guide craniofacial research. *Genesis* 2019;57(1):e23259.
- Meunier S, Shvedunova M, Van Nguyen N, Avila L, Vernos I, & Akhtar A. An epigenetic regulator emerges as microtubule minus-end binding and stabilizing factor in mitosis. *Nat Commun.* 2015;6(1):7889.
- Mitani T, Isikay S, Gezdirici A, Gulec EY, Punetha J, Fatih JM, Herman I, Akay G, Du H, Calame DG, *et al.* High prevalence of multilocus pathogenic variation in neurodevelopmental disorders in the Turkish population. *Am J Hum Genet* 2021;108(10):1981-2005.
- Oegema R, Barakat TS, Wilke M, Stouffs K, Amrom D, Aronica E, Bahi-Buisson N, Conti V, Fry AE, Geis T, *et al.* International consensus recommendations on the diagnostic work-up for malformations of cortical development. *Nat Rev Neurol* 2020;16(11):618-635.
- Oegema R, Cushion, TD, Phelps IG., Chung SK, Dempsey JC, Collins S, Mullins JG, Dudding T, Gill H, Green AJ, *et al.* Recognizable cerebellar dysplasia associated with mutations in multiple tubulin genes. *Hum Mol Genet.* 2015;24(18):5313-5325.
- Pavlova GA, Razuvaeva AV, Popova JV, Andreyeva EN, Yarinich LA, Lebedev MO, Pellacani C, Bonaccorsi S, Somma MP, Gatti M, *et al.* The role of Patronin in Drosophila mitosis. *BMC Molecular and Cell Biology* 2019;20(1):7.
- Pehlivan DY, Bayram N, Gunes Z, Coban Akdemir A, Shukla T, Bierhals B, Tabakci Y, Sahin, A. Gezdirici JM, Fatih EY, *et al.* The Genomics of Arthrogyryposis, a Complex Trait: Candidate Genes and Further Evidence for Oligogenic Inheritance. *Am J Hum Genet* 2019;105(1):132-150.
- Romaniello R, Arrigoni F, Fry AE, Bassi MT, Rees MI, Borgatti R, Pilz DT, & Cushion TD. Tubulin genes and malformations of cortical development. *Eur J Med Genet* 2018;61(12):744-754.
- Saneto RP, Lee IC, Koenig MK, Bao X, Weng SW, Naviaux RK and Wong LJ (2010). POLG DNA testing as an emerging standard of care before instituting valproic acid therapy for pediatric seizure disorders. *Seizure* 19(3): 140-146.

- Schüle R, Timmann D, Erasmus CE, Reichbauer J, Wayand M, van de Warrenburg B, Schöls L, Wilke C, Bevot A, Zuchner S *et al.* Solving unsolved rare neurological diseases—a SolveRD viewpoint. *Eur J Hum Genet.* 2021;29(9):1332-1336.
- Severino M, Geraldo AF, Utz N, Tortora D, Pogledic I, Klonowski W, Triulzi F, Arrigoni F, Mankad K, Leventer RJ, *et al.* Definitions and classification of malformations of cortical development: practical guidelines. *Brain* 2020;143(10):2874-2894.
- Smal I, Grigoriev I, Akhmanova A, Niessen WJ, & Meijering E. Microtubule dynamics analysis using kymographs and variable-rate particle filters. *IEEE Transactions on Image Processing.* 2010;19(7),1861-1876.
- Smith DS, Niethammer M, Ayala R, Zhou Y, Gambello MJ, Wynshaw-Boris A, & Tsai LH. Regulation of cytoplasmic dynein behaviour and microtubule organization by mammalian Lis1. *Nat Cell Biol* 2000;2(11):767-775.
- Sobreira N, Schiettecatte F, Valle D, & Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat* 2015;36(10):928-930.
- Splinter D, Razafsky DS, Schlager MA, Serra-Marques A, Grigoriev I, Demmers J, Keijzer N, Jiang K, Poser I, Hyman AA, *et al.* BICD2, dynactin, and LIS1 cooperate in regulating dynein recruitment to cellular structures. *Mol Biol Cell* 2012;23(21):4226-4241.
- Toya M, Kobayashi S, Kawasaki M, Shioi G, Kaneko M, Ishiuchi T, Misaki K, Meng W, & Takeichi M. CAMSAP3 orients the apical-to-basal polarity of microtubule arrays in epithelial cells. *Proc Natl Acad Sci U S A* 2016;113(2):332-337.
- Wong LJ, Naviaux RK, Brunetti-Pierri N, Zhang Q, Schmitt ES, Truong C, Milone M, Cohen BH, B, Wical B, Ganesh J, *et al.* Molecular and clinical genetics of mitochondrial diseases due to POLG mutations. *Hum Mutat* 2008;29(9): E150-172.
- Yamamoto M, Yoshimura K, Kitada M, Nakahara J, Seiwa C, Ueki T, Shimoda Y, Ishige A, Watanabe K, & Asou H. A new monoclonal antibody, A3B10, specific for astrocyte-lineage cells recognizes calmodulin-regulated spectrin-associated protein 1 (Camsap1). *J Neurosci Res* 2009;87(2):503-513.
- Yau KW, van Beuningen SF, Cunha-Ferreira I, Cloin BM, van Battum EY, Will L, Schätzle P, Tas RP, van Krugten J, Katrukha EA, *et al.* Microtubule minus-end binding protein CAMSAP2 controls axon specification and dendrite development. *Neuron* 2014;82(5):1058-1073.
- Yoshioka N, Asou H, Hisanaga S, & Kawano H. The astrocytic lineage marker calmodulin-regulated spectrin-associated protein 1 (Camsap1): phenotypic heterogeneity of newly born Camsap1-expressing cells in injured mouse brain. *J Comp Neurol* 2012;520(6):1301-1317.
- Zhou Z, Xu H, Li Y, Yang M, Zhang R, Shiraishi A, Kiyonari H, Liang X, Huang X, Wang Y, *et al.* CAMSAP1 breaks the homeostatic microtubule network to instruct neuronal polarity. *Proc Natl Acad Sci U S A* 2020;117(36):22193-22203.

4

SLC4A10 variants impair GABAergic transmission and CSF secretion causing a recognizable neurodevelopmental disorder in humans and mice.

James Fasham*, Antje K. Huebner*, Lutz Liebmann*, Reham Khalaf-Nazzal*,
Reza Maroofian, Nderim Kryeziu, Saskia B. Wortmann, Joseph S. Leslie, Nishanka
Ubeyratna, Grazia M.S. Mancini, Marjon van Slegtenhorst, Martina Wilke, Tobias
B. Haack, Hanan Shamseldin, Joseph G. Gleeson, Mohamed Almuhaizea, Imad
Dweikat, Bassam Abu-Libdeh, Muhannad Daana, Matthew N Wakeling, Lucy
McGavin, Peter D. Turnpenny, Fowzan S Alkuraya, Henry Houlden, Kai Kaila,
Andrew H. Crosby†, Emma L. Baple†, Christian A. Hübnert†

*Revised manuscript under review with **Brain**.*

4.1. Acknowledgements of co-authors and contributions to the paper

This study was undertaken as part of the “Stories of Hope” Palestinian translational genomic research programme, conceived, designed and led by Dr Reham Khalaf-Nazzal (Arab American University), Prof Peter Turnpenny, Prof Emma Baple and Prof Andrew Crosby.

Where specific experiments or analyses were performed by study collaborators or members of my supervisors’ group other than myself, then these are detailed below.

Clinical and genomic studies

Clinical data were obtained by local clinical care providers using a standardised proforma that I designed. Phenotype data for Family 1 was provided by Dr Reham Khalaf-Nazzal. Support with interpretation of clinical data was provided by Prof Baple. Dr Lucy McGavin (Derriford Hospital, Plymouth) reviewed available neuroradiological data and assisted me with interpretation of the findings.

Technical assistance with validation and cosegregation studies for the *SLC4A10* variants identified was provided by Mr Joseph Leslie and Mr Nishanka Ubeyratna (University of Exeter).

I performed analysis of genome data from Family 1. Analysis of exome data from families 2-5 was undertaken by our academic and clinical collaborators. I reviewed

the *SLC4A10* variants identified and any genomic variants that could not be excluded during those initial analyses.

Hanan Shamseldin and Mohamed Almuhaizea (King Faisal Specialist Hospital and Research Center) performed the RT-PCR experiments.

Cell and mouse studies

Cell and mouse studies were carried out in collaboration with Prof Christian Hübner (Jena University Hospital). Antje Huebner, Lutz Liebmann and Nderim Kryeziu undertook this work and I assisted with analyses of the results.

Manuscript draft and revision

I wrote and revised the manuscript with Prof Baple, Prof Crosby and Prof Hübner. All the co-authors provided comments and feedback prior to submission.

A total of 5 main and 11 supplementary figures were included in the submitted manuscript. Figures 4.3 - 4.5 were produced by Antje Huebner, Lutz Liebmann and Nderim Kryeziu, as were supplementary figures 4.S4, 4.S6 4.S10. Figure 4.S3 was produced by Hanan Shamseldin and Mohamed Almuhaizea.

4.2. Manuscript

Summary

SLC4A10 is a plasma-membrane bound transporter, which utilises the sodium gradient to drive cellular bicarbonate uptake, thus increasing intracellular pH. In the mammalian brain, *SLC4A10* is expressed in neurons and choroid plexus epithelial cells and *Slc4a10*^{-/-} mice display small lateral brain ventricles and mild behavioural abnormalities. Here we show that biallelic *SLC4A10* loss-of-function variants cause a neurodevelopmental disorder in humans, with ID, behavioural abnormalities, increased susceptibility to seizures and brain malformations including microcephaly, characteristic slit-like ventricles and corpus callosum abnormalities, paralleling findings in *Slc4a10*^{-/-} mice. Our molecular studies localise SLC4A10 to inhibitory presynapses and show that inhibitory neurotransmitter GABA release is compromised in *Slc4a10*^{-/-} mice, while excitatory neurotransmitter glutamate release is preserved. As GABAergic inhibitory interneurons gate signal flow and contribute to the timing and synchronisation of cortical oscillations, we propose that compromised synaptic inhibition stemming from SLC4A10 loss of function likely contributes to the pathophysiology of this disorder.

Introduction

A large variety of molecules involved in neuronal signalling, including ligand- and voltage-gated channels, show a remarkable sensitivity to changes in the intracellular and extracellular pH (Pasternack, Smirnov & Kaila 1996; Traynelis & Cull-Candy 1990; Tombaugh & Somjen 1996; Waldmann *et al.* 1997). As a rule, the excitability of neuronal networks is enhanced by alkalosis and suppressed by acidosis (Bocker *et al.* 2019; Chesler 2003; Feghhi *et al.* 2021; Ruffin *et al.* 2014; Sinning & Hübner 2013; Stawarski *et al.* 2020) which suggests a fundamental evolutionary role for pH as a neuromodulator during physiological and pathophysiological conditions. Numerous studies have provided evidence for mechanisms that control pH dynamics and actions in microdomains within (Burette *et al.* 2012; Sinning *et al.* 2011; Salameh, Hübner & Boron 2017) and outside (Stawarski *et al.* 2020) brain cells based on the heterogeneous spatial patterns of expression of both pH-sensitive and pH-regulatory proteins, including plasmalemmal Na^+/H^+ exchangers (Orlowski & Grinstein 2011), HCO_3^- transporters (Alper *et al.* 2013; Romero *et al.* 2013) as well as intra- and extracellular carbonic anhydrase isoforms (Pasternack, Smirnov & Kaila 1996).

In mammals, members of the SLC4 (Romero *et al.* 2013) and SLC26 (Alper *et al.* 2013) gene families have been identified as bicarbonate (HCO_3^-) transporters, many of which are associated with monogenic human diseases including distal renal tubular acidosis, haemolytic anaemia, corneal dystrophy, glaucoma and cataracts (Romero *et al.* 2013), as well as chondrodysplasia, chloride diarrhoea,

and hearing loss (Wang *et al.* 2020). The SLC4 family includes sodium-independent $\text{Cl}^-/\text{HCO}_3^-$ exchangers, electrogenic and electroneutral $\text{Na}^+-\text{HCO}_3^-$ cotransporters and sodium-driven $\text{Cl}^-/\text{HCO}_3^-$ exchangers that mediate HCO_3^- transport across the plasma membrane (**Table 4.S1**) (Romero *et al.* 2013). SLC4A10 utilises the transmembrane gradient of Na^+ to drive cellular net uptake of HCO_3^- , and thus increases intracellular pH. Both cytoplasmic and membrane-bound carbonic anhydrases are involved in the supply of HCO_3^- and may thus increase transport rates (McMurtrie *et al.* 2004). To which extent this is relevant for SLC4A10 mediated transport is yet unclear. Some controversy exists as to whether it acts as an electroneutral $\text{Na}^+-\text{HCO}_3^-$ cotransporter (NBCn2), or a sodium-coupled $\text{Cl}^-/\text{HCO}_3^-$ exchanger (NCBE) under physiological conditions (Damkier, Aalkjaer & Praetorius 2010; Parker *et al.* 2008). The expression of *SLC4A10* is predominantly neuronal, but it is also expressed in choroid plexus epithelia (Praetorius, Nejsum & Nielsen 2004; Jacobs *et al.* 2008) and in inner ear fibrocytes (Huebner *et al.* 2019). Mice deficient for SLC4A10 show a reduced brain ventricle size suggesting a role in transepithelial electrolyte transport and production of cerebrospinal fluid (CSF) (Jacobs *et al.* 2008). Although neuronal excitability was enhanced *in vitro* (Sinning *et al.* 2015), the experimental seizure threshold was paradoxically increased *in vivo* (Jacobs *et al.* 2008) and spontaneous seizures were not observed.

In humans, heterozygous genomic deletions comprising all or part of *SLC4A10* have been linked with autism spectrum disorder with additional features such as impaired motor and language skills or epilepsy (Belengeanu *et al.* 2014; Gurnett *et al.* 2008; Krepischi *et al.* 2010; Sebat *et al.* 2007). The causal relevance of these

genomic alterations is, however, unclear, as the interpretation of these findings is complicated by contiguous gene deletion. Here, we provide clinical, genetic, functional and mouse-model evidence to determine that autosomal recessive SLC4A10 loss of function results in ID with striking radiological abnormalities of the lateral ventricles, closely mirroring findings in *Slc4a10* knockout (KO) mice. As SLC4A10 localises to inhibitory presynapses and its disruption compromises γ -aminobutyric acid (GABA) release, we propose that alterations of the GABAergic system likely contribute to the pathomechanistic basis of this neurodevelopmental disorder.

Materials and Methods

Clinical studies

All families were recruited with written informed consent according to international guidelines, including the Declaration of Helsinki, and regional ethical approvals (Palestinian Health Research Council PHRC/HC/518/19, Technische Universität München, Muenchen Exome Seq.: 5360/12 S, King Faisal Specialist Hospital & Research Centre Research Advisory Council # 2121053, Erasmus Medical Centre Medical Ethics Review Committee 2012-387, IRB protocol number 150765). Affected individuals were examined and investigated by local clinicians according to routine clinical standards relevant to their clinical presentation.

Genetic studies

DNA and RNA were extracted from blood/buccal samples using standard techniques. In all five families whole genome sequencing (WGS) (Family 1) or whole-exome sequencing (WES) (Families 2-5) was undertaken to identify the cause of disease. Family pedigrees illustrating the relationships of affected and unaffected individuals in this study are shown in **Fig. 4.1**. Unless otherwise specified, genomic variants were filtered based on call quality, predicted consequence, segregation with the disease phenotype and allele frequency in population databases (variants with a frequency of >0.1% and/or present in >1 homozygous individual in gnomAD v2.1.1, v3.1.1 or in-house databases were excluded). Homozygous, compound heterozygous, X chromosome and *de novo* (when trio sequencing undertaken) variants present in exons or within ± 6 nucleotides in the intron that remained after filtering, were assessed for clinical correlation with the affected individual(s) phenotype.

In **Family 1**, WGS was performed (BGI, Hong Kong) on DNA from two affected individuals (Family 1; II:1 and II:2). Reads were aligned (BWA-MEM v0.7.15), mate-pairs fixed, and duplicates removed (Picard v2.7.1), InDel realignment/base quality recalibration (GATK v3.7.0), SNV/InDel detection (GATK HaplotypeCaller), annotation (Alamut v1.11), and read depth ascertained using an in-house pipeline. This conforms to GATK best practices. CNVs were detected using SavvyCNV (Laver *et al.* 2022).

In **Family 2**, DNA from the proband (II:1) and both unaffected parents underwent trio WES (Illumina) at Technical University München/Helmholtz Institute Neuherberg using the SureSelect50Mbv5 capture, as previously described (Wagner *et al.* 2019; Mayr *et al.* 2012).

In **Family 3**, trio WES of a single affected individual (II:2) and both parents was undertaken as previously described (ID: 17-4393) (Monies *et al.* 2019). In **Family 4** WES was performed on the two affected brothers at University of California San Diego (UCSD) using methods previously described (Novarino *et al.* 2014), with recessive variants within regions of homozygosity prioritised given the consanguineous nature of the family. In **Family 5**, trio WES was performed on both affected siblings and their two unaffected parents (four individuals in total) using Agilent SureSelect Target Clinical Research Exome V2 (Agilent Technologies, Santa Clara, CA, USA). Sequencing (paired-end 150 bp) was performed by the Illumina HiSeq 4000 platform (Illumina, San Diego, CA, USA, outsourced). Data were demultiplexed by Illumina Software CASAVA (Consensus Assessment of Sequence And Variation). Reads are mapped to the genome (build hg19/GRCh37) with the program BWA (reference: <http://bio-bwa.sourceforge.net>). Variants were detected with the Genome Analysis Toolkit (reference: <http://www.broadinstitute.org/gatk/>). Variants were filtered with the Cartagenia/Alissa Interpret software package (Agilent technologies) on quality (read depth ≥ 10), frequency in databases ($\geq 1\%$ in 200 alleles in dbSNP, ESP6500, the 1000 Genome project or the ExAC database) and location (within an exon or first/last 10 bp of introns).

In Family 1 unique primers for ddPCR (QX200 AutoDG Droplet Digital PCR System - Bio Rad, CA, USA) were designed for confirmation and cosegregation of the exon 5-11 *SLC4A10* deletion [NM_001178015: c.417_1341del]. In addition to two primers within the deletion (within exons 5 and 10), probes included an exon 5' to the deletion (Exon 4), an exon 3' to the deletion (Exon 11) and a housekeeping gene control (*RPP30*). Primer sequences are provided (**Fig. 4.S1**).

In Family 3 reverse transcription PCR (RT-PCR), using standard techniques, was undertaken on lymphoblast cell lines derived from affected and controls individuals to confirm the transcriptional outcome of the *SLC4A10* NM_001178015:c.2863-2A>C variant. RNA was extracted using RNeasy kit (QIAGEN-Catalogue # 74104) as per the manufacturer's protocol. cDNA was generated from 1 ug of RNA via iScript Select cDNA Synthesis kit (Bio-rad). Primers that cover exons 19-24 of *SLC4A10* transcript were used for RT-PCR to check for difference in splicing between affected and control lymphoblast.

In Families 2-4, dideoxy sequencing confirmation and cosegregation of single nucleotide *SLC4A10* variants was performed using standard techniques.

Cellular studies

Cloning

The human *SLC4A10* cDNA was cloned by PCR from a human cDNA library and subcloned into the pBI-CMV4 vector (Clon-tech #PT4443-5), a mammalian

bidirectional expression vector designed to constitutively express a protein of interest and DsRed2, a human codon-optimised variant of the *Discosoma sp.* red fluorescent protein. Disease associated SNVs were inserted by site-directed mutagenesis and verified by sequencing.

Cell Culture

N2a cells were cultured at 37°C with Dulbecco's Eagle's Minimum Essential Medium (DMEM) (Gibco #31966-021) supplemented with fetal bovine serum to a final concentration of 10% and 2% penicillin/streptomycin (Gibco). N2a cells were transfected with Lipofectamine 3000 (Invitrogen) according to the manufacturer's instructions.

For staining, cells were fixed with 4% PFA in PBS for 10 min and subsequently washed. Cells were stained with wheat germ agglutinin (WGA) coupled to Biotin (Biozol #B-1025) at a dilution of 1:500 and a polyclonal rabbit anti-SLC4A10 antibody (1:500) (Jacobs *et al.* 2008) at 4°C overnight. The secondary antibodies we used were a Streptavidin-Alexa Fluor 488 conjugate (1:1000, Invitrogen #S32354) and an Alexa Fluor 546-coupled goat anti-rabbit antibody (1:1000, Invitrogen). Analysis was done with a confocal microscope in the Airyscan mode (LSM 880, Zeiss). The plasma membrane region (PMR) was determined as the WGA-labelled cell rim.

Intracellular pH recordings

48h after transfection, the intracellular pH (pH_i) was measured using the ratiometric 2',7'-Bis(2-carboxyethyl)-5(6)-carboxyfluorescein (BCECF, Molecular Probes) fluorescent dye. Cells were washed with bicarbonate-buffered solution containing (in mM): 99 NaCl, 20 Na-gluconate, 5 KCl, 1 MgSO₄, 1.5 CaCl₂, 25 NaHCO₃ and 10 glucose. Coverslips were transferred to a heated perfusion chamber (Chamlide EC; Live Cell Instruments, 37 °C), which was mounted at an Axio Observer.Z1 microscope (Zeiss). An image was acquired for the red fluorescent protein (RFP) channel to identify transfected cells. Thereafter, BCECF-AM was added to a final concentration of 4 μ M and incubated for 10 min. The cells were superfused with bicarbonate-buffered solution at a linear flow rate of 2.5 ml/s. Emitted light of 510-535 nm was recorded after alternating excitation at 495 nm and 440 nm every 10 s and captured through a 10x objective with a charged-coupled device (CCD) camera (AxioCam MRm; Zeiss). The steady state pH_i was recorded for 5 min. Then 5 μ M 5-(N-ethyl-N-isopropyl) amiloride (EIPA) was added to the perfusion buffer to block Na⁺/H⁺ exchange activity and the pH recorded for another 5 min. The superfusion was then switched to bicarbonate-buffered solution containing 5 μ M EIPA and 20 mM sodium propionate instead of 20 mM Na⁺-gluconate for 5 min. After the propionate pulse cells were superfused again with the former used bicarbonate-buffered solution supplemented with 5 μ M EIPA. The cytoplasmic pH recovery was recorded during superfusion with 20 mM sodium propionate containing bicarbonate-buffered solution. For each coverslip more than 12 neighbouring transfected and non-transfected cells were analysed and data from

different coverslips were averaged. At the end of each experiment, a calibration was done with buffers between pH 6.5 and 7.5 (in mM: 135 KCl, 20 N-methyl-D-glucamine, 4 MgSO₄, 10 glucose, 30 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid [HEPES], 10 μM nigericin). A linear regression was calculated from the multipoint calibration curve, and F₄₉₅/F₄₄₀ ratio was converted into pH_i values.

Mouse studies

The generation of *Slc4a10*^{-/-} mice from a 129SvJ embryonic stem cell line was described previously (Jacobs *et al.* 2008). All experiments were conducted according to the German Law on the Protection of Animals and the corresponding European Communities Council Directive of November 24, 1986 (86/609/EEC) and were approved by the Thüringer Landesamt für Lebensmittelsicherheit und Verbraucherschutz (Thuringia State Office for Food Safety and Consumer Protection) under the registration number 02-001/13. Mice were group-housed on a 12-h light-dark cycle and fed with food and water *ad libitum*. If not indicated otherwise, the experiments started when the animals were 3 to 4 months old and weighed 25–35 g. Tests were performed during the light phase between 10:00 a.m. and 5:00 p.m.

The 2-object novel object recognition (NOR) task was used to evaluate recognition memory in rodents in 12-month-old wild-type and knockout mice of both sexes. During habituation, the animals were allowed to explore an open field arena on two

days with one day interval in between. One week after habituation, the animals were again exposed to the familiar arena but with two identical glass bottles with a blue cap placed at an equal distance. Four hours later, the mice were placed in the arena, after one glass bottle was replaced by a tower of yellow and green Lego bricks of the same height. Mice were recorded with a CCD camera (Panasonic) for ten minutes. The time spent exploring each object, the number of visits and the exploring time per visit were analysed off-line with Microsoft Windows Movie Maker. The difference score (time exploring novel object - time exploring familiar object) as well as the discrimination ratio (time exploring novel object / total time spent with both objects) were calculated.

Histology and immunohistochemistry

Hematoxylin and eosin (HE) staining followed standard protocols (Carl Roth, Germany). For immunofluorescence, brains of 2-to-3-month-old wild-type mice were prepared and fixed as described previously (Sinning *et al.* 2011). Free-floating cryosections (50 μm) were stained with a polyclonal rabbit anti-NeuN antibody (1:1000, Abcam, ab104225) or polyclonal rabbit anti-SLC4A10 antibody (Jacobs *et al.* 2008). For co-staining, the following primary antibodies were used: polyclonal guinea pig anti-vesicular GABA transporter (VGAT, 1:250, Synaptic Systems), polyclonal guinea pig anti-vesicular glutamate transporter 1 (VGLUT1, 1:500, Synaptic Systems). Alexa Fluor 488- and 546-coupled goat anti-rabbit and goat anti-guinea pig antibodies were used as secondary antibodies (1:1000, Invitrogen). Cell nuclei were stained by 4,6-diamidino-2-phenylindole (DAPI) (1

$\mu\text{g/ml}$, Sigma-Aldrich). Analysis was performed with a confocal microscope in the Airyscan mode (LSM 880, Zeiss). To quantify the degree of co-localisation, planes were selected with an optimised signal-to-noise ratio using the range indicator and adjusting it to the linear, non-saturated range. Images were taken randomly from the hippocampal CA1 region (stratum radiatum or stratum pyramidale $225 \times 225 \mu\text{m}$) of four different wild-type brains. The relative area of colocalisation was determined by scatter plot analysis using the colocalisation module of ZEN (Release 4.8.2, Zeiss) according to the Costes method (Dunn, Kamocka & McDonald 2011).

Slice preparation for electrophysiological recordings

Mice between two and three months of age were decapitated and the brain was removed from the skull and chilled (at $\sim 4^\circ\text{C}$) in artificial CSF (aCSF) containing (in mmol/L): 120 NaCl, 3 KCl, 5 MgSO_4 , 1.25 NaH_2PO_4 , 0.2 CaCl_2 , 10 d-glucose, and 25 NaHCO_3 , gassed with 95% O_2 -5% CO_2 . Horizontal brain slices ($350 \mu\text{m}$) including the hippocampus were prepared with a vibroslicer (Leica VT 1200S). Slices were stored at room temperature for at least 1h before use in recording aCSF containing (in mmol/L): 120 NaCl, 3 KCl, 1.3 MgSO_4 , 1.25 NaH_2PO_4 , 2.5 CaCl_2 , 10 d-glucose, and 25 NaHCO_3 , gassed with 95% O_2 -5% CO_2 , as described previously (Liebmann *et al.* 2008).

Patch clamp recordings

One slice at a time was placed in a recording chamber mounted on an upright microscope (Axio Examiner.A1; Zeiss) with differential interference contrast, $\times 40$ water-immersion objective, and $\times 10$ ocular to identify cells. The slices were continuously perfused with aCSF (flow rate 2–3 ml/min, room temperature, pH 7.3) consisting of (in mM): 120 NaCl, 3 KCl, 1.3 MgCl₂, 2.5 CaCl₂, 25 NaHCO₃, 1.25 KH₂PO₄, and 10 d-glucose.

For whole-cell recordings patch pipettes with an impedance of $\sim 3\text{--}4$ M Ω were pulled from borosilicate glass (OD 1.5 mm; Science Products) with a micropipette puller (P-97, Sutter Instrument) and filled with intracellular solutions for miniature excitatory postsynaptic currents (mEPSC) or miniature inhibitory postsynaptic currents (mIPSC) recordings, respectively.

Pyramidal neurons of the CA1 and CA3 were selected for recording if they displayed a pyramidal-shaped cell body. Patched cells were voltage clamped. Only cells with a resting membrane potential below -55 mV and an access resistance < 15 M Ω were included. Therefore, it was not necessary to compensate for the series resistance. Voltages were corrected for liquid junction potentials or series resistance. Signals were recorded using a patch-clamp amplifier (MultiClamp 700B; Axon Instruments). Responses were filtered at 5 kHz and digitised at 20 kHz (Digidata 1440A; Axon Instruments). All data were acquired, stored, and analysed on a personal computer using pClamp 10 (Axon Instruments).

mEPSCs and mIPSCs were recorded at a holding potential of -70 mV for at least 5 min in aCSF. mEPSCs were isolated by adding tetrodotoxin (0.5 μ M, Tocris Bioscience) to block action potential-induced glutamate release and bicuculline methiodide (20 μ M, Biomol) to block GABA_A responses. dl-APV [DL-2-Amino-5-phosphonopentanoic acid] (30 μ M) was added to suppress N-methyl-D-aspartate (NMDA) currents. The pipette solution contained the following (in mM): 120 CsMeSO₄, 17.5 CsCl, 10 HEPES, 5 1,2-bis(o-aminophenoxy)ethane-N,N,N',N'-tetraacetic acid (BAPTA), 2 Mg-ATP, 0.5 Na-GTP, 10 QX-314 [*N*-(2,6-dimethylphenylcarbamoylmethyl) triethylammonium bromide], pH 7.3, adjusted with CsOH.

Recordings of mIPSCs were performed using a CsCl-based intracellular solution (in mM): 122 CsCl, 8 NaCl, 0.2 MgCl₂, 10 HEPES, 2 egtazic acid (EGTA), 2 Mg-ATP, 0.5 Na-GTP, 10 QX-314 [*N*-(2,6-dimethylphenylcarbamoylmethyl)triethylammonium bromide], pH adjusted to 7.3 with CsOH. dl-APV (30 μ M), cyanquixaline (10 μ M) and tetrodotoxin (0.5 μ M) were added to the perfusate. Recordings of sIPSCs were performed in the absence of tetrodotoxin. In a subset of mIPSC experiments, 20 mM NaCl was substituted by the weak base trimethylamine chloride (TriMA; Sigma-Aldrich) to raise pH_i. In another subset of mIPSC experiments, 20 mM NaCl was substituted by the weak acid sodium propionate (Sigma-Aldrich) to lower pH_i. After a baseline recording of 5 min, the regular aCSF was replaced by aCSF with either TriMA or sodium propionate, and mIPSCs were recorded for further 5 min.

For mIPSC recordings in bicarbonate-free extracellular solution we used (in mM): 130 NaCl, 3 KCl, 1.3 MgSO₄, 1.25 NaH₂PO₄, 2.5 CaCl₂, 10 D-glucose, 10 HEPES, gassed with O₂, pH 7.3 with NaOH.

The following parameters of mEPSCs and mIPSCs/sIPSCs were determined: frequency, peak amplitude, time constant of decay (T_{decay}), half-width, and electrical charge transfer. Data analysis was performed off-line with the detection threshold levels set to 5 pA for mEPSCs and mIPSCs because of the peak-to-peak noise determined under AMPA/NMDA receptor and GABA_A receptor blockade.

Statistical analysis

Data are presented as mean +/- standard error of the mean (SEM) if not indicated otherwise. Data were tested for normal distribution using the Kolmogorov–Smirnov test (KS-test). Comparison of normally distributed data from two experimental groups were performed with the parametric two-tailed Student's t test. In experiments that included repeated measurements, differences between groups were tested by repeated-measures ANOVA. Cumulative distributions were tested using the Kolmogorov–Smirnov test. Significance was considered at p values <0.05.

Data availability

Full WGS and WES sequencing data are not available due to reasons of confidentiality; anonymised variant data will be made available on reasonable

request. The authors declare that all other data are contained within the manuscript and supplemental materials. *SLC4A10* variants have been deposited in ClinVar with submission number SUB11166749.

Results

Genetic analysis

We initially investigated the cause of disease in two male Palestinian siblings (aged 7 and 8 years) affected by a syndromic form of severe ID, with behaviours associated with autism spectrum disorder, slit ventricles and subtle craniofacial dysmorphism (Family 1). The younger child was microcephalic, with an OFC of 3.4 standard deviation scores below the mean (-3.4 SDS), whereas his older brother had an OFC of -2.3 SDS. To define the genetic cause of disease WGS was performed on DNA from both affected children (Family 1; II:1 and II:2). Filtering of WGS data using standard metrics described above identified a single standout candidate variant, a shared, homozygous out-of-frame deletion of exons 5-11 of *SLC4A10* [Chr2(GRCh38):g.161846109-161895992del;NM_001178015:c.417_1341del p.(Trp140Argfs*39)] clearly visible on genome sequencing (**Fig. 4.S1**) and predicted to result in nonsense-mediated decay and absence of the *SLC4A10* protein. The variant was confirmed using ddPCR as an orthologous method and found to cosegregate as expected for an autosomal recessive disorder (**Fig. 4.S2**).

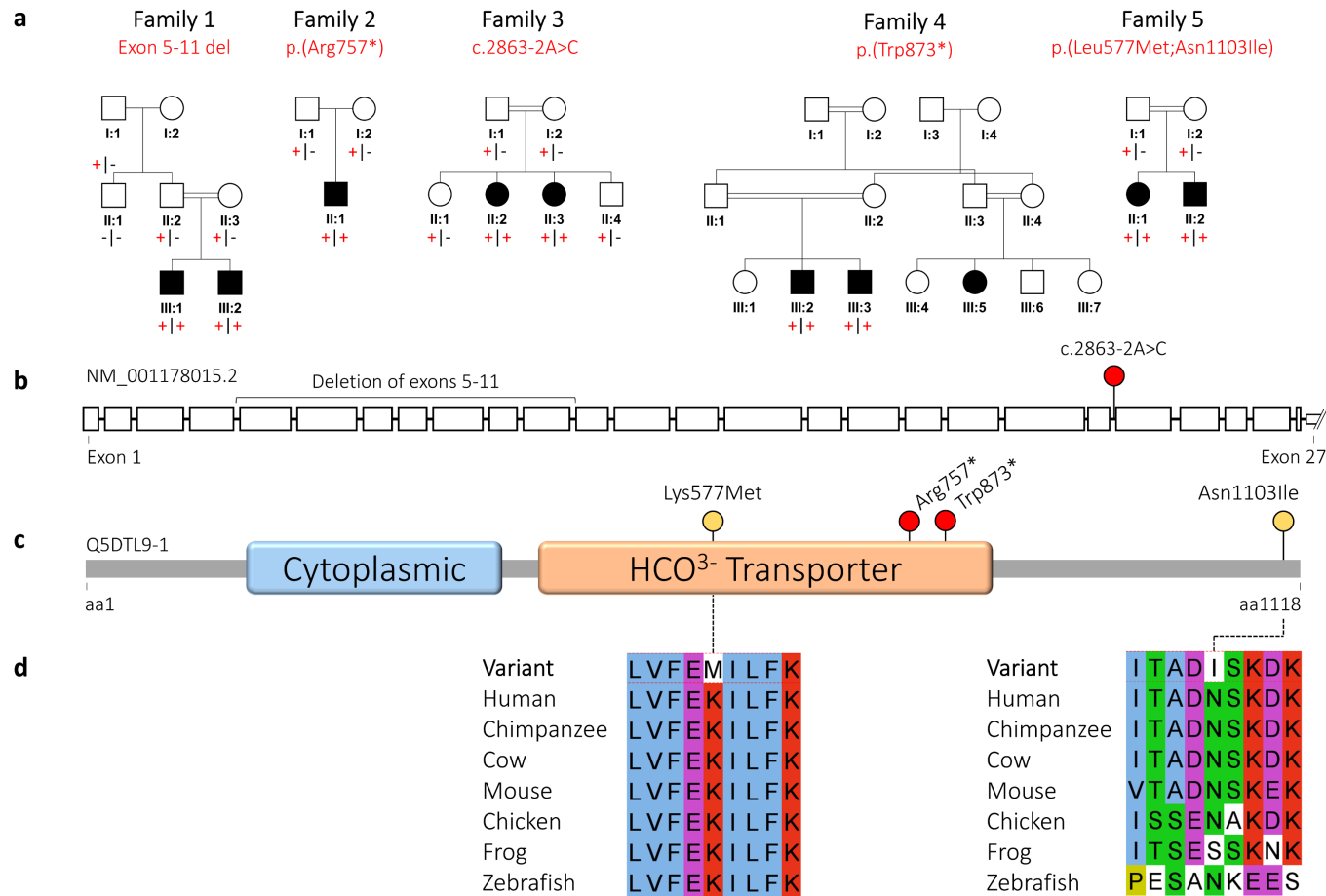


Figure 4.1: Family pedigrees and biallelic *SLC4A10* variants

a) Simplified family pedigrees for individuals affected with *SLC4A10*-related neurodevelopmental disorder, showing autosomal recessive segregation of *SLC4A10* variants. Co-segregation confirmed in other family members as indicated, in each case '+' indicating variant allele and '-' indicating wild type allele. b) Simplified *SLC4A10* exon structure (NM_001178015.2) showing location of the multi-exon deletion identified in Individuals III:1 and III:2 (Family 1) and the splicing variant (c.2863-2A>C). Only a part of the large, non-coding, UTR Exon 27 is shown. c) Simplified *SLC4A10* protein structure (Q5DTL9-1) showing location of missense (yellow) and predicted loss-of-function (red) variants in relation to the predicted domain architecture of *SLC4A10*. (Pfam domains- <https://www.ebi.ac.uk/interpro/>) of *SLC4A10*. Cytoplasmic: Band 3 cytoplasmic domain (PF07565); HCO₃⁻ transporter: HCO₃⁻ transporter family (PF00955). d) Multi-species alignments of *SLC4A10*, showing each of the missense variants identified in this study. Abbreviations: aa – amino acids

Individual	Family 1 III:1	Family 1 III:2	Family 2 II:1	Family 3 II:2	Family 3 II:3	Family 4 III:2	Family 4 III:3	Family 4 III:5	Family 5 II:1	Family 5 II:2
NM_001178015	homozygous deletion of exons 5-11	homozygous deletion of exons 5-11	homozygous p.(Arg757*)	homozygous c.2863-2A>C p.(Gln954_Phe955ins*13)	homozygous c.2863-2A>C p.(Gln954_Phe955ins*13)	homozygous p.(Trp873*)	homozygous p.(Trp854*)	homozygous p.(Trp854*)	homozygous p.(Lys577Met; Asn1103Ile)	homozygous p.(Lys577Met; Asn1103Ile)
Sex, Age	M, 8y10m	M, 7y8m	M, 4y8m	F, 8y	F, 4y	M, 10y	M, 6y3m	F, 17y5m	F, 11y	M, 6y
Ethnicity	Palestinian	Palestinian	European	Arab Saudi	Arab Saudi	Egyptian	Egyptian	Egyptian	Turkish	Turkish
Birth OFC	NK	NK	NK	normal	NK	34.2 [-0.8]	35 [-0.2]	33[-1.3]	NK	NK
OFC (cm)[SDS]	50.5 [-2.3]	48.5 [-3.4]	50 [-1.7]	47.5 [-4.5]	44.5 [-5.5]	45.5 [-5.6]	46.6 [-4.3]	48 [-5.4]	51.6 [-1.9]	46.7 [-4.2]
Height (cm)[SDS]	NK	NK	101.5 [-1.3]	125 [-0.4]	100 [-0.4]	123 [-2.5]	104 [-2.7]	150 [-2.2]	136 [-1.2]	111 [-1.0]
Weight (Kg)[SDS]	NK	NK	10.7 [-4.9]	19.8 [-1.8]	10.9 [-3.5]	23 [-2.3]	16 [-2.5]	45 [-1.8]	30.4 [-0.9]	17 [-1.7]
Feeding difficulties	NK	NK	✓	✓At birth	NK	✗	✗	✓	✓	✓
Intellectual disability	✓Severe	✓Severe	✓Severe	✓Profound	✓Profound	✓Severe	✓Severe	✓Severe	✓Moderate	✓Severe
Gross Motor	Walked >2y	Walked 5y	Rolling	Crawling	Not rolling	Walked 6y	Walked 6y	Walked 7y	Walked 2y	Walked 3y
Speech	Non-verbal	Non-verbal	Babbles	Babbles	Sounds	Non-verbal	Non-verbal	Non-verbal	Dysarthria	Non-verbal
Hyperactive	✓	✓	✗	✗	✗	✓	✓	✓	✗	✗
Autistic features	✓	✓	✗	✓	✗	✓	✓	✓	✗	✓
Seizures	✓	✓?	✗	✓GTCS	✗	Abn. EEG	✗	✗	✗	✗
Hearing loss	✗	✗	✗	NP	?	✗	✗	✗	✗	✗
Central tone	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Peripheral tone	↑	↑	↓	↑	↓	↓	↓	↓	↓	↑
Tendon reflexes	+++	+++	++	+++ & clonus	NK	++	++	++	++	+++
MRI brain										
Slit lateral ventricles	✓	✓	✓	NP	✓	NA	NA	NA	✓	✓
Dysmorphic CC	✓	✓	✓	NP	✓	NA	NA	NA	✗	✓
Fornix/SP	✓	✓	✓/✗	NP	✓	NA	NA	NA	✗	✓
Other findings					cranio-synostosis					

Table 4.1: Clinical findings in individuals with biallelic *SLC4A10* variants.

Abbreviations: ✓: feature is present, ✗: feature is absent, +++: exaggerated or brisk, ++: normal, Abn. EEG: abnormal Electroencephalogram, cm: centimetres, F: female, Fornix/SP: distorted configuration of fornix / septum pellucidum, GTCS: generalised tonic-clonic seizures, Kg: Kilogram, M: male, m: months, NK: not known, NP: not performed, OFC: occipitofrontal circumference, SDS: standard deviation scores from the mean, y: years.

Through collaborative studies (via GeneMatcher) we then identified eight additional affected individuals from four unrelated families (**Fig. 4.1**), in whom WES identified biallelic rare predicted loss-of-function *SLC4A10* variants (See **Fig. 4.1**, **Table 4.1**, **Supplemental case reports**, **Table 4.S2** and **Table 4.S3** for family pedigrees, clinical details, comprehensive case reports, *SLC4A10* variants and WGS/WES variant lists respectively). These individuals (aged 4-17 years) presented with clinical features overlapping those of the Palestinian children. In Family 2 trio WES (Individual II:1) identified a homozygous nonsense variant in exon 18/27 [Chr2(GRCh38):g.161949151C>T; NM_001178015:c.2269C>T p.(Arg757*)] also expected to undergo nonsense-mediated decay. Family 3 included two sisters with GDD identified as part of a large-scale study aiming to identify candidate new genetic causes of disease (Novarino *et al.* 2014). Trio WES of DNA from the older sister (Family 3, II:2) identified a homozygous canonical splice site variant Chr2(GRCh38):g.161964133A>C; NM_001178015:c.2863-2A>C, also confirmed to be homozygous in her affected sibling. RT-PCR revealed that the variant resulted in partial intron retention and a premature stop codon [r.(2772_2773ins2772+1_2772+175); r.(2773_2781del)]; p.(Gln954_Phe955ins*13) expected to result in nonsense-mediated decay (**Fig. 4.S3**). In Family 4, WES performed on DNA from two brothers, (Family 4, III:2 and III:3) identified a shared homozygous *SLC4A10* nonsense variant in exon 20/27 expected to result in nonsense-mediated decay [Chr2(GRCh38):g.161957066G>A NM_001178015.1:c.2619G>A p.(Trp873*)]. In Family 5, WES performed on DNA from two siblings and their parents identified a shared homozygous *SLC4A10*

haplotype comprising two missense variants, Chr2(GRCh38):g.161904888A>T NM_001178015:c.1730A>T p.(Lys577Met) and Chr2(GRCh38):g.161976840A>T NM_001178015:c.3308A>T; p.(Asn1103Ile), hereafter referred to as p.(Lys577Met;Asn1103Ile). p.(Lys577Met) affects an invariantly conserved residue within a helical transmembrane domain and is predicted deleterious by *in silico* tools Polyphen2 and SIFT with a high REVEL score (0.873), whereas p.(Asn1103Ile) affects a highly conserved residue but is predicted deleterious only by SIFT and benign by Polyphen with a low REVEL score (0.239) (**Table 4.S2**). All the *SLC4A10* variants identified in this study are absent from gnomAD v2.1.1 and v3.1.2; furthermore, there are no homozygous loss-of-function variants in canonical *SLC4A10* transcripts listed in publicly accessible genomic databases.

Clinical features of *SLC4A10*-related neurodevelopmental disorder

All ten affected individuals presented with hypotonia in infancy, with resultant significant feeding difficulties in 4/10. Psychomotor development was delayed in all individuals across all domains and ID was typically severe. Affected individuals were non-verbal, with one exception; although 7/10 children were ambulatory, walking was delayed in these children until between 2-7 years of age. There was no evidence of developmental regression and while hearing loss was noted in a *Slc4a10*^{-/-} mouse model (Huebner *et al.* 2019) it was not reported in any of the affected patients in this study. Seizures were reported in three individuals, but in two cases these were isolated episodes occurring in the first few years of life. In addition, an affected child from Family 4 displayed bitemporal epileptogenic discharges on EEG at age 5 years in the absence of overt clinical seizures, with spontaneous resolution thereafter.

Behavioural abnormalities were very commonly present and included features associated with autism spectrum disorder, such as anxiety and stereotyped movements (hand flapping, head nodding), hyperactivity and in some cases aggressive episodes. OFC was reported to be normal at birth, but recent measurements were below average in all cases (-1.7 SDS to -5.6 SDS) with 7/10 affected individuals meeting the criteria for microcephaly (<-3 SDS). Affected individuals were below average weight for their age, with height relatively preserved.

MRI neuroimaging findings were striking and consistent. Neuroradiological features of the *SLC4A10*-related neurodevelopmental disorder included microcephaly, with a relative preservation of brain volume compared to OFC and narrow sometimes slit-like lateral ventricles similar to those seen in the *Slc4a10*^{-/-} mouse model (**Fig. 4.S4**), even in those cases with less well-preserved cerebral volume. The corpus callosum was either normal, or dysmorphic (slightly thickened and blunted, flattened in a cranio-caudal direction and with sharply descending fornix) (**Fig. 4.2**, **Fig. 4.S5**). This is likely to be as a result of the small lateral ventricles displacing the fornix and septum pellucidum. There was an absence of cortical malformations and myelination was appropriate for age.

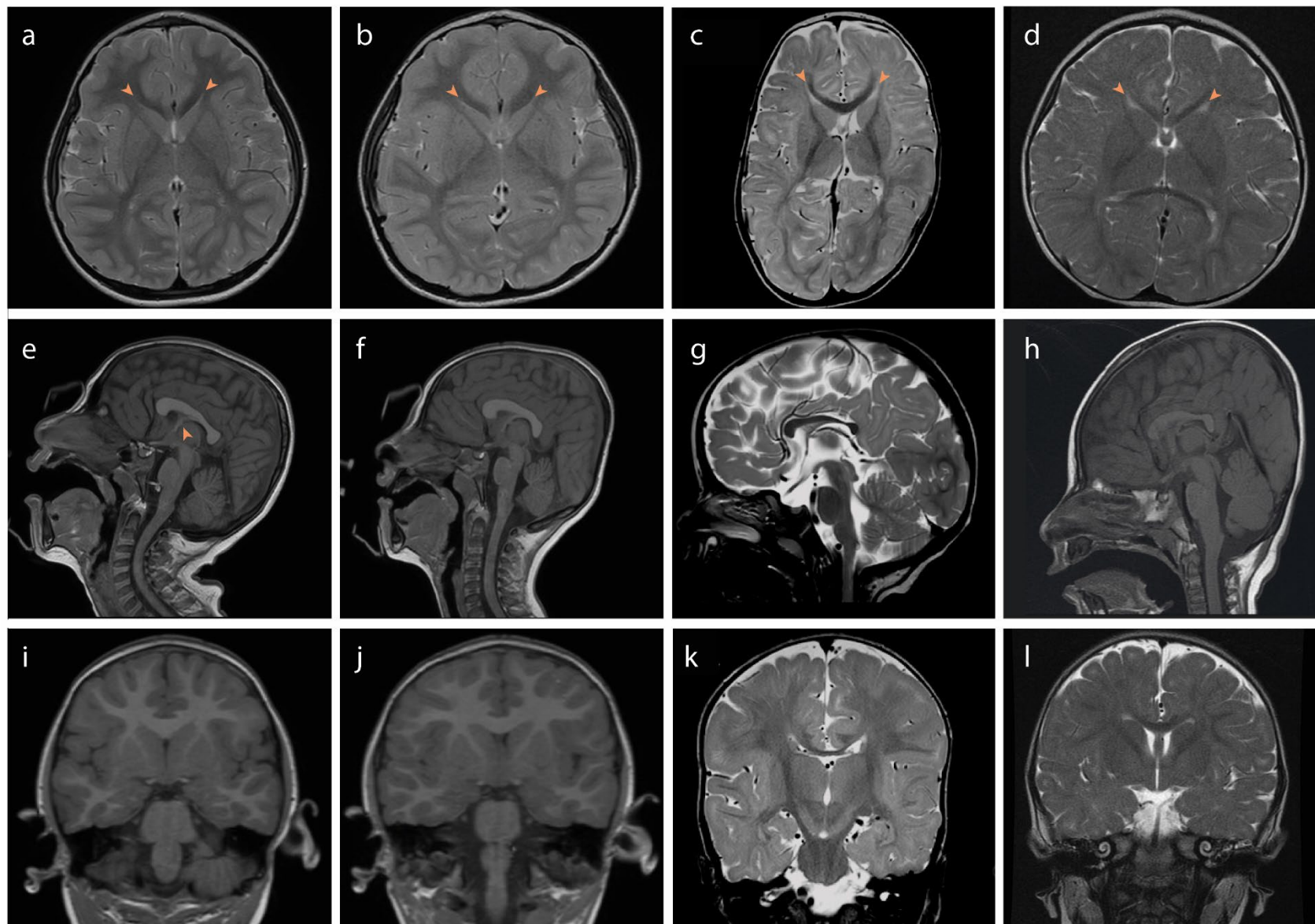


Figure 4.2: Neuroimaging from affected individuals with biallelic *SLC4A10* variants

a,e,j) Family 1, III:1. T2-weighted axial (**a**), T1-weighted sagittal and T1-weighted coronal (**i**) Magnetic resonance (MR) images of the patient at age 5 years

b,f,k) Family 1, III:2 T2-weighted axial (**b**), T1-weighted sagittal (**f**) and T1-weighted coronal (**j**) MR images of the patient at age 4 years

c,g,l) Family 2, II:1. T2-weighted axial (**c**), T2-weighted sagittal (**g**) and T2-weighted coronal (**k**) MR images of the patient at age 10 months

d,h,m) Family 3, II:1 T2-weighted axial (**d**), T1-weighted sagittal (**h**) and T2-weighted coronal (**l**) MR images at age 1 year 2 months.

In all cases lateral ventricles are small (**a-d** – arrowhead) with normal 4th ventricle (**e-h**), posterior fossa and external CSF spaces. In Family 1 the corpus callosum is dysmorphic, appearing thickened and flattened (**e,f**). This is associated with an unusual configuration of the fornix and septum pellucidum especially in Family 1, III:1 (**e** - arrowhead). In Families 2 and 3 it is hypoplastic (**g, h**). Myelination is complete or adequate for age in all cases.

Normal MRI brain images for comparison are available at <https://www.imaio.com/en/e-Anatomy/Brain/Brain-MRI-in-axial-slices> (adult) and <https://radiopaedia.org/cases/normal-mri-head-3-years-old-1?lang=gb> (three-year-old child)

Recovery from acidification is delayed in cells expressing disease-associated SLC4A10 variants.

We previously showed that acid extrusion is compromised in hippocampal neurons in acute brain slices from knockout mice (Jacobs *et al.* 2008). Here, we sought to provide insight into the functional consequences of the SLC4A10 missense variants using cellular studies. We first cloned wild-type and variant *SLC4A10* cDNAs into the mammalian expression vector pBI-CMV4. Two days post-transfection into the fast-growing mouse neuroblastoma cell line N2a, cells were fixed with 4% paraformaldehyde (PFA) and stained with an antibody directed against an N-terminal epitope of SLC4A10, as described previously, and with the lectin WGA to label glycan structures associated with the plasma membrane (Chazotte *et al.* 2011; Jacobs *et al.* 2008). As expected, cells transfected with the wild-type *SLC4A10* construct displayed a predominant labelling at the plasma membrane, whereas the SLC4A10 p.(Lys577Met;Asn1103Ile) variant protein

showed a predominant intracellular localisation (**Fig. 4.S6a**). The quantification of signal intensities for the interior of cells (not including the WGA labelled surface) as compared to the plasma membrane region (the WGA labelled surface) allowed us to calculate the ratio between cell surface and intracellular intensities, which was significantly increased for the *SLC4A10* p.(Lys577Met;Asn1103Ile) variant protein (**Fig. 4.S6b**).

While this outcome alone may explain the pathogenic mechanism of the *SLC4A10* p.(Lys577Met;Asn1103Ile) variant, we also assessed the impact on transport activity by BCECF fluorescence imaging in transfected N2a cells in bicarbonate-buffered salt solution with or without 5 μ M EIPA to block Na^+/H^+ exchange.

Representative single cell traces are shown in **Fig. 4.S7a**. Compared with untransfected cells, steady state pH_i was slightly more alkaline in cells transfected with the *SLC4A10* wild-type construct (**Fig. 4.S7b**). This shift in pH_i remained for both the p.(Lys577Met) and the p.(Asn1103Ile) variant proteins but was present to a lesser extent for the combined p.(Lys577Met;Asn1103Ile) variant (**Fig. 4.S7c**).

Bath application of 20 mM sodium propionate for 5 min induced an acid shift, the amplitude of which did not differ between wild-type and mutant constructs (**Fig. 4.S7d**). pH_i recovery during the propionate exposure was significantly faster for the wild-type construct compared to untransfected cells (transfected cells 163.1 ± 22.9 %, untransfected cells 100.0 ± 18.7 %, $n=7/7$, paired Student's-t-test $p=0.006$, **Fig. 4.S7e**) For p.(Lys577Met), p.(Asn1103Ile) and p.(Lys577Met;Asn1103Ile) the alkaline overshoot after propionate removal (which provides a quantification of net removal of acid during the propionate exposure) was significantly smaller

compared to wild type [one-way ANOVA $p < 0.0001$, $F = 9.434$, Newman-Keuls post-hoc tests: WT $287.5 \pm 18.2\%$, p.(Lys577Met) $169.2 \pm 14.0\%$, $n = 7/10$, $p < 0.0001$, WT $287.5 \pm 18.2\%$, p.(Asn1103Ile) $231.8 \pm 24.9\%$, $n = 7/10$, $p < 0.05$, WT $287.5 \pm 18.2\%$, p.(Lys577Met;Asn1103Ile) $139.6 \pm 18.6\%$, $n = 7/8$, $p < 0.0001$] (**Fig. 4.S7f**). Taken together, we conclude that intracellular pH homeostasis is significantly affected in cells expressing *SLC4A10* p.(Lys577Met) and p.(Asn1103Ile) variants alone and p.(Lys577Met;Asn1103Ile) *in cis*.

***Slc4a10*^{-/-} mice show behavioural abnormalities in the 2-object novel object recognition task and display grossly intact cortical structure.**

The identification of biallelic *SLC4A10* variants in affected individuals with cognitive impairment and behaviours associated with autism spectrum disorders prompted us to reanalyse the behaviour of *Slc4a10*^{-/-} mice. In our previous paper, we reported that motor functions including activity, locomotion and motor coordination, were not altered in *Slc4a10* knockout mice (Jacobs *et al.* 2008). Here, we used the 2-object NOR task to assess recognition memory, which is based on the spontaneous tendency of rodents to spend more time exploring a novel object than a familiar one (Grayson, Idris & Neill 2007). Interestingly, *Slc4a10* knockout mice clearly displayed a marked avoidance of the novel object (**Fig. 4.3a**).

As structural brain abnormalities were present in some of the affected patients, we also re-analysed the brain structure of *Slc4a10*^{-/-} mice. Overall, the brain was found to be smaller and the weight reduced in knockout mice compared to wild type

animals (**Fig. 4.3b**). As previously reported (Jacobs *et al.* 2008), we also noted smaller brain ventricles in *Slc4a10*^{-/-} mice, while the corpus callosum appeared intact (**Fig. 4.3c, Fig. 4.S4**).

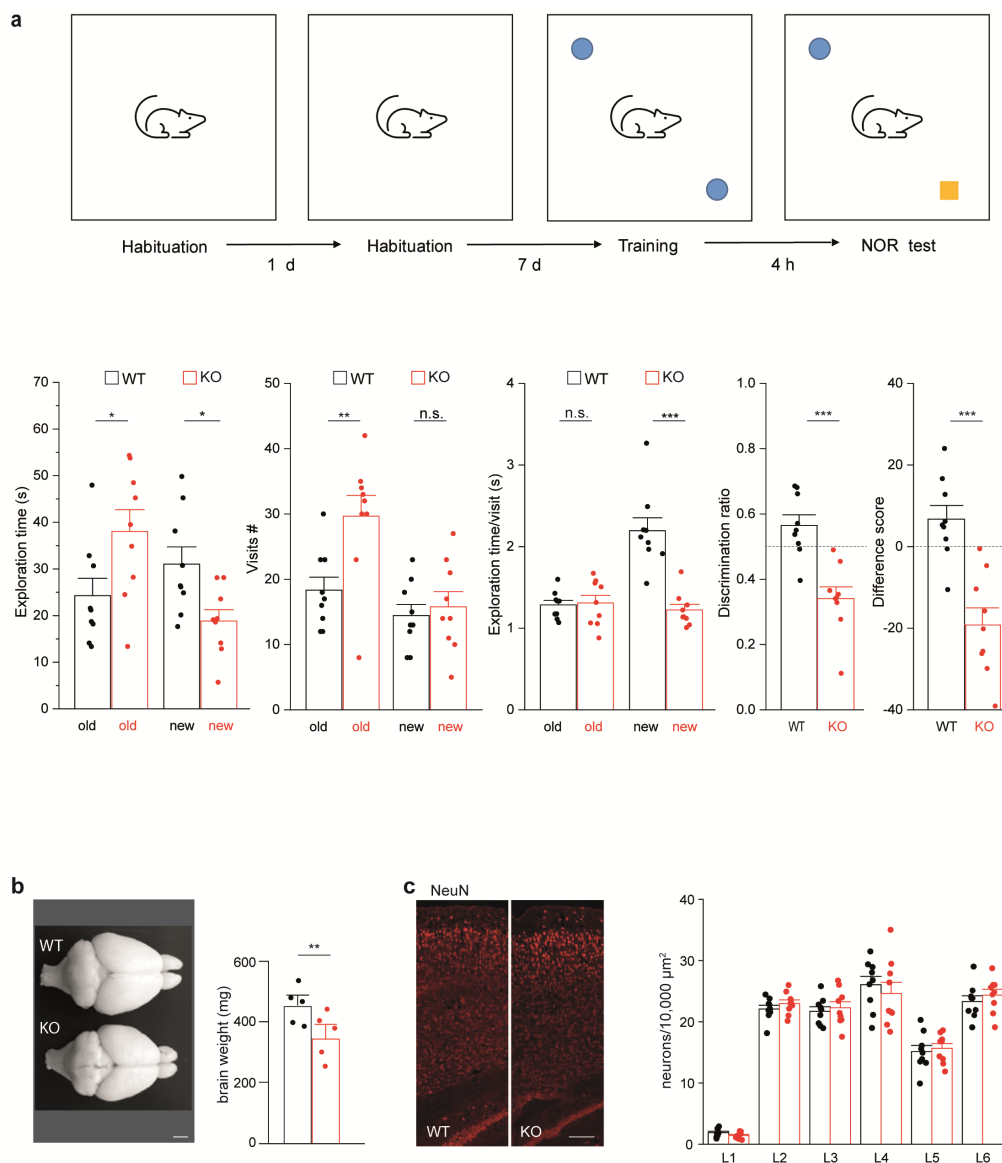


Figure 4.3: *Slc4a10*^{-/-} mice show behavioural abnormalities in the 2-object novel object recognition task and display grossly intact cortical architecture.

a) The recognition of the novel object is altered in knockout (KO) mice. **Upper:** Illustration of the 2-object novel object recognition (NOR) test. **Lower:** During the NOR test the exploration time, the number of visits for the old and the new, and the duration of these visits were quantified. A difference score (time exploring novel object - time exploring familiar object) and the discrimination ratio (time exploring the novel versus the familiar object) was calculated (9 mice per genotype, unpaired Student's t-test; * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$).

b) Top view of dissected brains from 12-month-old *Slc4a10* wild-type (WT) and KO mouse. The weight of perfused and fixed brains of KO mice was smaller compared to WT ($n=5$ mice per genotype; Student's t-test; ** $p < 0.01$). Scale bar: 2 mm. **c)** The gross architecture of the somatosensory cortex appeared intact in *Slc4a10* KO mice. Sagittal brain sections from 2-month-old *Slc4a10* WT and KO mice were stained for the pan neuronal marker NeuN and neurons counted layer wise ($n=3$ mice per genotype; unpaired Student's t-test). Scale bar: 75 μm . Quantitative data are presented as mean + SEM. L1-L6: cortical layers 1-6.

To test whether *Slc4a10*^{-/-} mice display abnormalities in cortex organisation, we also counted neurons labelled for the pan-neuronal marker NeuN (RBFOX3) (Munji *et al.* 2011) in sagittal sections of the motor and the somatosensory cortex of 2-month-old adult mice. Overall, the number of neurons per layer did not differ between genotypes (**Fig. 4.3d**), suggesting an absence of any gross cortical layering defect in *Slc4a10*^{-/-} mice.

SLC4A10 modulates GABAergic but not glutamatergic transmission.

To gain further insight regarding the role of SLC4A10 in neuronal functions, we co-stained mouse brain sections for SLC4A10 and either VGLUT1 (**Fig. 4.4a**), a presynaptic marker of excitatory synapses, or VGAT (**Fig. 4.4b**), a presynaptic marker of inhibitory synapses. We have previously published control staining on knockout tissues (Jacobs *et al.* 2008). Whereas the relative area of co-localisation was $6.4 \pm 0.5\%$ (n=28) for VGLUT1, it was $74.9 \pm 1.3\%$ (n=39) for VGAT (**Fig. 4.4c**). Pearson correlation coefficients (CC) between SLC4A10 and either VGLUT1 or VGAT signals after the Costes method (Dunn *et al.* 2011) are in agreement with a predominant localisation of SLC4A10 with GABAergic but not glutamatergic presynapses (**Fig. 4.4d**, CC VGLUT1/SLC4A10 [0.03] versus CC VGAT/SLC4A10 [0.5], 28 and 39 images each).

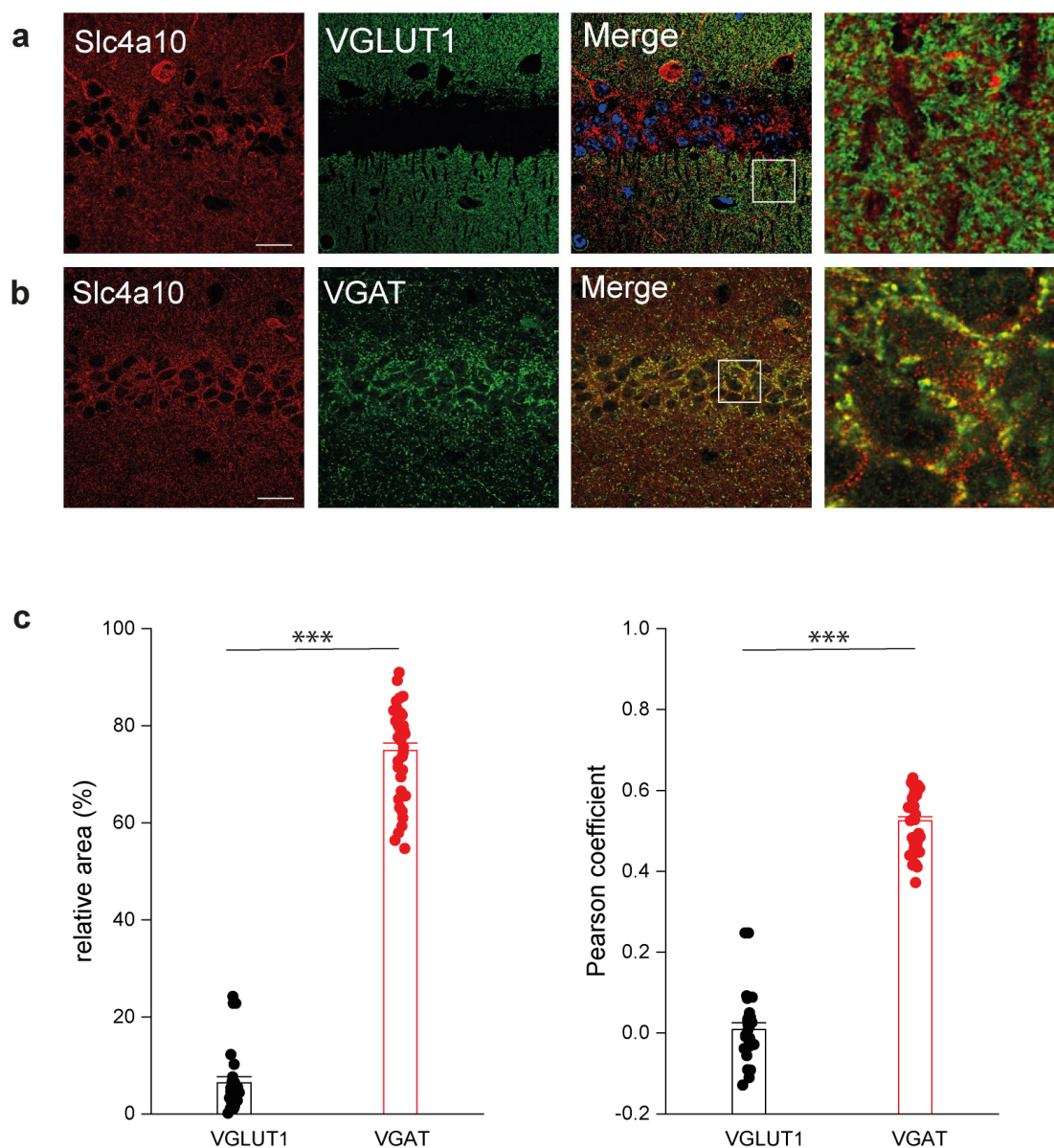


Figure 4.4: Localisation of SLC4A10 to GABAergic presynapses

Slc4a10 wild-type (WT) mouse brain sections (Scale bars: 20 μm , enhanced view of merged marker images also shown [boxed areas]). **a**) VGLUT1, a marker of excitatory presynaptic terminals, rarely co-localises with SLC4A10 in the CA1 region of the hippocampus (green: SLC4A10, red: VGLUT1). **b**) SLC4A10 and VGAT, a marker for GABAergic presynapses, co-localise in the CA1 region of the hippocampus (green: SLC4A10, red: VGAT). **c**) Quantitative analysis of co-localisation of SLC4A10 with either VGLUT1 or VGAT and calculation of Pearson correlation coefficients between these data in the CA1 region of the hippocampus (VGLUT1 $n=28$ and VGAT $n=39$ images each, Student's t-test; *** $p < 0.001$).

These data led us to next study whether neurotransmitter release is affected in the CA1 region of the hippocampus *Slc4a10*^{-/-} mice. Frequency, amplitude and kinetics of mEPSCs recorded in the presence of tetrodotoxin (TTX) did not differ between genotypes (**Fig. 4.5a-c** and **Table 4.S4**) suggesting that glutamate release is not affected by disruption of SLC4A10. In contrast, the frequency of mIPSCs in TTX, either recorded in CA1 (**Fig. 4.5d-f**, **Table 4.S4**) or CA3 (**Fig. 4.S8**), were significantly decreased in the presence of HCO₃⁻. While amplitudes were unaffected, T_{decay} and consequently the transferred electric charge per event were diminished in slices obtained from *Slc4a10*^{-/-} mice. As there is evidence that spontaneous and evoked neurotransmission are partially segregated at inhibitory synapses (Horvath *et al.* 2020), we also studied spontaneous postsynaptic currents (sIPSCs), the frequencies of which were reduced, while kinetics were unaffected (**Fig. 4.S9**).

As the disruption of the acid-extruder SLC4A10 is expected to decrease neuronal pHi (Jacobs *et al.* 2008), we tested whether the mIPSC frequency can be rescued by raising pHi. Indeed, 20 mM trimethyl ammonium (TriMA), which raises pHi without affecting extracellular pH (Eisner *et al.* 1989), increased the mIPSC frequency in preparations from *Slc4a10*^{-/-} mice (**Fig. 4.5d-f**), while the kinetics were not affected. Analogously, lowering pHi by replacing 20 mM NaCl by sodium propionate decreased mIPSC frequency in slices from wild-type mice (**Fig. 4.S10**). The effects of the disruption of *Slc4a10* on frequency and kinetics were eliminated under bicarbonate-free conditions in recordings performed in HEPES-buffered solution, arguing against structural defects or an altered subunit composition of postsynaptic GABA_A receptors

(**Fig. 4.5g**). Together, these data show that SLC4A10 modulates GABAergic synaptic transmission in a bicarbonate-dependent manner.

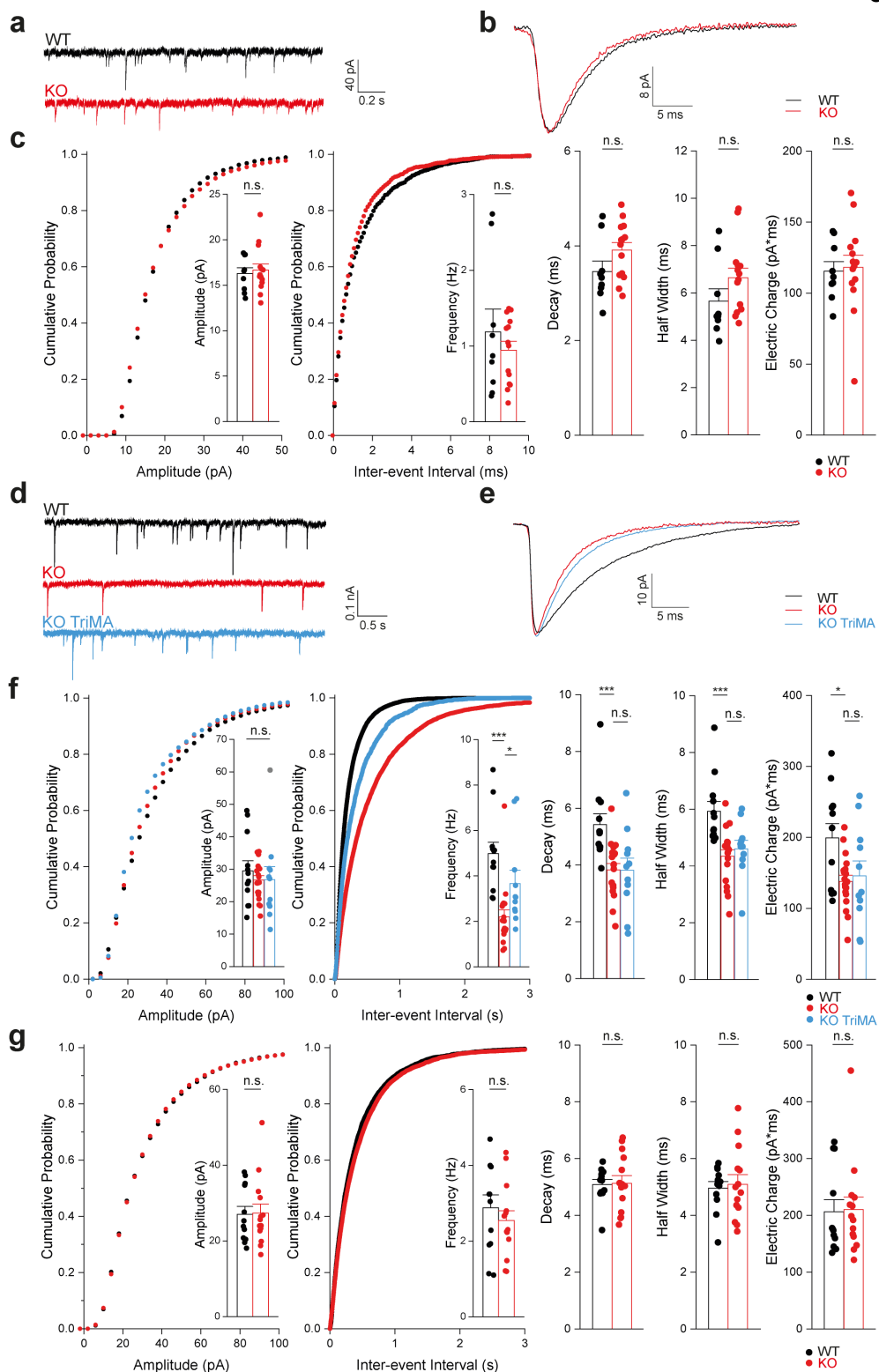


Figure 4.5: SLC4A10 acts on presynaptic pH_i to promote GABA release in CA1 pyramidal neurons.

Glutamatergic transmission is not impaired in CA1 neurons of *Slc4a10*^{-/-} mice (a-c).

- a)** Representative miniature excitatory postsynaptic current (mEPSC) recordings pyramidal neurons from *Slc4a10* wild-type (WT) and knockout (KO) mice.
- b)** Averaged mEPSCs show that the kinetics of mEPSCs are not affected by disruption of *Slc4a10*.
- c)** Cumulative plots and bar charts of different mEPSC properties. No significant differences were detected in mEPSC frequency, amplitude or kinetics (n=12/10).

The mIPSC frequency is diminished in *Slc4a10*^{-/-} mice in the presence of HCO₃⁻ (d-g).

- d)** Representative recordings of ongoing miniature inhibitory postsynaptic current (mIPSC) activity in pyramidal neurons from *Slc4a10* wild-type (WT) and knockout (KO) mice as well of pyramidal neurons from *Slc4a10* KO mice in the presence of 20 mM trimethylamine chloride (TriMA).
- e)** Averaged mIPSC recordings of pyramidal neurons from *Slc4a10* WT and KO mice to illustrate kinetics and amplitude.
- f)** Cumulative plots and bar charts of mIPSC properties (n=12/19/11; Student's t-test: * p<0.05; **p<0.01; ***p<0.001; n.s.: not significant). While no differences in the mean amplitudes of mIPSCs were observed, the frequency of mIPSCs was significantly diminished in cells derived from *Slc4a10* KO mice but could be partially rescued by application of TriMA. Diminished τ_{decay} and half-width of averaged mIPSCs in pyramidal neurons from *Slc4a10* KO mice compared with WT in bicarbonate-buffered aCSF were not affected by TriMA.
- g)** In HEPES-buffered nominally bicarbonate-free solution mIPSC frequencies and kinetics did not differ between genotypes (n=14/12; Student's t-test: * p<0.05; **p<0.01; ***p<0.001; n.s.: not significant). Quantitative data are shown as mean + standard error of the mean (SEM).

Discussion

Here we present clinical and genetic data from five unrelated families, alongside molecular and neurobiological findings in mice that define biallelic loss-of-function variants in *SLC4A10* as a cause of a severe neurodevelopmental disorder, frequently associated with microcephaly (<-3 SDS) and morphologically abnormal collapsed (slit) lateral ventricles. This slit-like appearance of the lateral ventricles appears to be characteristic of the disorder and mirrors findings in the *Slc4a10*^{-/-} mouse (Jacobs *et al.* 2008). SLC4A10 mediates sodium-dependent HCO₃⁻ transport at the basolateral side of choroid plexus epithelial cells (Jacobs *et al.* 2008). Thus, collapsed brain ventricles in knockout mice and in patients with *SLC4A10* biallelic loss-of-function alleles suggest that basolateral SLC4A10-

dependent Na⁺ uptake plays a key role for the apical sodium-coupled secretion of the CSF (Damkier, Brown & Praetorius 2013).

The four truncating *SLC4A10* variants identified are predicted to result in complete molecular loss of function [deletion of exons 5-11; p.(Trp140Argfs*39), p.(Arg757*), p.(Trp873*) and c.2863-2A>C; p.(Gln954_Phe955ins*13)]. Consistent with this, affected individuals homozygous for these variants have the most severe neurological outcomes. Additionally, while both the p.(Lys577Met), affecting the transmembrane region (**Fig. 4.S11**), and C-terminal p.(Asn1103Ile) variant proteins were each trafficked to the proximity of the plasma membrane individually, *SLC4A10* protein harbouring both p.(Lys577Met;Asn1103Ile) variants *in cis* was largely trapped intracellularly and intracellular pH was shown to be significantly affected (**Fig. 4.S6a,e,f**), strongly supportive of pathogenicity.

Previously a *de novo* balanced translocation disrupting *SLC4A10* was identified as a candidate cause of disease in a single individual described to have “mental retardation, progressive cognitive decline, and partial complex epilepsy” (Gurnett *et al.* 2008). However, a heterozygous *SLC4A10* variant causing a severe monogenic disease is not consistent with the autosomal recessive condition described here, given the unaffected parental / sibling carriers of loss-of-function *SLC4A10* variants, and the many heterozygous loss-of-function gene variants listed in gnomAD. While it remains unclear whether an undetected *SLC4A10* variant may have been present *in trans* with the disrupted *SLC4A10* allele, heterozygous loss

of SLC4A10 function due to the translocation event alone appears unlikely to be responsible for the neurological condition affecting this individual.

Our findings are also of note in light of recent genome-wide association studies (GWAS) which identify a highly statistically significant association between *SLC4A10* intronic or in *cis* regulatory-transcription binding region variants and neurological traits including cognitive function, educational attainment, brain and hippocampal volume and psychiatric morbidity (**Table 4.S5**) (Lee et al 2018; van der Meer *et al.* 2020; Ripke *et al.* 2014). Taken together with our present findings, these data provide compelling evidence for the importance of *SLC4A10* in normal neurological development and function and suggest a potential role for *SLC4A10* in traits mediated by oligo/polygenic inheritance.

Notably, *Slc4a10*^{-/-} mice show altered object discrimination with avoidance of the novel object thus resembling a mouse model of autism spectrum disorder (Kane *et al.* 2012). This prompted us to use our mouse model to further characterise the role of SLC4A10 in brain function. We previously showed that *SLC4A10* is broadly expressed in both principal cells and inhibitory interneurons and that its disruption impaired the recovery of neurons from an acid load in the somatodendritic compartment (Jacobs *et al.* 2008). Here, we show that SLC4A10 co-localises with a marker of GABAergic but not glutamatergic presynapses. In agreement with this localisation, GABA release was reduced, while glutamate release was not affected. This defect is characterised by a decrease of mIPSC frequency, while mIPSCs amplitudes remain unaltered. Intracellular alkalinisation with TriMA partially

rescued mIPSC frequency in brain slices from *Slc4a10*^{-/-} mice, while intracellular acidification induced a decrease of mIPSC frequency in wild-type mice, which further supports the conclusion that the difference in mIPSC frequency between the two genotypes are pH_i dependent. The knockout of a plasma membrane resident sodium-coupled anion exchanger such as SLC4A10 might also change the equilibrium potential for Na^+ (E_{Na}) and thus neuronal excitability. However, cellular Na^+ loading by pH-regulatory mechanisms typically requires blocking the Na-K ATPase (Kaila & Voipio 1987). Moreover, the relative permeability for Na^+ of a typical neuron at rest is very low compared to K^+ and Cl^- and thus exerts only a minor contribution to the resting membrane potential (Rutecki, Lebeda & Johnston 1985). In agreement, the resting membrane potential between wild-type and knockout principal neurons did not differ at steady state, thus excluding a major effect of SLC4A10 on E_{Na} and neuronal excitability.

In contrast to the decreased mIPSC frequency, mIPSC kinetics were only mildly changed, which can reflect alterations at the postsynaptic site such as changes in the receptor density or the composition of the receptor subunits (Farrant & Kaila 2007). However, the differences between genotypes were abolished under bicarbonate-free conditions which eliminates the activity of sodium-dependent HCO_3^- transporters arguing against this possibility. Because $GABA_A$ receptor function critically depends on extracellular pH (Dietrich & Morad 2010; Mozrzymas *et al.* 2003, Pasternack, Smirnov & Kaila 1996) changes in the kinetics rather suggest an increase in the pH of the synaptic cleft. Accordingly, TriMA, which only raises pH_i , but does not affect extracellular pH (Stenkamp *et al.* 2001), did not

change mIPSC kinetics. Notably, the disruption of the Na^+/H^+ exchanger NHE1/SLC9A1, another transporter expressed at inhibitory presynapses, also decreased mIPSC frequency and altered kinetics (Bocker *et al.* 2019).

Thus, both SLC4A10 and SLC9A1 seem likely to contribute to the regulation of pH_i at GABAergic nerve endings and, notably, biallelic variants in *SLC9A1* have been linked to a syndromic neurological disorder (Guissart *et al.* 2015). Furthermore, control of pH_i at glutamatergic presynapses is mediated by the combination of SLC9A1 (Bocker *et al.* 2019) and SLC4A8 (Grichtchenko *et al.* 2001), another sodium-dependent HCO_3^- transporter closely related to SLC4A10. Similar to the defect of GABA release upon disruption of SLC4A10, disruption of SLC4A8 affects glutamate release via its effect on presynaptic pH_i . These data show that changes in pH_i may affect the vesicle release machinery in both GABAergic and glutamatergic neurons in numerous ways. Presynaptic Ca^{2+} transients, which trigger synaptic vesicle exocytosis (Schneggenburger & Neher 2000), may be altered upon disruption of either *Slc4a8*, *Slc4a10* or *Slc9a1*, potentially because both Ca^{2+} influx via voltage-gated Ca^{2+} channels (VDCCs) (Doering & McRory 2007; Fraire-Zamora & González-Martínez 2004; Tombaugh & Somjen 1996) and Ca^{2+} release from intracellular stores (Ma *et al.* 1988; Tsukioka, Iino & Endo 1994) are strongly pH-dependent. Alternatively, H^+ may compete with Ca^{2+} at the binding site of synaptic vesicles, or may alter the function of proteins involved in vesicle release (Rizo & Xu 2015). Changes in pH_i might also affect the loading of GABA into synaptic vesicles, because VGAT operates as a GABA/ H^+ exchanger and critically depends on the H^+ electrochemical gradient generated by the vacuolar-

type H⁺ (Farsi *et al.* 2016; Egashira *et al.* 2016). However, the lack of effect of disruption of SLC4A10 on mIPSC amplitudes are evidence against such an effect (Frerking, Borges & Wilson 1995).

Consistent with a defect of GABA release, we previously reported an increased network excitability in acute brain slices obtained from *Slc4a10* knockout mice as evidenced by compromised paired-pulse facilitation and increased excitatory postsynaptic potential-spike coupling (E-S coupling) (Sinning *et al.* 2015). Changes in the production and composition of the CSF of *Slc4a10* knockout mice may have opposite effects on network excitability *in vivo*. Indeed, seizure susceptibility to pentylenetetrazole (PTZ) and hyperthermia-induced hyperventilation with respiratory alkalosis were diminished in *Slc4a10* knockout mice (Jacobs *et al.* 2008). Whether patients with SLC4A10-related disease are at increased risk of developing seizures, is as of yet unclear. In our study, only 2 out of 10 patients had a clear history of epilepsy. In order to more fully investigate seizure susceptibility, additional mouse models with either a disruption of *Slc4a10* in choroid plexus epithelial cells or in neurons will be desirable.

In summary, we present extensive genetic, clinical, functional and murine datasets confirming that biallelic SLC4A10 pathogenic loss-of-function gene variants cause a syndromic neurodevelopmental disorder. The abnormal slit-like brain ventricles characteristic of the disease likely reflect a reduced production of CSF in patients, which ordinarily forms a continuous fluid compartment with the interstitial fluid bathing all cells of the brain. While recent evidence suggests that the CSF contains

neuromodulators that strongly influence neuronal activity (Bjorefeldt *et al.* 2018), future studies are required to separate such systemic effects from local synaptic effects. Defects of GABAergic function are a recurrent finding in various neurodevelopmental and neuropsychiatric phenotypes such as ID, autism spectrum disorders, epilepsy and schizophrenia (Levitt, Eagleson & Powell 2004; Schmidt-Wilcke *et al.* 2018). Importantly, positive modulation of GABA_A receptors by diazepam and GABA_A receptor agonists have been shown to improve behavioural and neurophysiological defects in mouse models of fragile X syndrome (Heulens *et al.* 2012; Olmos-Serrano & Corbin 2011). Given this, it is tempting to hypothesise that enhancing inhibitory GABAergic transmission could be a possible therapeutic approach for ameliorating some of the neurological symptoms in patients with *SLC4A10*-related neurodevelopmental disorder.

4.3. Supplemental Material

Detailed clinical case descriptions

Family 1; III:1 and 2

Two male siblings, aged eight and seven, were born to unaffected first-cousin Palestinian parents. The elder brother (III:1) was the product of an uncomplicated pregnancy and delivery, but presented profoundly hypotonic at birth, requiring immediate and sustained respiratory support. He suffered a single afebrile seizure aged one year of age for which he has not required further treatment. His development was delayed, he bottom shuffled then walked after two years. Now, aged eight years and seven months, he can walk and run, feeds himself with his hands and is toilet trained but has only eight words. He displays behaviours suggestive of an autism spectrum disorder including hyperactivity, aggressive episodes and anxiety that have required residential care, and treatment with methylphenidate and risperidone. His younger brother (III:2), also the product of an uncomplicated pregnancy and delivery was initially discharge home where he suffered a cyanotic episode at 3 weeks of age, possibly a seizure, requiring urgent readmission. His development was profoundly delayed, walking at five years of age. Now aged seven years eight months he has no language, either spoken or receptive, poor fine motor skills and is not toilet trained. He is less aggressive than his sibling but exhibits stereotyped hand-flapping. Both brothers have been hospitalised on multiple occasions with upper and lower respiratory tract infections. Both are brachycephalic and microcephalic and with large, rotated ear lobes in keeping with this and hypotelorism. The younger child has inverted nipples. On neurological examination, both displayed axial hypotonia, peripheral spasticity and exaggerated deep tendon reflexes. MRI of both siblings showed slit lateral ventricles and a dysmorphic, thickened, flattened corpus callosum, which was associated with an unusual configuration of the midline structures (fornices and septum pellucidum).

Family 2; II:1

This boy was the only child born to unrelated Austrian parents. He was severely hypotonic in the neonatal period with poor suck, resulting in faltering growth and delayed acquisition of motor milestones, rolling first at three-and-a-half years. He has mild craniofacial dysmorphism with an elongated face, prominent metopic ridge, tent-shaped mouth with long philtrum, low-set large ears and tapering fingers. Skeletal anomalies include bilateral *coxa vara anteverta* and left developmental dysplasia of the hip requiring surgical remediation. At age four years and eight months he has a severe GDD (virtually non-ambulatory, babbles only, no speech), severe central and peripheral hypotonia, severe failure to thrive but no microcephaly. MRI, performed at 11 months

of age and reported by the local radiologist, showed slit lateral ventricles (first seen on neonatal cranial ultrasound) with generalised cerebral volume loss and a hypoplastic corpus callosum. Myelination appears appropriate for age. It was not possible to reevaluate the MRI at the point of inclusion in the study.

Family 3; II:2 and II:3

These are two sisters from Saudi Arabia, born at term who presented with profound hypotonia in infancy, microcephaly and GDD. The older sister (**II:2**), now 12 years of age, has severe neurodevelopmental delay – she did not crawl until at three years of age, and now sits only with support. Her fine motor skills are profoundly delayed and she still uses a palmar grasp. She is not reaching and shows no interest in play with toys. She has some behaviours associated with autism spectrum disorder, including hand flapping and head nodding with preference of holding papers and difficulty in bathing. She is affected by generalised tonic-clonic seizures, first apparent at age seven and now occurring approximately monthly and lasting from 2 to 5 minutes. Inter-ictal EEG performed aged 6 years showed background slowing. She was born with bilateral ankle contractures and bilateral foot deformities. Over time a left esotropia became apparent.

The younger sister (**II:3**) required nasogastric feeding in the first week of life and has progressed to severe GDD and cognitive impairment with microcephaly, but without apparent seizures. Additionally, lambdoid craniosynostosis and ankle contractures are present in the younger child. She is fully dependent on her caregiver. MRI of the younger sibling revealed small lateral ventricles with normal external CSF spaces and appropriate myelination. The corpus callosum and midline structures are within normal range (**Figure 4.2d,h,m**).

Previous unremarkable investigations include, very long chain fatty acids, acylcarnitine profile and serum amino acids in both. Additionally, included MRS, urine organic acids thyroid stimulating hormone, biotinidase levels and genetic tests for myotonic dystrophy and Prader Willi syndrome in the younger sibling.

Trio WES of the elder sister (II:2) was undertaken as part of a large-scale, first-tier clinical exome sequencing study (ID: 17-4393) using methods and variant filtering strategies previously described [Monies *et al.* 2017, Monies *et al.* 2019]. This identified a homozygous *SLC4A10* canonical splice acceptor site variant as the most likely cause of disease [Chr2(GRCh38):g.161964133A>C; NM_001178015:c.2863-2A>C]. The variant was confirmed by rtPCR of mRNA extracted from blood of both affected individuals, which identified an additional band representing mRNA with partial inclusion of intron 21 leading to premature stop after 13 additional amino acids

[p.(Gln954_Phe955ins*13)] (**Fig. 4.S3**), Individual II:3 was confirmed to be homozygous for the same *SLC4A10* variant using Sanger sequencing performed locally.

Family 4; III:2, III:3 and III:5

This extended Egyptian family consists of two affected brothers, ten years and six years, three months, and their affected female double first cousin (17 years, five months). All were born at term with normal birth weight and are now of proportionate short stature and low weight with disproportionately severe postnatal microcephaly (-4 to -6 SDS). All affected individuals have severe developmental impairment, first walking between six and seven years and two developed a broad-based gait (III:3 and III:5). All were non-verbal with extremely limited receptive language (cannot follow one-step commands or identify body parts). Behavioural abnormalities included hyperactivity and features associated with autism spectrum disorder. Neurological findings were consistent across all affected individuals and included global hypotonia, with normal deep tendon reflexes and plantar responses. No individuals suffered clinical seizures, but an EEG in Individual III:2 noted bitemporal epileptogenic discharges at the age of five years that resolved without treatment. Neuroimaging was not available for review. Ophthalmological examination, including visual evoked potentials and electroretinogram (ERG) was normal in Individual III:2 as was hearing assessment, measured by auditory brainstem response (ABR).

Family 5; II:1 and 2.

These two Turkish children presented with hypotonia, GDD and microcephaly. The elder sister (**II:1**) is more mildly affected; she walked aged two and at the age of 10-11 years her speech was dysarthric, she struggles to hold a pen and has an intelligence quotient (IQ) around 53. At the age of 15 years she attends a school for special needs education with her developmental level is comparable to a seven-year-old child and is described as quite sociable. She has mild dysmorphic feature (a prominent forehead and arched eyebrows) joint laxity, an accessory nipple, and fetal finger pads. The younger brother (**II:2**) is more severely affected; he is severely microcephalic, walked at aged three years, has no speech and cannot feed himself at six years of age. He has brachycephaly with a forehead upsweep, arched eyebrows with flat orbital ridges and a bulbous nasal tip and prognathism. He has central hypotonia but peripheral hypertonia with brisk reflexes and a positive Babinski's sign. MRI of the brother (**II:2**) identified a short and thick corpus callosum with hypoplastic pons and brainstem, normal myelination, lateral ventricles with collapsed anterior horns (**Fig. 4.S5d-f**). His sister's imaging was similar, without the dysmorphic appearance of the corpus callosum (**Fig. 4.S5a-c**). Both had mildly increased excretion of creatine, with other metabolic investigations reported normal.

Figure 4.S1: Genome sequencing data of affected individuals in Family 1, reveals a biallelic, multi-exon, deletion of *SLC4A10*.

IGV plots illustrating the homozygous deletions in *SLC4A10* in Family 1; III:1 [*top panel*] and III:2 [*middle panel*] shown alongside a wild-type control sequence obtained under the same conditions [*bottom panel* read alignment over the deleted region]. The deletion was confirmed using ddPCR (Fig. 4.S2).

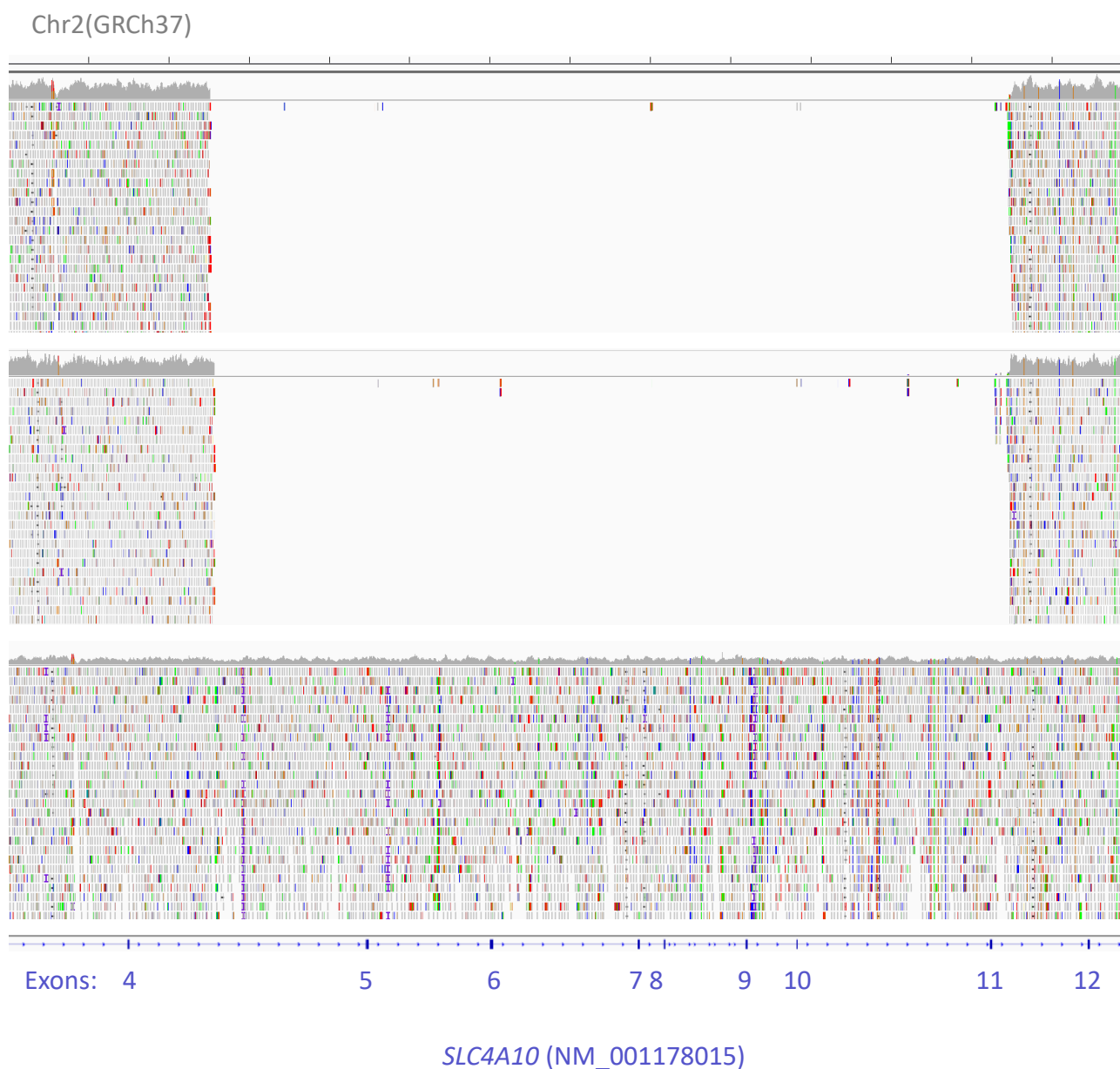
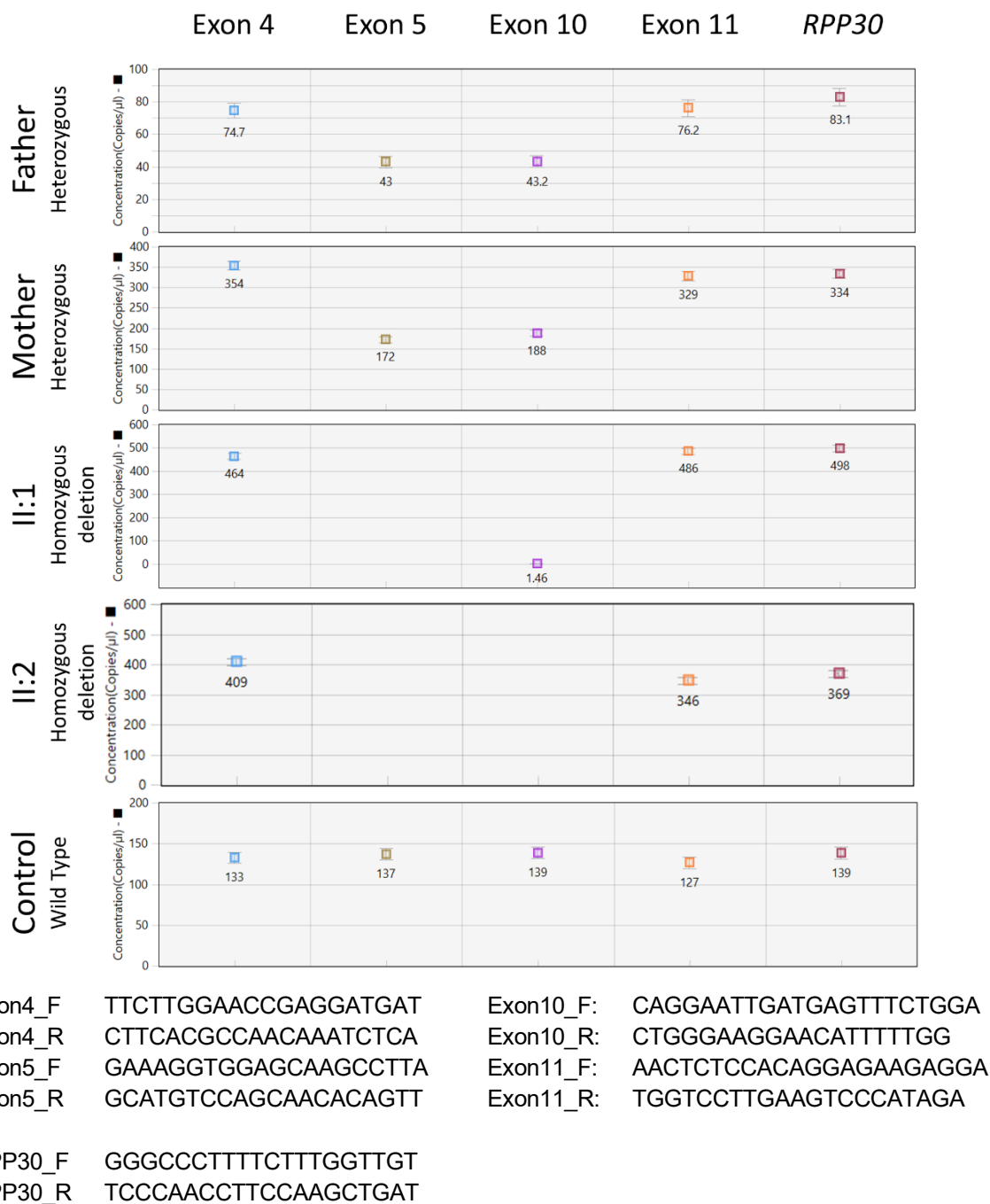


Figure 4.S2: ddPCR data confirms the presence of a multi-exon deletion of *SLC4A10* in individuals III:1 and III:2 (Family 1).

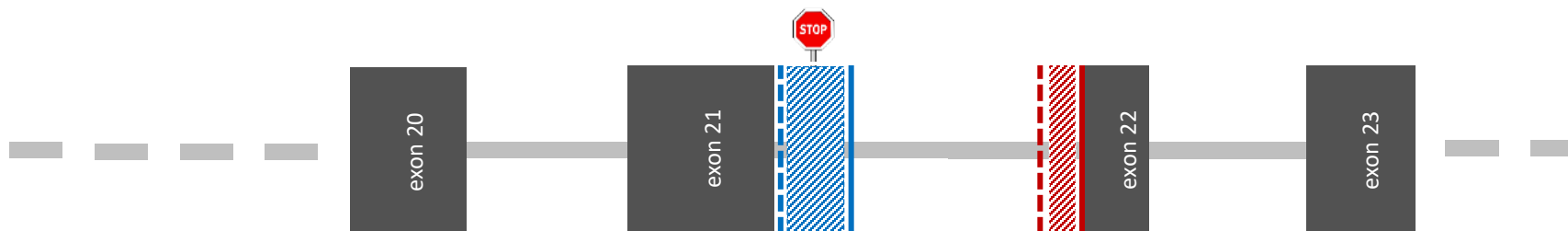


Abbreviations: F = forward, R = reverse

Conditions: 59 degrees

Figure 4.S3: RT-PCR demonstrates mRNA splicing effect of NM_001178015:c.2863-2A>C p.(Gln954_Phe955ins*13) variant (Family 3).

a) simplified gene intron-exon structure of *SLC4A10* showing splicing effect of the c.2863-2A>C variant. This variant weakens the splice acceptor site (red dashed line), leading to the preferential usage of a more 3' acceptor site (solid red line) with the resultant deletion of 9 bp from the start of exon 22 (red shaded region). This also affects the donor site, with an alternative more 3' site utilised (blue solid line) in preference to the canonical site (blue dashed line). The use of the alternative splice donor site results in the retention of 175 bp of intron 21 (shaded blue) resulting in 13 additional amino acids being included before a premature stop codon (red octagon).



b) *SLC4A10* mRNA sequence obtained from peripheral blood of an affected individual. Grey highlighting indicates nucleotides derived from exon 21. Blue highlighting indicates bases derived from the retention of intron 21. Red text in square brackets indicates the deleted sequence at the start of exon 22. Codon phase is shown with alternating text colour until the premature stop codon (red text, with square brackets).

```

TTTATTCCCATGCCAGTGCTATATGGAGTGTTTCTTTATATGGGTGCTTCATCTCTAAAGGGAATTCAGGTA AATTACTTTACAGTACAGTACAGGCACATCTGTGATGACTGACCTTAAGGTCTACTGATAAGTCATGTGACAGCTGAGAAAATGCCACCACCTGAGGAAACAGCTTTTAGACCACAATTA AATTTCTTCAAACCTTGTCAGAGTTACAAAAGTTAAAGAAGATTCTCTCCAGCATCT [TTCTTTGAT] AGGATAAAGCTCTTCTGGATGCCGGCAAACATCAACCAGATTTTATATACCTAAGGCACGTACCGCTTCGAAAAGTGCATCTCTTCACAATTATTTCAGATGAGTTGCCTTGGCCTTTTGTGGATAATAAAAGTTTCAAGAGCTGCTATTGTCTTTCCCATGATG

```

c) Chromatogram demonstrating the retention of part of intron 21 in an affected individual (blue dotted line).

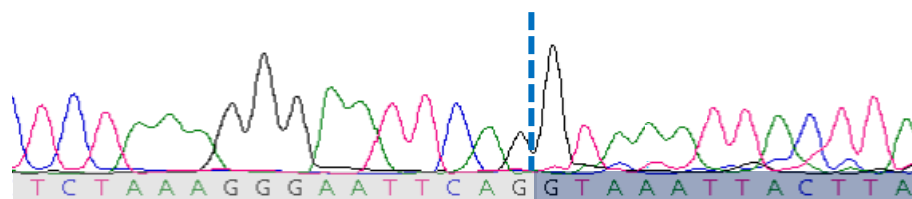


Figure 4.S4: T2-weighted sagittal MRI scan of wild-type and *Slc4a10* knockout mice

Whilst lateral ventricles (arrowheads) are seen in the wild-type mouse (left), these are not macroscopically apparent in the knockout mouse (right) suggesting a marked reduction in volume.

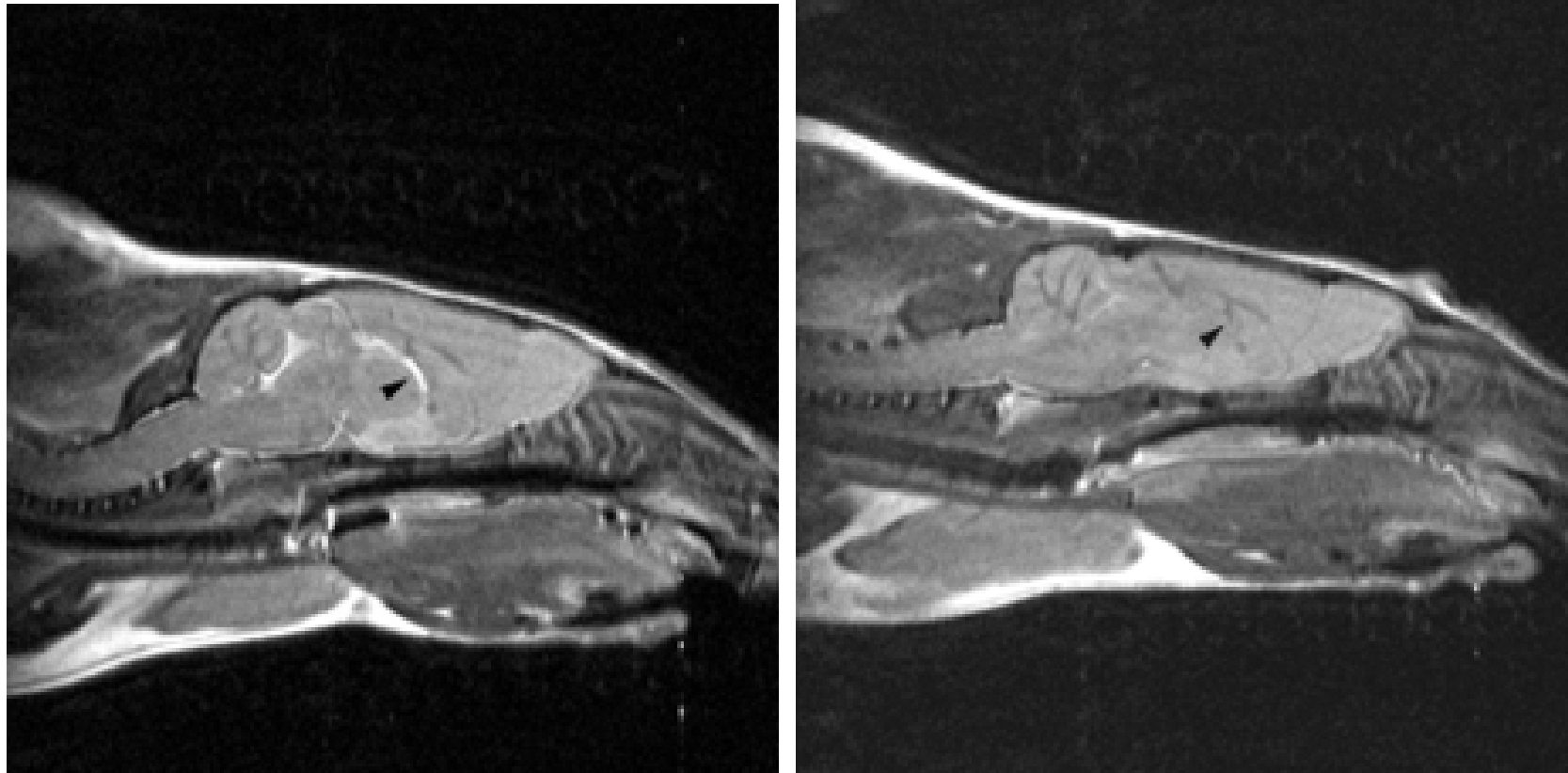


Figure 4.S5: Additional MRI images of individuals affected by *SLC4A10*-related disorder

All images made available to the authors are shown. **Top:** Family 5, Individual II:1 (female) 2 years 2 months, showing slit-like anterior horns of the lateral ventricles **a,b** T2 axial **c** T2 sagittal (midline) **Bottom:** Family 5, Individual II:2 (male) 2 years 7 months showing a dysplastic corpus callosum. **d,e** T1 axial **f** T1 sagittal (midline)

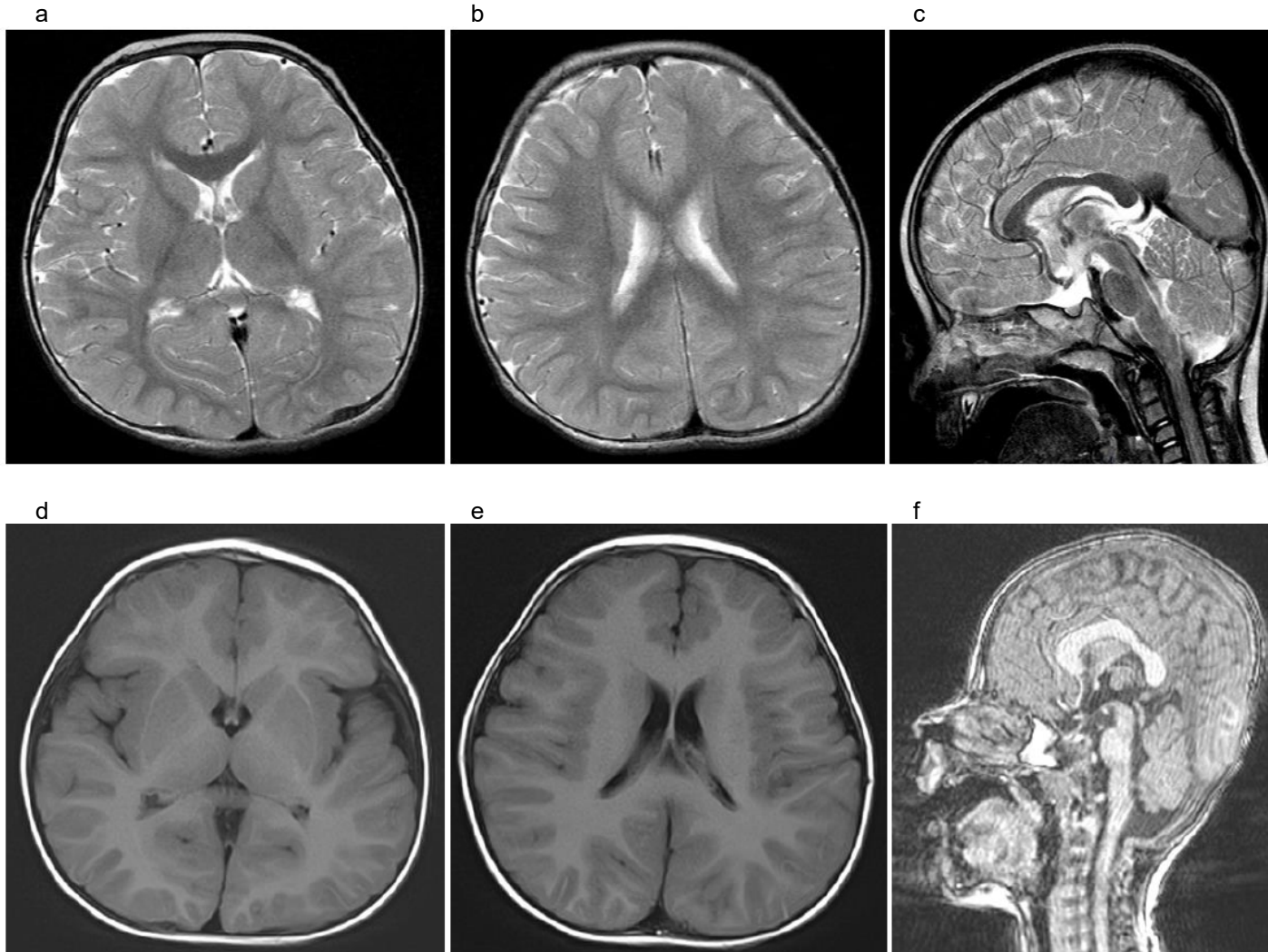
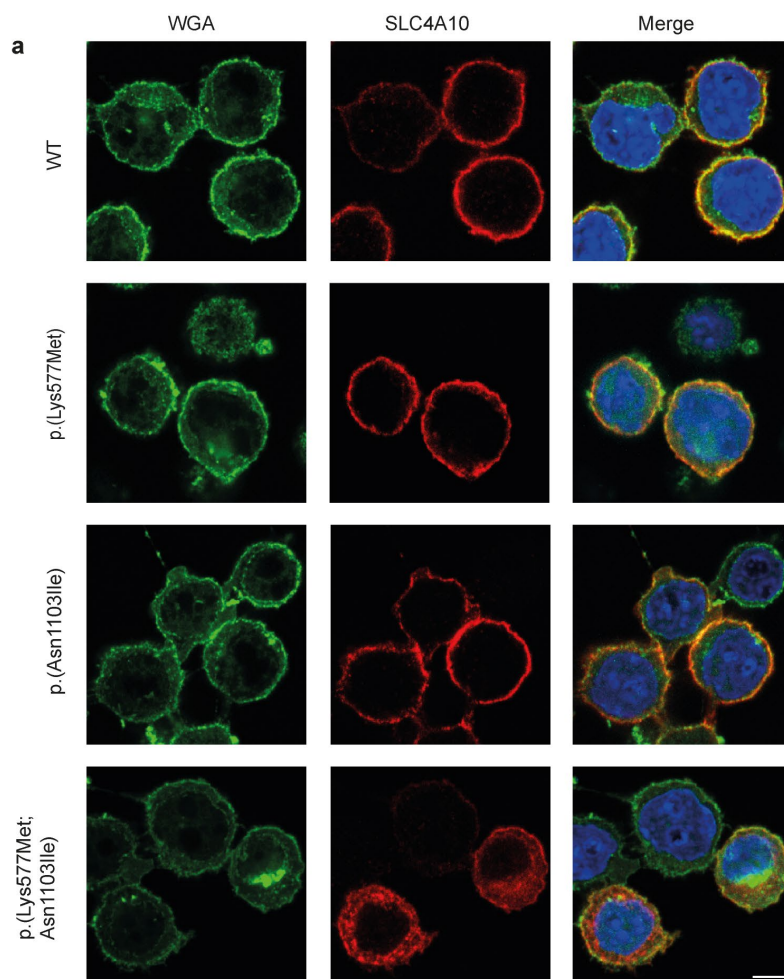
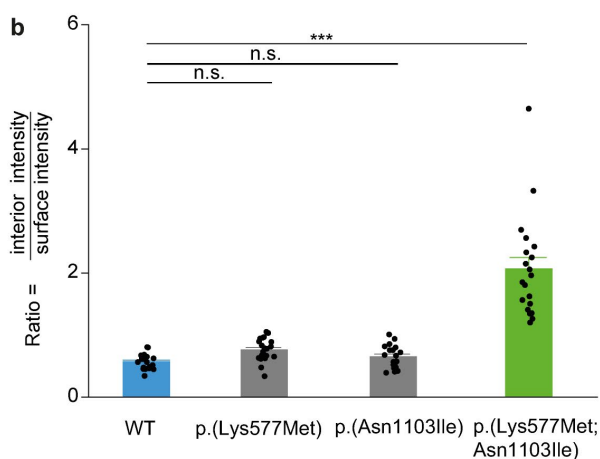


Figure 4.S6: Heterologous expression of SLC4A10 wild-type and disease associated missense variants in N2a cells

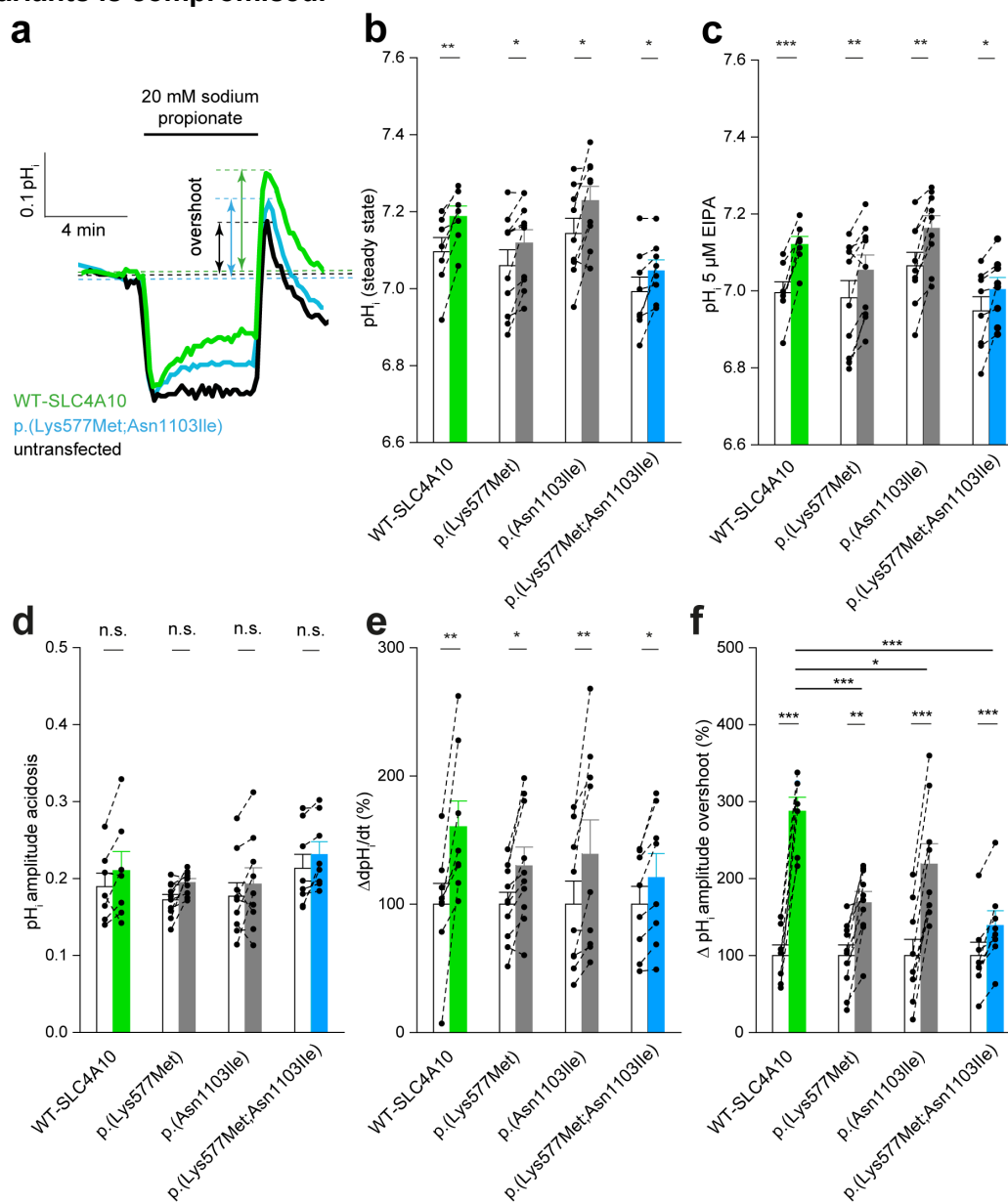


a) Two days post-transfection into the fast-growing mouse neuroblastoma cell line N2a, cells were fixed with 4% PFA and stained with an antibody directed against an N-terminal epitope of SLC4A10 and with the lectin wheat germ agglutinin (WGA) to label glycan structures associated with the Golgi apparatus or the plasma membrane. Cells transfected with the wild-type SLC4A10 construct display a predominant SLC4A10 labelling at the plasma membrane. While proteins containing either p.(Lys577Met) or p.(Asn1103Ile) were still targeted to the plasma membrane, the p.(Lys577Met;Asn1103Ile) variant was partially retained intracellularly. Scale bar: 1 μ m.



b) Quantification of the ratio of interior versus surface intensity of the SLC4A10 signal. n=12 cells each. (One-way ANOVA with Bonferroni post-hoc analysis. n.s. not significant, *** p<0.001).

Figure 4.S7: Acid extrusion by disease associated SLC4A10 missense variants is compromised.

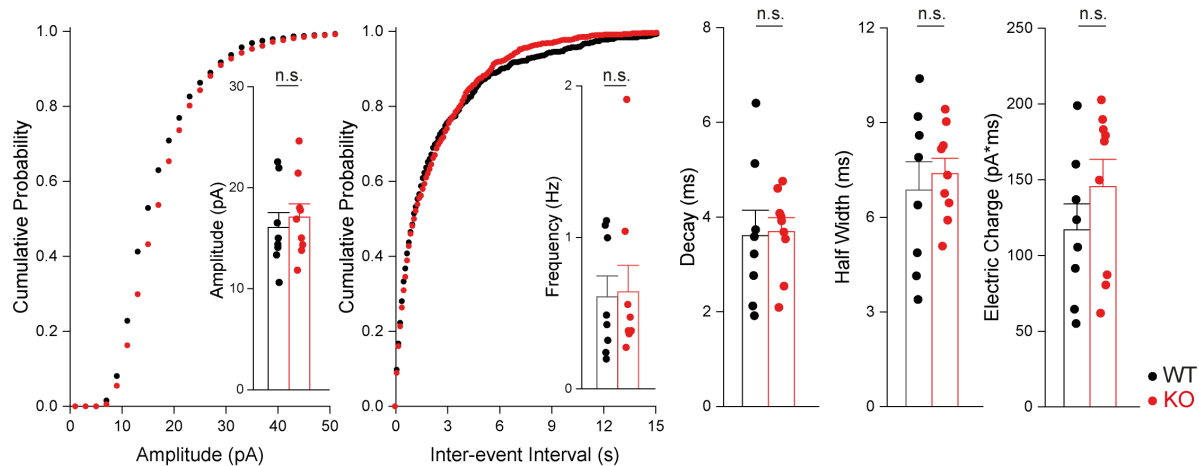


a) Representative single cell pH_i traces obtained in an untransfected N2a cell, a cell transfected with SLC4A10 WT and a cell transfected with the p.(Lys577Met;Asn1103Ile) variant superfused with bicarbonate-buffered solution with 5 μM EIPA to block Na^+/H^+ exchange. Cells were acidified by a 5 min 20 mM sodium propionate pulse. Calibration was performed with the high- $[K^+]_o$ /nigericin technique. **b,c**) The pH_i of transfected N2a cells construct was slightly more alkaline compared to untransfected cells at steady state (**b**) and in the presence of 5 μM EIPA (**c**). **d**) The amplitude of pH_i change in response to the sodium propionate pulse did not differ between transfected and untransfected cells. **e**) The recovery from the acid load did not differ between cells transfected with different SLC4A10 constructs tested. **f**) The amplitude of the overshoot was reduced for p.(Lys577Met), p.(Asn1103Ile) and the combination of both. Mean+SEM from 6 independent experiments with more than 60 cells analysed (unpaired Student's t-test; n.s.: not significant; * p<0.05; ** p<0.01; *** p<0.001)

Figure 4.S8: Disruption of *Slc4a10* reduces the mIPSC frequency in CA3 pyramidal neurons.

a) Glutamatergic transmission is not altered in CA3 neurons of *Slc4a10* KO mice. Cumulative plots and bar charts of mEPSC properties. No significant differences were detected in mEPSC frequency, amplitude or kinetics ($n=8/9$; Student's t-test). **b)** The mIPSC frequency is diminished in CA3 neurons of *Slc4a10* KO mice in the presence of HCO_3^- . Cumulative plots and bar charts of mIPSC properties ($n=6/7$; Mean+SEM; Student's t-test: * $p<0.05$; ** $p<0.01$; *** $p<0.001$; n.s.: not significant).

a mEPSCs recorded in CA3



b mIPSCs recorded in CA3

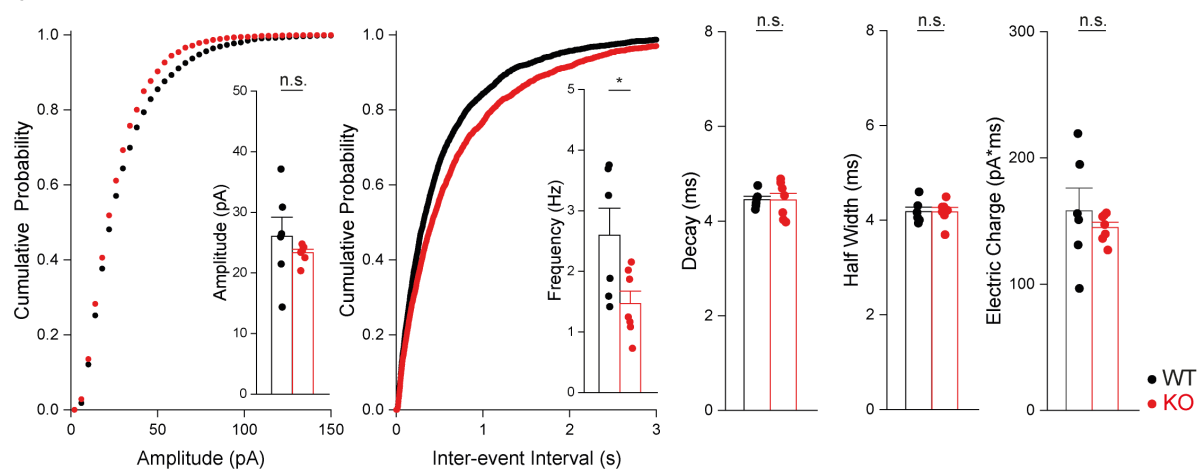


Figure 4.S9: Disruption of Slc4a10 reduces the sIPSC frequency in CA1 pyramidal neurons.

Cumulative plots and bar charts of different sIPSC properties. The sIPSC frequency was diminished in CA1 neurons of *Slc4a10* KO mice in the presence of HCO_3^- (n=21/22; Mean+SEM; Student's t-test: * p<0.05; **p<0.01; n.s.: not significant).

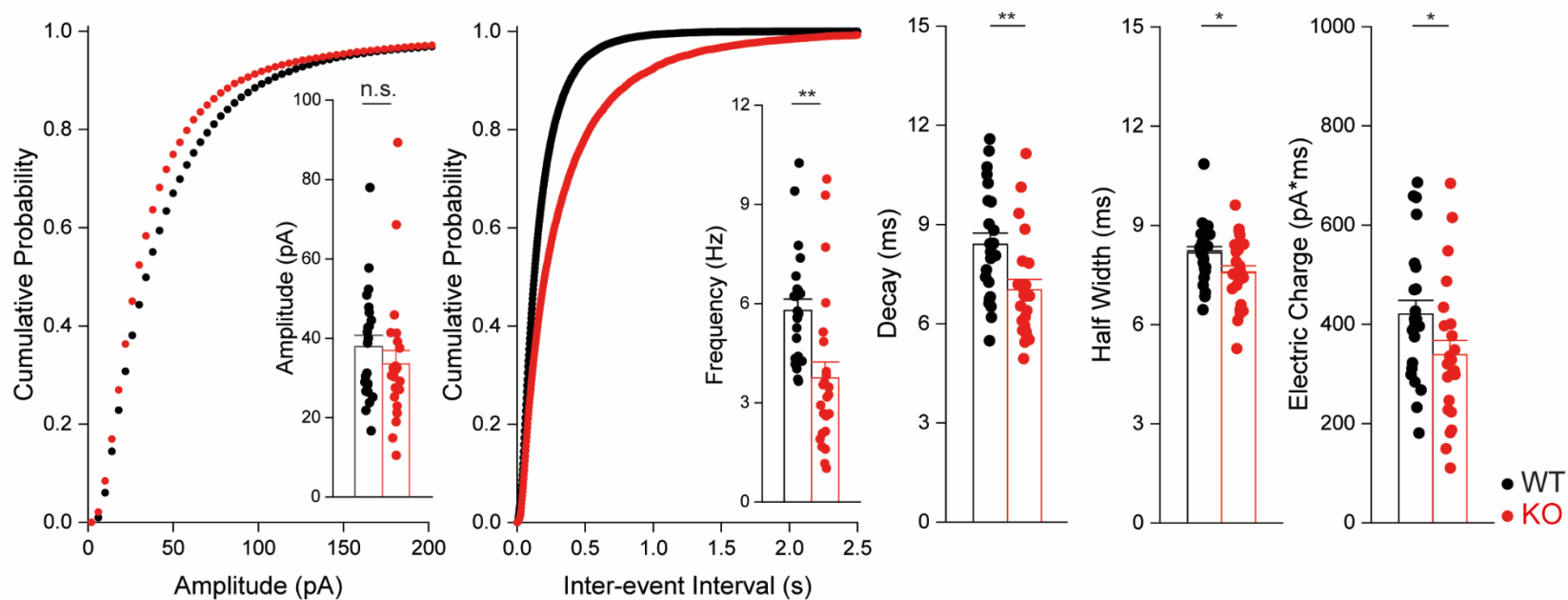


Figure 4.S10: Intracellular acidification with sodium propionate impairs GABA release.

a) In acute wild-type brain slices the mean mIPSC amplitude did not change upon substitution of 20 mM NaCl by sodium propionate.

b) mIPSC frequency was significantly diminished in the presence of 20 mM sodium propionate.

c) τ -decay, half-width and transported electric charge were not affected by sodium propionate. n=15 cells each; Mean+SEM; Student's t-test: * p<0.05; **p<0.01; ***p<0.001; n.s.: not significant.

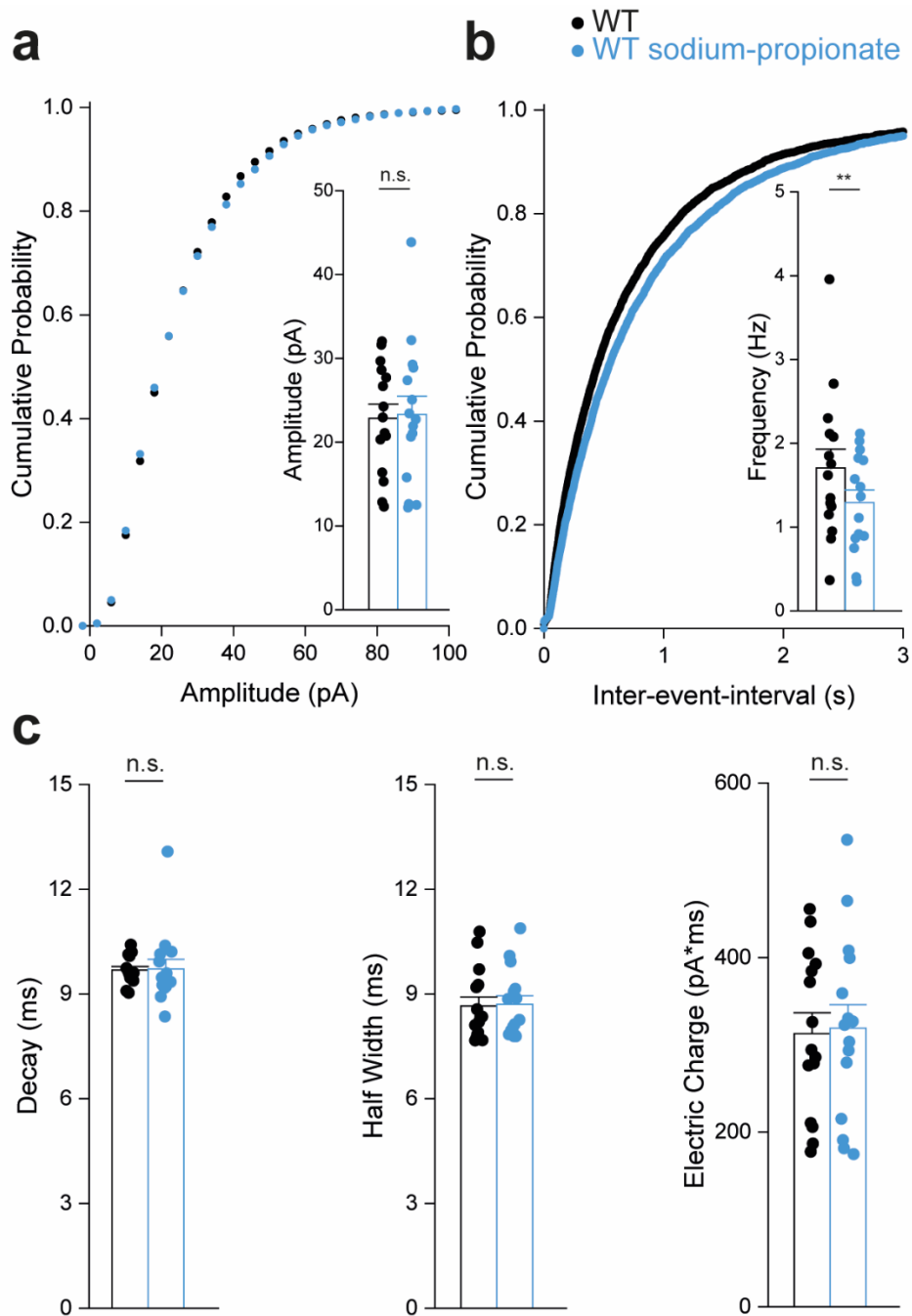
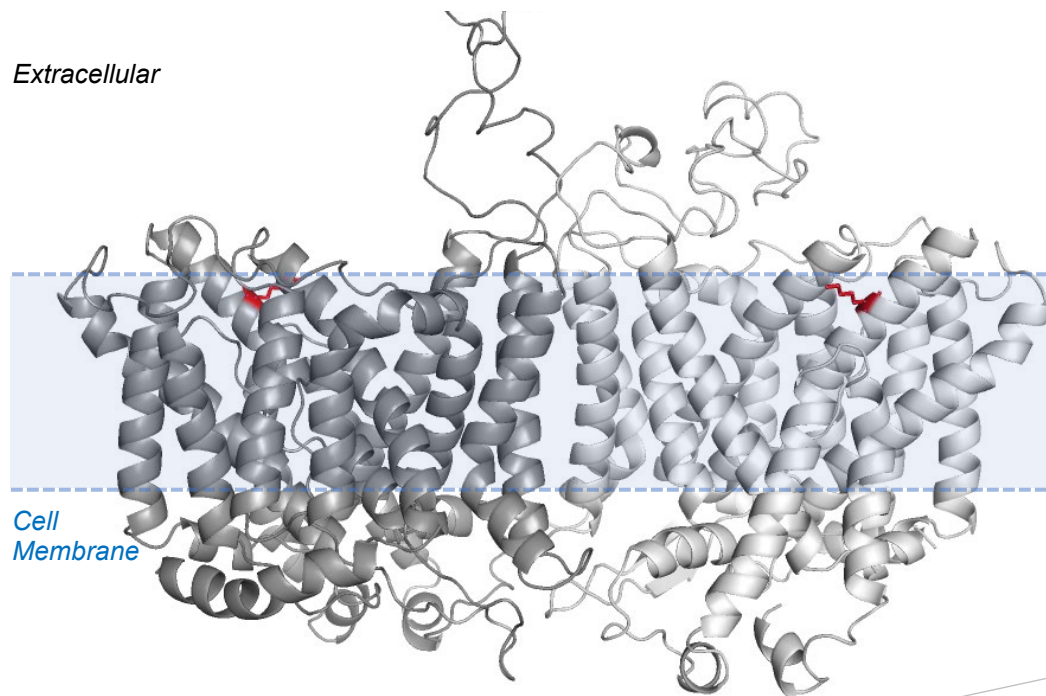


Figure 4.S11: Protein modelling of missense SLC4A10 variants

SLC4A10 p.(Lys577Met) [red] shown on 3.5 Å resolution, x-ray crystallographic structure of dimeric human SLC4A1 (4yzf)[residues 486 – 1037]. SLC4A1 is a paralogue of SLC4A10 with 41% sequence identity. The two dimeric chains are shown in light and dark grey. The change from a charged residue to a hydrophobic residue may impact positioning of the protein within the cell membrane [approximate position - blue].



Intracellular

The change from a charged residue to a hydrophobic residue may impact positioning of the protein within the cell membrane [approximate position - blue].

View from extracellular space

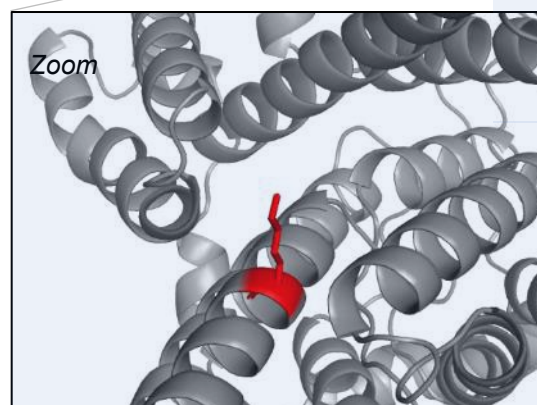


Table 4.S1: Human and murine disorders associated with other SLC4A class molecules.

Gene	Expression	Function	OMIM disorder	Inh	Human (not in OMIM)	Mouse
<i>SLC4A1</i> (<i>AE1</i>)	Blood, Kidney	mediates exchange of Cl ⁻ and HCO ₃ ⁻ across the phospholipid bilayer and plays a central role in respiration of carbon dioxide.	Cryohydrocytosis	AD		Mouse Severe spherocytosis and hemolysis (Mumtaz <i>et al.</i> , 2017; Peters <i>et al.</i> , 1996)
			Distal RTA 1	AD		
			Distal RTA 4 with haemolytic anaemia	AR		
			Ovalocytosis, SA type	AD		
			Spherocytosis, type 4	AD		
<i>SLC4A2</i> (<i>AE2</i>)	Many	nonerythroid anion exchanger	-	-	low phospholipid-associated cholelithiasis (Huynh <i>et al.</i> , 2019)	male, but not female ^{-/-} mice, infertile with histopathologic evidence of interruption in spermiogenesis (Medina <i>et al.</i> , 2003)
<i>SLC4A3</i> (<i>AE3</i>)	Heart, Ovary, others	electroneutral Cl ⁻ /HCO ₃ ⁻ exchanger	-	-	Short QT syndrome (Thorsen <i>et al.</i> , 2017) Idiopathic generalized epilepsy: increase in frequency of 867Asp variant in patients vs. controls (Sander <i>et al.</i> , 2002)	Reduced seizure threshold and increased seizure-induced mortality (Hentschke <i>et al.</i> , 2006) Double-knockout mice (with <i>Nkcc1</i> (SLC12A2)) showed impaired cardiac contractility (Prasad <i>et al.</i> , 2008)
<i>SLC4A4</i> (<i>NBC1</i>)	Brain, Pancreas, Kidney	transport of Na ⁺ and HCO ₃ ⁻ out of the corneal stroma and into the aqueous humour	Renal tubular acidosis, proximal, with ocular abnormalities (Igarashi <i>et al.</i> , 1999)	AR	-	-

<i>SLC4A5</i> (NBC4)	Thyroid, Testis (brain)	electrogenic, Cl ⁻ - independent, stilbene-inhibitable NaHCO ₃ transport	-	-	-	Abnormalities of choroid plexus and ↓ in CSF volume and pressure in lateral ventricles. CSF abnormal ion chemistry. ↓ seizure threshold, abnormal retinal architecture & ERG (Kao <i>et al.</i> , 2011)
<i>SLC4A7</i> (<i>SLC4A6</i> , NBC3)	Many (brain low)	Regulates intracellular pH and may play a role in HCO ₃ ⁻ salvage in secretory epithelia	-	-	Autism? (Satterstrom <i>et al.</i> , 2020)	Mice lacking NBC3 developed blindness and auditory impairment because of degeneration of sensory receptors in the eye and inner ear as in Usher syndrome (Bok <i>et al.</i> , 2003)
<i>SLC4A8</i> (<i>NDCBE</i> , <i>KNBC3</i>)	Brain Kidney	pH regulation in neurons	-	-	Autism? (Satterstrom <i>et al.</i> , 2020)	Reduced glutamate release in CA1 pyramidal layer + altered excitability (Sinning <i>et al.</i> , 2011) Effect on sodium reclamation in renal cortex collecting ducts (Leviel <i>et al.</i> , 2010)
<i>SLC4A9</i> (AE4)	Kidney	anion exchanger of the kidney cortex	-	-	-	Affects transepithelial absorption of NaCl by renal intercalated cells (Chambrey <i>et al.</i> , 2013)

Abbreviations: RTA = Renal tubular acidosis

Table 4.S2: SLC4A10 variants identified in affected individuals in this study

	GRCh38 g.	GRCh37 g.	nucleotide change	NM_001178 015c.	Exon /27	SIFT (<0.05)	Polyphen (prob.)	REVEL	gnomAD v2.1.1	gnomAD v3.1.1
Trp140Argfs*39	2:161846109- 161895992	2:162702619- 162752502	del	417_1341d el	5-11	NA	NA	NA	-	-
Arg757*	2:161949151	2:162805661	C>T	2269C>T	18	NA	NA	NA	-	-
Gln954_Phe955 ins*13	2:161964133	2:162820643	A>C	2863-2A>C	Intron 21/26	NA	NA	NA	-	-
Trp873*	2:161957066	2:162813576	G>A	2619G>A	20	NA	NA	NA	-	-
Lys577Met	2:161904888	2:162761398	A>T	1730A>T	14	damaging 0.048	probably damaging 0.985	0.873	-	-
Asn1103Ile	2:161976840	2:162833350	A>T	3308A>T	25	damaging 0.002	benign 0.197	0.239	-	-

Abbreviations - = Absent, NA = not applicable

Table 4.S3: Variants identified by exome / genome sequencing

Individual(s)	Variant (GRCh37)	Zygosity	Inh	Gene	Expression	OMIM phenotype	gnomAD v2.1.1 (AF)	REVEL	ClinVar	Interpretation
Family 1 III:1 & 2	Chr2(GRCh37):g.168105649G>A NM_152381.5:c.7747G>A; p.(Val2583Met)	hom	NK	<i>XIRP2</i>	Muscle; Low in brain	-	-	0.017	-	No disease association, low brain expression predicted benign
	Chr2(GRCh37):g.171256775_ 171256780del NM_138995.4:c.1869_1874del; p.(His624_Gln625del)	hom	NK	<i>MYO3B</i>	Low in brain	-	105 hets	-	-	No disease association, low brain expression
	Chr2(GRCh37):g.170010978A>G NM_004525.2:c.12287T>C; p.(Ile4096Thr)	hom	NK	<i>LRP2</i>	Kidney; Thyroid, Low in brain	Donnai-Barrow syndrome, 222448	109 hets	0.255	VUS	Published phenotype is absent, predicted benign
	Chr5(GRCh37):g.179160382C>G NM_014757.4:c.269C>G; p.(Pro90Arg)	hom	NK	<i>MAML1</i>	Widespread	-	-	0.066	-	No disease association, predicted benign
	Chr5(GRCh37):g.195286G>A NM_001080478.2:c.1363G>A; p.(Ala455Thr)	hom	NK	<i>LRRC14B</i>	Muscle; Low in brain	-	-	0.048	-	No disease association, low brain expression predicted benign
	Chr6(GRCh37):g.90333750C>T NM_014942.4:c.1192C>T; p.(Arg398Trp)	hom	NK	<i>ANKRD6</i>	Widespread	-	11 hets	0.397	-	No disease association
	Chr11(GRCh37):g.33566490A>C NM_012194.2:c.2060A>C; p.(Asn687Thr)	hom	NK	<i>KIAA1549L</i>	Brain	-	-	0.179	-	No disease association, predicted benign
	Chr11(GRCh37):g.118769236G>A NM_182557.2:c.4388C>T p.(Pro1463Leu)	hom	NK	<i>BCL9L</i>	Widespread	-	2 hets	0.272	-	No disease association
	Chr11(GRCh37):g.34144027A>G NM_024662.2:c.802A>G; p.(Ile268Val)	hom	NK	<i>NAT10</i>	Widespread	-	5 hets	0.204	-	No disease association, predicted benign
	Chr11(GRCh37):g.124056686C>G NM_001355213.1:c.710C>G; p.(Ala237Gly)	hom	NK	<i>OR10D3</i>	Testis; Low in brain	-	4 hets	0.235	-	No disease association, low brain expression
Chr11(GRCh37):g.108385518C>G NM_015065.2:c.716G>C; p.(Arg239Thr)	hom	NK	<i>EXPH5</i>	Skin and cerebellum	Epidermolysis bullosa, nonspecific,	137 hets	0.106	-	Published phenotype is absent	

Chapter 4

Family 1 III:1 & 2	Chr11(GRCh37):g.108409802A>T NM_015065.2:c.392T>A; p.(Phe131Tyr)	hom	NK			autosomal recessive, 615028	138 hets	0.125	-	Published phenotype is absent
	Chr12(GRCh37):g.55356369G>T NM_001098815.2:c.1313C>A; p.(Ala438Glu)	hom	NK	<i>TESPA1</i>	Brain	-	-	0.152	-	No disease association
	Chr12(GRCh37):g.49445208_ 49445234del NM_003482.3:c.2232_2258del; p.(Arg755_Pro763del)	hom	NK	<i>KMT2D</i>	Widespread	Kabuki syndrome 1, 147920	22 hets	-	VUS	inheritance is AD; parents are unaffected
	Chr22(GRCh37):g.32289717G>A NM_001136029.2:c.4156G>A; p.(Ala1386Thr)	hom	NK	<i>DEPDC5</i>	Widespread	Epilepsy, familial focal, with variable foci 1, 604364	18 hets	0.251	-	inheritance is AD; parents are unaffected
	ChrX(GRCh37):g.18283700C>G NM_006089.2:c.948+5G>C; p.?	hemi	NK	<i>SCML2</i>	Testis / Ovary, Low in brain	-	-	-	-	No disease association, low brain expression
	ChrX(GRCh37):g.47920225C>G NM_001037735.3:c.115G>C; p.(Glu39Gln)	hemi	NK	<i>ZNF630</i>	Widespread	-	-	0.205	-	No disease association
	ChrX(GRCh37):g.100401154T>C NM_006733.3:c.1714T>C; p.(Tyr572His)	hemi	NK	<i>CENPI</i>	Low in brain	-	-	0.076	-	No disease association, low brain expression, predicted benign
	Chr8(GRCh37):g.107749806G>A NM_001198532.1:c.2018G>A; p.(Arg673His)	het	NK	<i>OXR1</i>	Widespread	Cerebellar hypoplasia/atro phy, epilepsy, and global developmental delay, 213000	4 hets	0.398	-	All described cases have had biallelic loss- of-function variants
	Chr8(GRCh37):g.107763087C>G NM_001198532.1:c.2543C>G; p.(Ser848Cys)	het	NK				6 hets	0.298	-	
	Chr7(GRCh37):g.151845190G>A NM_170606.3:c.13822C>T; p.(Arg4608Cys)	het	NK	<i>KMT2C</i>	Widespread	Kleefstra syndrome 2, 617768	4 hets	0.217	-	Published phenotype is absent, one variant predicted benign
Chr7(GRCh37):g.151946979C>A NM_170606.3:c.1795G>T; p.(Asp599Tyr)	het	NK	39 hets				0.321	Not defined		
Chr14(GRCh37):g.105609114G>A NM_145159.2:c.3521C>T; p.(Pro1174Leu)	het	NK	<i>JAG2</i>	Widespread	-	90 hets	0.169	-	No disease association, one variant predicted benign	
Chr14(GRCh37):g.105612780G>A NM_145159.2:c.2537C>T; p.(Ser846Phe)	het	NK				-	-	0.398		-

Family 2 II:1	Chr3(GRCh37):g.148928067T>C NM_000096.4:c.494A>G; p.(Gln165Arg)	het	NK	<i>CP</i>	Liver	Hemosiderosis, systemic, due to acerulo- plasminemia, 604290	-	0.423	-	Late onset disorder
Family 3 II:2 & 3	Chr2(GRCh37):g.182386959A>G NM_000885:c.1964A>G; p.(Lys655Arg)	hom	NK	<i>ITGA4</i>	Lymphocytes	-	-	0.208	-	Mouse KO exhibit embryonic lethality
	Chr2(GRCh37):g.225376319A>G NM_003590:c.655-20T>C	hom	NK	<i>CUL3</i>	Testis	Neurodevelopm ental disorder with or without autism or seizures, 619239	4 hets	-	-	OMIM disorder is AD SpliceAI predicts no splicing change (0.01)
	Chr13(GRCh37):g.21099923G>T NM_015974:c.11C>A; p.(Ser4Tyr)	hom	NK	<i>CRYL1</i>	Liver, Kidney	-	11 hets	0.125	-	No disease association, predicted benign
Family 4 III:2 & III:3	Chr3(GRCh37):g.33661190C>T NM_001207044.1:c.523G>A; p.(Ala175Thr)	hom	NK	<i>CLASP2</i>	Widespread	-	2 hets	0.477	-	No disease association
Family 5 II:1 & 2	Chr1(GRCh37):g.19166140C>T NM_152232.2:c.2473G>A, p.(Ala825Thr)	hom	Bipar- ental	<i>TAS1R2</i>	Low, skin only	-	49 hets	0.242	-	Taste receptor No disease association

Abbreviations: -, absent; AD, autosomal dominant; hemi; hemizygous, het, heterozygous; hets, heterozygous individuals; hom, homozygous; NK, not known; ProbD Probably damaging; PossD, possibly damaging; VUS, Variant of uncertain significance.

Table 4.S4: Summary of electrophysiological recording from acute brain slices of *Slc4a10* WT and knock-out mice

	WT (n)	KO (n)	KO+TriMA (n)	t-test	one-way ANOVA with Newman-Keuls posthoc	KS-Test
capacitance (pF)	25.7±1.0 (22)	26.0±0.7 (31)		p=0.79		
input resistance (MΩ)	60.6±2.5 (22)	60.9±1.6 (31)		p=0.91		
<i>mEPSCs</i>						
amplitude (pA)	16.3±0.6 (10)	15.9±0.8 (12)		p=0.76		p>0.05
frequency (Hz)	1.2±0.3 (10)	0.9±0.1 (12)		p=0.40		p>0.05
T _{decay} (ms)	3.5±0.2 (10)	3.9±0.7 (12)		p=0.16		
charge transfer (pA*ms)	118.1±10.0 (10)	115.4±6.8 (12)		p=0.83		
<i>mIPSCs</i>						
amplitude (pA)	29.5±3.1 (12)	26.6±1.4 (19)	26.8±4.0 (11)		F=0.36, p=0.70; KO vs. WT p>0.05; WT vs. KO+TriMA: p>0.05	p>0.05
frequency (Hz)	5.0±0.5 (12)	2.2±0.3 (19)	3.7±0.59 (11)		F=10.99, p=0.0002; KO vs. WT p<0.001; WT vs. KO+TriMA p>0.05	p<0.0001
T _{decay} (ms)	5.4±0.4 (12)	3.8±0.2 (19)	3.8±0.4 (11)		F=7.60, p=0.002; WT vs. KO p<0.001; KO vs. KO+TriMA p>0.05	
charge transfer (pA*ms)	199.2±20.1 (12)	143.3±9.8 (19)	145.8±20.9 (11)		F=4.58, p=0.016; KO vs. WT p<0.05; KO vs. KO+TriMA: p>0.05	
<i>mIPSCs (HEPES)</i>						
amplitude (pA)	27.0±2.1 (12)	27.3±2.4 (14)		p=0.93		
frequency (Hz)	2.9±0.3 (12)	2.5±0.3 (14)		p=0.43		
T _{decay} (ms)	5.1±0.2 (12)	5.1±0.3 (14)		p=0.91		p>0.05
charge transfer (pA*ms)	206.4±21.5 (12)	210.2±22.0 (14)		p=0.90		p>0.05

Table 4.S5: SLC4A10 GWAS associations

A) associations with cognitive ability: MTAG: multi-trait analysis of genome-wide association studies (GWAS)

GRCh38 Chr:Pos	RsID	Risk allele	Site	P-Value	Effect size (β)	Trait	Ref
2:161962111	rs4500960	T	Intronic	2×10^{-37}	0.021 unit ↓	Highest math class taken (MTAG)	[1]
2:161962111	rs4500960	T	Intronic	3×10^{-30}	0.015 unit ↓	Educational attainment (MTAG)	[1]
2:161962111	rs4500960	T	Intronic	7×10^{-26}	0.024 unit ↓	Cognitive performance (MTAG)	[1]
2:161962111	rs4500960	T	Intronic	1×10^{-20}	0.013 unit ↓	Educational attainment (years of education)	[1]
2:161962111	rs4500960	T	Intronic	3×10^{-11}	6.637 z-score ↓	General cognitive ability	[2]
2:161962111	rs4500960	T	Intronic	3×10^{-10}	0.014 unit ↓	Educational attainment (years of education)	[3]
2:161962111	rs4500960	C	Intronic	2×10^{-20}	0.021 unit ↑	Highest math class taken	[1]
2:161971491	rs4664442	A	Intronic	5×10^{-17}	0.025 unit ↓	Intelligence (MTAG)	[4]
2:161971491	rs4664442	?	Intronic	1×10^{-13}	0.015 unit ↓	Cognitive ability, years of educational attainment or schizophrenia (pleiotropy)	[5]
2:161945674	rs11693702	A	Intronic	9×10^{-13}	0.021 unit ↓	Cognitive performance	[1]
2:161957297	rs10221808	A	Intronic	3×10^{-10}	0.010 unit ↓	Educational attainment (years of education)	[1]
2:161957297	rs10221808	?	Intronic	7×10^{-8}	-	Educational attainment (years of education)	[6]
2:161988766	rs2098526	A	Downstream	2×10^{-8}	0.0241 unit ↓	Educational attainment (years of education)	[1]

B) associations with brain volume: DS: Downstream, TF site: Transcription factor site, CRE: cis-regulatory element

GRCh38 Chr:Pos	RsID	Risk allele	Site	P-Value	Effect size (β)	Trait	Ref
2:161989055	rs1861979	T	TF site (DS)	9×10^{-22}	11.92 mm ³ ↑	Hippocampal tail volume	[7]
2:161989055	rs1861979	T	TF site (DS)	5×10^{-13}	39.54 mm ³ ↑	Total hippocampal volume	[7]
2:161989055	rs1861979		TF site (DS)	2×10^{-10}	6.42 mm ³ ↑	Dentate gyrus molecular layer volume	[7]
2:161989929	rs2909443	G	Downstream	3×10^{-13}	6.11 mm ³ ↑	Hippocampal tail volume (corrected for total hippocampal volume)	[7]

2:161986568	rs2909455	T	CRE (DS)	6×10^{-11}	5.4 mm ³ ↑	Subiculum volume	[7]
-------------	-----------	---	----------	---------------------	-----------------------	------------------	-----

C) associations with mental health outcomes: ADHD: attention-deficit hyperactivity disorder, CBT: cognitive behavioural therapy

GRCh38 Chr:Pos	RsID	Risk allele	Site	P-Value	Effect size	Trait	Ref
2:161989345	rs2909457	G	Intronic	5×10^{-8}	OR 1.06	Schizophrenia	[8]
2:161989345	rs2909457	G	Intronic	6×10^{-8}	OR 1.06	Schizophrenia	[9]
2:161989345	rs2909457	?	Intronic	1×10^{-7}		Schizophrenia	[10]
2:161972220	rs34685708	?	Intronic	4×10^{-7}		Schizophrenia	[11]
2:161719475	rs56037433	?	Intronic	9×10^{-7}		Bipolar disorder or ADHD	[12]
2:161587424	rs79996792	A	Upstream	3×10^{-6}	0.019 unit ↓	Neuroticism	[13]
2:161476919	rs12468729	G	Upstream	4×10^{-6}	OR 1.07	Bipolar disorder	[14]
2:161443775	rs13432654	?	Upstream	8×10^{-6}		Anxiety disorder, CBT	[15]

References for Table 4.S5:

- [1] Lee, J.J., *et al.*, Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet.* 2018. 50(8), 1112-1121.
- [2] Davies, G., *et al.*, Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat Commun.* 2018. 9(1), 2098.
- [3] Okbay, A., *et al.*, Genome-wide association study identifies 74 loci associated with educational attainment. *Nature.* 2016. 533(7604), 539-42.
- [4] Hill, W.D., *et al.*, A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol Psychiatry.* 2019. 24(2), 169-181.
- [5] Lam, M., *et al.*, Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways. *Am J Hum Genet.* 2019. 105(2), 334-350.
- [6] Kichaev, G., *et al.*, Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet.* 2019. 104(1), 65-75.
- [7] van der Meer, D., *et al.*, Brain scans from 21,297 individuals reveal the genetic architecture of hippocampal subfield volumes. *Mol Psychiatry.* 2020. 25(11), 3053-3065.
- [8] Ripke, S., *et al.*, Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014. 511(7510): p. 421-427.
- [9] Goes, F.S., *et al.*, Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am J Med Genet B Neuropsychiatr Genet.* 2015. 168(8), 649-59.

- [10] Periyasamy, S., *et al.*, Association of Schizophrenia Risk with Disordered Niacin Metabolism in an Indian Genome-wide Association Study. *JAMA Psychiatry*. 2019. 76(10), 1026-1034.
- [11] Pardiñas, A.F., *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet*. 2018. 50(3), 381-389.
- [12] van Hulzen, K.J.E., *et al.*, Genetic Overlap Between Attention-Deficit/Hyperactivity Disorder and Bipolar Disorder: Evidence From Genome-wide Association Study Meta-analysis. *Biol Psychiatry*, 2017. 82(9), 634-641.
- [13] Okbay, A., *et al.*, Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet*. 2016. 48(6), 624-33.
- [14] Stahl, E.A., *et al.*, Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat Genet*. 2019. 51(5), 793-803.
- [15] Coleman, J.R., *et al.*, Genome-wide association study of response to cognitive-behavioural therapy in children with anxiety disorders. *Br J Psychiatry*, 2016. 209(3), 236-43.

4.4. Further findings and future work

During the process of completing these investigations two additional families with biallelic *SLC4A10* gene variants and neurodevelopmental disorders were identified (Families A & B, **Appendix 7.17** and **Appendix 7.18**). In each case there was insufficient evidence to definitively ascribe the disease to the *SLC4A10* gene variants and so these were not included in the submitted manuscript. However, functional analyses of these variants [NM_001178015:c.667C>T; p.(His223Tyr), NM_001178015:c.2894C>T; p.(Pro965Leu)] may provide some additional information about the pathophysiology of the *SLC4A10*-related disorder. Both additional families have apparently autosomal recessive mild-severe ID with microcephaly (OFC -2.6 SDS to -4.2 SDS), consistent with individuals described above with the *SLC4A10*-related neurodevelopmental disorder. Additionally, two affected individuals in Family A share mild gait ataxia and dysarthria, not described in other individuals in the *SLC4A10*-related disorder cohort. Affected individuals from Family A and Family B do not display the typical neuroradiological features of the *SLC4A10*-related neurodevelopmental disorder (slit lateral ventricles and abnormalities of the corpus callosum), although missense alleles associated with monogenic conditions may have a less deleterious (i.e. hypomorphic) effect than loss-of-function alleles. We see some evidence of this in the *SLC4A10*-related disorder already, with Family 5 [NM_001178015: p.(Lys577Met;Asn1103Ile)] where the neuroradiological phenotype is less pronounced than in individuals with predicted loss-of-function variants. Instead, the neuroradiology for Family A shows mild delayed myelination and in Family B

there is periventricular heterotopia, suggestive of a neuronal migration disorder (**Appendix 7.17d**). These findings also raise questions about the co-occurrence of an additional genetic diagnosis, known to occur more frequently in populations with high homozygosity (Posey *et al.* 2017; Smith *et al.* 2019). Both identified missense variants are absent in population databases and predicted damaging by *in silico* tools (**Appendix 7.19**). The p.(His223Tyr) variant affects a highly conserved residue within the conserved cytoplasmic domain of SLC4A10 (**Appendix 7.20**), but no additional topological information is available as there is no reliable protein crystal structure available for this region and AlphaFold (Deepmind, London) does not predict this region with confidence. The p.(Pro955Leu) variant affects a highly conserved proline residue within an extracellular loop of the bicarbonate-transporter domain of SLC4A10. This region shows significant regional missense constraint with a complete absence of potentially benign missense variants present in a homozygous state in gnomAD (shown as green lollipops). Protein modelling (**Appendix 7.21**) demonstrates that this variant is located towards the centre of the dimeric complex suggesting a potential role in transport or dimerisation. Additional functional studies show abnormal pH overshoot, also seen with the null allele, suggestive of abnormal HCO₃⁻ transporter function (**Appendix 7.22**). In the case of the p.(His223Tyr) variant this may be partly explained by the abnormal localisation of the protein away from the cell surface, similar to the null allele.

Our findings highlight pharmacological modulation of GABA as a possible therapy for the *SLC4A10*-related neurodevelopmental disorder. These agents are potentially appealing as a number of GABA_A receptor agonists, such as benzodiazepines, baclofen and progabide, have already demonstrated a

favourable safety profile in trials and achieved licenses for the treatment of spasticity, dyskinesias, epilepsy and sleep disorders (Alabed *et al.* 2018). Dysregulated GABAergic signalling has also been noted in fragile X syndrome, likely the most common monogenic cause of developmental delay worldwide. In particular, studies of an *Fmr1*^{-/-} mouse model of fragile X syndrome demonstrated that the δ -subunit-selective, extra-synaptic GABA_A receptor agonist gaboxadol normalised aberrant hyperactive and anxiety-like behaviour behaviours (Cogram *et al.* 2019). Following these, phase 2a studies (performed in a small number of human subjects to assess the safety tolerability and efficacy and optimum dose) in patients with moderate-to-severe neurobehavioral phenotypes in fragile X syndrome have recently reported efficacy for the same agent (Budimirovic *et al.* 2021), with a further study planned. Given possible pathomechanistic similarities between Fragile X and the *SLC4A10*-related neurodevelopmental disorder these findings are encouraging and suggest a possible therapeutic avenue. To further develop the rationale for this, and eventually development of a pharmacological treatment for patients, studies in an *Slc4a10*^{-/-} mouse model would be valuable to provide evidence of efficacy, such as improvement in the Novel Object Recognition test suggesting cognitive improvement.

From a mechanistic point of view, it will be important to try and unpick the relative contributions of CSF dysregulation, probably mediated by a previously described role at the choroid plexus, and neuronal GABAergic dysfunction to the neurodevelopmental disorder phenotype. A choroid-plexus-specific conditional knockout of *Slc4a10* in a mouse model would be informative with the Cre/lox system being the most common method of achieving this. This involves

coupling an inducible Cre recombinase a choroid specific promotor, inducing recombination of two introduced loxP (locus of x-over, P1) sites around the gene of interest (Kim *et al.* 2018). At least nine mouse lines targeting the choroid plexus using this approach have been described (Jang *et al.* 2022).

4.5. References

- Alabed S, Latifeh Y, Mohammad HA, Bergman H. Gamma-aminobutyric acid agonists for antipsychotic-induced tardive dyskinesia. *Cochrane Database Syst Rev*. 2018;4(4):Cd000203.
- Alper SL, Sharma AK. The SLC26 gene family of anion transporters and channels. *Mol Aspects Med*. 2013;34(2-3):494-515.
- Belengeanu V, Gamage TH, Farcas S, Stoian M, Andreescu N, Belengeanu A, Frengen E, & Misceo D. A de novo 2.3 Mb deletion in 2q24.2q24.3 in a 20-month-old developmentally delayed girl. *Gene* 2014;539(1):168-172.
- Bjorefeldt A, Illes S, Zetterberg H, Hanse E. Neuromodulation via the Cerebrospinal Fluid: Insights from Recent in Vitro Studies. Review. *Front Neural Circuits*. 2018;12(5)
- Bocker HT, Heinrich T, Liebmann L, Hennings JC, Seemann E, Gerth M, Jakovčevski I, Preobraschenski J, Kessels MM, Westermann M, *et al*. The Na⁺/H⁺ Exchanger Nhe1 Modulates Network Excitability via GABA Release. *Cereb Cortex*. 2019;29(10):4263-4276.
- Bok D, Galbraith G, Lopez I, Woodruff M, Nusinowitz S, BeltrandelRio H, Huang W, Zhao S, Geske R, Montgomery C, *et al*. Blindness and auditory impairment caused by loss of the sodium bicarbonate cotransporter NBC3. *Nat Genet* 2003;34(3):313-319.
- Budimirovic DB, Dominick KC, Gabis LV, Adams M, Adera M, Huang L, Ventola P, Tartaglia NR, Berry-Kravis E. Gaboxadol in Fragile X Syndrome: A 12-Week Randomized, Double-Blind, Parallel-Group, Phase 2a Study. *Front Pharmacol*. 2021;12:757825.
- Burette AC, Weinberg RJ, Sassani P, Abuladze N, Kao L, Kurtz I. The sodium-driven chloride/bicarbonate exchanger in presynaptic terminals. *J Comp Neurol*. 2012;520(7):1481-92.
- Chambrey R, Kurth I, Peti-Peterdi J, Houillier P, Purkerson JM, Leviel F, Hentschke M, Zdebik AA, Schwartz GJ, Hübner CA, *et al*. Renal intercalated cells are rather energized by a proton than a sodium pump. *Proc Natl Acad Sci U S A* 2013;110(19):7928-7933.
- Chazotte B. Labeling membrane glycoproteins or glycolipids with fluorescent wheat germ agglutinin. *Cold Spring Harb Protoc*. 2011;2011(5)
- Chesler M. Regulation and modulation of pH in the brain. *Physiol Rev*. 2003;83(4):1183-221.
- Cogram P, Deacon RMJ, Warner-Schmidt JL, von Schimmelmann MJ, Abrahams BS, During MJ. Gaboxadol Normalizes Behavioral Abnormalities in a Mouse Model of Fragile X Syndrome. *Front Behav Neurosci*. 2019;13:141.
- Damkier HH, Aalkjaer C, Praetorius J. Na⁺-dependent HCO₃⁻ import by the slc4a10 gene product involves Cl⁻ export. *J Biol Chem*. 2010;285(35):26998-7007.
- Damkier HH, Brown PD, Praetorius J. Cerebrospinal fluid secretion by the choroid plexus. *Physiol Rev*. 2013;93(4):1847-92.

- Dietrich CJ, Morad M. Synaptic acidification enhances GABAA signaling. *J Neurosci*. 2010;30(47):16044-52.
- Doering CJ, McRory JE. Effects of extracellular pH on neuronal calcium channel activation. *Neuroscience*. 2007;146(3):1032-43.
- Dunn KW, Kamocka MM, McDonald JH. A practical guide to evaluating colocalization in biological microscopy. *American Journal of Physiology-Cell Physiology*. 2011;300(4):C723-C742.
- Egashira Y, Takase M, Watanabe S, Ishida J, Fukamizu A, Kaneko R, Yanagawa Y, & Takamori S. Unique pH dynamics in GABAergic synaptic vesicles illuminates the mechanism and kinetics of GABA loading. *Proc Natl Acad Sci U S A*. 2016;113(38):10702-10707.
- Eisner DA, Kenning NA, O'Neill SC, Pocock G, Richards CD, Valdeolmillos M. A novel method for absolute calibration of intracellular pH indicators. *Pflugers Arch*. 1989;413(5):553-8.
- Farrant M, Kaila K. The cellular, molecular and ionic basis of GABA(A) receptor signalling. *Prog Brain Res*. 2007;160:59-87.
- Farsi Z, Preobraschenski J, van den Bogaart G, Riedel D, Jahn R, Woehler A. Single-vesicle imaging reveals different transport mechanisms between glutamatergic and GABAergic vesicles. *Science*. 2016;351(6276):981-4.
- Fegghi T, Hernandez RX, Stawarski M, *et al*. Computational modeling predicts ephemeral acidic microdomains in the glutamatergic synaptic cleft. *Biophysical Journal* 2021;120(24):5575-5591.
- Fraire-Zamora JJ, González-Martínez MT. Effect of intracellular pH on depolarization-evoked calcium influx in human sperm. *Am J Physiol Cell Physiol*. 2004;287(6):C1688-96.
- Frerking M, Borges S, Wilson M. Variation in GABA mini amplitude is the consequence of variation in transmitter concentration. *Neuron*. 1995;15(4):885-895.
- Grayson B, Idris NF, Neill JC. Atypical antipsychotics attenuate a sub-chronic PCP-induced cognitive deficit in the novel object recognition task in the rat. *Behav Brain Res*. 2007;184(1):31-8.
- Grichtchenko, II, Choi I, Zhong X, Bray-Ward P, Russell JM, Boron WF. Cloning, characterization, and chromosomal mapping of a human electroneutral Na(+)-driven Cl-HCO₃ exchanger. *J Biol Chem*. 2001;276(11):8358-63.
- Guissart C, Li X, Leheup B, Drouot N, Montaut-Verient B, Raffo E, Jonveaux P, Roux AF, Claustres M, Fliegel L, *et al*. Mutation of SLC9A1, encoding the major Na⁺/H⁺ exchanger, causes ataxia-deafness Lichtenstein-Knorr syndrome. *Hum Mol Genet*. 2015;24(2):463-70.
- Gurnett CA, Veile R, Zempel J, Blackburn L, Lovett M, Bowcock A. Disruption of sodium bicarbonate transporter SLC4A10 in a patient with complex partial epilepsy and mental retardation. *Archives of neurology*. 2008;65(4):550-3.
- Hentschke M, Wiemann M, Hentschke S, Kurth I, Hermans-Borgmeyer I, Seidenbecher T, Jentsch TJ, Gal A, & Hübner CA. Mice with a targeted disruption of the Cl⁻/HCO₃⁻ exchanger AE3 display a reduced seizure threshold. *Mol Cell Biol* 2006;26(1):182-191.

- Heulens I, D'Hulst C, Van Dam D, De Deyn PP, Kooy RF. Pharmacological treatment of fragile X syndrome with GABAergic drugs in a knockout mouse model. *Behav Brain Res*. 2012;229(1):244-9.
- Horvath PM, Piazza MK, Monteggia LM, Kavalali ET. Spontaneous and evoked neurotransmission are partially segregated at inhibitory synapses. *eLife*. 2020;9:e52852.
- Huebner AK, Maier H, Maul A, *et al*. Early Hearing Loss upon Disruption of Slc4a10 in C57BL/6 Mice. *J Assoc Res Otolaryngol*. 2019;20(3):233-245.
- Huynh MT, Nguyen TT, Grison S, Lascols O, Fernandez E, & Barbu V. Clinical characteristics and genetic profiles of young and adult patients with cholestatic liver disease. *Rev Esp Enferm Dig* 2019;111(10):775-788.
- Igarashi T, Inatomi J, Sekine T, Cha SH, Kanai Y, Kunimi M, Tsukamoto K, Satoh H, Shimadzu M, Tozawa F, *et al*. Mutations in SLC4A4 cause permanent isolated proximal renal tubular acidosis with ocular abnormalities. *Nature genetics* 1999;23(3):264-266.
- Jacobs S, Ruusuvuori E, Sipilä ST, *et al*. Mice with targeted Slc4a10 gene disruption have small brain ventricles and show reduced neuronal excitability. *Proc Natl Acad Sci U S A*. 2008;105(1):311-6.
- Jang A, Lehtinen MK. Experimental approaches for manipulating choroid plexus epithelial cells. *Fluids and Barriers of the CNS*. 2022;19(1):36.
- Kaila K, Voipio J. Postsynaptic fall in intracellular pH induced by GABA-activated bicarbonate conductance. *Nature*. 1987;330(6144):163-5.
- Kane MJ, Angoa-Peréz M, Briggs DI, Sykes CE, Francescutti DM, Rosenberg DR, & Kuhn DM. Mice genetically depleted of brain serotonin display social impairments, communication deficits and repetitive behaviors: possible relevance to autism. *PloS one*. 2012;7(11):e48975.
- Kao L, Kurtz LM, Shao X, Papadopoulos MC, Liu L, Bok D, Nusinowitz S, Chen B, Stella SL, Andre M, *et al*. Severe neurologic impairment in mice with targeted disruption of the electrogenic sodium bicarbonate cotransporter NBCe2 (Slc4a5 gene). *J Biol Chem* 2011;286(37):32563-32574.
- Kim H, Kim M, Im SK, Fang S. Mouse Cre-LoxP system: general principles to determine tissue-specific roles of target genes. *Lab Anim Res*. 2018;34(4):147-159.
- Krepischi AC, Knijnenburg J, Bertola DR, Kim CA, Pearson PL, Bijlsma E, Suzhai K, Kok F, Vianna-Morgante AM, & Rosenberg C. Two distinct regions in 2q24.2-q24.3 associated with idiopathic epilepsy. *Epilepsia*. 2010;51(12):2457-60.
- Laver TW, De Franco E, Johnson MB, Patel KA, Ellard S, Weedon MN, Flanagan SE, & Wakeling MN. SavvyCNV: Genome-wide CNV calling from off-target reads. *PLOS Computational Biology*. 2022;18(3):e1009940.
- Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, Nguyen-Viet TA, Bowers P, Sidorenko J, Karlsson Linnér R, *et al*. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet*. 2018;50(8):1112-1121.
- Leviel F, Hübner CA, Houillier P, Morla L, El Moghrabi S, Brideau G, Hassan H, Parker MD, Kurth I, Kougioumtzes A, *et al*. The Na⁺-dependent chloride-bicarbonate exchanger

- SLC4A8 mediates an electroneutral Na⁺ reabsorption process in the renal cortical collecting ducts of mice. *J Clin Invest* 2010;120(5):1627-1635.
- Levitt P, Eagleson KL, Powell EM. Regulation of neocortical interneuron development and the implications for neurodevelopmental disorders. *Trends Neurosci.* 2004;27(7):400-6.
- Liebmann L, Karst H, Sidiropoulou K, van Gemert N, Meijer OC, Poirazi P, & Joëls M. Differential effects of corticosterone on the slow afterhyperpolarization in the basolateral amygdala and CA1 region: possible role of calcium channel subunits. *J Neurophysiol.* 2008;99(2):958-68.
- Ma J, Fill M, Knudson CM, Campbell KP, Coronado R. Ryanodine receptor of skeletal muscle is a gap junction-type channel. *Science.* 1988;242(4875):99-102.
- Mayr JA, Haack TB, Graf E, *et al.* Lack of the mitochondrial protein acylglycerol kinase causes Sengers syndrome. *Am J Hum Genet.* 2012;90(2):314-20.
- McMurtrie HL, Cleary HJ, Alvarez BV, Loiselle FB, Sterling D, Morgan PE, Johnson DE, & Casey JR. The bicarbonate transport metabolon. *J Enzyme Inhib Med Chem.* 2004;19(3):231-6.
- Medina JF, Recalde S, Prieto J, Lecanda J, Saez E, Funk CD, Vecino P, van Roon MA, Ottenhoff R, Bosma PJ, *et al.* Anion exchanger 2 is essential for spermiogenesis in mice. *Proc Natl Acad Sci U S A* 2003;100(26):15847-15852.
- Monies D, Abouelhoda M, AlSayed M, Alhassnan Z, Alotaibi M, Kayyali H, Al-Owain M, Shah A, Rahbeeni Z, Al-Muhaizea MA, *et al.* The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes. *Hum Genet.* 2017. 136(8): 921-939.
- Monies D, Abouelhoda M, Assoum M, Moghrabi N, Rafiullah R, Almontashiri N, Alowain M, Alzaidan H, Alsayed M, Subhani S, *et al.* Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a Highly Consanguineous Population. *Am J Hum Genet.* 2019;104(6):1182-1201.
- Mozrzymas JW, Żarmowska ED, Pytel M, Mercik K. Modulation of GABA_A Receptors by Hydrogen Ions Reveals Synaptic GABA Transient and a Crucial Role of the Desensitization Process. *J Neurosci.* 2003;23(22):7981-7992.
- Mumtaz R, Trepiccione F, Hennings JC, Huebner AK, Serbin B, Picard N, Ullah A, Păunescu TG, Capen DE, Lashhab RM, *et al.* Intercalated Cell Depletion and Vacuolar H⁽⁺⁾-ATPase Mistargeting in an Ae1 R607H Knockin Model. *Journal of the American Society of Nephrology : JASN* 2017;28(5):1507-1520.
- Munji RN, Choe Y, Li G, Siegenthaler JA, Pleasure SJ. Wnt signaling regulates neuronal differentiation of cortical intermediate progenitors. *J Neurosci.* 2011;31(5):1676-87.
- Novarino G, Fenstermaker AG, Zaki MS, Hofree M, Silhavy JL, Heiberg AD, Abdellateef M, Rosti B, Scott E, Mansour L, *et al.* Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. *Science.* 2014;343(6170):506-511.
- Olmos-Serrano JL, Corbin JG. Amygdala regulation of fear and emotionality in fragile X syndrome. *Dev Neurosci.* 2011;33(5):365-78.
- Orlowski J, Grinstein S. Na⁺/H⁺ exchangers. *Compr Physiol.* 2011;1(4):2083-100.

- Parker MD, Musa-Aziz R, Rojas JD, Choi I, Daly CM, Boron WF. Characterization of human SLC4A10 as an electroneutral Na/HCO₃ cotransporter (NBCn2) with Cl⁻ self-exchange activity. *J Biol Chem*. 2008;283(19):12777-88.
- Pasternack M, Smirnov S, Kaila K. Proton modulation of functionally distinct GABA_A receptors in acutely isolated pyramidal neurons of rat hippocampus. *Neuropharmacology*. 1996;35(9-10):1279-88.
- Peters LL, Shivdasani RA, Liu SC, Hanspal M, John KM, Gonzalez JM, Brugnara C, Gwynn B, Mohandas N, Alper SL, *et al*. Anion exchanger 1 (band 3) is required to prevent erythrocyte membrane surface loss but not to form the membrane skeleton. *Cell* 1996;86(6):917-927.
- Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, Walkiewicz M, Bi W, Xiao R, Ding Y, *et al*. Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N Engl J Med* 2017 376(1), 21-31.
- Praetorius J, Nejsum LN, Nielsen S. A SCL4A10 gene product maps selectively to the basolateral plasma membrane of choroid plexus epithelial cells. *Am J Physiol Cell Physiol*. 2004;286(3):C601-10.
- Prasad V, Bodi I, Meyer JW, Wang Y, Ashraf M, Engle SJ, Doetschman T, Sisco K, Nieman ML, Miller ML, *et al*. Impaired cardiac contractility in mice lacking both the AE3 Cl⁻/HCO₃⁻ exchanger and the NKCC1 Na⁺-K⁺-2Cl⁻ cotransporter: effects on Ca²⁺ handling and protein phosphatases. *J Biol Chem* 2008;283(46):31303-31314.
- Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, *et al*. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511(7510):421-427.
- Rizo J, Xu J. The Synaptic Vesicle Release Machinery. *Annu Rev Biophysics*. 2015;44(1):339-367.
- Romero MF, Chen AP, Parker MD, Boron WF. The SLC4 family of bicarbonate (HCO₃⁻) transporters. *Mol Aspects Med*. 2013;34(2-3):159-82.
- Ruffin VA, Salameh AI, Boron WF, Parker MD. Intracellular pH regulation by acid-base transporters in mammalian neurons. *Front Physiol*. 2014;5:43.
- Rutecki PA, Lebeda FJ, Johnston D. Epileptiform activity induced by changes in extracellular potassium in hippocampus. *Journal of neurophysiology*. 1985;54(5):1363-74
- Salameh AI, Hübner CA, Boron WF. Role of Cl⁻-HCO₃⁻ exchanger AE3 in intracellular pH homeostasis in cultured murine hippocampal neurons, and in crosstalk to adjacent astrocytes. *J Physiol*. 2017;595(1):93-124.
- Sander T, Toliat MR, Heils A, Leschik G, Becker C, Rüschenhoff F, Rohde K, Mundlos S, & Nürnberg P. Association of the 867Asp variant of the human anion exchanger 3 gene with common subtypes of idiopathic generalized epilepsy. *Epilepsy Res* 2002;51(3):249-255.
- Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An JY, Peng M, Collins R, Grove J, Klei L, *et al*. Large-Scale Exome Sequencing Study Implicates Both

- Developmental and Functional Changes in the Neurobiology of Autism. *Cell* 2020;180(3):568-584.e523.
- Schmidt-Wilcke T, Fuchs E, Funke K, Vlachos A, Müller-Dahlhaus F, Puts NAJ, Harris RE, & Edden RAE. GABA-from Inhibition to Cognition: Emerging Concepts. *Neuroscientist*. 2018;24(5):501-515.
- Schneggenburger R, Neher E. Intracellular calcium dependence of transmitter release rates at a fast central synapse. *Nature*. 2000;406(6798):889-93.
- Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, *et al*. Strong association of de novo copy number mutations with autism. *Science* 2007;316(5823):445-9.
- Sinning A, Hübner CA. Minireview: pH and synaptic transmission. *FEBS Lett* 2013;587(13):1923-8.
- Sinning A, Liebmann L, Kougioumtzes A, Westermann M, Bruehl C, Hubner CA. Synaptic glutamate release is modulated by the Na⁺-driven Cl⁻/HCO₃⁻ exchanger Slc4a8. *J Neurosci* 2011;31(20):7300-11.
- Sinning A, Liebmann L, Hübner CA. Disruption of Slc4a10 augments neuronal excitability and modulates synaptic short-term plasticity. *Front Cell Neurosci*. 2015;9:223.
- Smith ED, Blanco K, Sajan SA, Hunter JM, Shinde DN, Wayburn B, Rossi M, Huang J, Stevens CA, Muss C, *et al*. A retrospective review of multiple findings in diagnostic exome sequencing: half are distinct and half are overlapping diagnoses. *Genet Med* 2019;21(10):2199-2207.
- Stawarski M, Hernandez RX, Fegghi T, *et al*. Neuronal Glutamatergic Synaptic Clefs Alkalinize Rather Than Acidify during Neurotransmission. *J Neurosci* 2020;40(8):1611-1624.
- Stenkamp K, Palva JM, Uusisaari M, Schuchmann S, Schmitz D, Heinemann U, & Kaila K. Enhanced temporal stability of cholinergic hippocampal gamma oscillations following respiratory alkalosis in vitro. *J Neurophysiol* 2001;85(5):2063-9.
- Thorsen K, Dam VS, Kjaer-Sorensen K, Pedersen LN, Skeberdis VA, Jurevičius J, Treinys R, Petersen I, Nielsen MS, Oxvig C, *et al*. Loss-of-activity-mutation in the cardiac chloride-bicarbonate exchanger AE3 causes short QT syndrome. *Nat Commun* 2017;8(1):1696.
- Tombaugh GC, Somjen GG. Effects of extracellular pH on voltage-gated Na⁺, K⁺ and Ca²⁺ currents in isolated rat CA1 neurons. *J Physiol* 1996;493(3):719-32.
- Tsukioka M, Iino M, Endo M. pH dependence of inositol 1,4,5-trisphosphate-induced Ca²⁺ release in permeabilized smooth muscle cells of the guinea-pig. *J Physiol* 1994;475(3):369-75.
- Traynelis SF, Cull-Candy SG. Proton inhibition of N-methyl-D-aspartate receptors in cerebellar neurons. *Nature* 1990;345(6273):347-50.
- van der Meer D, Rokicki J, Kaufmann T, Córdova-Palomera A, Moberget T, Alnæs D, Bettella F, Frei O, Doan NT, Sønderby IE, *et al*. Brain scans from 21,297 individuals reveal the

genetic architecture of hippocampal subfield volumes. *Mol psychiatry* 2020;25(11):3053-3065.

Wagner M, Berutti R, Lorenz-Depiereux B, Graf E, Eckstein G, Mayr JA, Meitinger T, Ahting U, Prokisch H, Strom TM, *et al.* Mitochondrial DNA mutation analysis from exome sequencing-A more holistic approach in diagnostics of suspected mitochondrial disease. *J Inherit Metab Dis* 2019;42(5):909-917.

Waldmann R, Champigny G, Bassilana F, Heurteaux C, Lazdunski M. A proton-gated cation channel involved in acid-sensing. *Nature* 1997;386(6621):173-177.

Wang J, Wang W, Wang H, Tuo B. Physiological and Pathological Functions of SLC26A6. *Front Med (Lausanne)*. 2020;7:618256.

5

No association between *SCN9A* and monogenic human epilepsy disorders

James Fasham, Joseph S Leslie, Jamie W Harrison, James Deline,
Katie B Williams, Ashley Kuhl, Jessica Scott Schwoerer, Harold E
Cross, Andrew H Crosby*, Emma L Baple*

PloS Genet. 2020;16(11):e1009161.

5.1. Acknowledgements of co-authors and contributions

This study was undertaken as part of the “Windows of Hope” Amish translational genomic research programme, conceived, designed and led by Prof Andrew Crosby, Prof Emma Baple and Prof Harold Cross (University of Arizona).

Where specific experiments or analyses were performed by study collaborators or members of my supervisors’ group other than myself, then these are detailed below.

Clinical and genomic studies

Clinical data were obtained by local clinical care providers using a standardised proforma that I designed.

Technical assistance with the *SCN9A* cosegregation studies was provided by Mr Joseph Leslie (University of Exeter).

Dr Jamie Harrison (University of Exeter) obtained allele frequency data for the *SCN9A* variants in UK Biobank performed the rare variant burden testing with my assistance.

Manuscript draft and revision

I wrote and revised the manuscript with Prof Baple and Prof Crosby. All the co-authors provided comments and feedback prior to submission.

5.2. Abstract

Many studies have demonstrated the clinical utility and importance of epilepsy gene panel testing to confirm the specific aetiology of disease, enable appropriate therapeutic interventions, and inform accurate family counselling. Previously, *SCN9A* gene variants, in particular a c.1921A>T p.(Asn641Tyr) substitution, have been identified as a likely autosomal dominant cause of febrile seizures/“febrile seizures plus” and other monogenic seizure phenotypes indistinguishable from those associated with variants in *SCN1A*, leading to inclusion of *SCN9A* on epilepsy gene testing panels. Here we present serendipitous findings of genetic studies that identify the *SCN9A* c.1921A>T p.(Asn641Tyr) variant at high frequency in the Amish community in the absence of such seizure phenotypes. Together with findings in UK Biobank these data refute an association of *SCN9A* with epilepsy, which has important clinical diagnostic implications.

5.3. Introduction

The clinical utility and importance of extended gene panel testing for clinically and genetically heterogeneous disorders such as epilepsy is undisputed. Knowledge of the precise genetic aetiology of a patient's epilepsy can significantly alter therapeutic management and understanding the inheritance pattern of the genetic subtype informs genetic counselling for both the patient and the wider family (Kearney *et al.* 2019). However, the inclusion of inappropriate genes or omission of relevant genes can potentially lead to false-positive or missed diagnoses, respectively. The lack of consensus and international guidance for curation and inclusion of a gene in panels means that existing diagnostic panels often contain genes of research interest or those with historical (sometimes incomplete or inaccurate) evidence (Strande *et al.* 2017). Although Genomics England PanelApp (Martin *et al.* 2019), ClinGen (Strande *et al.* 2017) and other similar initiatives are trying to address this issue, a review of current seizure panels offered by clinical laboratory studies revealed that this remains a significant concern. This is particularly important in epilepsy disorders where certain medications are either more effective in controlling seizures or contraindicated in patients with particular monogenic causes of disease (Balestrini & Sisodiya 2018). Notable examples of this include the *SCN1A*-associated seizure disorders, where commonly used sodium-channel-blocking medications carbamazepine, vigabatrin and lamotrigine should be avoided because they may worsen the condition by inducing and/or prolonging seizures (Wilmshurst *et al.* 2015). Monoallelic variation of the alpha subunit of the sodium channel (*SCN1A*) gene is a well-established cause of a spectrum of seizure disorders that include simple febrile seizures, "febrile seizures plus"

(FS+) and genetic epilepsy with febrile seizures plus (GEFS+). *SCN1A* variants also account for 70-80% of Dravet syndrome (DS), a debilitating autosomal dominant infantile-onset epileptic encephalopathy (Wheless, Fulton & Mudigoudar 2020).

In 2009, Singh *et al.* described a large Utah family comprising 21 individuals affected by febrile/afebrile seizures clinically indistinguishable from the seizure phenotypes associated with variants in *SCN1A* (Singh *et al.* 2009). Their genetic studies identified a missense variant in the sodium channel protein type 9 subunit alpha (voltage-gated sodium channel $Na_v1.7$; *SCN9A* NM_002977:c.1921A>T; p.(Asn641Tyr)), thereafter referred to as *SCN9A* p.(Asn641Tyr), which cosegregated with the disease in all but one individual who did not have a history of seizures. Eleven variant carriers displayed a typical febrile convulsion history, with no seizures reported after age six years. In the remaining ten, afebrile seizures followed typical febrile convulsions, the majority of which resolved by age 16 years with only two progressing to intractable epilepsy. Inherited autosomal dominant forms of familial febrile seizures typically show reduced penetrance, with between 10 and 30% of individuals inheriting the familial gene variant remaining seizure free (Bonanni *et al.* 2004; Mantegazza *et al.* 2005), whereas the penetrance of the seizure phenotype in this family was 95%. The authors then investigated a series of 92 unrelated patients with a personal history of febrile seizures with or without a family history of seizures, and identified a further five rare missense variants in *SCN9A*, with current allele frequencies ranging from 0 to 1.8% in the Genome Aggregation database (GnomAD v2.1.1, **Table 5.1**).

GRCh38 reference (rs number)	Genotype (NM_002977)	Phenotype	gnomAD ¹ AC (Hom.) AF	SCN9A variant familial segregation	Reference
SCN9A Variants proposed in Singh et al. 2009					
Chr2:166284506T>A (rs121908918)	c.1921A>T p.(Asn641Tyr)	FS, AFS, TLE	3 (0) 0.001%	Utah family variant	Singh et al. 2009
Chr2:166311573T>C (rs121908920)	c.184A>G p.(Ile62Val)	FS	6 (0) 0.003%	none stated	Singh et al. 2009
Chr2:166306531G>T (rs121908921)	c.446C>A p.(Pro149Gln)	FS	0	none stated	Singh et al. 2009
Chr2:166286469C>T (rs58022607)	c.1469G>A p.(Ser490Asn)	complex FS	4023 (198) 1.8%	none stated	Singh et al. 2009
Chr2:166281786T>C (rs121908919)	c.1964A>G p.(Lys655Arg)	FS, GSW, IGE GEFS+	428 (0) 0.2%	none stated	Singh et al. 2009
Chr2:166280452T>C (rs182650126)	c.2215A>G p.(Ile739Val)	FS, IGE	472 (1) 0.2%	Inherited from an unaffected parent. none stated	Alves et al. 2019 Singh et al. 2009
SCN9A variants proposed in subsequent publications					
Chr2:166311728T>C (rs267607030)	c.29A>G p.(Gln10Arg)	GEFS+	25 (1) 0.01%	Inherited from an affected parent and present in an affected sibling.	Cen et al. 2017
Chr2:166307014A>G	c.319T>C p.(Tyr107His)	FS	0	Inherited from an affected parent.	Banfi et al. 2020
Chr2:166303195G>T (rs201743233)	c.796C>A p.(Leu266Met)	GEFS+	2 (0) <0.001%	Inherited from an unaffected parent.	Mulley et al. 2013
Chr2:166293358C>T (rs765818027)	c.980G>A p.(Gly327Glu)	BECTS GEFS+	11 (0) 0.005%	Inherited from an unaffected parent, identified in affected sibling. Inherited from an affected parent.	Liu et al. 2019 Yang et al. 2018
Chr2:166198900del (rs1353037253)	c.5702_5706del p.(I1901fs)	GEFS+	10 (0) 0.005%	Inherited from an affected parent	Yang et al. 2018
Chr2:166198733T>C (rs761742207)	c.5873A>G p.(Tyr1958Cys)	GEFS+	2 (0) <0.001%	Inherited from an affected parent and identified in a further affected individual and one individual of unknown affection.	Zhang et al. 2020

Table 5.1: SCN9A variants proposed as a monogenic causes of epilepsy.

While some of these variants are rare or novel, a number [p.(Gln10Arg), p.(Ser490Asn), p.(Lys655Arg), p.(Ile739Val)] are present at notable allele frequency in heterozygous as well as homozygous state, inconsistent with them being causative of a monogenic seizure disorder. Additionally, the other reported variants were mostly defined in a single affected individual or small nuclear families in which limited, or no wider cosegregation studies could be performed. ¹ gnomAD v2.1.1 non-neuro cohort. Abbreviations: AC, Allele count; AF Allele frequency; AFS, Afebrile seizures; BECTS, benign partial epilepsy of childhood with centrotemporal spikes; FS, Febrile Seizures; GEFS+, generalised epilepsy with febrile seizures plus; GSW, generalised spike wave; Hom. Homozygous individuals; IGE, idiopathic generalised epilepsy; TLE, temporal lobe epilepsy

The studies undertaken in the Utah family included genome-wide linkage analysis, which identified a ~10cM region of Chr2q24 that cosegregated with the disease followed by targeted analysis of candidate genes. The Chr2q24 linked region encompassed ~65 genes including the known epilepsy gene cluster (*SCN1A*, *SCN2A* and *SCN3A*). Dideoxy sequencing of *SCN1A*, *SCN2A*, *SCN3A*, *SCN7A*, *KCNH7* and *SLC4A10* was performed, alongside Agilent Human Genome Comparative Genomic Hybridization (CGH) Microarray 4x44K (Agilent, Santa Clara, CA) analysis of the distal 10 Mb of the Chr2q24 region (*SCN1A*, *SCN2A*, *SCN3A*, *SCN7A* and *SCN9A*) and multiplex amplicon quantification (MAQ) CNV analysis of *SCN1A*, none of which revealed any candidate causative variants. However, as certain difficult to identify genomic variants (e.g. deep intronic, structural variants or CNVs) may evade detection using these techniques (Møller *et al.* 2008), it remains possible that an undiscovered pathogenic variant in one of the known epilepsy associated genes within the Utah family locus, in particular *SCN1A* with which there is a close phenotypical fit, may be responsible for the condition.

Intending to confirm the role of *SCN9A* in seizure susceptibility, the authors next generated an *Scn9a*-Asn641Tyr knock-in mouse model. This revealed a significant increase in susceptibility to electrically-induced seizures in homozygous, but not heterozygous, animals (Oakley *et al.* 2009). Following this study, other rare/novel *SCN9A* variants (**Table 5.1**) were reported as putative monogenic causes of disease in individuals and small families with familial febrile seizures (Banfi *et al.* 2020; Yang *et al.* 2018), FS+ and GEFS+ (Alves *et al.* 2019; Cen *et al.* 2017; Liu *et al.* 2019; Mulley *et al.* 2013; Zhang *et al.* 2020)

and as a modifier of Dravet syndrome, in some cases in the presence of accompanying *SCN1A* pathogenic variants (Mulley *et al.* 2013; Singh *et al.* 2009) (**Table 5.1**). These studies, stemming from those of Singh and colleagues, have widely led to the inclusion of *SCN9A* on epilepsy gene testing panels.

Here we describe the serendipitous identification of the *SCN9A* p.(Asn641Tyr) variant within the Wisconsin Amish community, in which it is present at notable frequency in individuals with no personal or family history of febrile seizures. This Amish founder variant was identified as part of an ongoing study to characterise the genetic causes of inherited neurodevelopmental disorders present amongst the Wisconsin Amish and Mennonite communities (University of Arizona IRB (10-0050-01)), in which we investigated an Amish male infant presenting with dysmorphic facial features and GDD (**Fig. 5.1**. Kinship 1, X:9). Trio whole-exome sequencing identified a likely pathogenic *de novo* missense variant in *CHD4*, compatible with the child's phenotype and the likely cause of the child's syndromic presentation. In addition to the *CHD4* variant, our genomic studies also identified the *SCN9A* p.(Asn641Tyr) (Chr2(GRCh38):g.166284506T>A, NM_002977:c.1921A>T) variant inherited from the healthy mother, who reported no history of febrile or afebrile seizures. Family extension studies were undertaken to investigate the relevance of the variant, in addition to cross-referencing these findings with our in-house Amish exome database alongside the comprehensive genealogical records of the Amish. These studies identified the *SCN9A* p.(Asn641Tyr) variant in a total of seven nuclear families (**Fig. 5.1**) including three nuclear families (C,D and E) that closely interlink.

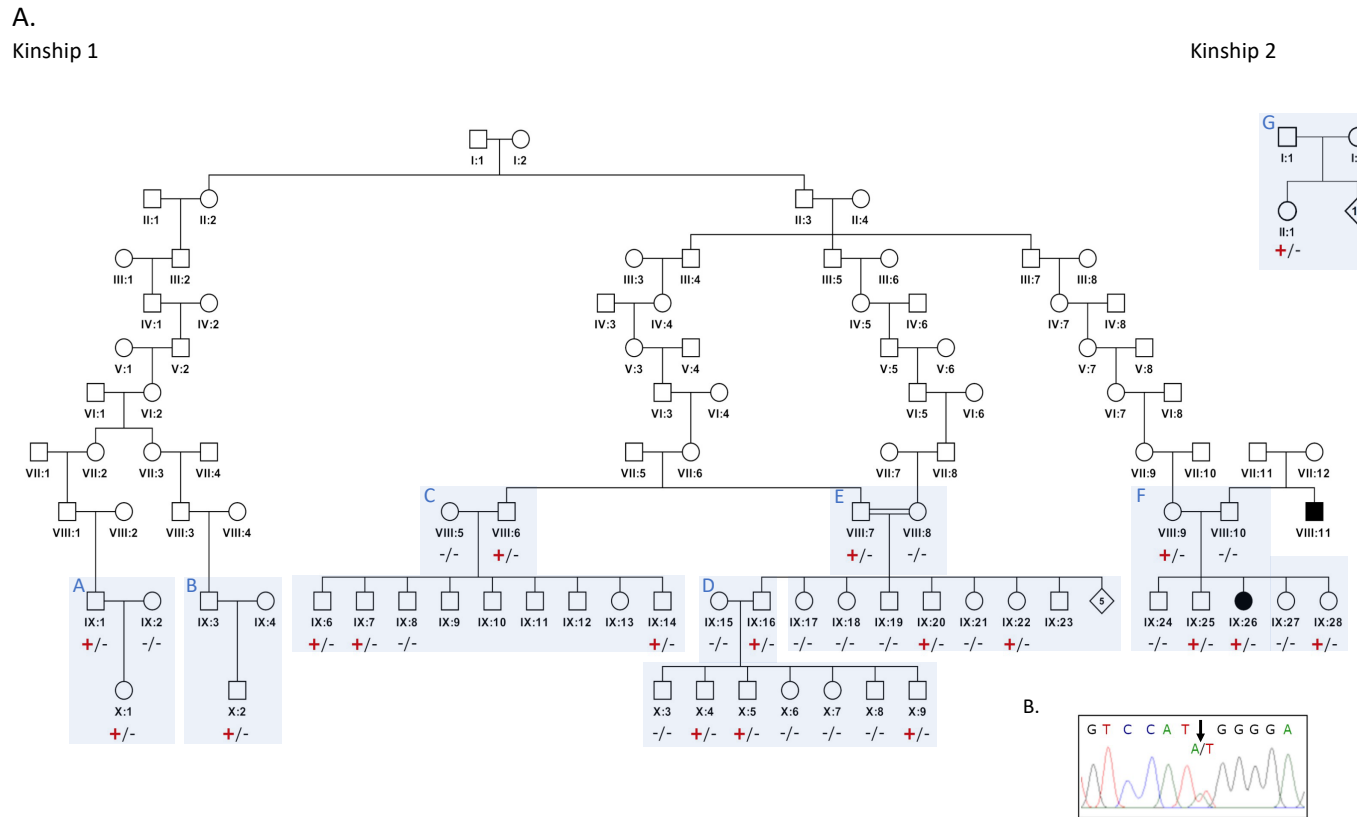


Figure 5.1: Family pedigrees showing *SCN9A* NM_002977:c.1921A>T p.(Asn641Tyr) genotype data

(A). Kinship 1: A simplified pedigree of the extended Amish family investigated, showing relationships between 18 distantly related individuals found to be heterozygous for the *SCN9A* p.(Asn641Tyr) variant. Individuals shaded black are affected with afebrile seizures. **Kinship 2:** An additional Amish family comprising of one individual with no history of seizures for whom exome data were available, found to be heterozygous for the *SCN9A* p.(Asn641Tyr) variant (II:1). Kinships 1 and 2 likely share common ancestry, but could not be connected through available records. Genotype is shown under individuals (variant: +, wild type: -). **(B).** Electropherogram showing the DNA sequence at the position of *SCN9A* c.1921A>T in a heterozygous individual.

In these families, further genetic studies confirmed the presence of the variant in 11 unaffected family members. Three other families (A,B and F) comprised of seven additional confirmed *SCN9A* p.(Asn641Tyr) variant carriers, all of whom interlink with the first three families through a 9th generation ancestral couple. A final nuclear family (G) could not be linked with available ancestral data. From these studies it is evident that the variant was transmitted through an additional minimum of 22 constitutive gene carrier parental couples and consequently will inevitably have been transmitted to hundreds of their offspring in whom genetic studies are not possible. The proband originally investigated in our study remains seizure free at two years and five months of age. Additionally, careful inspection of the available medical records across the wider extended Amish family and careful questioning of each individual carrier of the *SCN9A* p.(Asn641Tyr) variant and/or their parents, identified only one individual who carried the variant with a history of seizures. Importantly, this seven-year-old child (Kinship 1 IX:26) had a two-year history of left sided focal seizures not associated with loss of consciousness, normal development and no history of febrile convulsions. Metabolic testing and magnetic resonance imaging of the brain and spinal cord were unremarkable. While in a proportion of patients with focal seizures the disorder is associated with focal cortical dysplasia and/or monoallelic variants in the genes encoding the mTOR inhibitory GATOR1 complex (Iffland *et al.* 2019), neither *SCN9A* nor *SCN1A* have been associated with this seizure type.

5.4. Discussion

The high frequency of the *SCN9A* p.(Asn641Tyr) variant in the Amish involving hundreds of variant carriers with no history of seizure phenotypes is clearly inconsistent with it representing a highly penetrant cause of these conditions (Singh *et al.* 2009). The benign nature of this variant is also consistent with our investigations in UK Biobank in which the variant was identified in two heterozygous carriers, neither of whom display a history of seizures. Further comparison of Amish *SCN9A* c.1921A>T p.(Asn641Tyr) carriers with those in UK Biobank shows that all eight Amish individuals for whom exome data were available (see methods in **Supplemental Text** and **Fig. 5.1** legend) and both UK Biobank *SCN9A* c.1921A>T p.(Asn641Tyr) carriers, share a rare synonymous variant in a closely linked gene; *TTC21B* (Chr2(GRCh38): g.165883959A>G NM_024753.4:c.3519T>C p.(Thr1173=), rs115504901) situated ~400kb from *SCN9A*. The *TTC21B* variant has a European allele frequency of 0.4% (in gnomAD (v2.1.1) and a higher allele frequency of 5% in the Amish (in-house data). The complete co-occurrence of these two rare variants in the Anabaptist and UK Biobank datasets in all individuals with *SCN9A* p.(Asn641Tyr) indicates that they likely occur *in cis*, and potentially all derive from an individual mutagenic event occurring in a single ancestral European founder in whom the *SCN9A* variant arose on the *TTC21B* haplotype.

The gnomAD database v.2.1.1 (non-neuro) currently identifies three non-Finnish European (NFE) heterozygous carriers of the *SCN9A* c.1921A>T p.(Asn641Tyr) variant at an overall allele frequency of 1.4×10^{-5} with a further five NFE carriers in gnomAD v3.0. GnomAD is an aggregated database of exome

and genome data from unrelated individuals sequenced as part of various disease-specific and population genetic studies, it serves as a useful proxy population control dataset for severe early onset paediatric diseases and is utilised as an aide to genomic variant interpretation by research and diagnostic laboratories worldwide (Karczewski *et al.* 2020; Lek *et al.* 2016). While it is not possible to draw meaningful conclusions about the pathogenicity of the *SCN9A* p.(Asn641Tyr) variant from gnomAD, these data confirms that it represents a low frequency variant present throughout the European population.

In the Amish, and many other community settings worldwide, particular genetic variants may become enriched and increase in allele frequency due to ancestral genetic bottleneck events, geographical isolation, community marriage patterns and large family sizes. This includes both pathogenic and benign variants for which increased allele frequency allows improved annotation and interpretation of pathogenicity (Abouelhoda *et al.* 2016; Jung *et al.* 2020). The data presented here are a good example of this, repudiating the proposed autosomal dominant association of the p.(Asn641Tyr) *SCN9A* gene variant with seizure disorders.

The *SCN9A* variants identified as potentially pathogenic subsequent to p.(Asn641Tyr), include a number ([p.(Gln10Arg), p.(Ser490Asn), p.(Lys655Arg), p.(Ile739Val)]) which are present at population allele frequencies inconsistent with them being causative of a monogenic seizure disorder (**Table 5.1** and **Table 5.S1**). Additionally, where this information is reported, these variants were all defined in either a single affected individual or small nuclear families, in which limited or no wider cosegregation studies could be performed and in which genomic studies were mostly relatively limited (**Table 5.1** and **Table**

5.S2). Further evidence against an association between *SCN9A* and monogenic epilepsy is provided by a recent study of 31,058 parent-offspring trios, in which ~25% of probands had epilepsy/history of seizures. This study found no significant enrichment of *de novo* variants in *SCN9A*, with no history of epilepsy reported in the two *de novo* *SCN9A* variant carriers where this information was available (Kaplanis *et al.* 2020). Further to this, independent GWAS studies that show multiple significant associations between SNPs associated with *SCN1A/SCN2A* and epilepsy and/or febrile seizures, fail to do so for *SCN9A* (Buniello *et al.* 2019). Additionally, our own *SCN9A* rare (allele frequency <1%) variant burden analysis, of individuals with epilepsy of Northern European ancestry versus ethnically matched controls in UK Biobank exome data, defined no enrichment of plausibly causative rare *SCN9A* variants (**Table 5.S3**; $p = 0.398$), nor a disease association with any single variant (after correction for multiple testing).

SCN9A is unlike the epilepsy-related voltage-gated sodium channel alpha (VGSC- α) subunit molecules *SCN1A*, *SCN2A*, *SCN3A* and *SCN8A* (Berkovic *et al.* 2004; Mulley *et al.* 2005; Vanoye *et al.* 2014) each of which is expressed primarily in brain, whereas *SCN9A* is expressed primarily in peripheral nerves (**Fig. 5.S1**). These primarily brain-expressed genes are also constrained for missense alterations and disease associated missense variants are primarily clustered over the functionally important ion-transporter domains; neither of these scenarios is applicable to *SCN9A* (**Fig. S2**) Thus, a role for *SCN9A* in sensory perception and pain is more congruous with our findings. Indeed, multiple studies document *SCN9A* gene variants associated with neuropathic pain syndromes including primary erythromelalgia and small fibre neuropathy

(MIM: 133020) and congenital insensitivity to pain (MIM: 243000) (Cox *et al.* 2006; Faber *et al.* 2012; Michiels *et al.* 2005; Yang *et al.* 2004).

The publications identifying an association between *SCN9A* and epilepsy have led to its widespread incorporation into monogenic inherited seizure disorder diagnostic testing panels (ClinGen 2018), including Athena Diagnostics, USA (2020); Blueprint Genetics, Finland (2020); Centogene, Germany (2020); Dynacare, Canada (2020); EGL Genetics, USA (2020); Invitae, USA (2020); Mayo Clinic Labs, USA (2019). The presence of *SCN9A* on these panels and its currently widely accepted status as an epilepsy disease gene, clearly presents a substantial risk of misdiagnosis to patients. This is of particular concern for genetic epilepsies in which a precise molecular diagnosis informs drug choice and a genetic misdiagnosis may have devastating and sometimes lethal consequences (Banfi *et al.* 2020; Helbig & Ellis 2020; Williams v. Quest Diagnostics 2018). Thus given our findings, we consider ClinGen and other expert groups reappraisal of the evidence regarding the role of *SCN9A* in monogenic seizure phenotypes to be of extreme importance and urgency, so as to refute this association and mitigate future harms.

5.5. Supplemental material

Supplemental methods

Genomic studies

Blood/buccal samples were obtained with informed consent (University of Arizona IRB - 1000000050). DNA was extracted using standard techniques. Whole-exome sequencing (WES), was performed on nine individuals (VIII:9, VIII:10, IX:6, IX:8, IX:15, IX:16, IX:26, X1, X:9) using Agilent SureSelect Whole Exome v6 / Twist Human Core Exome targeting. Individual X:2 underwent clinical exome sequencing, using Illumina TruSight targeting.

In all cases read alignment (BWA-MEM (v0.7.17) was performed, mate-pairs fixed and duplicates removed (Picard v2.15.0), InDel realignment/base quality recalibration (GATK v3.7.0), SNV/InDel detection (GATK HaplotypeCaller), annotation (Alamut Batch v1.8 or v1.10), and read depth (GATK DepthOfCoverage). CNVs were detected using both ExomeDepth (Plagnol *et al.* 2012) and SavvyCNV (Laver *et al.* 2022). This conforms to GATK Best Practices.

Dideoxy sequencing confirmation of the *SCN9A* NM_002977 c.1921A>T p.(Asn641Tyr) variant was undertaken using standard techniques. The *SCN9A* NM_002977 c.1921A>T p.(Asn641Tyr) variant was submitted to ClinVar (www.ncbi.nlm.nih.gov/clinvar, accession SCV001371862).

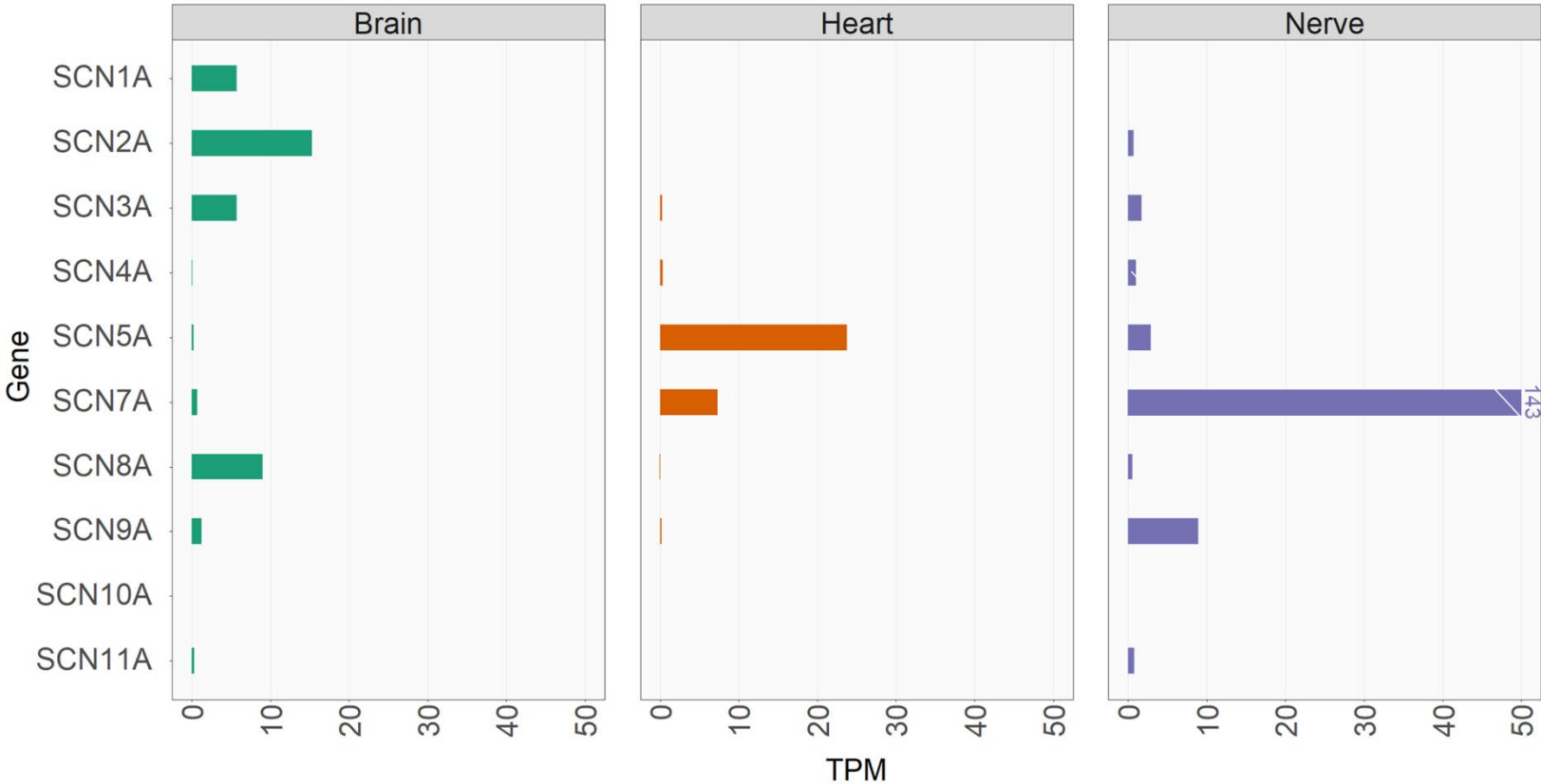
UK Biobank rare variant burden analysis

The UK Biobank data for 49,953 participants with available exome sequence

data called by the Regeneron Seal Point Balinese (SPB) pipeline, were used (Van Hout *et al.* 2019). This analysis was limited to those participants of Northern European ancestry (41,249). The epilepsy phenotype was generated in Stata V16 (StataCorp 2019) including as a case anyone with an International Classification of Diseases, Ninth Revision or International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-9/10) primary/secondary code containing any evidence of epilepsy (ICD-9: 34509, 34510, 34519, 3452, 3453, 3454, 34550, 34559, 3457, 3459; ICD-10: G40.0-9, G41.0, G41.1, G41.2, G41.8, G41.9) or any self-reported history of epilepsy.

The UK Biobank exome data were annotated with SnpEff 4.3t (Cingolani *et al.* 2012). All exonic and canonical splice site variants predicted to alter the SCN9A amino acid sequence, annotated in transcript NM 002977.3, with a frequency of <1% in both UK Biobank and GnomAD (Karczewski *et al.* 2020) were extracted and manually curated. Total raw variant numbers were compared in cases and control using a two-sided Fisher's exact test, as previously described (Cohen *et al.* 2004). In addition, variant frequencies were calculated, and exome wide association tests performed using Plink V1.9 (Purcell *et al.* 2007).

Figure 5.S1: Sodium voltage-gated channel alpha subunit gene family expression data



Expression data reproduced from the Genotype-Tissue Expression (GTEx) portal (www.gtexportal.org) for nine of the ten VGSC- α genes. Unlike the epilepsy-related VGSC- α genes (*SCN1A*, *SCN2A*, *SCN3A* and *SCN8A*), *SCN9A* is expressed primarily in peripheral nerves. Abbreviations: TPM = transcripts per million

Figure 5.S2: Variant clustering in *SCN9A* alongside other VGSC- α genes (*SCN1A* and *SCN3A*) associated with epilepsy

VGSC- α Domains		Ion Transporter 1	Ion Transporter 2	Ion Transporter 3	Ion Transporter 4					
SCN1A	Positions	1-126	127-434	435-766	767-1002	1003-1216	1217-1493	1494-1513	1514-1797	1798-2009
	Epilepsy variants (Per a.a)	21 (0.17)	88 (0.29)	9 (0.03)	59 (0.25)	5 (0.02)	64 (0.23)	8 (0.38)	70 (0.25)	20 (0.09)
SCN3A	Positions	1-126	127-435	436-758	759-994	995-1204	1205-1478	1479-1524	1525-1782	1783-2000
	Epilepsy variants	0	1	0	2	0	2	0	2	0
SCN9A	Positions	1-124	125-412	413-742	743-978	979-1190	1191-1468	1469-1513	1514-1771	1772-1987
	Pain variants	1	3	1	6	2	4	4	2	0
	Epilepsy variants	2	1	3	0	2	1	0	0	0

Top of genogram (VGSC- α domains) shows a representative domain structure of VGSC- α genes, with four ion transporter Pfam domains (blue boxes) separated by interspersed disordered regions (beige boxes).

Beneath, *SCN9A* is shown alongside the two VGSC- α genes that are both associated with epilepsy and have Pfam annotated ion channel domains (*SCN1A* and *SCN3A*), with the amino acid position of each domain indicated in each molecule ('positions' row). The number of ClinVar pathogenic annotated variants (with at least one pathogenic / likely pathogenic annotation) in each is shown below each domain. The four domains with the highest variant density are indicated by a darker shade.

For *SCN1A* (which has a large number of disease-associated variants described) this is also calculated per amino acid to aid interpretation of the mutation load relative to the size of each domain. Variants in *SCN9A* are separated into those associated with pain syndromes, and putative variants proposed to be associated with seizures; those associated with both are recorded twice, and those with no phenotypic description have been excluded.

This shows that the epilepsy-related VGSC- α genes (*SCN1A*, *SCN3A*) display significant spatial clustering of putative disease-associated variants within regions known to be crucial for molecular function, in particular the ion transporter domain regions (shown by dark blue shading) and the interlinking regions between transporter domains 3 and 4 (shown by dark beige shading). For *SCN9A*, the expected clustering of variants in functionally important regions is only seen for pain-associated (erythromelalgia and paroxysmal pain) phenotypes with variants proposed to be associated with seizure disorder phenotypes displaying no spatial clustering, consistent with a benign nature.

Table 5.S1: UK Biobank allele frequencies for the *SCN9A* variants in Table 5.1

Allele frequencies are calculated from Biobank exome data (SPB pipeline), available for 49,959 individuals at the time of publication.

GRCh38 reference	GRCh37 reference	(rs number)	Cases - variant alleles (/1198)	Controls - variant alleles (/98720)	Cases - variant allele frequency	Controls - variant allele frequency
<i>SCN9A variants proposed in Singh et al. 2009</i>						
2:166284506 T>A p.(Asn641Tyr)	2:167141016 T>A	rs121908918	0	2	0	<0.01%
2:166311573 T>C p.(Ile62Val)	2:167168083 T>C	rs121908920	0	0	0	0
2:166306531 G>T p.(Pro149Gln)	2:167163041 G>T	rs121908921	0	1	0	<0.01%
2:166286469 C>T p.(Ser490Asn)	2:167142979 C>T	rs58022607	5	618	0.42%	0.63%
2:166281786 T>Cp.(Lys655Arg)	2:167138296 T>C	rs121908919	1	276	0.08%	0.28%
2:166280452 T>Cp.(Ile739Val)	2:167136962 T>C	rs182650126	7	447	0.58%	0.45%
<i>SCN9A variants proposed in subsequent publications</i>						
2:166311728 T>Cp.(Gln10Arg)	2:167168238 T>C	rs267607030	0	1	0	<0.01%
2:166307014 A>G p.(Tyr107His)	2:167163524 A>G	-	0	0	0	0
2:166303195 G>T p.(Leu266Met)	2:167159705 G>T	rs201743233	0	0	0	0
2:166293358 C>T p.(Gly327Glu)	2:167149868 C>T	rs765818027	0	0	0	0
2:166198900 del p.(I1901fs)	2:167055409 del	rs1353037253	0	0	0	0
2:166198733 T>C p.(Tyr1958Cys)	2:167055243 T>C	rs761742207	0	0	0	0

Variant frequencies are separated by presence (case) or absence of epilepsy (control) as defined in our methods (above). All *SCN9A* variants examined are more frequent in controls than cases except p.(Ile739Val), where there is no significant difference between cases and controls (two-sided Fisher's exact test $p = 0.51$)

Table 5.S2: Heterozygous *SCN9A* variants proposed as a monogenic cause of seizure disorders in subsequent publications, including the testing methodology employed

Genotype (NM_002977)	Phenotype	gnomAD ¹ AC (Hom.) AF	<i>SCN9A</i> variant familial segregation	Genetic testing strategy	Additional variants not excluded	Reference
c.29A>G p.(Gln10Arg)	GEFS+	25 (1) 0.01%	Inherited from an affected parent and present in an affected sibling	Proband-only NGS panel: 480 epilepsy-related genes (including <i>SCN1A</i>)		Cen <i>et al.</i> 2017
c.319T>C p.(Tyr107His)	FS	0	Inherited from an affected parent	Proband-only targeted NGS panel: Cardiac and channelopathy-related genes, karyotype and aCGH	de novo 1.3 Mb duplication, <i>POLG</i> and <i>AKAP9</i> variants	Banfi <i>et al.</i> 2020
c.796C>A p.(Leu266Met)	GEFS+	2 (0) <0.001%	Inherited from an unaffected parent	Dideoxy sequencing: <i>SCN1A/B</i> , <i>GABRG2</i> , <i>PCDH19</i>		Mulley <i>et al.</i> 2013
c.980G>A p.(Gly327Glu)	BECTS	11 (0) 0.005%	Inherited from an unaffected parent and identified in an affected sibling	Trio WES		Liu <i>et al.</i> 2019
	GEFS+		Inherited from an affected parent	Dideoxy sequencing: <i>SCN1A</i> and common epilepsy genes		Yang <i>et al.</i> 2018
c.1964A>G p.(Lys655Arg)	GEFS+	428 (0) 0.2%	Inherited from an unaffected parent	Trio WES and virtual gene panel analysis: 21 epilepsy-related genes (including <i>SCN1A</i>) and aCGH	<i>ANKRD11</i> heterozygous nonsense	Alves <i>et al.</i> 2019
c.5702_5706del p.(11901fs)	GEFS+	10 (0) 0.005%	Inherited from an affected parent	Dideoxy sequencing: <i>SCN1A</i> and common epilepsy genes		Yang <i>et al.</i> 2018
c.5873A>G p.(Tyr1958Cys)	GEFS+	2 (0) <0.001%	Inherited from an affected parent and identified in a further affected individual and one individual of unknown affection.	Trio WES: <i>SCN1A</i> variants examined		Zhang <i>et al.</i> 2020

Abbreviations: AC = Allele count; aCGH = array comparative genomic hybridisation; AF = Allele frequency; BECTS = benign partial epilepsy of childhood with centrottemporal spikes; FS = Febrile Seizures; GEFS+ = generalised epilepsy with febrile seizures plus; Hom.= Homozygous individuals; NGS = Next-generation sequencing; TLE = temporal lobe epilepsy. ¹ gnomAD v2.1.1 non-neuro cohort.

Table 5.S3: Rare variant burden analysis in UK Biobank

	Case Alleles	Control Alleles	Total Alleles
Variant alleles	30	2,835	2,865
Wild type alleles	465,996	37,149,705	37,615,701
Total alleles	466,026	37,152,540	37,618,566

Allele frequencies of *SCN9A* variants predicted to have an impact on *SCN9A* amino acid sequence were compared between cases and controls, with a small but not significant increase observed in controls (two-sided Fisher's exact test $p = 0.398$).

5.6. Further findings and future work

This work was directly cited in the ClinGen review of the association between *SCN9A* and epilepsy, which resulted in a status change from “Limited” to “Refuted” (<https://search.clinicalgenome.org/kb/genes/HGNC:10597>: accessed 09/03/2021). Since diagnostic laboratories typically use ClinGen to inform the content of their panels, this will directly reduce genetic misdiagnoses. Whilst this development is extremely positive there remain further challenges: The *SCN9A*-epilepsy association is currently still listed by Illumina and Ambry Genetics (See GenCC <https://search.thegencc.org/genes/HGNC:10597> accessed 18/10/2022) and Genomics England PanelApp (Martin *et al.*, 2019) (<https://panelapp.genomicsengland.co.uk/panels/90/gene/SCN9A/> accessed 18/10/2022) albeit with “limited” evidence noted in each case. This is despite publication of this work more than a year ago, and national (the UK national neurogenetics meeting 2021) and international presentations (ESHG 2021). Additionally, review articles (Bayat *et al.*, 2021) and case studies published since the beginning of 2021 (Albaradie *et al.*, 2021; Ma *et al.*, 2021) continue to associate *SCN9A* variants with epilepsy, highlighting the need to further raise awareness. Thus, a follow-up publication is planned.

This study has been highly impactful and widely commended (winning prizes at the Exeter Annual Research Event, Clinical Genetics Society (CGS) meeting 2021 and the European Society of Human Genetics (ESHG) meeting 2021) prompting questions about whether other disease-gene associations with limited evidence could be investigated through focused studies originating in genetically isolated communities. There is potential to further utilise this

approach, but only where a majority of evidence for the disease-gene association is associated with one or a small number of rare variants, as was the case in the *SCN9A* variant in the Amish. Additionally, studies would be opportunistic, rather than targeted to a gene or variant. These studies would also be reliant on aggregation of genomic information (exomes and genome) across genetically isolated populations. As part of this project, I have performed aggregation of our population exome and genome datasets (>300 Amish samples, >140 Palestinian samples, >140 Pakistani samples). In the future it will be possible to use this resource to correctly classify genetic variants within these populations annotated as pathogenic by HGMD or ClinVar, which are in fact not associated with genetic disease.

5.7. References

- Abouelhoda M, Faquih T, El-Kalioby M, Alkuraya FS. Revisiting the morbid genome of Mendelian disorders. *Genome biology*. 2016;17(1):235
- Albaradie R, Baig DN & Bashir S. Sodium voltage-gated channel alpha subunit 9 mutation in epilepsy. *Eur Rev Med Pharmacol Sci*. 2021;25(24):7873-7877.
- Alves RM, Uva P, Veiga MF, Oppo M, Zschaber FCR, Porcu G, Porto HP, Persico I, Onano S, Cuccuru G, *et al*. Novel ANKRD11 gene mutation in an individual with a mild phenotype of KBG syndrome associated to a GEFS+ phenotypic spectrum: a case report. *BMC Med Genet*. 2019;20(1):16
- Athena Diagnostics. Epilepsy Advanced Sequencing and CNV Evaluation 2020. Available from: <https://www.athenadiagnostics.com/view-full-catalog/e/epilepsy-advanced-sequencing-and-cnv-evaluation>. Last accessed: 21/10/2020
- Balestrini S, Sisodiya SM. Pharmacogenomics in epilepsy. *Neurosci Lett*. 2018;667:27-39
- Banfi P, Coll M, Oliva A, Alcalde M, Striano P, Mauri M, Princiotta L, Campuzano O, Versino M, Brugada R. Lamotrigine induced Brugada-pattern in a patient with genetic epilepsy associated with a novel variant in SCN9A. *Gene*. 2020;754:144847
- Bayat A, Bayat M, Rubboli G, & Møller RS. Epilepsy Syndromes in the First Year of Life and Usefulness of Genetic Testing for Precision Therapy. *Genes*. 2021;12(7):1051.
- Berkovic SF, Heron SE, Giordano L, Marini C, Guerrini R, Kaplan RE, Gambardella A, Steinlein OK, Grinton BE, Dean JT, *et al*. Benign familial neonatal-infantile seizures: characterization of a new sodium channelopathy. *Ann Neurol*. 2004;55(4):550-7
- Blueprint Genetics. Comprehensive Epilepsy Panel. Available from: <https://blueprintgenetics.com/tests/panels/neurology/comprehensive-epilepsy-panel/> Last accessed: 21/10/2020
- Bonanni P, Malcarne M, Moro F, Veggiotti P, Buti D, Ferrari AR, Parrini E, Mei D, Volzone A, Zara F, *et al*. Generalized epilepsy with febrile seizures plus (GEFS+): clinical spectrum in seven Italian families unrelated to SCN1A, SCN1B, and GABRG2 gene mutations. *Epilepsia*. 2004;45(2):149-58
- Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, *et al*. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*. 2019;47(D1):D1005-D12
- Cen Z, Lou Y, Guo Y, Wang J, Feng J. Q10R mutation in SCN9A gene is associated with generalized epilepsy with febrile seizures plus. *Seizure*. 2017;50:186-8

Centogene. Epilepsy Panel. Available from:

<https://www.centogene.com/science/centopedia/ngs-panel-genetic-testing-for-generalized-epilepsy-with-febrile-seizures.html> Last accessed: 21/10/2020

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X & Ruden DM. 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* (Austin). 2012;6(2):80-92.

ClinGen. SCN9A - epilepsy 2018. Available from:

<https://search.clinicalgenome.org/kb/genes/HGNC:10597>. Last accessed: 11/08/2020

Cohen JC, Kiss RS, Pertsemliadis A, Marcel YL, McPherson R & Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004;305(5685):869-872.

Cox JJ, Reimann F, Nicholas AK, Thornton G, Roberts E, Springell K, Karbani G, Jafri H, Mannan J, Raashid Y, *et al.* An SCN9A channelopathy causes congenital inability to experience pain. *Nature*. 2006;444(7121):894-8

Dynacare. Neurosure Epilepsy Gene Panel: Comprehensive (Ontario). Available from:

<https://www.dynacare.ca/specialpages/secondarynav/find-a-test/nat/neurosure%C2%A0epilepsy%C2%A0gene%C2%A0panel-%C2%A0comprehensive.aspx?sr=ONT&st=>. Last accessed: 21/10/2020

EGL Genetics. Epilepsy and Seizure Disorders Panel: Sequencing and CNV Analysis. Available from: <https://www.egl-eurofins.com/tests/MEPI1>. Last accessed: 21/10/2020

Faber CG, Hoeijmakers JG, Ahn HS, Cheng X, Han C, Choi JS, Estacion M, Lauria G, Vanhoutte EK, Gerrits MM, *et al.* Gain of function Nav1.7 mutations in idiopathic small fiber neuropathy. *Ann Neurol*. 2012;71(1):26-39

Helbig I, Ellis CA. Personalized medicine in genetic epilepsies – possibilities, challenges, and new frontiers. *Neuropharmacology*. 2020:107970

Iffland PH, 2nd, Carson V, Bordey A, Crino PB. GATORopathies: The role of amino acid regulatory gene mutations in epilepsy and cortical malformations. *Epilepsia*. 2019;60(11):2163-73

Invitae. Invitae Epilepsy Panel. Available from:

<https://www.invitae.com/en/physician/tests/03401/>. Last accessed: 21/10/2020

Jung KS, Hong KW, Jo HY, Choi J, Ban HJ, Cho SB, Chung M. KRADB: the large-scale variant database of 1722 Koreans based on whole genome sequencing. *Database*. 2020;2020

Kaplanis J, Samocha KE, Wiel L, Zhang Z, Arvai KJ, Eberhardt RY, Gallone G, Lelieveld SH, Martin HC, McRae JF, *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* 2020;586(7831):757-762

- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–443.
- Kearney H, Byrne S, Cavalleri GL, Delanty N. Tackling Epilepsy With High-definition Precision Medicine: A Review. *JAMA Neurol.* 2019;76(9):1109-1116
- Laver TW, De Franco E, Johnson MB, Patel KA, Ellard S, Weedon MN, Flanagan SE, & Wakeling MN. (2022). SavvyCNV: Genome-wide CNV calling from off-target reads. *PLOS Computational Biology*, 2022;18(3):e1009940
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-91
- Liu Z, Ye X, Qiao P, Luo W, Wu Y, He Y, Gao P. G327E mutation in SCN9A gene causes idiopathic focal epilepsy with Rolandic spikes: a case report of twin sisters. *Neurological Sciences*. 2019;40(7):1457-60
- Ma, H., Guo, Y., Chen, Z., Wang, L., Tang, Z., Zhang, J., Miao, Q., & Zhai, Q. (2021). Mutations in the sodium channel genes SCN1A, SCN3A, and SCN9A in children with epilepsy with febrile seizures plus(EFS+). *Seizure*. 2021;88:146-152.
- Mantegazza M, Gambardella A, Rusconi R, Schiavon E, Annesi F, Cassulini RR, Labate A, Carrideo S, Chifari R, Canevini MP, *et al.* Identification of an Nav1.1 sodium channel (SCN1A) loss-of-function mutation associated with familial simple febrile seizures. *Proc Natl Acad Sci U S A*. 2005;102(50):18177-82
- Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, Leong IUS, Smith KR, Gerasimenko O, Haraldsdottir E, *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019;51(11):1560-5
- Mayo Clinic Labs. Targeted Genes and Methodology Details for Epilepsy/Seizure Genetic Panels 2019. Available from: https://www.mayocliniclabs.com/it-mmfiles/Targeted_Genes_and_Methodology_Details_for_Epilepsy_Genetic_Panels.pdf. Last accessed: 21/10/2020
- Michiels JJ, te Morsche RHM, Jansen JBMJ, Drenth JPH. Autosomal Dominant Erythralgia Associated With a Novel Mutation in the Voltage-Gated Sodium Channel α Subunit Nav1.7. *Archives of neurology*. 2005;62(10):1587-90
- Møller RS, Schneider LM, Hansen CP, Bugge M, Ullmann R, Tommerup N, Tümer Z. Balanced translocation in a patient with severe myoclonic epilepsy of infancy disrupts the sodium channel gene SCN1A. *Epilepsia*. 2008;49(6):1091-4
- Mulley JC, Hodgson B, McMahon JM, Iona X, Bellows S, Mullen SA, Farrell K, Mackay M, Sadleir L, Bleasel A, *et al.* Role of the sodium channel SCN9A in genetic epilepsy with febrile seizures plus and Dravet syndrome. *Epilepsia*. 2013;54(9):e122-e6

- Mulley JC, Scheffer IE, Petrou S, Dibbens LM, Berkovic SF, Harkin LA. SCN1A mutations and epilepsy. *Hum Mutat.* 2005;25(6):535-42
- Oakley JC, Kalume F, Yu FH, Scheuer T, Catterall WA. Temperature- and age-dependent seizures in a mouse model of severe myoclonic epilepsy in infancy. *Proc Natl Acad Sci U S A.* 2009;106(10):3994-9
- Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics.* 2012;28(21):2747-2754.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D., Maller J, Sklar, P, de Bakker PIW, Daly MJ *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.* 2007;81(3):559–575.
- Singh NA, Pappas C, Dahle EJ, Claes LR, Pruess TH, De Jonghe P, Thompson J, Dixon M, Gurnett C, Peiffer A *et al.* A role of SCN9A in human epilepsies, as a cause of febrile seizures and as a potential modifier of Dravet syndrome. *PLoS genetics.* 2009;5(9):e1000649
- StataCorp. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC.
- Strande NT, Riggs ER, Buchanan AH, Ceyhan-Birsoy O, DiStefano M, Dwight SS, Goldstein J, Ghosh R, Seifert BA, Sneddon TP, *et al.* Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am J Hum Genet.* 2017;100(6):895-906
- Van Hout CV, Tachmazidou I, Backman JD, Hoffman JX, Ye B, Pandey AK, Gonzaga-Jauregui C, Khalid S, Liu D, Banerjee N, *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *BioRxiv.* 2019;572347
- Vanoye CG, Gurnett CA, Holland KD, George AL, Jr., Kearney JA. Novel SCN3A variants associated with focal epilepsy in children. *Neurobiol Dis.* 2014;62:313-22
- Wheless JW, Fulton SP, Mudigoudar BD. Dravet Syndrome: A Review of Current Management. *Pediatric neurology.* 2020;107:28-40
- Williams v. Quest Diagnostics, Inc.: United States District Court for the District Of South Carolina Columbia Division; 2018. p. 432.
- Wilmshurst JM, Gaillard WD, Vinayan KP, Tsuchida TN, Plouin P, Van Bogaert P, Carrizosa J, Elia M, Craiu D, Jovic NJ, Nordli D, *et al.* Summary of recommendations for the management of infantile seizures: Task Force Report for the ILAE Commission of Pediatrics. *Epilepsia.* 2015;56(8):1185-97

Yang Y, Wang Y, Li S, Xu Z, Li H, Ma L, Fan J, Bu D, Liu B, Fan Z, *et al.* Mutations in SCN9A, encoding a sodium channel alpha subunit, in patients with primary erythralgia. *J Med Genet.* 2004;41(3):171-4

Yang C, Hua Y, Zhang W, Xu J, Xu L, Gao F, Jiang P. Variable epilepsy phenotypes associated with heterozygous mutation in the SCN9A gene: report of two cases. *Neurological Sciences.* 2018;39(6):1113-5

Zhang T, Chen M, Zhu A, Zhang X, Fang T. Novel mutation of SCN9A gene causing generalized epilepsy with febrile seizures plus in a Chinese family. *Neurol Sci.* 2020; 41(7): 1913–1917.

S1 Table: UK Biobank allele frequencies for the SCN9A variants in Table 1

GRCh38 reference	GRCh37 reference	(rs number)	Cases - variant alleles (/1198)	Controls - variant alleles (/98720)	Cases - variant allele frequency	Controls - variant allele frequency
<i>SCN9A variants proposed in Singh et al. 2009^[7]</i>						
2:166284506 T>A p.(Asn641Tyr)	2:167141016 T>A	rs121908918	0	2	0	<0.01%
2:166311573 T>C p.(Ile62Val)	2:167168083 T>C	rs121908920	0	0	0	0
2:166306531 G>T p.(Pro149Gln)	2:167163041 G>T	rs121908921	0	1	0	<0.01%
2:166286469 C>T p.(Ser490Asn)	2:167142979 C>T	rs58022607	5	618	0.42%	0.63%
2:166281786 T>C p.(Lys655Arg)	2:167138296 T>C	rs121908919	1	276	0.08%	0.28%
2:166280452 T>C p.(Ile739Val)	2:167136962 T>C	rs182650126	7	447	0.58%	0.45%
<i>SCN9A variants proposed in subsequent publications</i>						
2:166311728 T>C p.(Gln10Arg)	2:167168238 T>C	rs267607030	0	1	0	<0.01%
2:166307014 A>G p.(Tyr107His)	2:167163524 A>G	-	0	0	0	0
2:166303195 G>T p.(Leu266Met)	2:167159705 G>T	rs201743233	0	0	0	0
2:166293358 C>T p.(Gly327Glu)	2:167149868 C>T	rs765818027	0	0	0	0
2:166198900 del p.(I1901fs)	2:167055409 del	rs1353037253	0	0	0	0
2:166198733 T>C p.(Tyr1958Cys)	2:167055243 T>C	rs761742207	0	0	0	0

Allele frequencies are calculated from Biobank exome data (SPB pipeline), available for 49,959 individuals at the time of publication. Variant frequencies are separated by presence or absence of epilepsy as defined in our methods (above). All *SCN9A* variants examined are more frequent in controls than cases except p.(Ile739Val), where there is no significant difference between cases and controls (two sided Fisher's exact test $p = 0.51$)

S2 Table: Heterozygous *SCN9A* variants proposed as a monogenic cause of seizure disorders in subsequent publications, including the testing methodology employed

Genotype (NM_002977)	Phenotype	gnomAD ¹ AC (Hom.) AF	<i>SCN9A</i> variant familial segregation	Genetic testing strategy	Additional variants not excluded	Reference
c.29A>G p.(Gln10Arg)	GEFS+	25 (1) 0.01%	Inherited from an affected parent and present in an affected sibling	Proband only NGS panel: 480 epilepsy-related genes (including <i>SCN1A</i>)		Cen <i>et al.</i> 2017 [14]
c.319T>C p.(Tyr107His)	FS	0	Inherited from an affected parent	Proband only targeted NGS panel: Cardiac and channelopathy-related genes, karyotype and aCGH	de novo 1.3 Mb duplication, <i>POLG</i> and <i>AKAP9</i> variants	Banfi <i>et al.</i> 2020[13]
c.796C>A p.(Leu266Met)	GEFS+	2 (0) <0.001%	Inherited from an unaffected parent	Dideoxy sequencing: <i>SCN1A/B</i> , <i>GABRG2</i> , <i>PCDH19</i>		Mulley <i>et al.</i> 2013 [18]
c.980G>A p.(Gly327Glu)	BECTS	11 (0) 0.005%	Inherited from an unaffected parent and identified in an affected sibling	Trio WES		Liu <i>et al.</i> 2019 [15]
	GEFS+		Inherited from an affected parent	Dideoxy sequencing: <i>SCN1A</i> and common epilepsy genes		Yang <i>et al.</i> 2018 [12]
c.1964A>G p.(Lys655Arg)	GEFS+	428 (0) 0.2%	Inherited from an unaffected parent	Trio WES and virtual gene panel analysis: 21 epilepsy-related genes (including <i>SCN1A</i>) and aCGH	<i>ANKRD11</i> heterozygous nonsense	Alves <i>et al.</i> 2019 [17]
c.5702_5706del p.(I1901fs)	GEFS+	10 (0) 0.005%	Inherited from an affected parent	Dideoxy sequencing: <i>SCN1A</i> and common epilepsy genes		Yang <i>et al.</i> 2018 [12]
c.5873A>G p.(Tyr1958Cys)	GEFS+	2 (0) <0.001%	Inherited from an affected parent and identified in a further affected individual and one individual of unknown affection.	Trio WES: <i>SCN1A</i> variants examined		Zhang <i>et al.</i> 2020 [16]

Abbreviations: AC, Allele count; aCGH, array comparative genomic hybridization AF Allele frequency; BECTS, benign partial epilepsy of childhood with centrotemporal spikes; FS, Febrile Seizures; GEFS+, generalised epilepsy with febrile seizures plus; Hom. Homozygous individuals; NGS, Next-generation sequencing; TLE, temporal lobe epilepsy. ¹ gnomAD v2.1.1 non-neuro cohort.

6

Concluding comments

This study forms part of a wider initiative aiming to delineate and characterise neurodevelopmental disorders within Palestinian and Anabaptist communities and understand the relevance of rare genetic variants to health and disease. This thesis details studies undertaken as part of this project entailing the clinical, genetic and molecular delineation of two novel monogenic forms of neurodevelopmental disorder, associated with biallelic variants in two genes; *CAMSAP1* and *SLC4A10*. Additionally, the work undertaken has enabled the clinical relevance of sequence variants in *SCN9A*, previously considered to be associated with febrile seizures and other monogenic seizure types, to be re-evaluated and confirmed not to be a cause of this group of disorders. This has had an immediate and important impact on the composition of gene panels for genetic epilepsies. Taken together, the work described in this thesis has notably advanced medical-scientific understanding of the development, physiology and pathophysiology of the nervous system, while also providing important clinical benefits for the Anabaptist and Palestinian communities and rare disease patients worldwide.

An additional discovery stemming from these studies involves the identification of biallelic variants in the succinate dehydrogenase subunit D (*SDHD*) gene in an extended Palestinian family associated with mitochondrial complex II deficiency (MIM: 619167), a severe neurodevelopmental disorder with developmental regression (Lin *et al.*, 2021), confirming variants in this gene as a cause of this condition (published open-access in the *European Journal of Human Genetics*).

The impacting findings made during this project in part derive from the unique genetic architectures of these communities. For both *CAMSAP1* and *SLC4A10* the initial discovery and disease delineation stemmed from findings in a single nuclear family within a genetically isolated Palestinian community; in each case the disease-causing variant was serendipitously increased in frequency due to cultural marriage patterns and geographical isolation. Subsequent investigations in large UK data sets comprising a large number of individuals with neurodevelopmental disorders (the DDD study and the 100,000 Genomes Project) (Smedley *et al.*, 2021; Wright *et al.*, 2018) did not identify any further individuals with biallelic loss-of-function variants in each gene. This indicates that these discoveries were unlikely to have been made in such admixed cohorts of Northern European ancestry, confirming the importance and value of studies in genetically isolated communities. Additionally, Chapter 5 highlights that within genetically isolated communities it is uniquely possible to identify large numbers of individuals with otherwise extremely rare gene variants, such as *SCN9A* c.1921A>T p.(Asn641Tyr). This genetic power provides an extremely useful means by which to correctly delineate the clinical relevance of a genetic variant, especially when associated with incomplete penetrance and/or a clinically variable phenotype. In the case of the *SCN9A* variant in the Amish, this provided sufficient evidence to correct a long-held variant-disease and gene-disease association. The incorrect disease-gene association in this case is unlikely to be exceptional among those described in the era before the widespread availability of worldwide genomic population databases. This study further highlights the value of the unique genetic architecture which makes up

particular communities globally to medical-scientific understanding of rare genetic variants.

One facet of genetic architecture can be visualised by comparing the patterns and degrees of autozygosity in different communities, reflecting their different ancestries and marriage patterns (**Figure 6.1**).

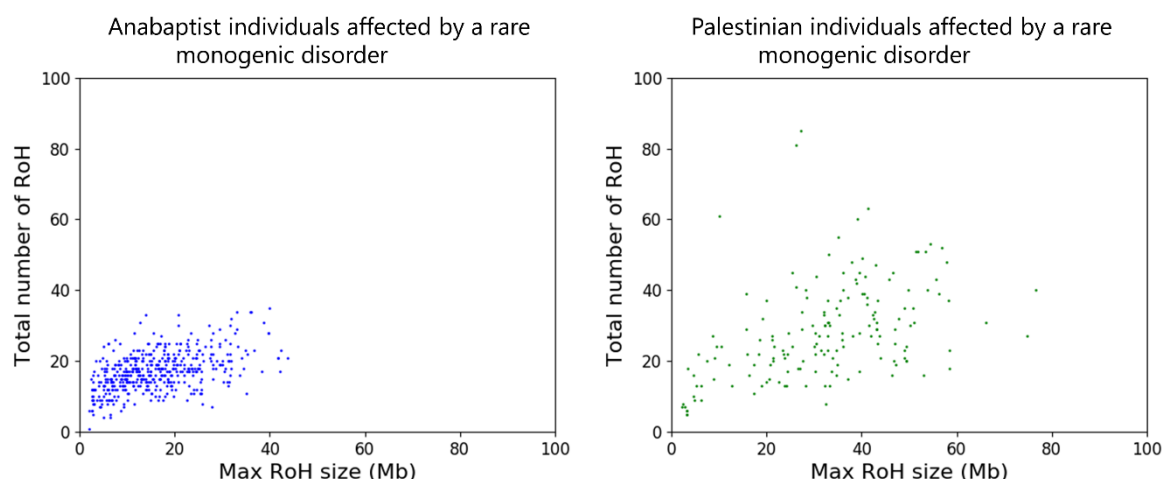


Figure 6.1: Two isolated communities show different patterns of autozygosity

Original work, unpublished data. Abbreviations: Mb, Megabase; RoH, region(s) of homozygosity

Whilst further genomic studies are required, our data thus far indicates that degrees of homozygosity in Palestinian communities closely resembles those of Pakistani communities, with numerous larger genomic regions of homozygosity, suggesting recent unions of close relatives. The absence to date of the disease-causing *CAMSAP1* and *SLC4A10* variants in other Palestinian families within our project is consistent with this. However, our studies so far involve modest sample sizes and it is probable that many variants are more widely distributed in each region, which will likely become apparent as the project continues with more families being enrolled. Recent unions of close relatives predispose to the

accumulation of otherwise rare or “private” (those not seen outside a small community or family) genetic variants, originating in recent ancestry. While again dependent on the specific community involved, many genetic variants present in Anabaptist communities typically entail ‘older’ variants, likely derived from long-deceased relatives many of whom were resident in Northern Europe before Anabaptist migrations to America (Example in forthcoming publication - “A genotype-phenotype analysis of 30 individuals with biallelic pathogenic variants in KPTN”). These variants are present in modern European populations and are hence also relevant to the diagnosis and treatment of rare neurodevelopmental disorders arising there.

Detailed knowledge of specific pathogenic variants in a community, such as the Amish, is beneficial to that community. It enables the development of targeted, rapid and cost-effective genetic approaches, such as simple dideoxy sequencing assays, performed before next-generation sequencing and multiplex PCR-based assays (Crowgey *et al.*, 2019), or chip-based methods (PlexSeq diagnostics, Cleveland, Ohio, see **Section 1.11**) to maximise diagnostic yield. Additionally, the knowledge acquired can also drastically improve diagnostic yield for rare genetic disease in a community - with our work contributing to a diagnostic uplift in the Amish from ~5% to ~70% since 2000 (*personal communication*), a much greater increase than could be attributed to improvements in sequencing alone. Pre-conception carrier screening is an established approach in some genetically isolated communities such as Ashkenazi Jews (Davidov *et al.*, 2022), where it has greatly reduced the prevalence of Tay-Sachs disease through changing marriage patterns (Singer & Sagi-Dain, 2020). Additionally, where there is evidence that an early

intervention alters the course of a condition there may be a justification for newborn screening. Historically this has only been relevant to metabolic disorders, since testing is usually performed primarily using tandem mass spectrometry. However, researchers are currently considering whether genetic methodologies (exome / genome sequencing) could enable screening for a greater number of conditions, whilst maintaining current levels of sensitivity and specificity (Adhikari *et al.*, 2020; Bick *et al.*, 2022).

Detailed knowledge of benign variants in a community is also important. This project has generated exome or genome data for 142 Palestinian individuals (130 families), almost certainly more Palestinian exome and genome data has been aggregated through this project than is publicly available in gnomAD (see **Section 1.15**). This represents a potentially valuable resource for the community which begins to address the genomic inequality that they currently face. Work in the Amish and other Anabaptist communities, carried out as part of the Windows of Hope and other related projects, has enabled the generation of the Anabaptist Variant Server (AVS - see **Section 1.15**), which has greatly increased the worldwide knowledge of genetic variation in these groups. Considering this model, we are actively seeking options for appropriate sharing of our accumulated de-identified Palestinian genetic data. This work, aiming to improve healthcare outcomes in a low / lower middle income country (LMIC) is in line with a key commitment of a number of funders, including the National Institute for Health and Care Research (NIHR, 2022), to fund research “for the direct and primary benefit of people” in those communities.

From a research perspective, while there are many advantages to investigating genetic disorders present in families from genetically isolated communities, their distinct genetic architecture also poses challenges. The high homozygosity typical of individuals in the community, and in particular the underrepresentation of these communities in population databases such as gnomAD, means that each individual will typically have between 50 and 100 candidate autosomal recessive variants that cannot be easily excluded (*unpublished observation*). While this challenge can often be overcome by careful variant evaluation and the identification of other families with the same genetic condition globally through international collaborative sequencing programs and GeneMatcher platforms, this clearly highlights the need not only for better inclusion of genetically isolated communities within such databases, but also for increased community-focussed clinical-genetic programmes (such as is the case in Anabaptist communities) to better understand the clinical relevance of rare genetic variants present in each region. Additional challenges remain as individuals with high homozygosity are also significantly more likely to have multiple genetic diagnoses, sometimes called multiple potentially relevant findings (MPRF). These were present in 22% of affected individuals from consanguineous families vs 8% in non-consanguineous families in one study (Smith *et al.*, 2019). This multi-locus inheritance can create a compound phenotype that can be challenging to disentangle and can obscure a potential diagnosis (Posey *et al.*, 2017).

As described earlier in this chapter, the studies described in this thesis have identified two novel monogenic neurodevelopmental disorders, associated with biallelic pathogenic variants in *CAMSAP1* and *SLC4A10*. These have been

grounded in the highly translational Windows of Hope project (www.WOHproject.com), based amongst the Amish communities of Ohio, and the Stories of Hope project more recently established in Palestinian communities in the West Bank. Both projects have involvement of patients and their clinicians at all stages of planning and execution to maximise impact and potential patient and family benefit with all results being fed back to local clinicians. The impact on the seventeen affected individuals identified in these studies and their families will be transformative; it will reduce their burden of genetic and non-genetic investigations, highlight potential complications that could be sought with screening investigations and brings long-term hope for future treatment for individuals with the *SLC4A10*-related neurodevelopmental disorder. Even where there is as yet no treatment, these findings will enable families to form supportive communities, previously shown to be an important source of emotional and practical support (Delisle *et al.*, 2017), and where these communities exist (<https://landonsleague.com> for *CAMSAP1*) it will enable them to grow. These findings will also potentiate genetic counselling to inform family planning, including the ability to give a family a specific risk of having a further affected child (likely 25% for each child if each parent is heterozygous for a pathogenic variant). Strategies, such as prenatal diagnosis and preimplantation genetic diagnosis, are then available to couples. The pathogenic variants in *CAMSAP1*, *SLC4A10* and *SDHD* could also be founder variants, with relevance to a village, region, or even Palestinians worldwide, enabling the screening techniques discussed above.

The benefits of this work also reach beyond the Palestinian and Amish communities who are involved. Our studies detailing discoveries of *CAMSAP1*-

and *SLC4A10*-related disorders included eight further families, mostly of Northern European heritage, who receive diagnoses and will benefit from them in a similar manner. This highlights the importance of this work, and as our previous studies (Fasham *et al.*, 2021) have shown, once conditions are identified in genetically isolated communities, they almost always lead to diagnoses being provided for other families globally. By reaching out to existing disease-specific cohorts [e.g. unsolved neuronal migration disorder cases for *CAMSAP1* (Baldassari *et al.*, 2019)] and disease-agnostic cohorts [e.g. SOLVE-RD for *SLC4A10* (Schüle *et al.*, 2021)] we expect a greater number of affected individuals will be identified, receiving the benefits described above and improving our understanding of the pathomechanism and genetic and phenotypic spectrum of these disorders.

A specific aim of this study was to facilitate precision medicine for patients and families affected by rare neurodevelopmental disorders. Currently diagnostic testing for such conditions (with ID as an example) in England is carried out using genome sequencing (NHS England, 2022). A virtual gene panel approach is employed to limit the clinical analysis workload, especially important with singleton analyses (Turnbull *et al.*, 2018). Neither *CAMSAP1* nor *SLC4A10* are currently present on neurodevelopmental disorder panels and even after publication of our work there is likely to be a delay in their inclusion. To evidence this, despite description of the original gene-disease association more than two years ago *SDHD* remains absent from ClinGen's mitochondrial panel (clinicalgenome.org/affiliation/50027, accessed 19/10/2022). Delays between initial description and gene panel inclusion allows careful curation of these data by experts. They may, therefore, be justifiable to prevent incorrect attributions

and diagnoses, such as the association that we refute between *SCN9A* and epilepsy. However, these delays are also a principal reason for missed diagnoses in large exome and genome sequencing projects (Smedley *et al.*, 2021) and for these reasons should be minimised. In a follow-up project we are exploring how this this can be improved using data from the 100,000 Genomes Project.

6.1. References

- Adhikari AN, Gallagher RC, Wang Y, Currier RJ, Amatuni G, Bassaganyas L, Chen F, Kundu K, Kvale M, Mooney SD, *et al.* The role of exome sequencing in newborn screening for inborn errors of metabolism. *Nat Med.* 2020;26(9):1392-1397.
- Baldassari S, Ribierre T, Marsan E, Adle-Biassette H, Ferrand-Sorbets S, Bulteau C, Dorison N, Fohlen M, Polivka M, Weckhuysen S, *et al.* Dissecting the genetic basis of focal cortical dysplasia: a large cohort study. *Acta Neuropathol.* 2019;138(6):885-900.
- Bick D, Ahmed A, Deen D, Ferlini A, Garnier N, Kasperaviciute D, Leblond M, Pichini A, Rendon A, Satija A, *et al.* Newborn Screening by Genomic Sequencing: Opportunities and Challenges. *Int J Neonatal Screen.* 2022;8(3):40.
- Crowgey EL, Washburn MC, Kolb EA, & Puffenberger EG. Development of a Novel Next-Generation Sequencing Assay for Carrier Screening in Old Order Amish and Mennonite Populations of Pennsylvania. *J Mol Diagn.* 2019;21(4):687-694.
- Davidov B, Levon A, Volkov H, Orenstein N, Karo R, Fatal Gazit I, Magal N, Basel-Salmon L, & Golan Mashiach M. Pathogenic variant-based preconception carrier screening in the Israeli Jewish population. *Clin Genet.* 2022;101(5-6):517-529.
- Delisle VC, Gumuchian ST, Rice DB, Levis AW, Kloda LA, Körner A, & Thombs BD. Perceived Benefits and Factors that Influence the Ability to Establish and Maintain Patient Support Groups in Rare Diseases: A Scoping Review. *Patient.* 2017;10(3):283-293.
- Fasham J, Lin S, Ghosh P, Radio FC, Farrow EG, Thiffault I, Kussman J, Zhou D, Hemming R, Zahka K, *et al.* Elucidating the clinical spectrum and molecular basis of HYAL2 deficiency. *Genet Med.* 2021;24(3):631-644.
- Lin S, Fasham J, Al-Hijawi F, Qutob N, Gunning A, Leslie JS, McGavin L, Ubeyratna N, Baker W, Zeid R, *et al.* Consolidating biallelic SDHD variants as a cause of mitochondrial complex II deficiency. *Eur J Hum Genet.* 2021;29(10):1570-1576.
- NHS England. (11/08/2022). National genomic test directory for rare and inherited disease. Retrieved from <https://www.england.nhs.uk/publication/national-genomic-test-directories/> on 19/10/2022.
- NIHR. (2022). Global health research. Retrieved from <https://www.nihr.ac.uk/explore-nihr/funding-programmes/global-health.htm> on 19/10/2022.
- Posey JE., Harel T, Liu P, Rosenfeld JA., James RA., Coban Akdemir ZH., Walkiewicz M, Bi W, Xiao R, Ding Y *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N Engl J Med.* 2017;376(1):21-31.
- Schüle R, Timmann D, Erasmus CE, Reichbauer J, Wayand M, van de Warrenburg B, Schöls L, Wilke C., Bevot A, Zuchner S, *et al.* Solving unsolved rare neurological diseases-a Solve-RD viewpoint. *Eur J Hum Genet.* 2021;29(9):1332-1336.
- Singer A, & Sagi-Dain L. Impact of a national genetic carrier-screening program for reproductive purposes. *Acta Obstetrica et Gynecologica Scandinavica.* 2020;99(6), 802-808.
- Smedley D, Smith KR, Martin A, Thomas EA., McDonagh EM., Cipriani V, Ellingford JM, Arno G, Tucci A., Vandrovcova J, *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med.* 2021;385(20):1868-1880.
- Smith ED, Blanco K, Sajan SA, Hunter JM, Shinde DN, Wayburn B, Rossi M, Huang J, Stevens CA, Muss C, Alcaraz W, *et al.* A retrospective review of multiple findings in diagnostic

exome sequencing: half are distinct and half are overlapping diagnoses. *Genet Med.* 2019;21(10):2199-2207.

Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, Halai D, Baple E, Craig C, Hamblin A, *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ.* 2018;361:k1687.

Wright, CF, McRae JF, Clayton S, Gallone G, Aitken S, FitzGerald TW, Jones P, Prigmore E, Rajan D, Lord J, *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med.* 2018;20(10):1216-1223.

7

Appendix

7.1. Peer-reviewed manuscripts arising from this project

ORCID ID: 0000-0002-7614-9202

All open access, ordered by contribution then by date

First Author (6)

Fasham J*, Huebner AK*, Liebmann L*, Khalaf-Nazzal R*, *et al.* SLC4A10 mutation impairs GABAergic transmission causing a recognisable neurodevelopmental disorder.

Revision under review with *Brain*.

Khalaf-Nazzal R*, Fasham J*, Biallelic CAMSAP1 variants cause a clinically recognizable neuronal migration disorder. *Am J Hum Genet.* 2022;109(11):2068-2079

Fasham J*, Lin S*, Ghosh P*, Radio FC*, *et al.* Elucidating the clinical spectrum and molecular basis of HYAL2 deficiency, *Genet Med.* 2022;24(3):631-644

Lin S*, Fasham J*, Al-Hijawi F*, *et al.* Consolidating biallelic SDHD variants as a cause of mitochondrial complex II deficiency. *Eur J Hum Genet.* 2021;29(10):1570-1576

Khalaf-Nazzal R*, Fasham J*, Ubeyratna N, *et al.* Final Exon Frameshift Biallelic PTPN23 Variants Are Associated with Microcephalic Complex Hereditary Spastic Paraplegia. *Brain Sci.* 2021;11(5):614

Fasham J, Leslie JS, Harrison JW, *et al.* No association between SCN9A and monogenic human epilepsy disorders. *PLoS Genet.* 2020;16(11):e1009161

Co-author (8)

Leslie JS*, Hjeij R*, Vivante A*, Bearce EA, Dyer L, Wang J, Rawlins L, Kennedy J, Ubeyratna N, Fasham J, *et al.*, Biallelic DAW1 variants cause a motile ciliopathy characterized by laterality defects and subtle ciliary beating abnormalities, **Genet Med.** 2022;24(11):2249-2261

Tábara LC*, Al-Salmi F*, Maroofian R*, Al-Futaisi AM, Al-Murshedi F, Kennedy J, Day JO, Courtin T, Al-Khayat A, Galedari H, Mazaheri N, Protasoni M, Johnson M, Leslie JS, Salter CG, Rawlins LE, Fasham J, *et al.* TMEM63C mutations cause mitochondrial morphology defects and underlie hereditary spastic paraplegia, **Brain.** 2022;145(9):3095-3107

Ma Y, Wang X, Shoshany N, Jiao X, Lee A, Ku G, Baple EL, Fasham J, *et al.* CLCC1 c.75C>A Mutation in Pakistani Derived Retinitis Pigmentosa Families Likely Originated with a Single Founder Mutation 2,000–5,000 Years Ago, **Front Genet.** 2022;22:13:804924

Salter CG*, Cai Y*, Lo B*, Helman G*, Taylor H, McCartney A, Leslie JS, Accogoli A, Zara F, Traverso M, Fasham J, *et al.* Biallelic PI4KA variants cause neurological, intestinal and immunological disease, **Brain.** 2021;144(12):3597-3610

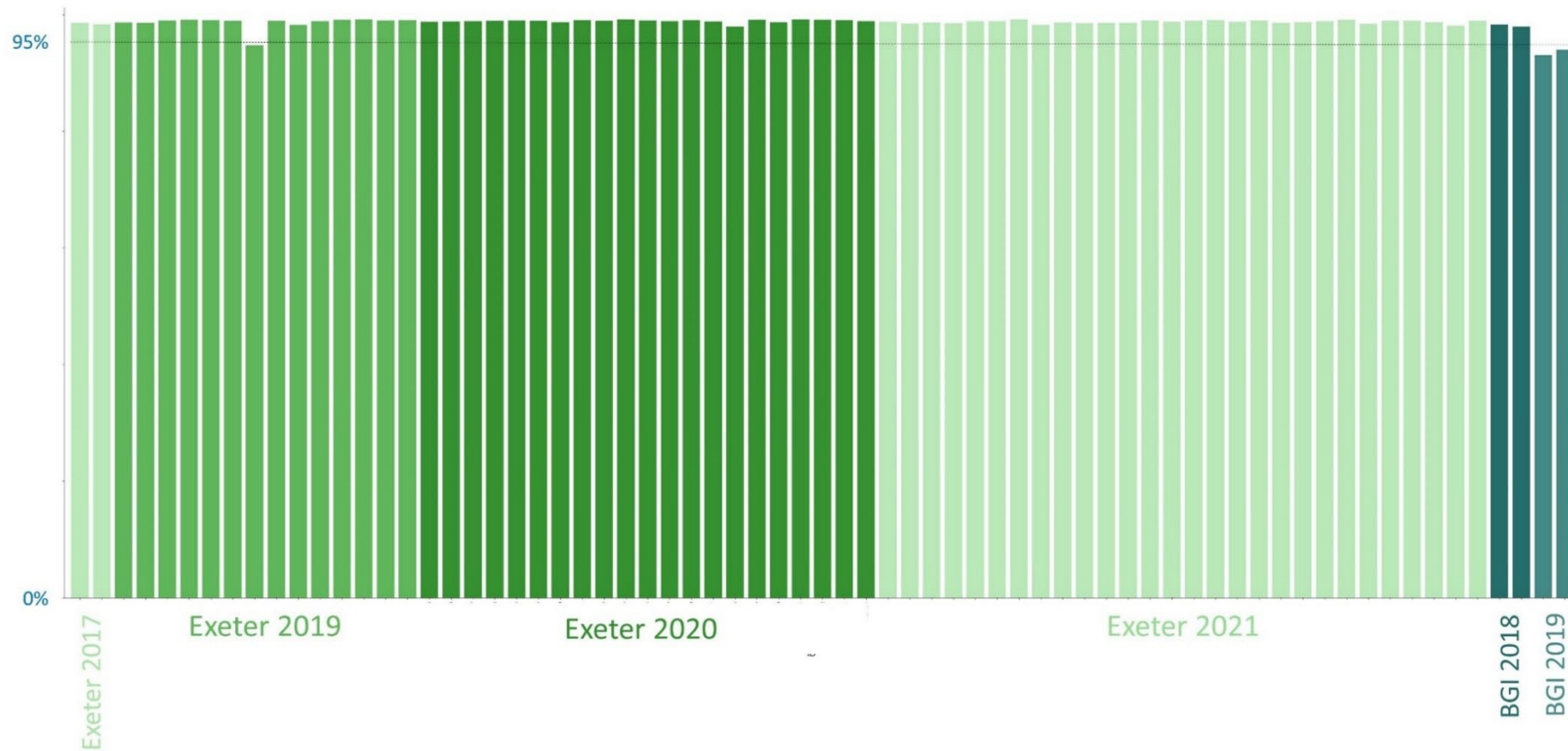
Gunning AC*, Fryer V*, Fasham J, *et al.* Assessing performance of pathogenicity predictors using clinically relevant variant datasets **J Med Genet.** 2021;58:547-555.

Rickman OJ*, Salter CG*, Gunning AC, Fasham J, *et al.* Dominant mitochondrial membrane protein-associated neurodegeneration (MPAN) variants cluster within a specific C19orf12 isoform **Parkinsonism Relat Disord.** 2020;82:84-86

Leslie JS., Rawlins LE, Chioza BA, Olubodun OR, Salter CG, Fasham J, *et al.* MNS1 variant associated with situs inversus and male infertility. **Eur J Hum Genet.** 2020;28:50–55

Akbar A, Prince C, Payne C, Fasham J, *et al.* Novel nonsense variants in SLURP1 and DSG1 cause palmoplantar keratoderma in Pakistani families. ***BMC Med Genet.*** 2019; 20(1):145.

7.2. Example QC metrics – Coverage @ 20X for Palestinian exome data



7.3. QC.py – QC metric aggregator and plotter

```

# searches all folders for QC information
# requires folders to be set up in usual manner as per Grace
# requires a pre-existing batch list
"/mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/Scripts/QC/batch_list.tsv"

# Imports required software libraries (may need to be installed if not in your VE)
import csv
import fnmatch
import matplotlib.pyplot as plt
import os
import pandas as pd
import seaborn as sns
import sys

# command line argument 1 is the top-level directory you want to analyse without end / (e.g.
~/CrosbyBapleGroup/Amish, Palestine)
directory = sys.argv[1]
# Creates a list of all of the subdirectories within this (these should be the individual exomes)
directory_list = os.listdir(directory)
print(directory_list)
# batch_list is a tsv with two columns - ID and Batch.
batch_list = "/mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/Scripts/QC/batch_list.tsv"
# All batches that are used in batch_list must be present in the batches list below
batches = ['Otogenerics','Exeter 2016','Exeter 2017','Exeter 2018','Exeter 2019','Exeter 2020','Exeter 2021',
'Baylor 2017','BGI 2018','BGI 2019','BGI May 2019','BGI May 2019 repeat','BGI 2020','UW','Pittsburg']

output_tuples_list = []
failed_samples = []

# Set up two possible position dictionaries to identify location of each metric within the metrics files
# (older hs_metrics files have fewer columns)
old_position_dict = {"total_reads":5, "pf_reads":6, "pct_pf_unique_reads_aligned":11, "fold_enrichment":24,
"pct_target_bases_2x":27, "pct_target_bases_10x":28, "pct_target_bases_20x":29,
"pct_target_bases_30x":30, "at_dropout":41, "gc_dropout":42}
new_position_dict = {"total_reads":5, "pf_reads":6, "pct_pf_unique_reads_aligned":11,
"fold_enrichment":27, "pct_target_bases_2x":36, "pct_target_bases_10x":37, "pct_target_bases_20x":38,
"pct_target_bases_30x":39, "at_dropout":50, "gc_dropout":51}

# Function which updates Global variables with information from the Picard HS metrics file for this Sample
# Takes HS metrics file location and dictionary (old_position_dict, new_position_dict) as arguments
def extract_hs_metrics(hs_metrics_file, dict):
    # Will fail here if there is no HS metrics file
    with open(hs_metrics_file) as f:
        # Defines global variable that can be used outside this function
        global total_reads, pf_reads, pct_pf_unique_reads_aligned, fold_enrichment, pct_target_bases_2x,
        pct_target_bases_10x, pct_target_bases_20x, pct_target_bases_30x, at_dropout, gc_dropout
        reader = csv.reader(f, delimiter='\t')
        i = 0
        #Skips the header
        while i < 6:
            i += 1
            next(reader)
        # Extracts metrics, uses the specific dictionary given (old_position_dict, new_position_dict) to find the
        correct column
        for row in reader:
            if row[0] == "metrics":
                total_reads = row[dict["total_reads"]]
                pf_reads = row[dict["pf_reads"]]
                pct_pf_unique_reads_aligned = row[dict["pct_pf_unique_reads_aligned"]]
                fold_enrichment = row[dict["fold_enrichment"]]
                pct_target_bases_2x = row[dict["pct_target_bases_2x"]]

```

```

pct_target_bases_10x = row[dict["pct_target_bases_10x"]]
pct_target_bases_20x = row[dict["pct_target_bases_20x"]]
pct_target_bases_30x = row[dict["pct_target_bases_30x"]]
at_dropout = row[dict["at_dropout"]]
gc_dropout = row[dict["gc_dropout"]]

def file_len(fname):
    with open(fname) as f:
        for i, l in enumerate(f):
            pass
    return i + 1

# reads each subdirectory (sample) one by one
for dir in directory_list:
    # trigger = 1 is a switch that can be turned off to stop this process if failures
    trigger = 1
    DirPath = directory + '/' + dir + '/'
    path1 = directory + '/' + dir + '/' + "r01_metrics" + '/'

    # Reset all the variables from the previous Sample:
    pf_reads, pct_pf_unique_reads_aligned, fold_enrichment, pct_target_bases_2x, pct_target_bases_10x,
    pct_target_bases_20x, pct_target_bases_30x, at_dropout, gc_dropout, E, S, total_homozygosity =
    (0,)*12

    # Check for the location of the HS metric files
    # Try first using the new_position_dict positions (Route 1)
    try:
        for file in os.listdir(path1):
            if fnmatch.fnmatch(file, '*S_metrics.csv'):
                # Looking for HS_metrics.csv, This will fail if there is no HS metrics file (and eventually exit with
                "Exception as e:" line)
                path2 = directory + '/' + dir + '/' + "r01_metrics" + '/' + file
                # This will fail if HS metrics file is in the old_position_dict configuration, this is addressed with
                except
                block below
                extract_hs_metrics(path2,new_position_dict)
                # Prints to the shell - Sample ID and that it is taking Route 1
                print("1 - " + dir)
                # extract homozygosity data
                if fnmatch.fnmatch(file, 'homozygosity.csv'):
                    path3 = directory + '/' + dir + '/' + "r01_metrics" + '/' + "homozygosity.csv"
                    with open(path3) as tsvfile:
                        reader = csv.reader(tsvfile, delimiter='\t')
                        for row in reader:
                            total_homozygosity = row[1]

            except:
                # Alternative try using old_position_dict positions (Route 2)
                try:
                    for file in os.listdir(path1):
                        if fnmatch.fnmatch(file, '*S_metrics.csv'):
                            path2 = directory + '/' + dir + '/' + "r01_metrics" + '/' + file
                            # Prints to the shell - Sample ID and that it is taking Route 2
                            print("2 - " + dir)
                            extract_hs_metrics(path2,old_position_dict)

                except Exception as e:
                    # Errors usually imply that the directory is not a Sample (e.g. project directories) or not fully
                    processed
                    (normally no r01_metrics/ directory)
                    print(path1)
                    # the full error is printed along with the Sample
                    print(e)
                    # trigger = 0 highlights a failed Sample for no further processing
                    trigger = 0

    # if there is a HS metrics file (trigger still =1) in the subdirectory (sample), get CNV data then outputs the

```

variables for this directory into a tuple

if trigger == 1:

```

# Get CNV data from ExomeDepth and SavvyCNV files
# Define two blank variable S and E
S = E = ""
for i in os.listdir(DirPath + "r01_exome_depth" + '/'):
    # Find the exome_depth fakealamut file and save its path as E
    if fnmatch.fnmatch(i, "fakealamut.txt"):
        path2 = directory + '/' + dir + '/' + "r01_exome_depth" + '/' + i
        E = str(file_len(path2))
try:
    for j in os.listdir(DirPath + "r01_savvycnv" + '/'):
        # Find the savvycnv fakealamut file and save its path as S
        if fnmatch.fnmatch(j, "fakealamut.txt"):
            pathB = directory + '/' + dir + '/' + "r01_savvycnv" + '/' + j
            S = str(file_len(pathB))
except:
    S = ""

# Output all variables for this Sample to list as tuples
output_tuple = (dir, pf_reads, pct_pf_unique_reads_aligned,
                fold_enrichment, pct_target_bases_2x, pct_target_bases_10x, pct_target_bases_20x,
                pct_target_bases_30x, at_dropout, gc_dropout, E, S, total_homozygosity)
# Append this samples list of tuples to the final "all samples" list
output_tuples_list.append(output_tuple)

else:
    # If trigger = 0, mark as a failed samples
    failed_samples.append(dir)

```

```

columns_list=['pf_reads','pct_pf_unique_reads_aligned','fold_enrichment','pct_target_bases_2x',
'pct_target_bases_10x','pct_target_bases_20x','pct_target_bases_30x','at_dropout',
'gc_dropout','ExomeDepth','SavvyCNV','total_homozygosity']

```

Creates dataframe from results (list of tuples)

```
df = pd.DataFrame(output_tuples_list, columns = ['ID'] + columns_list)
```

For each column change strings to numeric

```
for i in columns_list:
```

```
    df[i] = pd.to_numeric(df[i], errors='coerce')
```

Merges batch data from separate txt file

this file has been moved from a previous version

```
df_batch = pd.read_csv(batch_list, sep='\t', header=0, dtype={'chrom': str})
```

```
df = df.merge(df_batch, how='left', left_on='ID', right_on='ID')
```

defines batches as an ordered categorical variable and sort databases by batch

```
df['Batch'] = pd.Categorical(df['Batch'], categories=batches, ordered=True)
```

```
df.sort_values(by=['Batch','ID'], inplace=True)
```

Output the dataframe to excel

```
df.to_excel("output.xlsx")
```

Creates bar plots

```
def plot_bar(metric):
```

```
    # Define a list of consecutive colours for each batch on the bar plot
```

```
    colors_list =
```

```
    ['grey','grey','grey','#D9EEED','#A4CFCE','#6EA8A6','#458987','#256D6A','black','#BAE8B8',
    '#369231','#60B75B','#8FD58B','#BAE8B8','#E2F7E1','grey']
```

```
    # Define the figure size
```

```
    plt.figure(figsize=(50,20))
```

```
    # for each batch
```

```
    for i in batches:
```

```
        # This subplot is defined by batch column ID
```

```
        subplotdf = df[df['Batch'] == i]
```

```
        # Choose the next colour in the list
```

```
        colour = colors_list.pop()
```

```
plt.bar(subplotdf['ID'], subplotdf[metric],color=colour)
plt.xlabel('ID',fontsize=18)
plt.xticks(rotation=90)
plt.tight_layout()
fig_out_path = metric + ".png"
plt.savefig(fig_out_path)
plt.close()
```

```
plot_bar('pf_reads')
plot_bar('pct_pf_unique_reads_aligned')
plot_bar('pct_target_bases_10x')
plot_bar('pct_target_bases_20x')
plot_bar('at_dropout')
plot_bar('ExomeDepth')
plot_bar('SavvyCNV')
```

```
print(df.head)
print(failed_samples)
print(str(len(failed_samples)) + '/' + str(len(directory_list)))
```


7.4. Additional scripts developed as part of this thesis

Rationale for development and uses

Virtual gene panels

In families where only a single individual underwent next-generation sequencing (singleton exome or genome) and where the family history did not strongly suggest a single form of inheritance the number of potentially causative variants was typically very high (500 to 1,000 after Trio filtering step). In some cases, it was necessary to further filter these variants to include only those known to be associated with a specific phenotype to maximise the chances of obtaining a known diagnosis. This approach, referred to as a virtual panel, is known to be an effective method to reduce interpretation workload but maintain diagnostic rates for proband-only (singleton) exome sequencing for rare disorders (Molina-Ramírez *et al.*, 2021). It is particularly helpful when the disorder has a clear and distinct phenotype and a small list of genetic causes (e.g. Cornelia de Lange syndrome – four genes associated) and less helpful when the phenotype is non-specific with a large number of genetic causes (e.g. non-specific ID – >2,500 genes associated).

It is not possible to achieve filtering by multiple parameters (e.g. a gene list) in Microsoft Excel so a Python script was devised (**Appendix 7.5**) that took the Alamut output file as its input. An added advantage of using the Alamut output file was that it is generated before final filtering steps, and thus includes some variants that will later be filtered out by the Trio script, such as potentially

mosaic variants with unusual allelic ratios and higher frequency, potentially hypomorphic, variants. To allow for flexibility, the gene panel could either be defined manually, as a list of tab-separated values or automatically by specifying a PanelApp panel and confidence level (red, yellow, green) and utilising the PanelApp Application Programming Interface (API) (Martin *et al.*, 2019).

“Second Hit” script

In some cases, it was important to identify all variants in a gene with the minimal amount of filtering possible. This is particularly relevant when a phenotype that is clearly present has only one known genetic cause (e.g. Mowat-Wilson syndrome – variants in *ZEB2*). The need also arises frequently when a single variant is detected in a known recessive disease gene with a compatible phenotype. To address this problem a bash script (**Appendix 7.6**) was written to extract unfiltered single nucleotide variants (SNVs), CNV and chromosomal breakpoint data from the relevant files.

7.5. VirtualPanel.py – A virtual panel creator

```

# needs the OMIM file genemap2.txt in the same folder
# needs to run with alamut.py input_file
# source ~/env/bin/activate
# python3 alamut190412.py WE_EX1617121.alamut.txt _panelsettings.xlsx

# Imports required software libraries (may need to be installed if not in your VE)
import time
from pathlib import Path
import sys
import requests
import math
import numpy as np
import pickle
import xlswriter
from GetGeneListModule import get_gene_list
import pandas as pd
from pandas import ExcelWriter

# File to be filtered is command line argument 1
input_file = Path(sys.argv[1])
# Choice of panel (PanelApp panel ID and confidence level) is command line argument 2
panel = str(sys.argv[2])
# Name of output file is command line argument 3
outfile_name = str(input_file)

# Function creates a list of genes from a PanelApp panel ID and confidence level (stored as "panel"
variable)
# Uses another self-contained script, "get_gene_list" - written by me, to do this
def gene_list_from_panel_ID(panel):
    panel_number = str(panel.split("-")[0])
    min_confidence_level = str(panel.split("-")[1])
    Genes_and_alias_List = get_gene_list(panel_number,min_confidence_level)
    genes_list = Genes_and_alias_List[0]
    alias_list = Genes_and_alias_List[1]
    print("Gene Panel =")
    print(genes_list)
    print("Aliases =")
    print(alias_list)
    combined_list = genes_list + alias_list
    return(combined_list)

## IMPORT FILE TO BE FILTERED ##
# tab separated, header row, low_memory=False allows assignment of dtypes
print()
print("starting file import")
df = pd.read_csv(input_file, sep='\t', header=0, dtype={'chrom': str}, low_memory=False, encoding='latin-1')
print("done")
print()

# Imports Gene-phenotype and Gene-inheritance dictionaries (stored as pickle file)
print("starting annotation with PanelApp data")
f = open("/home/eem/jf488/CrosbyBapleGroup/Scripts/Filter/PanelAppGeneData", "rb")
PanelAppPhenotypesDict = pickle.load(f)
PanelAppInheritanceDict = pickle.load(f)
f.close()

df["phenotypes (PanelApp)"] = df["gene"].map(PanelAppPhenotypesDict)
df["Inh (PanelApp)"] = df["gene"].map(PanelAppInheritanceDict)
print("done")
print()

# Annotation - OMIM
print("starting OMIM annotation")

```

```

OMIMdf = pd.read_csv("genemap2.txt", sep='\t', header=3, low_memory=False)
OMIMdf = OMIMdf.iloc[:,8, 12]]
df = df.merge(OMIMdf, how='left', left_on="gene", right_on="Approved Symbol")
print("done")
print()

# Transcript compression – drop transcripts where genomic position, nucleotide change, amino acid
change and slicing effect are the same
df = df.sort_values(["chrom","inputPos","transLen"], ascending=[True, True, True])
df = df.drop_duplicates(subset = ["chrom","inputPos","inputRef","inputAlt","gene",
"wtAA_1","varAA_1","localSpliceEffect"], keep='last')
print("done")
print()

# Create gene list - see "gene_list_from_panel_ID" function, defined above
print("starting making gene list")
combined_list = gene_list_from_panel_ID(panel)
print("done")
print()

# Filter dataframe using gene list
print("starting virtual panel filter")
dfVP = df[df["gene"].isin(combined_list)]
print("Virtual Panel (" + str(len(combined_list)) + " genes): " + str(dfVP.shape[0]) + " rows")

# reorder columns
cols = list(df)
first_columns = ['gene','Phenotypes','phenotypes (PanelApp)','Inh
(PanelApp)','varType','codingEffect','varLocation','gNomen','cNomen','pNomen','SpliceAI','gnomadAltCount
_all','gnomadHomCount_all','ExeterExomes_het','ExeterExomes_hom','hgmdSubCategory','clinVarClinSig
nifs']
first_columns.reverse()
for i in first_columns:
    cols.insert(0, cols.pop(cols.index(i)))
dfVP = dfVP.reindex(columns= cols)
dfVP.rename(columns={'varType': 'Type', 'hgmdSubCategory': 'HGMD', 'clinVarClinSignifs': 'ClinVar'},
inplace=True)
print()

Excel_file = outfile_name + "." + panel + ".VPanel.xlsx"

# Create a Pandas Excel writer using XlsxWriter as the engine.
writer = pd.ExcelWriter(Excel_file, engine='xlsxwriter')
dfVP.to_excel(writer, sheet_name='1')
writer.save()

print("done - file saved to " + Excel_file)

```

7.6. SecondHit.sh - a tool to extract all unfiltered variants for a particular gene

```
# usage example: bash ~/CrosbyBapleGroup/Scripts/SecondHit2022.sh TTN PL0598

gene=$1 # must be HUGO gene symbol
gene2=",${gene}," # prevents repeated matching (e.g. TTN-AS1)
dir=$2 # Patient ID
type=$3 # enter Genome if genome, else blank (file structure is different)
geneRef=/mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/Scripts/UCSC_genes37.csv
# from USCS table browser, UCSC genes - knownCanonical add hg19.kgXref fields geneSymbol

# extracts chr and coordinates for gene - required for the raw VCF. all other files filtered by gene name
chr=`grep $gene2 $geneRef | awk -F ',' '{print $1; exit}' | cut -c 4-`
start=`grep $gene2 $geneRef | awk -F ',' '{print $2; exit}'`
end=`grep $gene2 $geneRef | awk -F ',' '{print $3; exit}'`

# For genome files
if [[ $type == "Genome" ]]
then

# (1) Searches for unannotated SNVs in unfiltered vcf
awk -v chr=$chr -v start=$start -v end=$end '$1 == chr && $2 >= start && $2 <= end {print}'
/mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/Genomes/*/r03_vcfs/$dir.vcf >
~/SecondHit/$dir.$gene.SecondHit.txt

# (2) Searches for SNVs in unfiltered Alamut file
grep $gene /mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/Genomes/*/r03_alamut/
$dir.alamut.txt >> ~/SecondHit/$dir.$gene.SecondHit.txt

# (3) Searches for CNVs in unfiltered SavvyCNV file
grep $gene /mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/Genomes/*/r03_savvycnv/
$dir.fakealamut.txt >> ~/SecondHit/$dir.$gene.SecondHit.txt

# (4) Searches in unfiltered scramble file
grep $gene /mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/Genomes/*/r03_scramble/
$dir.fakealamut.txt >> ~/SecondHit/$dir.$gene.SecondHit.txt

# (5) Searches for predicted chromosome breakpoints within this region
grep $gene
/mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/Genomes/*/r03_breakpoints/$dir*.txt
>> ~/SecondHit/$dir.$gene.SecondHit.txt

# For exome files (process is the same, but file locations differ)
else

# (1)
awk -v chr=$chr -v start=$start -v end=$end '$1 == chr && $2 >= start && $2 <= end {print}'
/mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/*/dir/r01_vcfs/$dir.vcf >
~/SecondHit/$dir.$gene.SecondHit.txt

# (2)
grep $gene /mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/*/dir/r01_alamut/
$dir.alamut.txt >> ~/SecondHit/$dir.$gene.SecondHit.txt

# (3)
grep $gene /mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/*/dir/r01_savvycnv/
$dir.fakealamut.txt >> ~/SecondHit/$dir.$gene.SecondHit.txt

# (4)
grep $gene /mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/*/dir/r01_scramble/
$dir.fakealamut.txt >> ~/SecondHit/$dir.$gene.SecondHit.txt

# (5)
grep $gene /mnt/Data8/exome_sequencing/external/CrosbyBapleGroup/*/dir/r01_breakpoints/
$dir*.txt >> ~/SecondHit/$dir.$gene.SecondHit.txt

fi
```

7.7. databaseQuery.sh - a tool to query a database of variants

```

# Prints a header to the terminal explaining current numbers contained within the database and warnings
regarding duplicates
echo "Total unique NGS samples =
"$(($AmishExomes+$AmishGenomes+$PalestineExomes+$PalestineGenomes+$PakistaniExomes+Oma
niExomes+$OmaniGenomes))
echo "updated 20/05/2022"
echo ""
echo "$(($AmishExomes+$AmishGenomes))" Amish:"
echo $AmishExomes" Amish Exomes +7 duplicates with Baylor:
A1262,A1278,A1279,A1288,A1366,A1395,A1410"
# note A2223 and A2235 are not also in the singletons directory, but are not included as we have
genomes
echo $AmishGenomes" Amish Genomes +13 duplicates with Exome WG0946-7 [2], WG1011-20
[10], WG1176"
echo ""
echo "$(($PalestineExomes+$PalestineGenomes))" Palestinian:"
echo $PalestineExomes" Palestine Exomes " #find
~/CrosbyBapleGroup/Palestinian*/r01_alamut/*.alamut.txt | grep -v "temp_" | wc -l
echo $PalestineGenomes" Palestine Genomes +4 duplicates: PL0043, PL0091, PL0046, PL0336,
PL0468"
echo ""
echo $PakistaniExomes" Pakistani Exomes Includes Trusight,"
echo ""
echo $OmaniExomes" Omani Exomes "
echo $OmaniGenomes" Omani Genomes WG0950, WG1048"

# Settings – deleted using command line argument 2
# The default setting (no argument 2) is verbose, outputting the whole line from the alamut file
# "ID" setting, a concise output simply listing the ID numbers of those with the variant with no other details
if [[ $2 == "ID" ]]; then
    command="grep -l"
else
    command="grep -H"
fi
# "hom" setting includes only individuals with homozygous variant calls ("1/1"), the output is verbose
if [[ $2 == "hom" ]]; then
    homcheck="grep 1/1"
else
    homcheck="cat"
fi

# Search within exome files and output according to Settings
find ~/CrosbyBapleGroup/Amish*/r01_alamut/*.alamut.txt | grep -v "temp_" >> ~/exome.tmp
find ~/CrosbyBapleGroup/Palestinian*/r01_alamut/*.alamut.txt | grep -v "temp_" >> ~/exome.tmp
find ~/CrosbyBapleGroup/Pakistani*/r01_alamut/*.alamut.txt | grep -v "temp_" >> ~/exome.tmp
find ~/CrosbyBapleGroup/Omani*/r01_alamut/*.alamut.txt | grep -v "temp_" >> ~/exome.tmp

echo ""
echo "## Exomes ##"

if [[ $2 != "ID" ]]; then
    final="cat"
else
    final="cut -d / -f7"
fi

cat ~/exome.tmp | xargs $command $1 | $final | $homcheck
$command $1 ~/CrosbyBapleGroup/Palestinian/Multiple/PL0180,PL0181/r01_alamut/*.alamut.txt | cut -d '/'
-f 10- | cut -d '.' -f 1
# these are the only 2 files not also in the singletons directory
rm ~/exome.tmp

# Search within genome files and output according to Settings

```

```
echo ""
echo "## Genomes ##"

find ~/CrosbyBapleGroup/Genomes/WG*/r03_alamut/WG*.alamut.txt >> ~/genome.tmp

if [[ $2 != "ID" ]]; then
    final="cat"
    final2="cat"
else
    final="cut -d / -f 9-"
    final2="cut -d . -f 1"
fi

cat ~/genome.tmp | xargs $command $1 | $final | $final2 | $homcheck
rm ~/genome.tmp

echo ""
```

7.8. Research proposal: DDD (CAP330)

Proposal #	330
Proposed by	Dr Emma Baple
Research Group	Prof. Andrew Crosby Prof. Caroline Wright Dr Julia Rankin Dr James Fasham Joseph Leslie
Contact details	ebaple@nhs.net jamesfasham@nhs.net
Additional DDD researchers involved from other centres	
Date	03/09/2020
Title of project	Elucidating the phenotype associated with mutation of Calmodulin-regulated spectrin-associated proteins (CAMSAPs)
Brief description of project	CAMSAP molecules bind with microtubules via the C-terminal CKK domain and stabilises the negative end. MARK2 is thought to be a key kinase regulating the binding of CAMSAPs to microtubules. This project will investigate whether there are human phenotypes associated with mutation of these molecules.
Please specific how the patient subset should be identified	HGNC Gene Names: CAMSAP1, CAMSAP2, CAMSAP3 and MARK2 HPO Phenotypes (term and code): All Patient Info Question:
What data do you require?	Option 2. Unreported variants – we can provide you with VCF files, phenotypes and extended clinical data via sFTP (<i>requires a completed Data Access Agreement</i>)
Will the project require additional local ethical approval?	No
Will data be shared with investigators outside DDD?	No
If yes, what type of data and with whom?	
Additional comments	

7.9. Research proposal: 100,000 Genomes project (RR349)

Available at <https://research.genomicsengland.co.uk/research-registry>
(accessed 01/02/2022)

Research registry ID: RR349 **Date submitted:** 30/03/2020

Project lead: James Fasham

Title: Novel insights into rare inherited disorders

Primary domains Enhanced interpretation

Lay summary:

Rare inherited disorders can affect the way that the body and the brain grows, develops and functions. These disorders occur as a result of changes to the genetic code which may be passed down through families. The causes of these disorders remain poorly understood, presenting an immense healthcare burden worldwide. This means that affected families often undergo multiple investigations in search of a diagnosis.

Studies of rare inherited diseases that occur more frequently within certain communities, such as the Amish, make a significant contribution to the scientific understanding of human growth, development and function. The unique genetic make-up of these communities makes it easier for scientists to understand the genetic causes of inherited disease. This project aims to look at the Genomics England genome sequencing data alongside community sequencing data to identify new genetic causes of human inherited disease. Studying the biological function of these genes will help us understand more about human health and the medical problems that arise when the genes involved do not function properly. This knowledge is important to ultimately develop better treatments for these disorders and crucially this work will also provide diagnoses for patients and their families worldwide.

7.10. Biallelic.sh – for extracting 100,000 Genomes project data

```
module load bio/BCFtools/1.9-foss-2018b
```

```
# Reference file containing gene start and end coordinates in GRCh38
```

```
ref_file=/public_data_resources/genes_and_regions/ensemble_genes_coordinates_96_grch38_2019-04-12.tsv
```

```
log=$gene/log.txt
```

```
match=`awk -F '/t' -v gene=$gene '$4 == gene {print; exit}' $ref_file`
```

```
chromosome=`awk -F '/t' -v gene=$gene '$4 == gene {print $1; exit}' $ref_file`
```

```
startpos=`awk -F '/t' -v gene=$gene '$4 == gene {print $2; exit}' $ref_file`
```

```
endpos=`awk -F '/t' -v gene=$gene '$4 == gene {print $3; exit}' $ref_file`
```

```
#create a working directory and copies command, reference line file and time to log file
```

```
mkdir $gene/
```

```
echo `date` > $log
```

```
echo $0 >> $log
```

```
echo "match line in reference file is:" >> $log; echo $match >> $log
```

```
# defines the region to extract and copies as to the log file
```

```
region38=$chromosome":"$startpos"-"$endpos
```

```
echo "region is:" >> $log; echo $region >> $log
```

```
# The aggregate call file is separated into chunks with arbitrary start and end positions
```

```
# There is a standard VCF format file containing individual level genotype information and a VEP annotated VCF format file covering the same variants but without genotype data.
```

```
# Each chunk is 150 to 200 GB compressed.
```

```
# This section identifies which chunk contains the region of interest using a pre-generated reference files agg_VEP.csv and agg_calls.csv
```

```
VepFile=`awk -v chr="$chromosome" -v pos="$startpos -F,' '$2 == chr && $3 <pos && $4 >pos {print $1}' < reference/agg_VEP.csv`
```

```
callFile=`awk -v chr="$chromosome" -v pos="$startpos -F,' '$2 == chr && $3 <pos && $4 >pos {print $1}' < reference/agg_calls.csv`
```

```
# This crops the region of interest from the VEP-annotated file and filters it by protein coding variants
```

```
Bcftools view -r $region38 $VepFile | egrep
```

```
"#CHROM|missense|stop_lost|stop_gained|inframe|frameshift|splice_donor|splice_acceptor" >
```

```
$gene/$gene.VEP.filt.csv
```

```
variants=$(wc -l $gene/$gene.VEP.filt.csv)
```

```
echo "there are $variants protein-coding variants in $gene in 100,000 genomes data"
```

```
# This crops the region of interest from the VCF aggregate call file – this takes greater than 40 minutes and generate a file of greater than 17Gb
```

```
Bcftools view -r $region38 $callFile > $gene/$gene.call.csv
```

```
# Create a header for the merged file (the original header will be lost in the merge)
```

```
head -3000 $gene/$gene.call.csv | egrep "^#CHROM" > $gene/$gene.call.header.tsv
```

```
egrep "^#CHROM" $gene/$gene.VEP.filt.csv > $gene/$gene.VEP.header.tsv
```

```
paste $gene/$gene.call.header.tsv $gene/$gene.VEP.header.tsv > $gene/$gene.header.tsv
```

```
# For both files create a unique variant ID from the core VCF data in the first column and sort the files using this value then join them (left merge)
```

```
# defines a temp directory for the sort (this generates approximately 11,000 temporary files)
```

```
awk '{print $1"-"$2"-"$3"-"$4"-"$5"0"}' $gene/$gene.VEP.filt.csv | sort -T $gene/VEP_temp -k1,1 >
```

```
$gene/call_temp.txt
```

```
awk '{print $1"-"$2"-"$3"-"$4"-"$5"0"}' $gene/$gene.call.csv | sort -T $gene/call_temp -k1,1 >
```

```
$gene/call_temp.txt
```

```
join -j1 | sort -T $gene/VEP_temp.txt $gene/call_temp.txt > $gene/$gene.out.txt
```

```
# Remove temporary files
```

```
rm $gene/$gene.call.csv
```

```
rm $gene/$gene.call.header.tsv $gene/$gene.call_temp.txt $gene/$gene.VEP.filt.csv
```

```
$gene/$gene.VEP.header.tsv $gene/$gene.VEP_temp.txt
```

7.11. Biallelic.py – for filtering 100,000 Genomes project data

```

# imports required libraries
import numpy as np
import pandas as pd
import csv
import re
from pathlib import Path

# imports a database of relevant phenotype information for all individuals
labkey_df = df = pd.read_CSV("/re_gecip/enhanced_interpretation/JFasham/labkey_phenotype.tsv",
sep='\t', low_memory=False)

#Takes command line arguments and processes these to identify input files and define output files
input_file = Path(sys.argv[1])
header_file = Path(sys.argv[2])
mono_out_file = str(Path(sys.argv[3])) + ".final.mono.tsv"
bi_out_file = str(Path(sys.argv[3])) + ".final.bi.tsv"
family_out_file = str(Path(sys.argv[3])) + ".final.fam.tsv"
transcript = sys.argv[4]

#Reads in the header file and formats it
pre_header = list(csv.reader(open(header_file, 'tr'), delimiter='\t'))
header = ["Unique"] + pre_header[0]

#Reads in a file of variants to exclude if these are provided
try:
    with open('exclude.tsv') as f:
        exclude_list = []
        for row in csv.reader(f):
            exclude_list.append(row[0])
        print(exclude_list)
except:
    print("no exclusions file found")

#Reads in the data file
df_raw = pd.read_csv(input_file, sep=' ', low_memory=False, header=None)
df_raw.columns = header

#Creates a copy of the data file - this is necessary to get the raw data past the numerise function intact
df_copy = pd.read_csv(input_file, sep=' ', low_memory=False, header=None)
df_copy.columns = header

#The numerise function uses regular expressions to in-place replace heterozygous calls with 1 and
homozygous calls with 2.
def numerise(x):
    print("numerising")
    x.drop(x.columns[1:17], axis=1, inplace=True)
    x.Replace(re.compile('0/0:*'), 0, inplace=True)
    x.replace(re.compile('0/1:*'), 1, inplace=True)
    x.replace(re.compile('1/1:*'), 2, inplace=True)

    x = x.apply(pd.to_numeric, errors='coerce')

    return x

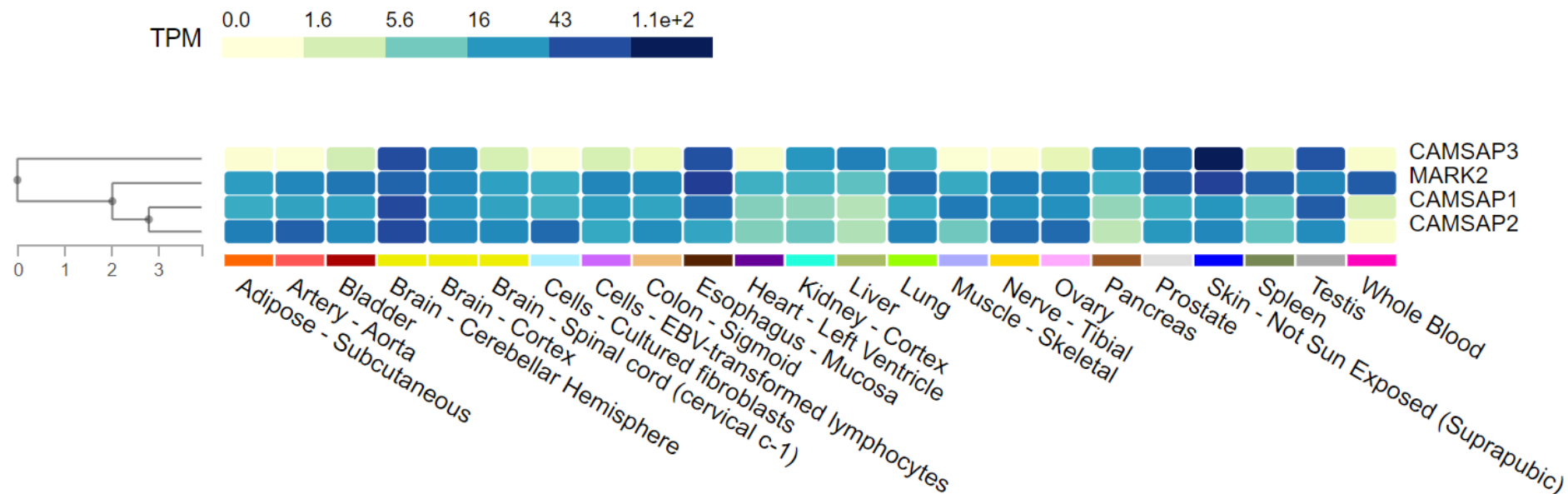
```

7.12. CAMSAP1 study data collection proforma

	Individual 1	Individual 2
CAMSAP1 variants (NM_015447.4)		
Ethnicity		
Sex		
Age at last assessment		
Gestation (weeks)		
Birth weight kg (SDS)		
Height cm (SDS)		
Weight kg (SDS)		
Head circumference cm (SDS)		
DEVELOPMENT		
Developmental Delay/intellectual impairment		
Vision		
Speech and language		
Hearing		
Gross motor		
Fine motor		
NEUROLOGICAL		
Central tone		
Peripheral tone		
Reflexes		
Abnormal movements (please describe)		
Seizures		
Age of onset		
Medication		
EEG findings		
CRANIOFACIAL FEATURES		
Prominent metopic suture		
Prominent ears		
High arched palate		
Retrognathia		
Wide nasal bridge		
NEUROIMAGING FINDINGS		
Neuronal migration disorder (Lissencephaly/Pachygyria, please describe)		
Agenesis of corpus callosum		
Cerebellar hypoplasia (please describe)		
Other features (please describe)		
OTHER CLINICAL FEATURES (please describe)		
Hirsutism		
OTHER INVESTIGATIONS (please describe)		

7.13. Relative expression of CAMSAPs and MARK2 in human cells

Data were obtained from the GTEx database (<https://gtexportal.org>) last accessed 30/09/2022



7.14. Identified individuals with variants in *CAMSAP3*

NM_020902	GnomAD (v2.1.1; v3.1)	Age	Sex	GDD	Microcephaly	Seizures	SNHL	Growth restriction	Tone	MRI	Other
<i>Biallelic</i>											
<i>in trans</i> c.1050-2A>G paternal <i>de novo</i>	0;0	4y	NK	✓ Sev.	✓	NK	NK	NK	NK	volume loss	involuntary writing movements
p.(Pro774Alafs*95) mother is carrier	0;0										
biallelic predicted LoF	NK	NK	NK	✓ Sev.	✓	✓ infantile spasms	NK	NK	NK	↓ white matter, thin cc.	Autism spectrum disorder
Homozygous p.(Gln1047Arg) Highly constrained	0;1 het	2y	F	Y	Y	NK	NK	NK	↓	NK	Fulminant Hepatitis
Homozygous p.(Val1065Met)	65;6 het	2.5y	M	Y	15%	no	NK	Tall	↓	NK	Aortic dilatation Arachnodactyly Arthrogryposis Ligamentous laxity Pectus carinatum Strabismus
<i>Monoallelic</i>											
heterozygous: unknown		6y	M	Sev.	✓	NK	NK	Y	↑	NK	
Heterozygous <i>de novo</i> missense		NK	NK	NK	NK	NK	Y	NK	NK	NK	multiple aneurysms, hypoplasia of the left half of the body L cerebral hemisphere hypoplasia tooth anomalies cutis aplasia on the left hand
Heterozygous <i>de novo</i> missense		NK	NK	✓	✓	NK	NK	NK	NK	NK	neoplasm - also second candidate
Heterozygous <i>de novo</i> LoF		NK	NK	NK	NK	NK	NK	NK	NK	NK	"severe neurodevelopmental phenotype"
Heterozygous <i>de novo</i> p.(Arg1040Trp)	0;1 het	NK	NK	✓	✓	NK	NK	NK	NK	NK	Abnormalities of outer ear, nose and male genitalia
Heterozygous <i>de novo</i> Arg1145Cys in CKK domain	0;0	8y	M	✓	No	NK	NK	NK	NK	normal	

7.15. Identified individuals with variants in MARK2

	DDD	ClinVar	ClinVar	DDD	DDD	Decipher	DDD	100K	DDD /100K	ClinVar
NM_001039469.2	p.(Gln95fs)	p.(Arg302*)	p.(Thr374fs)	p.(Phe554fs)	p.(Gln747*)	p.(*789Leufs*190)	p.(Ala80Val)	p.(Glu100Lys)	p.(His167Gln; Gln168Lys)	p.(Phe194Ser)
<i>de novo</i>	Yes	Yes	Yes	NK	Yes	NK	Yes	Yes	Yes	Yes
gnomAD v2.1.1	-	-	-	-	-	-	-	-	-	-
Protein domain / outcome	pLoF			pLoF	Escape nonsense - mediated decay	Extends the protein by 190aa	Protein kinase	Protein kinase	Protein kinase	
Gender	Male	NK	NK	Female	Male	Male	Male	Female	Female	NK
Age	6y10m			4y8m	3y4m		2y		10y11m	
birthweight (SDS)	-0.61	IUGR	NK	NK	-0.95	NK	-0.04	IUGR	-0.92	NK
birth_ofc (SDS)	NK	NK	NK	NK	NK	NK	NK	NK	-1.66	NK
height (SDS)	~50th%	NK	NK	-3.31	~25th%	NK	NK	NK	+2.33	NK
weight	NK	NK	NK	-1.3	~75th%	NK	NK	NK	+2.01	NK
OFC (SDS)	~25th%	NK	NK	-1.09	-0.39	NK	-1.51	NK	-3.86	NK
GDD	✓	✓	NK	✓	✓	✓	✓	✓	✓	✓
Smiled	8w	NK	NK	NK	NK	NK	NK	NK	NK	NK
Sat	6m	NK	NK	9m	7m	NK	NK	NK	1y	NK
Walked	11m	NK	NK	3-4y	13m	NK	1y4m	NK	2y	NK
First words	NK	NK	NK	Not yet	2.5-3y	NK	Not yet	NK	2y6m	NK
Seizures	✗			✗	✗	NK	✗	✓Infantile spasms	✗	✓
Brain MRI	Normal	NK	NK	NK	NK	NK	Abnormal	NK	NK	NK

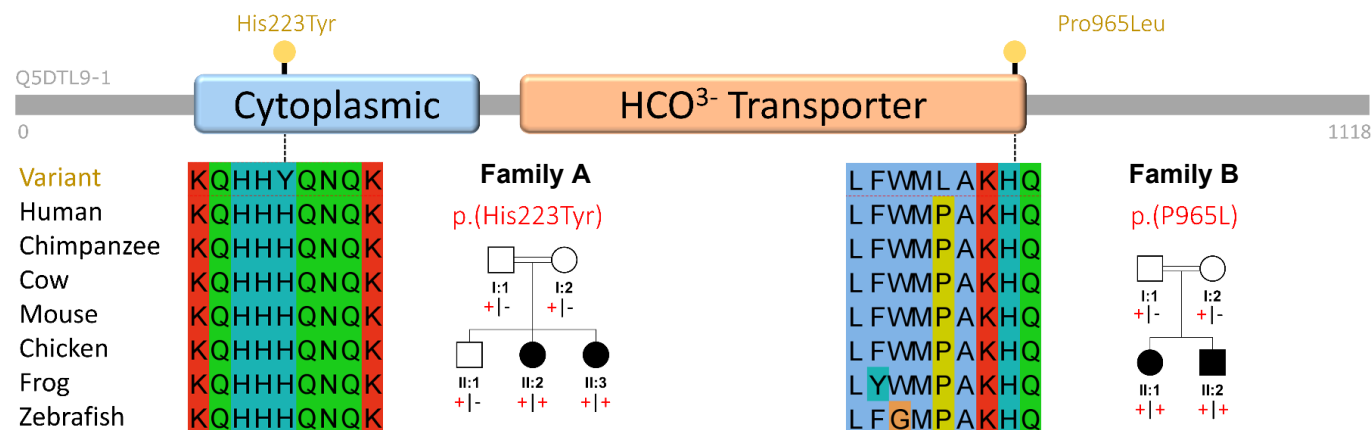
Abbreviations: - : absent from gnomAD, aa: amino acids, DDD pLoF: predicted loss of function, GDD: global developmental delay, OFC: occipitofrontal circumference SDS: standard deviation scores

7.16. SLC4A10 study data collection proforma

	Individual 1	Individual 2
SLC4A10 variant(s) genomic position (with build)		
SLC4A10 Transcript and effect(s)		
Discovered by... (e.g. Clinical Trio exome [provider])		
Family structure (no of affected / unaffected siblings)		
Segregation findings		
Other variants of note		
Ethnicity		
Nationality		
Sex		
Year of birth		
Age at last assessment		
Height at last assessment		
Birthweight		
Weight at last assessment		
OFC at last assessment		
Birth OFC		
Feeding difficulties reported		
Global developmental delay / impairment (degree)		
Developmental Regression?		
Features of Autism (specify)		
Gross motor abilities / milestones		
Fine Motor abilities / milestones		
Speech & Language abilities / milestones		
Behaviour		
Vision		
Hearing loss		
Central tone		
Peripheral tone		
Deep tendon reflexes		
Seizures		
EEG findings		
Cranial imaging (modality, age, findings)		
Radiology images available		
Cardiac abnormalities		
Skeletal abnormalities		
Facial dysmorphisms		
Clinical photos available		
Recurrent infections		
Other findings		
Pregnancy complications		
Birth complications		
Gestation		
Neonatal period		

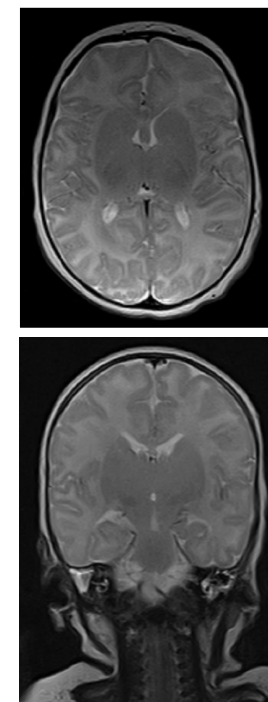
7.17. Additional individuals were identified with biallelic missense variants in *SLC4A10*

Two additional families with biallelic variants in *SLC4A10* were identified during these studies. Details are given here.



Family A: These Iranian sisters of Baloch descent, aged 7 and 5 years. Both have severe postnatal microcephaly (-3.9 to -4.2 SD), disproportionate to slightly below average other growth parameters. Neurodevelopmental delay is mild to moderate, with both sisters walking at two years and with mild speech delay. Neurological findings include mild hypotonia, with normal reflexes, ataxic gait and dysarthria. The older sister has a normal cranial MRI (right) and the younger child's MRI was under myelinated and was noted to have mesial temporal sclerosis in the absence of clinical seizures by local radiologist. WES in Individual II:2 identified a homozygous *SLC4A10* missense variant [Chr2(GRCh38):g.161862963C>T; NM_001178015:c.667C>T; p(His223Tyr)], as a candidate cause of disease.

Family B: Two siblings born to consanguineous Iraqi parents. The seven-year-old sister walked at 33 months and now has moderate ID and is unable to speak in sentences. The brother, aged six years, walked at 18 months and has mild ID with features suggestive of autism, speaking in simple sentences and following simple commands. Both siblings are microcephalic with central hypotonia and without seizures. Cranial MRI in both siblings (not shown) identified hCC and grey matter heterotopia along the lateral ventricular margins. In addition, the younger brother has possible cortical dysplasia involving medial frontal lobes. WES in one sibling identified a homozygous *SLC4A10* missense variant [Chr2(GRCh38):g.1619641666C>T NM_001178015:c.2894C>T; p.(Pro965Leu)] as a candidate cause of disease.



MRI brain for Individual II:3 from Family A

7.18. Clinical features of all individuals identified with missense variants in *SLC4A10*

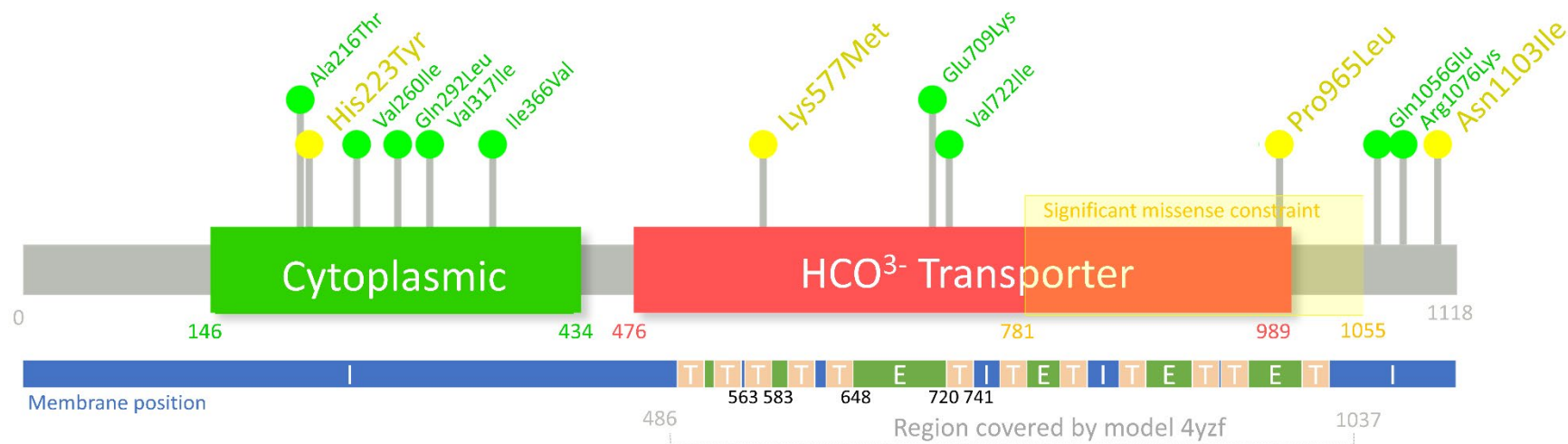
Individuals not included in the manuscript are highlighted in yellow. OFC = occipitofrontal circumference, SDS = Standard deviation scores

Individual	Family 4; II:1	Family 4; II:2	Family A; II:1	Family A; II:1	Family B; II:1
NM_001178015	homozygous c.1730A>T p.Lys577Met & c.3308A>T p.Asn1103Ile	homozygous c.1730A>T p.Lys577Met & c.3308A>T p.Asn1103Ile	homozygous c.667C>T p.His223Tyr	homozygous c.667C>T p.His223Tyr	homozygous c.2894C>T; p.Pro965Leu
Ethnicity Sex, Age at last assessment	Turkish F, 11y	Turkish M, 6y	Balooch F, 7y	Balooch F, 5y	Arab F, 7y
Growth parameters <i>Birth OFC</i> <i>OFC (cm)[SDS]</i> <i>Height (cm)[SDS]</i> <i>Weight (Kg)[SDS]</i> Feeding difficulties	NK 51.6 [-1.9] 136 [-1.2] 30.4 [-0.9] ✓	NK 46.7 [-4.2] 111 [-1.0] 17 [-1.7] ✓	33 47.5[-4.2] 118 [-0.6] 21 [-0.6] ✗	33 47 [-3.9] 97 [-2.7] 15 [-1.6] ✗	NK 49.5 [-2.6] 129.6 [+1.6] 44 [+3.4] ✗
DD/ID Neurology <i>Central tone</i> <i>Peripheral tone</i> <i>DTRs</i> <i>Seizures</i> Deafness	✓Mild-Mod ↓ ↓ ++ ✗ ✗	✓Mod-Sev ↓ ↑ brisk, bilateral Babinski ✗ ✗	✓Mild-Mod Mild ↑ ↔ ++ ✗ ✗	✓Mild-Mod Mild ↑ ↔ ++ ✗ ✗	✓Moderate ↓ ↑ + ✗ ✗
MRI Brain <i>Myelination</i> <i>Slit lateral ventricles</i> <i>hCC</i> <i>PV heterotopia</i> <i>Other</i>	Normal ✓ ✗ ✗	Normal ✗ ✓ ✗	delayed ✗ ✗ ✗	delayed ✗ ✗ ✗	Normal ✗ ✓ ✓ Mesial temporal sclerosis
Other	prominent forehead, arched eyebrows	brachycephaly upsweep on forehead bulbous nasal tip flat orbital ridges arched eyebrows prognathism	mild gait ataxia & dysarthria	mild gait ataxia & dysarthria	Strabismus, autism spectrum disorder, attentional and behavioural difficulties

7.19. Database frequency and *in silico* predictions of all missense variants identified in *SLC4A10*

	GRCh38 g.	GRCh37 g.	nucleotide change	NM_001178015 c.	Exon /27	SIFT (<0.05)	Polyphen (prob.)	REVEL	gnomAD v2.1.1	gnomAD v3.1.1
His223Tyr	2:161862963	2:162719473	C>T	667C>T	6	damaging 0	probably damaging 0.958	0.813	Absent	Absent
Lys577Met	2:161904888	2:162761398	A>T	1730A>T	14	damaging 0.048	probably damaging 0.985		Absent	Absent
Pro965Leu	2:161964166	2:162820676	C>T	2894C>T	22	damaging 0	probably damaging 1		Absent	Absent
Asn1103Ile	2:161976840	2:162833350	A>T	3308A>T	25	damaging 0.002	benign 0.197	0.239	Absent	Absent

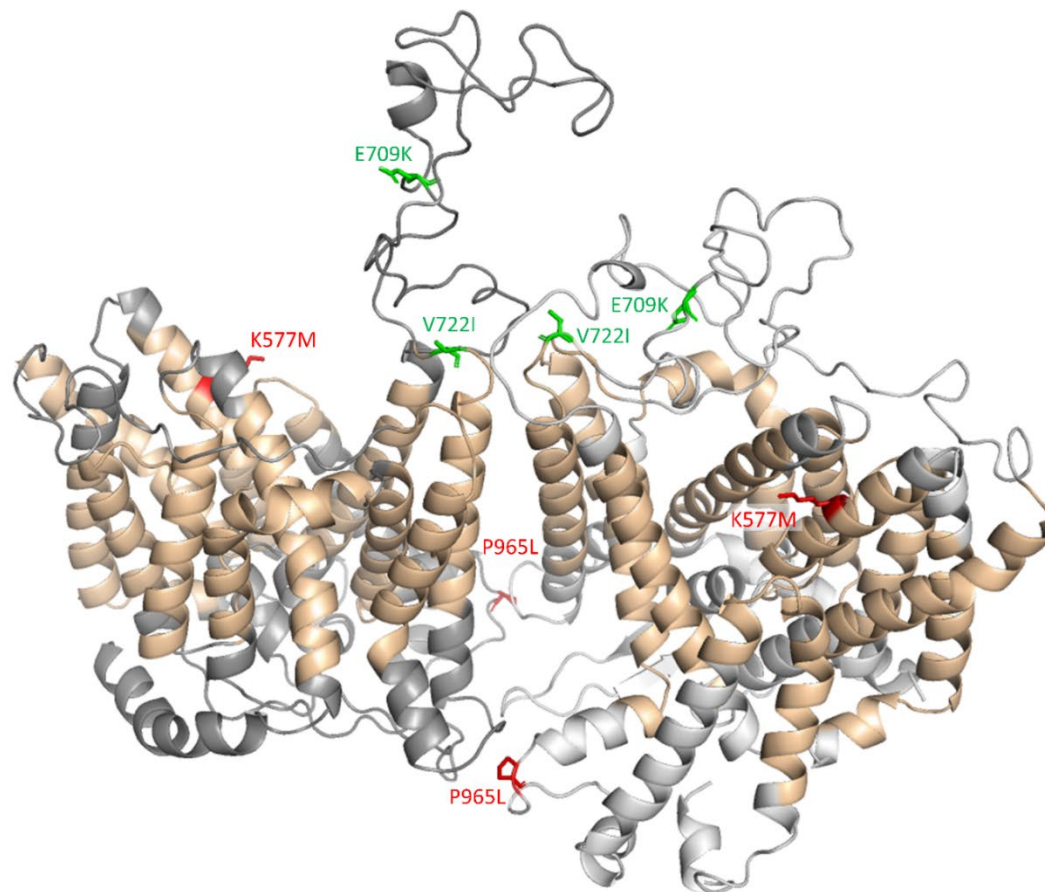
7.20. Distribution of *SLC4A10* missense variants with regard to protein domains and intra/extracellular location of the *SLC4A10* protein



Likely benign variants (green) present in homozygous state in gnomAD v2.1.1 or GnomAD v3.1, and disease-associated variants (yellow) identified in this study. The position of each variant with respect to domain/transmembrane architecture is shown: I = Intracellular, T = Transmembrane E = Extracellular. A region of statistically significant missense constraint is shown between residues 781 and 1055 ($p = 6.87 \times 10^{-8}$). Figure created using Lollipop: Jay JJ, Brouwer C (2016) Lollipop in the Clinic: Information Dense Mutation Plots for Precision Medicine. PLoS ONE 11(8): e0160519.

7.21. Additional protein modelling of missense variants in *SLC4A10*

Protein modelling demonstrating the three-dimensional orientation of disease-associated variants in dimeric human *SLC4A1*, using a published X-ray crystallography-derived structure (41% sequence identity). Residues 486-1037 are included in this model. Transmembrane helices are shaded beige, with other residues in grey. The likely pathogenic alteration Lys577Met (red) is shown to be located within a predicted transmembrane helix. Pro965Leu is intracellular, but in a region of extremely high missense constraint **Appendix 7.20**). The likely benign (homozygous in gnomAD) variants Glu709Lys and Val722Ile within the HCO₃⁻ transporter domain are predicted to lie outside of the membrane.



7.22. Functional studies of additional missense variants in *SLC4A10*

a) Heterologous expression of *SLC4A10* WT and disease-associated variants in N2a cells. Variant p.(H223T) is predominantly localised intracellularly in a similar manner to the null allele (R757*). Scale bar: 10 μm .
b,c) Both additional missense variants show abnormal pH overshoot suggesting lower transport activity than the wild-type (WT) protein. Representative single cell pH_i traces obtained for untransfected N2a cells (black) and cells transfected with the WT (green) or the p.(N1103I) (blue) construct superfused with bicarbonate-buffered solution containing 5 μM EIPA to block Na⁺/H⁺ exchange. Cells were acidified by a 5 min sodium propionate pulse. Calibration was performed with the high-[K⁺]_o/nigericin technique. Quantitative data are shown as mean + SEM from 6 independent experiments with more than 60 cells analysed (unpaired Student's t-test. n.s.: not significant; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

